Tests and sensitivity analysis for Frescalo output

Mark Hill, April 2011

Introduction and summary of results

The methods outlined in the paper 'Local frequency as a key to interpreting species occurrence data when recording effort is not known' were implemented as a Fortran program called Frescalo (FREquency SCAling LOcal). Although some components of Frescalo were written in summer 2010, the full version was not completed until January 2011. The program takes species occurrence data in the form

[location species date]

and applies neighbourhood weighting to infer (a) the probability of species presence and (b) changes in species frequency over time. Neighbourhood weights are calculated by two supporting programs, one to calculate weights based on spatial distance, the other to calculate weights based on floristic distance. These can be downloaded with Frescalo. The programs to calculate weights are merely supporting programs for Frescalo, and the user is free to define neighbourhoods in other ways if these seem better.

This supplementary paper investigates four questions:

1. How well can Frescalo estimate species richness?
2. How well does Frescalo correct for variations in recording intensity over time?
3. How sensitive is Frescalo to variation in the benchmark limit R*?
4. What are the characteristics of the smoothing kernel used to define neighbourhoods?

The outcome of tests to answer question 1 was mixed. Estimates were only as good as the neighbourhoods used to estimate species probabilities. One case, investigated in detail, showed that a training set used to define the neighbourhood of hectad TL67 had included plant records from before the Second World War. After the war, there was rapid land-use change and loss of habitat due to construction of an air force base. Because of the time discrepancy, the calculated neighbourhood of TL67 included habitats that were no longer present when its bryophytes were sampled during 1960-2009. If the method is to be used for detailed evaluation of species absences, then more careful attention must be paid to the attributes of the site at the time of sampling.

Frescalo was highly effective at correcting for variations in recording intensity. A test dataset was divided into two time periods. Records of the liverwort *Microlejeunea ulicina* were 30% fewer as a proportion of all records in the second time period than the first. Results from Frescalo, however, indicated a 23% increase. This result was correct. The district where the liverwort grows was relatively poorly sampled in the second time period. In Frescalo, recording intensity is estimated for each location and time. Results are corrected accordingly.

The benchmark limit R* had very little effect on relative change between time periods. It is one of the least sensitive parameters of the method. Raising the value of R* increased time factors by a constant multiplier. In an example, the mean proportion of benchmark species found per hectad per quinquennium fell from 27% to 16% when R* was raised from 0.14 to 0.41. This fall was was the same at each time period, with the result that the estimated change in species frequency was unaffected.

The characteristics of the smoothing kernel used to define neighbourhoods are briefly set out. However, the first test, of how well Frescalo can estimate species richness had demonstrated that the size and homogeneity of neighbourhoods can have a large effect, especially in transitional and coastal areas. Effects of the smoothing kernel are relatively small and were therefore not investigated in detail.

Dataset for testing

Frescalo was developed and first tried out on a large dataset with at least moderately good species lists for the great majority of hectads (10-km squares of the National Grid) in Britain. For subsequent tests, a subset of this dataset was used, called Test_B. This consisted of all bryophyte data in the British Bryological Society database for which the following conditions apply

1. The hectad is in eastern England, i.e. in England east of a line from Scarborough on the Yorkshire coast to Bognor Regis on the Sussex coast. There are 390 such hectads, and they are identified by having the letter T as the first their Ordnance Survey identifier, e.g. TL45, the hectad within which Cambridge is located.
2. The exact calendar year of the record is known, and is in the 50-year interval 1960-2009.

There are 110,133 such records, corresponding to an average of 5.65 records per hectad per year. The average hectad had 97.5 species recorded from it, but recording was uneven, with several northern hectads having less than 10 records. The richest hectad, with 288 species, was TQ14, containing Leith Hill in Surrey, which is the highest point in the region (294 m).

Test 1. How well can Frescalo estimate species richness?

For this test, two sizes of neighbourhood were set up

1. N200, which is as described in the main paper, comprising the 200 closest hectads to a target hectad, from which the 100 floristically most similar hectads were selected, with weights depending on physical proximity and floristic similarity.
2. N50, which is 4 times as small, comprising the 50 closest hectads, from which 25 were selected, with the same weighting.

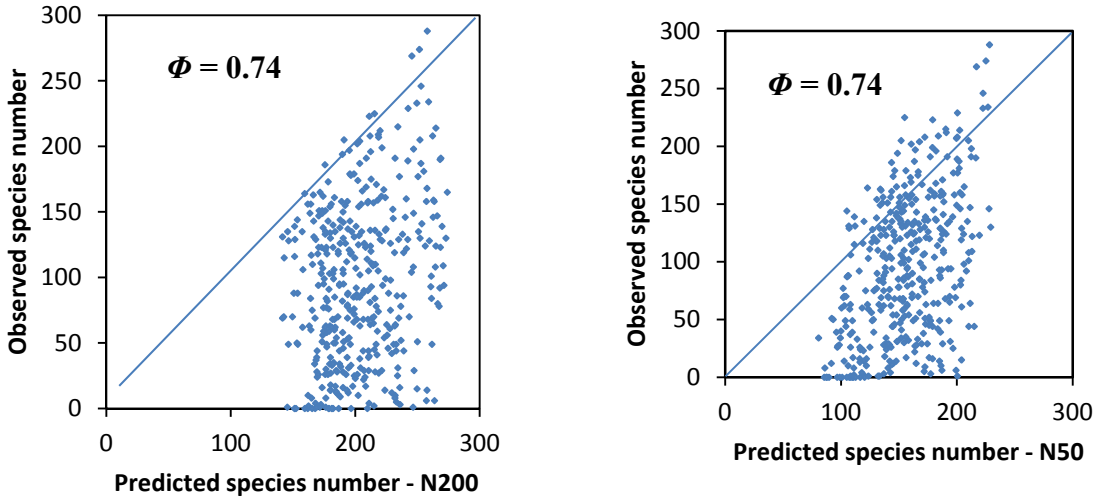

Figure S1. Observed and estimated species richness for hectads in eastern England, based on larger neighbourhoods (N200) and smaller neighbourhoods (N50).

With smoothing from the wider area N200, the upper boundary of the predicted species number showed the pattern that would be expected if at least some of the hectads were in fact well recorded (Fig. S1). A few hectads were richer than the predicted total, but most were markedly less rich. With the smaller size of neighbourhood N50, species numbers were underestimated when Φ = 0.74. Smaller neighbourhoods are in general more homogeneous, so that within them species' local mean frequency should be higher. Systematic underestimation can largely be eliminated by setting Φ for N50 to 0.8 (Fig. S2), but for some hectads, there remain substantial discrepancies. A hectad showing such a discrepancy is investigated below.

TQ97 (Fig. S2) was selected because it had a large discrepancy between the two estimates. Estimated species richness was 182 (N200) and 94 (N50). 67 bryophyte taxa have been recorded from it. This hectad, of which 43% is land, includes Sheerness on the Isle of Sheppey, Kent. Sheppey lies within the Thames Gateway development area and is strongly urbanized with some coastal marsh and arable farming. The hectad is at present under-recorded, with some common urban and suburban bryophytes such as *Lunularia cruciata*, *Pseudocrossidium hornschuchianum*, *Rhytidiadelphus squarrosus* and *Schistidium crassipilum* not listed. It has several features that make estimation of totals difficult: rapid urban development, air pollution, coastal position. Coastal and fragmentary hectads often have their species total overestimated when using N200, because some non-coastal hectads are included in the neighbourhood. An educated guess at the true total for TQ97 is 120 species, which is closer to the N50 estimate.

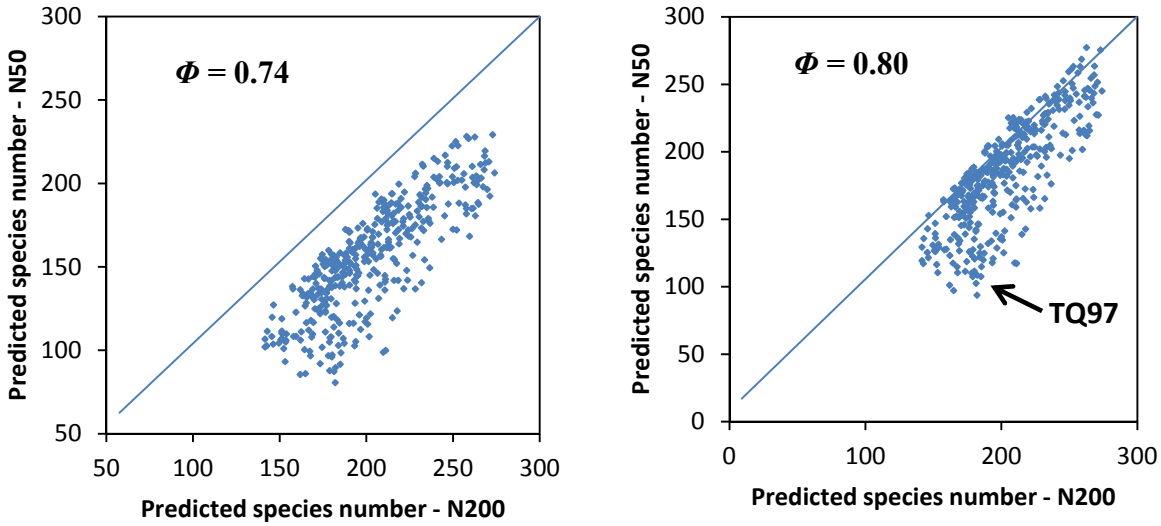

Figure S2. Predicted species numbers per hectad for differing values of Φ applied to neighbourhoods of size N50. Both sets of predictions for size N200 were based on Φ = 0.74.

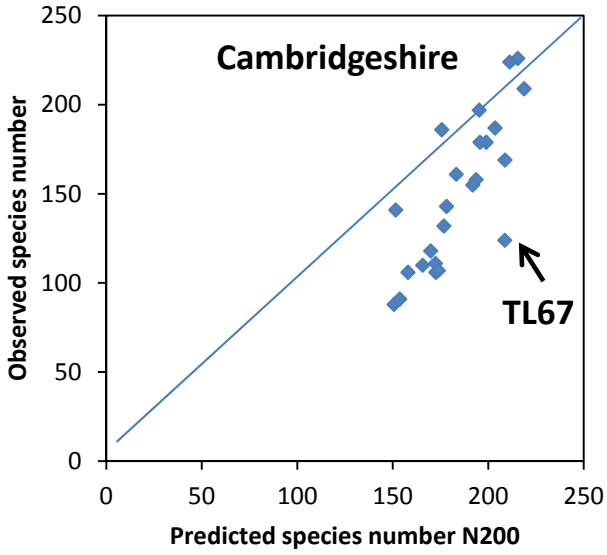

Figure S3. Predicted and observed numbers of species for Cambridgeshire hectads. Predictions for neighbourhood size N200 were based on Φ = 0.74.

It is instructive also to consider the Frescalo estimates (Fig. S3) for Cambridgeshire, which is one of the best-worked counties in eastern England. Undoubtedly the northern part of the county, which consists of fenland, mostly now in intensive arable cultivation, is less well worked than the bryologically richer south. This largely explains why in the more species-poor hectads, fewer species were actually found than were expected. One hectad, TL67, stands out as having an exceptionally low observed total. This hectad (partly in Suffolk) consists mainly of intensive arable fenland. Part of it is occupied by a large US Air Force base at Mildenhall. For this hectad, 209 species were predicted (N200) while only 124 were observed. (The N50 Φ = 0.80 prediction was even higher, with 225 species). Part of the discrepancy is due to habitat deterioration. If a longer time period is considered, including the 1950s, the observed total rises to 143 taxa. There are in my opinion about 24 additional species that are currently likely to be found in the hectad but have never been recorded there. That still leaves 42 species that are now highly unlikely.

The vascular plant database that was used to select and weight the neighbourhoods was not time-limited but included records back to the 19th century. Before the Second World War, there were numerous species of open grassland such as *Thymus pulegioides* and *T. serpyllum* in the hectad, as well as a fair number of species of long-established woodland ('ancient woodland species') including *Paris quadrifolia*, which was found in 1932. The time discrepancy between the vascular-plant training set for the neighbourhood and the dataset used for bryophytes (1960-2009) therefore resulted in leakage of grassland and ancient-woodland bryophytes into the estimated flora of the hectad. Many of these bryophytes may have been present in the 19th century, but there is no good record from that far back.

This result shows that estimates of species numbers are only as good as the neighbourhoods used to define them. Hectad TL67 is marginal to the East Anglian Fens and has suffered major land use change. The vascular-plant training set should not have included records from before 1960.

Test 2. How well does Frescalo correct for variations in recording intensity over time?
For this analysis, two reduced datasets, Test_B1 and Test_B2 were compiled. Time factors calculated by Frescalo were compared with those derived from simple proportions of all records. The datasets were defined as follows.

1. Test_B1 – the same as Test_B, but with just two time-periods, 1960-1984 and 1985-2009
2. Test_B2 – as for Test_B1, but with data from hectads in the 100-km square TQ of the British National Grid excluded if they fell in an even year (1986, 1988, ... , 2006, 2008) in the second period.

The purpose of setting up Test_B2 is that during 1985-2009 TQ should have only half the recording intensity that it did in Test_B1. This 100-km square includes the Weald of Kent and Sussex, as well as the highest point in eastern England, Leith Hill, mentioned above. It has a substantially larger liverwort flora than the rest of the region, and several liverwort species are much more frequent there than in the rest of the region (Table S1).

The test consists of comparing time factors for these liverworts between datasets Test_B1 and Test_B2. The results are set out in table S2. There are several clear results and some remarkable differences. One of these is that the epiphyte *Microlejeunea ulicina* is interpreted to be increasing by Frescalo while decreasing as a proportion of total records. It has undoubtedly increased in much of eastern and midland England, following reduced air pollution and higher winter temperatures. Its apparent decrease using the proportion of total records is spurious, and is the result of a strong imbalance in where the records were made. In the earlier period, 47% of the individual records were from the 100-km square TQ, whereas in the later period the proportion was 14%. If recording had been even over the region, the proportion should have been 25% (Table S1).

NAME	TQ	Others	Total
Diplophyllum albicans	38	23	61
Lophozia ventricosa	24	11	35
Microlejeunea ulicina	38	5	43
Nardia scalaris	24	5	29
Scapania nemorea	21	4	25
Solenostoma gracillimum	27	12	39
Hectads with land in region	97	293	390

Table S1. Selected liverworts that were more frequent in 100-km square TQ than elsewhere in eastern England.

	1960-1984	1985-2009	$\log_e$ ratio	1960-1984	1985-2009	$\log_e$ ratio	difference
	Test_B1			Test_B2			
(a) Frescalo							
Diplophyllum albicans	0.556	0.366	-0.418	0.530	0.356	-0.398	0.020
Lophozia ventricosa	0.427	0.237	-0.589	0.398	0.240	-0.506	0.083
Microlejeunea ulicina	0.520	0.639	0.206	0.489	0.609	0.219	0.013
Nardia scalaris	0.606	0.119	-1.628	0.566	0.108	-1.656	-0.029
Scapania nemorea	0.400	0.227	-0.567	0.408	0.162	-0.924	-0.357
Solenostoma gracillimum	0.359	0.306	-0.160	0.372	0.285	-0.266	-0.107
Mean							-0.063
(b) Summarized records							
Diplophyllum albicans	2225	1329	-0.515	2225	1208	-0.611	-0.095
Lophozia ventricosa	1335	598	-0.803	1335	569	-0.853	-0.050
Microlejeunea ulicina	1668	1163	-0.361	1668	960	-0.552	-0.192
Nardia scalaris	1446	233	-1.826	1446	178	-2.095	-0.269
Scapania nemorea	1057	366	-1.061	1057	213	-1.602	-0.541
Solenostoma gracillimum	1168	764	-0.424	1168	604	-0.659	-0.235
Mean							-0.230
(c) Individual years							
Diplophyllum albicans	2718	1008	-0.992	2718	746	-1.293	-0.301
Lophozia ventricosa	1221	478	-0.938	1221	359	-1.224	-0.286
Microlejeunea ulicina	2351	737	-1.160	2351	511	-1.526	-0.366
Nardia scalaris	1466	103	-2.656	1466	69	-3.056	-0.401
Scapania nemorea	1008	181	-1.717	1008	83	-2.497	-0.780
Solenostoma gracillimum	1191	349	-1.227	1191	249	-1.565	-0.338
Mean							-0.412

Table S2. Changes in frequency of liverworts in experiment where the intensity of recording in a liverwort-rich part of the region was dropped by a factor of two in the second time period (Test_B2), compared with the values in the database (Test_B1). The methods under comparison were (a) Frescalo with standard parameter settings (N200, Φ = 0.74), (b) proportion ($\times 10^6$) of total hectad records where presence in a hectad is registered for the whole period 1960-1984 or 1985-2009, and (c) proportion ($\times 10^6$) of total hectad records for individual years. Log ratios are logs to base e of the ratio of frequency for 1985-2009 to that in 1960-1984. The column 'difference' is the difference between this value for Test_B2 and Test_B1.

This fact alone shows how unreliable an estimate of frequency based on a simple proportion of total records can be. A similar point emerges, though less dramatically, from the test itself. Converting the mean log differences to percentage change, Test_B2 reduced the average frequency of these liverworts by 6% (Frescalo), 21% (lumping hectad presences in each time period together) and 28% (treating each hectad occurrence per year as a separate record). Treating the differences for the 6 species in (a) and (b) in a paired comparison, $t_5 = 3.36$ (almost exactly 99% significant in a one-tailed test). Given that the mean frequency of the six selected liverworts in hectads within 100-km square TQ was 1.77 per hectad, whereas in the rest of the region it was 0.20 per hectad, then the test Test_B2 would be expected to reduce the mean number of liverwort records per hectad in the second period from (0.14 × 1.77 + 0.86 × 0.20) to (0.14 × 0.5 × 1.77 + 0.93 × 0.20), a reduction of 26%. This is in adequate general agreement with the observed mean reduction of 28%.

Test 3. How sensitive is Frescalo to variation in the benchmark limit R*?

The analysis presented in the main paper (Fig. 6) was recalculated using the full British and Irish dataset but with varying R* (Fig. S4). The outcome of the result is reported briefly in the main paper. The result shows clearly that although the absolute values of the time factors depend strongly on R*, the proportions are hardly affected at all. By default, the value of R* was set to 0.27. The effect of increasing R* is to increase the number of benchmark species by including less frequent ones. The less frequent species are less likely to be found, so that the proportion of benchmark species recorded at any particular place and time will, on average, fall when R* is increased. This proportion is used to calculate time factors. In the notation of the main paper, the time factor is approximated by the equation

$$x_{jt} \approx \sum_i P_{ijt} / \sum_i Q_{ijt} \approx \text{Number of occurrences at time } t \,/\, \sum_i s_{it} f'_{ij}$$

The denominator can be expressed as a product

$$\sum_i s_{it} f'_{ij} = \sum_i f'_{ij} \cdot (\sum_i s_{it} f'_{ij} / \sum_i f'_{ij}) = \text{Probability sum for species } j \times \text{mean } s_{it}$$

where the mean of s_{it} is a weighted mean, weighted by f'_{ij} which is the probability of occurrence of species *j* at site *i*. The probability sum is simply the number of dots on the map if the smoothed data were sampled and plotted as a dot-map of occurrences, independent of time. Substituting,

$$x_{jt} \approx \text{Number of occurrences at time } t \,/\, (\text{Probability sum for species } j \times \text{mean } s_{it}) \qquad \text{(S1)}$$

As R* increases, so the mean s_{it} falls, and x_{jt} rises in inverse proportion. Note that the mean s_{it} is taken over all locations where species *j* is likely to occur. If there has been little sampling at time t, then there will be proportionately fewer occurrences, and the mean s_{it} falls in proportion.

Based on the values shown in Fig. S4, the average proportion of benchmark species found per sampling occasion, fell from 0.27 to 0.16 as R* increased from 0.14 to 0.41. Indeed, with a nearly three-fold increase in R* from 0.14 to 0.41, there was merely a 76% increase in the average number of benchmark species recorded (Table S3). The commonest species are indeed commoner, and for the most part much more likely to be found on short visits.

As a final check on variation in recording intensity for different species, mean s_{it} values were calculated for seven species with differing distributions (Table S4). Summing equation S1 over all time periods and rearranging, mean recording intensity is calculated as

$$\text{mean } s_{it} \approx \text{Number of occurrences} \,/\, (\text{Probability sum for species } j \times \text{sum of time factors})$$

Note that the recording intensity for a species is a weighted mean, weighted by the probability of finding the given species. The smallest value of s_{it} was 0.11 for *Myurium hochstetteri*, which is found mainly on the Outer Hebrides and whose sites are inaccessible and visited rarely. An additional factor is that *M. hochstetteri* is perhaps the most charismatic moss in the British flora, so that even non-bryologists will send in records of it, without recording the accompanying benchmark species. The largest value was 0.55 for *Scapania ornithopodioides*, which is found in the very richest bryophyte sites in Britain and Ireland. Its localities have been much visited over the past 50 years (and indeed also in the 1950s), with the result that on average more than half the benchmark species were found per quinquennium in hectads where it is likely to grow.

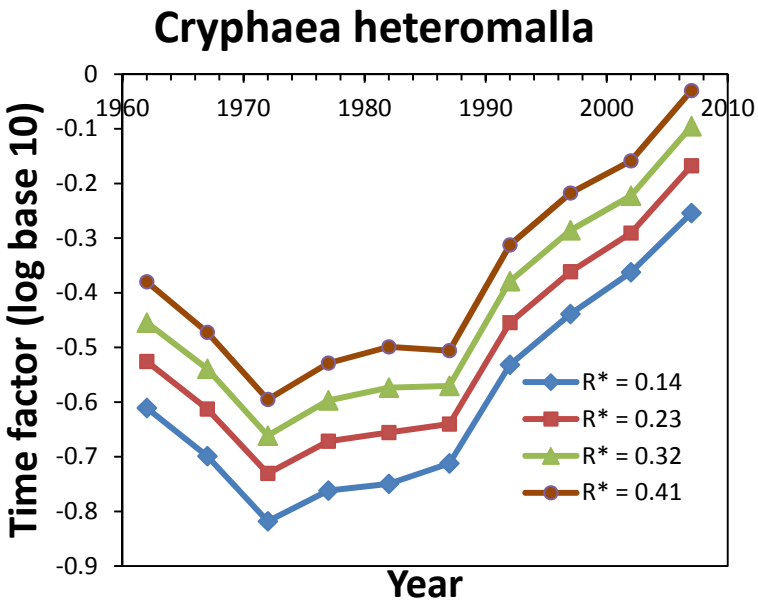

Figure S4. Time factors for frequency of the epiphyte moss *Cryphaea heteromalla*, estimated by Frescalo with differing benchmark limits R*.

R*	Prob. sum	Sum of time factors	Number of records	Mean s_{it}	Benchmark number (%)
0.14	2299	2.79	1729	0.27	3.8
0.23	2299	3.37	1729	0.22	5.1
0.32	2299	3.99	1729	0.19	6.0
0.41	2299	4.65	1729	0.16	6.6

Table S3. Summed time factors and benchmark species proportions for four values of the benchmark limit R* (data as in Figure S4). Number of records is the total for *Cryphaea heteromalla* over the period 1960-2009. The mean value of s_{it} is the mean proportion of benchmark species found at sites where *Cryphaea heteromalla* would be expected to occur. Benchmark number is expressed as a proportion % of the total estimated number of species, calculated as R* × mean s_{it} .

Species	Prob. sum	Sum of time factors	Total records	Mean s_{it}
Cryphaea heteromalla	2299	4.65	1729	0.16
Cyclodictyon laetevirens	129	3.61	105	0.22
Dialytrichia mucronata	419	2.02	187	0.22
Myurium hochstetteri	118	3.94	51	0.11
Ptilium crista-castrensis	632	2.99	338	0.18
Scapania ornithopodioides	272	0.90	135	0.55
Tortula vahliana	54	2.31	35	0.28

Table S4. Approximate mean benchmark frequency s_{it} for seven bryophyte species with the benchmark limit setting R* = 0.41.

A note on the smoothing kernel used to define neighbourhoods reported in the main paper

The smoothing kernel used in the main paper is

$$w_{ii'} = \left(1 - \frac{(k-1)^2}{200^2}\right)^4 \left(1 - \frac{(l-1)^2}{100^2}\right)^4$$

This is a product of two terms, one for physical distance and the other for floristic distance. Both of these terms have the form

$y = (1 - x^2)^4$ K24

If we denote this kernel by the name K24 (referring to the two exponents), then it may be compared with kernel K33, which is commonly used in LOESS smoothing

$y = (1-x^3)^3$ K33

As can be seen in Fig. S5, K24 decreases from the origin more rapidly than K33. Indeed, it strongly resembles a Gaussian curve and could not readily be distinguished in practice. When a hectad is situated inland, the number of neighbours rises in proportion to the square of the distance from the target hectad roughly in proportion

$k = \pi r^2 / A$

where r is the distance from the target and A is the area of a hectad (100 km^2). The smoothing surface will therefore be very similar to a bivariate Gaussian one.

The advantage of a more compact kernel such as K33 is that the same weight can be achieved with a smaller radius. The disadvantage is that in neighbourhoods with very low sampling density, smoothing will be less smooth, resulting in more erratic estimates of species probability. We saw in Test 1 that the size of smoothing neighbourhood can have a marked effect on estimates of species richness. This is a much larger effect than the more subtle difference between kernels K24 and K33, which has not been investigated here.

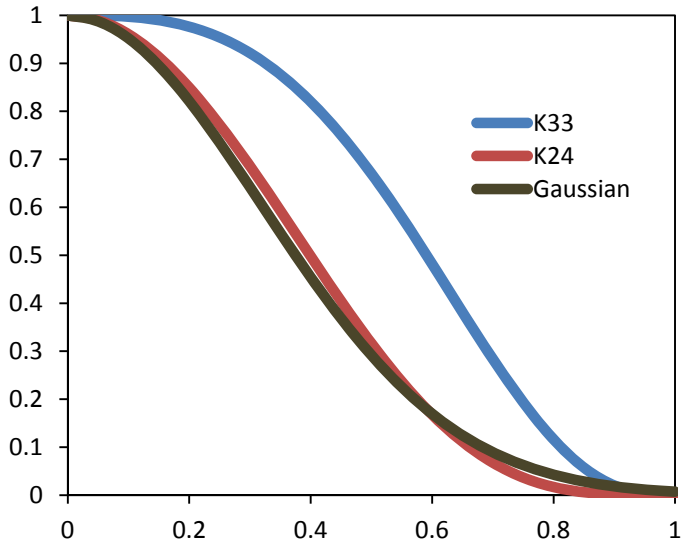

Figure S5. Two polynomial smoothing kernels K33 and K24 compared with a Gaussian kernel.