Appendix S1. Assessing the shape and slope of the species-area relationship from the species-occupancy distribution

Arnošt L. Šizling and David Storch

Center for Theoretical Study, Charles University, Jilská 1, 110 00-CZ Praha 1, Czech Republic,

sizling@cts.cuni.cz; storch@cts.cuni

Introduction

Species richness patterns are inevitably linked to the patterns of species spatial distribution as the number of species in a site is given by the number of species ranges that overlap there. However, these kinds of patterns have been studied mostly separately from each other, with only few attempts to make an explicit connection between them. The most prominent example of such interrelated patterns concerns the species-area relationship (SAR, i.e. the relationship between species number and area on which the number has been counted), and the frequency distribution of species occupancies (hereafter species-occupancy distribution). Although both patterns have been studied from the beginning of 20th century (Raunkiaer 1910, Arrhenius 1921), and although species relative occupancies apparently affect the slope of the SAR at least in the extreme cases (if all species occurred everywhere, the number of species would not increase with area, whereas if all species occupied only one site, mean species number would increase almost linearly with area), the exact connections between them have remained unexplored. The reason is that the formal theory connecting both patterns was either unrealistic (Ney-Nifle and Mangel 1999, Maurer 1999) or missing.

The SAR can be often well expressed as a power-law, which indicates scale invariance or self-similarity (Gisiger 2001). This has led to the formulation of a theory explicitly relating the power-law to the self-similarity at the community level (Harte et al. 1999). Although Harte et al. (2001) and Lennon et al. (2002) claimed that this is not compatible with the self-similarity revealed at the level of spatial distribution of individual species, Šizling and Storch (2004) have shown that within finite areas the power-law can be actually attributed to the self-similarity in individual species distributions, and that this effect is responsible for the slope and shape of the SAR in central European birds. Here we show that assuming the self-similarity of species spatial distributions, the slope and shape of the SAR can be derived using only the distribution of species relative occupancies.

Our following explorations are based on the *finite area model* of the SAR (Šizling and Storch 2004), which comes out from the knowledge that the mean number of species within an area can be calculated by summing species occupancy probabilities, p_{occ}, for area A. In self-similarly distributed species these probabilities increase approximately linearly with area in the log-log scale, up to the

point A_{sat} where $p_{occ}=1$. The A_{sat} represents the "area of saturation", i.e. the minimum area of a study plot that is necessarily occupied by the species, regardless of its location. The A_{sat} therefore depends on the area and shape of the largest distributional gap (see Figure 1 in Šizling and Storch 2004), and thus on the number of occupied sites and their spatial arrangement. Then the species number can be calculated according to the formula

$$\overline{S}[A]=\sum_{i=1}^{S_{tot}} p_{occ,i}[A]=\sum_{i=S_{sat}[A]+1}^{S_{tot}} \pi_i A^{z_i}+S_{sat}[A] \qquad \text{(S1)}$$

where $\overline{S}[A]$ is the mean number of species observed within a sample plot of area A randomly placed within the total area A_{tot} (i.e. the area of the whole study plot within which the sample plots can be laid), S_{tot} is the total number of species occurring within the A_{tot}, and $S_{sat}[A]$ is the number of species whose relationship between p_{occ} and A has reached saturation (i.e. the number of species with $A_{sat} \le A$). Parameters π_i and z_i correspond to the probability of occupancy in $A=1$ and to the rate of increase of p_{occ} with area, respectively.

According to the model (Figure S1), three parameters for each species spatial distribution (π, A_{sat} and z) are required to predict the resulting SAR. However, we will show that these parameters are so closely related to each other that the SAR can be ultimately predicted just by one of them.

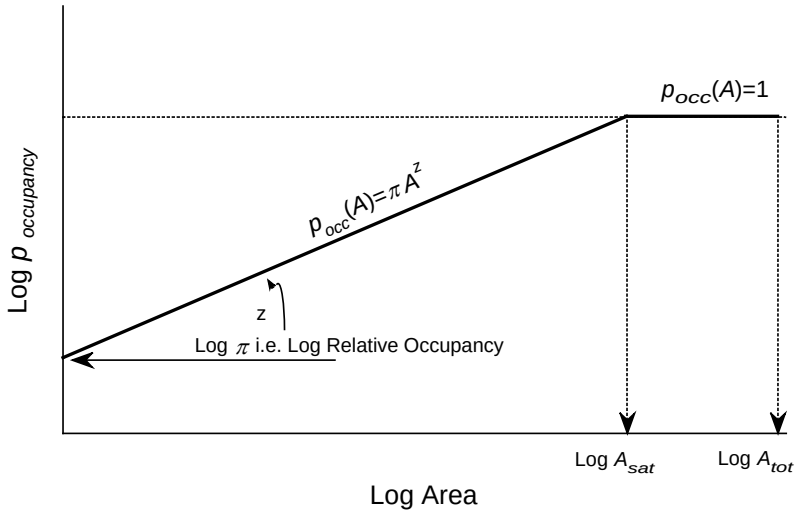

Fig S1: Graphical representation of the simple finite-area model. Log p_{occ} increase linearly with log *A*, up to the A_{sat} when $p_{occ}=1$. The slope *z* is determined by the A_{sat} and π (probability of occupancy of the unit area).

The interdependence between parameters of the finite area model

As A_{sat} and π represent two points defining a line, and *z* is the slope of the line (Figure S1), it is clear that one of the parameters is redundant. The relationship between them follows formula

$$z_i = -\ln \pi_i / \ln A_{sat\,i} \qquad \text{(S2)}$$

In the following text we will therefore deal only with the relationship between the parameters A_{sat} and π, since – assuming that the self-similarity is accurately captured by the finite area model - these are sufficient for characterizing species spatial distributions, and thus the resulting SAR.

The A_{sat} and π are not dependent on each other in a strict sense (as they would be if just one value of A_{sat} could be assigned to each π), but they constrain each other in the following way. Imagine a spatial distribution of a species represented by a lattice with some occupied cells (Figure S2). The π can be estimated as the proportion of the total number of cells occupied, and A_{sat} is given by the maximum possible gap, i.e. by the largest possible square that does not contain any occupied cell. The possible range of A_{sat} is therefore determined by the number and potential arrangements of *unoccupied* cells. The minimum and maximum possible A_{sat} (let us call them the *geometric constraints* of A_{sat}, $A_{sat\,MinG}$ and $A_{sat\,MaxG}$) can be calculated as follows:

Minimum possible A_{sat} can be obtained in the case of regular spatial distribution, simply because in that case changing a location of any occupied cell cannot make the A_{sat} smaller (Figure S2). As the shape of the sample plot is square, the size of minimum A_{sat} follows the formula

$$A_{sat} \geq A_{sat\,MinG} = \mathrm{Trunc}^2\left(\sqrt{A_{tot}} \Big/ \left(\mathrm{Trunc}\left(\sqrt{A_{occ}}\right)+1\right)\right) \qquad \text{(S3)}$$

where Trunc is the function that truncates an argument to the integer, A_{tot} is the total area (see Figure S2 where $A_{tot} = 5 \times 5$ grid cells), and A_{occ} is the occupied area ($A_{occ} = \pi A_{tot}$).

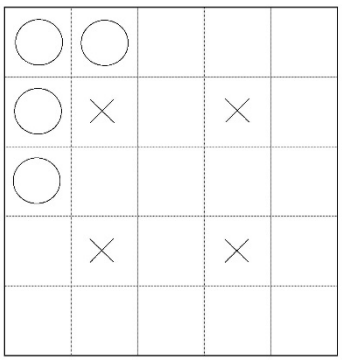

Fig S2: An example of the spatial distributions of a species occupying 4 cells within the grid of 25 cells (its relative occupancy is $4/25 = 0.16$). The circles represent the distribution with maximum possible A_{sat}, the sharps refer to the distribution with minimum possible A_{sat}.

Maximum possible A_{sat}, on the other hand, cannot be higher than the A_{unocc}, i.e. $A_{sat} \leq A_{tot} - A_{occ}$. The exact value of A_{sat} depends on the shape of the sample plot and on the spatial distribution of occupied cells within A_{tot}, and the highest possible A_{sat} is apparently reached when all the occupied cells are located along the edge of the A_{tot} (Figure S2). For the square-shaped sample plots we can then write

$$A_{sat} \leq A_{sat\,MaxG} = \mathrm{Trunc}^2\left(\sqrt{A_{tot} - A_{occ}}\right) \qquad \text{(S4)}$$

Note that for high π the interval $\left\lfloor A_{sat\,MinG}; A_{sat\,MaxG} \right\rfloor$ is quite narrow, as the dependence of A_{sat} on the location of occupied cells is relatively weak, whereas for small π the A_{sat} strongly depends on the location of occupied cells within the A_{tot}, and thus the interval of possible A_{sat} is relatively wide.

These constraints are generated by simple geometric logic and emerge without any consideration of internal spatial structure of species distribution. But both extreme structures (i.e. the regular distribution and the distribution confined to the edge of the sampled area) are apparently far from self-similar. Although these cases can be in fact considered as extreme realizations of random self-similar distribution (random fractals, see Hastings and Sugihara 1993), the probability of such realizations is very small. The A_{sat} for respective π will thus most probably lie within much narrower interval than that given by simple geometric constraints. Let us call these new probabilistic constraints, imposed upon A_{sat} due to the assumption of self-similarity, the *self-similar constraints*. The effect of variation of A_{sat} within them on the resulting SAR must be evaluated numerically.

Empirical evaluation of the sensitivity of SAR on possible A_{sat} variation

To evaluate the sensitivity of the SAR to the distribution of π and to the variation of A_{sat} between its two constraints, we have conducted a series of simulations. For the purpose of these simulations we used data on bird distribution in central Europe (Storch and Šizling 2002) from which the distribution of π was extracted (see Figure S3b). Then we calculated the constraints imposed upon A_{sat} by each π and tested how the resulting SAR is sensitive to the variation of A_{sat} within these constraints.

The data on the distribution of birds in central Europe comprise two scales of resolution, that of basic grid cell size of 11.1×12 km (Czech Republic, hereafter CR; Šťastný et al. 1996) and

a) **b)**

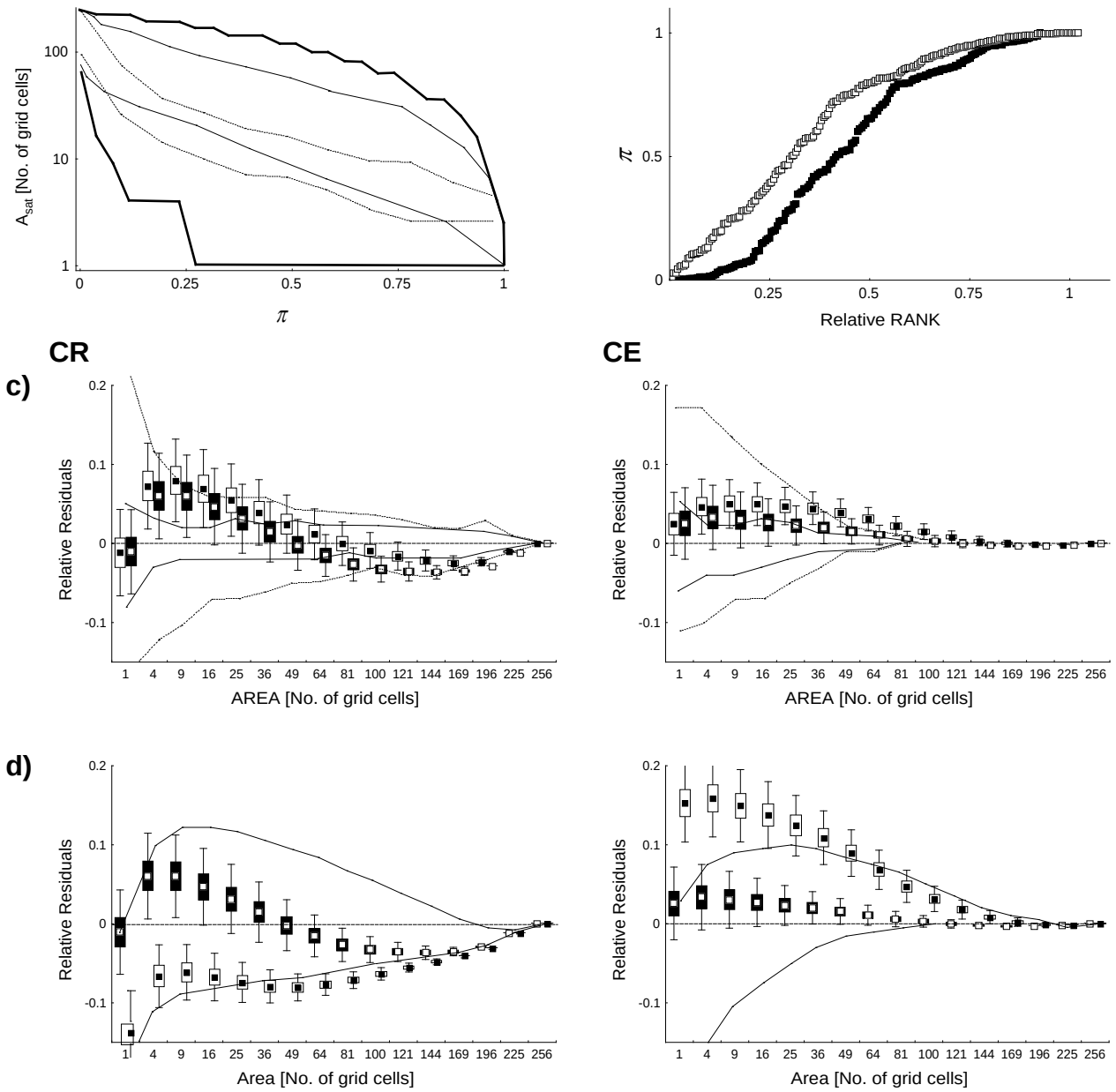

Fig S3: The settings and results of the tests concerning the sensitivity of the SAR on the distribution of π and the variation of A_{sat}. (a) The two types of constraints imposed on A_{sat} (thick line – geometric constraints, thin line – self-similar constraints), and the 95% confidence interval of A_{sat} for the case of random spatial distribution (dotted line). Note that all observed A_{sat} fell to the interval between the self-similar constraints, indicating that the real species distributions were indeed close to the self-similarity. (b) Distributions of π for the Czech Republic (black squares) and central Europe (white squares). (c) The relative residuals from the observed SAR for SARs constructed by the random drawing of A_{sat} from the interval given by the geometric constraints (white boxes) and self-similar constraints (black boxes) of A_{sat}. The dashed line refers to mean observed number of species, and dotted and full lines represent 95% and 50% confidence intervals of observed species numbers. The bias for sampled areas $\geq 10\times 10$ grid cells occurring in the case of CR is an

artefact of the fixation of A_{sat}, which diminished when we used different procedures of calculating *z*. (d) The comparison between the relative residuals obtained using the procedure described above (black boxes) and those that used inappropriate distributions of π for the prediction of species numbers (i.e. the distribution of π from CE was used for the prediction of the SAR for the CR and vice versa, white boxes). In both cases the self-similar constraints were used. Full lines refer to maximum and minimum species numbers obtained within geometric constraints of A_{sat} for appropriate distributions of π – note that using the inappropriate distributions (white boxes), which are only slightly different (Figure S3b), leads to predictions that occur outside of these hard boundaries.

---

that of basic grid cell size of 50×50 km (central Europe, hereafter CE; Hagemeijer and Blair 1997). Both data sets consist of 16×16 grid cells (see Figure 3 in Šizling and Storch 2004), containing the information about probable or confirmed breeding of each bird species within each cell (see Storch and Šizling 2002).

For each species, π_i was calculated as the intercept of the regression line of the relationship between log A_i and log $p_{occ,i}$, within the range in which the dependency $p_{occ,i}[A]$ was increasing. This line was fixed in the point of $A_{sat,i}$, so that the regression line had only one free parameter. The $A_{sat,i}$ was set as the middle point between the minimum square-shaped area which necessarily contained an occupied cell and the maximum empty square-shaped area.

The possible ranges of variation of A_{sat} for each π were constructed in two ways (Figure S3a):

1. **Geometric constraints** of A_{sat}, calculated using equations S3 and S4.
2. **Self-similar constraints.** Here we constructed self-similar distributions according to the procedure described in Šizling and Storch (2004; Appendix 2). We performed 500 simulations for the fractal dimension $FD=0.1$, then 500 simulations for $FD=0.2$, etc., up to $FD=1.9$ (note that $FD=2.0$ means that the species occupies the whole area). For each simulation we calculated π and A_{sat} as described above, and set the boundaries for A_{sat} as the 95% nonparametric confidence interval of the obtained results, i.e. the area within π - A_{sat} biplot that contained 95% of simulation results for each respective π. The confidence of the reliability of these intervals is higher than 99.9% ($\beta>0.95, \gamma<0.001$; Wilks 1941, Jílek 1988).

We then performed 500 simulations of the SAR, randomly varying A_{sat} within the constraints. In each simulation, (1) π was drawn from the respective distribution of π in number $N=S_{tot}$, (2) for

each π, A_{sat} was randomly drawn from the interval within the calculated boundaries, and (3) after calculating respective z_i for each pair of π_i and $A_{sat,i}$ (equation S2), mean species number $S_{estimated}$ was obtained using equation S1. This was performed for both types of constraints on A_{sat}. For the comparison of predicted and observed species numbers for each area we used relative residuals calculated as $\left(S_{observed} - S_{estimated}\right)/S_{tot}$. These residuals are equal to the mean of ε_i, $\bar{\varepsilon}$, used in the previous paper (Šizling and Storch 2004).

The residuals were low for all simulated SARs, for both CR and CE (Figure S3c). As expected, higher residuals were generally produced by the model where A_{sat} could vary more widely within the geometric constraints. However, even in this case the predicted species numbers did not differ from the observed numbers by more than 10% of S_{tot}. Note that the systematic deviation between observed and predicted species numbers for sampled areas $\leq 6\times 6$ grid cells has been shown to be attributable entirely to the approximative nature of equation S1, which does not represent an accurate expression of self-similarity for small areas (Šizling and Storch 2004).

On the other hand, the predicted species numbers were strongly dependent on the respective distribution of π. When we performed the same simulations as described above (using the self-similar constraints), but taking π from the other distribution (i.e. taking π from the distribution for CE in S_{tot} equal to the species number of CR and comparing the predicted species numbers with the observed numbers for CR, and vice versa), the deviations between predicted and observed species numbers were much higher than the deviations calculated from the appropriate distribution of π (Figure S3d). They were even higher than the maximum deviations that would be obtained if all species had the extreme spatial arrangement of occupied cells, i.e. the regularly distributed cells and cells located along the edge of the A_{tot}.

These results indicate that the SAR is not substantially sensitive to the variation of A_{sat} within the constraints imposed on it by the distribution of π, but are very sensitive to the exact distribution of π. Relative species occupancies therefore directly affect the shape and slope of the SAR.

Relationship between the species-occupancy distribution and slope and shape of the SAR

According to Šizling and Storch (2004), the slope of the SAR in logarithmic space can be calculated using equation

$$Z = \ln\left(S_{tot}/\sum \pi_i\right)/\ln\left(A_{tot}\right) \qquad \text{(S5)}$$

where A_{tot} is the total number of grid cells, which refers to the coarseness of the grid. Therefore, for given A_{tot}, *z* is determined by the mean value of the species relative occupancies $\bar{\pi}$ ($\bar{\pi} = \sum \pi_i / S_{tot}$), so that $z = -\ln(\bar{\pi})/\ln(A_{tot})$. Thus, the higher the mean species relative occupancy, the lower the slope of the SAR, bounded at $z = 0$ when mean species relative occupancy is equal to 1. However, the SAR is not necessarily precisely linear on the log-log scale, and for highly nonlinear cases it does not make sense to take *z* as a reliable descriptor of the SAR. It is thus necessary to explore also the effect of the distribution of π on the shape of the SAR.

For this purpose we generated 10 000 random distributions of π ($S_{tot} = 200$), constructed as rank-π relationships expressed by random third-order polynomials (i.e. three kinds of distributions - regular, unimodal, and bimodal - were allowed), keeping mean π per species such that *z* = 0.2. For each distribution we calculated standard deviation, skewness and kurtosis (which correspond to the second, third, and fourth central moments of the distributions) and constructed the SAR according to the procedure described above (with the self-similar constraints of A_{sat}). Then we analysed the effects of these parameters on the curvilinearity of the SAR (hereafter *CL*). The *CL* was calculated using the sum of squares of distances from the line defined by the two extreme points of the SAR (the maximum ($A = A_{tot}; \bar{S} = S_{tot}$) and minimum ($A = 1; \bar{S} = \sum \pi_i$); the slope of the line is equal to *Z* (equation S5)). The squared distances were calculated for all points of A_{sat} in the log-log space and then averaged (Šizling and Storch 2004).

The *CL* depends strongly and negatively on the standard deviation of π ($r = -0.87$, $p < 0.0001$, see inset in Figure S4). The other parameters also have significant, but smaller, effects on *CL* ($r = -0.55$ for skewness and 0.30 for kurtosis; $p < 0.0001$ for both variables). These effects imply that the SAR is closer to the power-law in the case of bimodal (which leads to increasing standard deviation and decreasing kurtosis) and/or right-skewed (increasing both standard deviation and skewness) distributions (Figure S4). Note that the distribution of π is bounded by zero and one, and so the standard deviation cannot be elevated by a simple increase in the range of values, but only by increasing the right-skew or bimodality.

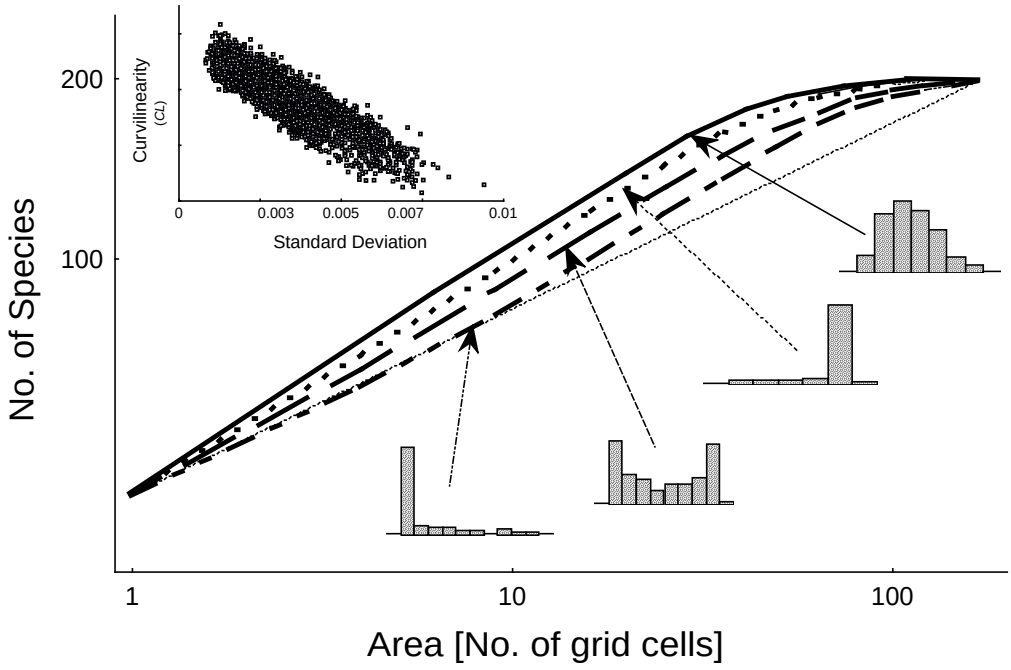

Fig S4: The relationship between different types of species-occupancy distribution and the curvilinearity of the SAR (see text for the details of the construction). The strongly right-skewed and bimodal distributions have larger standard deviations and produce SARs which are very close to the power-law. The resulting SARs were obtained using the mean of 500 simulations for each distribution. The inset shows the relationship between the standard deviation of π and the curvilinearity of the SAR, *CL* (N=10,000).

Discussion

The species-area relationship is strongly sensitive to the distribution of species relative occupancies, whereas its sensitivity to particular spatial structure of species distribution is much lower. Note, however, that we have shown this only for self-similar spatial distributions, because this is the only case in which the SAR can be predicted using the finite area model (Šizling and Storch 2004; Equation S1). Therefore, our results do not mean that the SAR is directly dependent on the species-occupancy distribution regardless on the spatial structure; they say simply that the shape and slope of the SAR are not dependent on the particular realizations of self-similar spatial distribution given that the distribution of species relative occupancies does not change.

The shape of the SAR is close to the power-law if the species occupancy distribution is either bimodal or strongly right-skewed. These types of occupancy distribution are actually those most commonly observed in nature (Hanski 1982, Gaston and Blackburn 2000, Storch and Šizling 2002) and thus it is not surprising that the SAR is also commonly expressed as a power-law. However, the species-occupancy distribution is not scale invariant – if we considered very fine resolution where the size of the basic grid cell was comparable to the average home range of individuals of the given

taxon, the occupancy distribution would be close to the distribution of species abundances which is unimodal, albeit still left-skewed (Preston 1962, May 1975, Hubbell 2001). This could potentially affect the shape of the SAR on small scales. Indeed, there is some evidence that at very small scales the slope of the SAR changes (Crawley and Harral 2001, Hubbell 2001), and the SAR becomes curvilinear in a log-log space (Harte et al. 2009). It is therefore probable that our model works only within particular spatial scales. Only over this range of scales will the assumption of self-similarity be valid, allowing the derivation of the SAR from the species-occupancy distribution. We have evidence that for birds these scales comprise grids of cells larger than ca 10×10 km, but it is probable that this scale will differ among different taxa. Thus the shape of the SAR may be taxon-dependent (Crawley and Harral 2001, Marquet et al. 2004).

Until now we have dealt with purely geometric considerations, showing that the shape and slope of the SAR are related to the distribution of species relative occupancies. This finding implies that if we want to explain the shape and slope of the SAR in terms of the mechanisms producing it, we have to look for the processes generating also the species-occupancy distribution. It is not a coincidence that the same processes have been proposed as explanations for both patterns. We can distinguish three major groups of explanations for both patterns: (1) sampling effect (the result of random location of individuals across space according to the distribution of species abundances, see Preston 1960, Coleman 1981, Nee et al. 1991), (2) habitat heterogeneity (the effect of the spatial distribution of habitats preferred by individual species, see Rosenzweig 1995, Storch and Šizling 2002) and (3) spatial population dynamics which leads to spatial aggregation not attributable solely to habitat aggregation (Hanski and Gyllenberg 1997, Storch and Šizling 2002, Storch et al. 2003). In the case of central European birds we have already shown that neither species-occupancy distributions (Storch and Šizling 2002) nor the SAR (Storch et al. 2003) can be attributed only to sampling effect or habitat heterogeneity, and that spatial aggregation is significantly higher than expected solely from these effects.

Regardless on the relative contribution of the effects of habitat heterogeneity and spatial population dynamics, the ultimate cause of the highly unequal occupancy distribution as well as the shape and slope of the SAR is spatial aggregation on various scales. This is in accord with previous findings concerning the importance of spatial aggregation for diversity patterns (Plotkin et al. 2000, He and Legendre 2002). Our approach extends these notions by explicitly relating these effects to the observed patterns of self-similarity of species distribution and the power-law approximation of the SAR. However, two questions remain open: (1) what generates the self-similarity, i.e. the similar pattern of spatial aggregation on various scales of resolution (Storch et al. 2008), and (2) which processes affect mean species occupancies responsible for the slope of the SAR. Regardless of the

responsible processes, the species-occupancy distribution and the species-area relationship are ultimately caused by the same biological phenomenon, the spatial aggregation within many spatial scales.

Acknowledgement

We thank Ethan White, Jiří Reif, Marco Patausso and Kevin Gaston for helpful comments.

Literature Cited

Arrhenius, O. 1921. Species and area. Journal of Ecology **9**: 95-99.

Coleman, D. B. 1981. On random placement and species-area relations. Mathematical Biosciences **54**: 191-215.

Crawley, M. J., and J. E. Harral. 2001. Scale dependence in plant biodiversity. Science **291**: 864-868.

Gaston, K. J., and T. M. Blackburn. 2000. Pattern and Process in Macroecology. Blackwell Science, Oxford.

Gisiger, T. 2001. Scale invariance in biology: coincidence or footprint of a universal mechanism? Biological Reviews **76**: 161-209.

Hagemeijer, W. J. M., and M. J. Blair. 1997. The EBCC Atlas of European Breeding Birds. T. & A.D. Poyser, London.

Hanski, I., and M. Gyllenberg. 1997. Uniting two general patterns in the distribution of species. Science **275**: 397-400.

Hanski, I. 1982. Dynamics of regional distribution: the core and satellite species hypothesis. Oikos **38**: 210-221.

Harte, J., T. Blackburn, and A. Ostling. 2001. Self-similarity and the relationship between abundance and range size. American Naturalist **157**: 374-386.

Harte, J., A. Kinzig, and J. Green. 1999. Self-Similarity in the Distribution and Abundance of Species. Science **284**: 334-336.

Harte, J., B. Smith, and D. Storch. 2009. Biodiversity scales from plots to biomes with a universal species–area curve. Ecology Letters **12**: 789–797.

Hastings, H. M., and G. Sugihara. 1993. Fractals, a User's Guide for the Natural Sciences. Oxford University Press, Oxford.

He, FL., and P. Legendre. 2002. Species diversity patterns derived from species-area models. Ecology **85**: 1185-1198.

Hubbell, S. P. 2001. A Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press, Princeton, NJ.

Jílek, M. 1988. Statistical and Tolerance Limits. SNTL, Praha, (in Czech).

Lennon, J. J., W. E. Kunin, and S. Hartley. 2002. Fractal species distributions do not produce power-law species area distribution. Oikos **97**: 378-386.

Marquet, P.A., M. Fernández, S. A. Navarrete, and C. Valdovinos. 2004. Diversity emerging: Towards a deconstruction of biodiversity patterns. *in* M. Lomolino and L. R. Heaney, editors. New directions in the geography of nature. Frontiers of Biogeography, Cambridge University Press, Cambridge.

Maurer, B. A. 1999. Untangling ecological complexity: The macroscopic perspective. University of Chicago Press, Chicago.

Nee, S., R. D. Gregory, and R. M. May. 1991. Core and satellite species: theory and artefacts. Oikos **62**: 83-87.

Ney-Nifle, M. and M. Mangel. 1999. Species-area curves based on geographic range and occupancy. Journal of Theoretical Biology **196**: 327-342.

Plotkin, J. B., M. D. Potts, N. Leslie, N. Manokaran, J. LaFrankie, and P. S. Ashton. 2000. Species-area curves, spatial aggregation, and habitat specialization in tropical forests. Journal of Theoretical Biology **207**: 81-89.

Preston, F. W. 1960. Time and space and the variation of species. Ecology **29**: 254-283.

Preston, F. W. 1962. The canonical distribution of commonness and rarity. Ecology **43**: 185-215, 410-432.

Raunkiaer, C. 1910. Investigations and statistics of plant formations. Botanisk Tidsskrift 30.

Rosenzweig, M. L. 1995. Species Diversity in Space and Time. Cambridge University Press, Cambridge.

Šizling, A. L., and D. Storch. 2004. Power-law species-area relationships and self-similar species distributions within finite areas. Ecology Letters **7**: 60-68.

Šťastný, K., V. Bejček, and K. Hudec. 1996. Atlas of Breeding Bird Distribution in the Czech Republic 1985-1989. Nakladatelství a vydavatelství H&H, Jinočany, (in Czech).

Storch, D., and A. L. Šizling. 2002. Patterns in commoness and rarity in central European birds: Reliability of the core-satellite hypothesis. Ecography **25**: 405-416.

Storch, D., A. L. Šizling, and K. J. Gaston. 2003. Geometry of the species-area relationship in central European birds: testing the mechanism. Journal of Animal Ecology **72**: 509-519.

Storch, D., A. L. Šizling, J. Reif, J. Polechová, E. Šizlingová, and K. J. Gaston. 2008. The quest for a null model for macroecological patterns: geometry of species distributions at multiple spatial scales. Ecology Letters **11**: 771–784.

Wilks, S. S. 1941. Determination of sample size for setting tolerance limits. Annals of Mathematical Statistics **12**: 91-96.