**The Statistical Machinery of Alien Species**

M G Bowler    Department of Physics, University of Oxford  
               Keble Road, Oxford OX1 3RH, UK (m.bowler1@physics.ox.ac.uk)

C K Kelly     Department of Zoology, University of Oxford  
               South Parks Road, Oxford OX1 3PS, UK (colleen.kelly@zoo.ox.ac.uk)

**Abstract.** Many species of plants are found in regions to which they are alien and their global distribution has been found to exhibit several remarkable patterns, characterised by exponential functions of the kind that could arise through versions of MacArthur's broken stick. We show here that these various patterns are all quantitatively reproduced by a simple algorithm, in terms of a single parameter – a single stick to be broken. This algorithm admits a biological interpretation in terms of niche structures fluctuating with time and productivity; with sites and species highly idiosyncratic. Technically, this is an application of statistical mechanics to ecology quite different from the familiar application to species abundance distributions.

**Keywords** naturalized species, niche dynamics, statistical mechanics, community assembly, idiosyncracy.

1. **Introduction**

Large numbers of species are found naturalised in geographical sites remote from the area to which they are native. This paper is concerned with some remarkable observations of regularities in the distribution over the globe of species alien to the locations in which they are found. The principal observations are that the number of species found at $n$ sites falls off exponentially with $n$ and that the number of pairs of sites sharing $n$ species alien to both is also an exponential function of that $n$. These striking patterns had not, as far as we know, been observed before and the origin of these exponential distributions and the possible relationship between them seems to us to be of considerable interest and of relevance to the rules of community assembly. A tentative hypothesis for the machinery generating the exponential distribution of species over sites was advanced in the paper presenting the data, Kelly et al (2011) but the distribution of pairs and triplets of sites as a function of the number of species shared remained somewhat mysterious. Here we address in much greater detail the possible machinery for generating the exponential distribution of species over sites and show that an extension of this picture accounts for all of the exponentials encountered in the data. Given the number of sites and the number of alien species there found, an excellent description is generated by a simple algorithm in terms of a single free parameter. The existence of such a simple and precise description is of interest in itself, even apart from the biological interpretation of the algorithm that we propose.
The available data, from 16 geographical sites (on the scale of nation states) were collated in Kelly et al (2011) and comprised 5350 plant species found at sites to which they are alien with 10409 establishments; thus on average an alien species is found at \( \sim 2 \) different sites. It should be emphasised that these species are better described as 'naturalised' than by the word 'invasive' – very few alien species are invasive in the sense of 'pestiferous'. Analysis of this substantial sample revealed the remarkable features to which we alluded briefly above. The first is that the number of species found alien at \( n \) sites is, for \( n > 1 \), exponentially distributed with

\[
S(n) = S_0 \exp(-\beta n)
\]  

(1)

where the parameter \( \beta \) has a value of 0.52. Thus there are 873 species found alien to two sites (such as Wyoming and New Zealand) and 43 species alien to 8 sites and there found. No species was found at more than 13 sites. The exponential is a very particular function and while it is no surprise to find very many species at one or two sites to which they are alien and fewer at many, an exponential distribution must reflect some particular mechanism of community assembly - and there are other exponential distributions encountered in these data.

These concern pairs and triplets of sites and the number of alien species that they share – for example, Wyoming and New Zealand have in common 164 species alien to both. The number of site pairs sharing \( p \) species is exponentially distributed in \( p \). Similarly the number of triplets of sites sharing \( t \) alien species is exponential in \( t \) (and the pattern repeats even for quartets). These correlations depend not only on the distribution of species over sites but also on the distribution of sites over species. There are 560 triplets of sites and 120 pairs; the exponential distributions are well defined. That their origin is related to that of the exponential \( S(n) \) is plausible, but not obvious.

For only 16 sites, the distribution of the number of sites \( M(s) \) at which \( s \) species are found cannot look anything like a continuous distribution. An ideogram (Fig. 2, lower panel) shows individual site occupancy widely spread, but tending to cluster at low occupancy; these data are consistent with having been drawn from an underlying exponential distribution (Kelly et al 2011).

The final remarkable observation contained in the data of Kelly et al (2011) is that the number of species shared pairwise has no relationship to the distance on the surface of the globe separating the sites. In Fig. 2 of Kelly et al (2011) each pair of sites is plotted as a point in the plane of number of species in common versus geographical separation. The distribution of points with respect to the distance between site pairs is strongly clustered; the distribution of points with respect to the number of species shared falls off with that number. This fall off is exponential, as stated above, Fig. 3A of Kelly et al (2011). Yet there is no correlation between the number of naturalised species common to a pair of sites and the separation of those sites – pairs with comparatively few species in common are found in all geographical clusters and pairs
with very many species in common are found at very large separations as frequently as at small separations. (The lack of any significant correlation was established quantitatively with a Mantel test.) As argued in detail in Kelly et al (2011), the lack of dependence of the number of species shared on the distance separating members of the paired sites rules out any attempt to interpret the exponential falloff in the number of species at \( n \) sites as an exponential attenuation with distance; alien species must be around everywhere all the time. The corollary is that the exponential form must be associated with the availability of sites for alien species and the explanation must be sought in such terms. The exponential \( S(n) \) might suggest to the ecologist apportionment of suitable resource for alien species according to MacArthur's broken stick (MacArthur 1960), where a fixed resource sufficient to support a specified number of alien establishments is divided at random among a fixed number of species. To the physicist that exponential suggests a version of the microcanonical ensemble in statistical mechanics, where a conserved quantity (such as energy for a simple gas) is distributed over a fixed number of agents. For this problem, the fixed resource (or conserved quantity) is the total number of alien establishments (the alien footprint of Kelly et al 2011) and the agents are the alien species. These two suggestions are equivalent and this is the hypothetical explanation for the exponential distribution of the number of species found at \( n \) sites put forward in Kelly et al (2011). It left unexplained the exponential multiplet correlations.

2. Indicators of biological machinery

Having conjectured that the combination of an exponential distribution \( S(n) \) and an exponential singlet distribution \( M(s) \) might generate the exponentially distributed multiplets, we considered processes, in the context of statistical mechanics, that result in these two exponentials and how they might be combined. We sought an algorithm yielding the contents of every cell in a matrix \( N(J, K) \) in which rows represent the 16 sites (from Chile to Wyoming) and columns the many thousands of alien species involved (species 5350 is \( Zygophyllum Fabago \)). From such a matrix the pair and other multiplet distributions can be constructed. An algorithm containing the observed multiplet distributions must exist, for Nature has found it, but Nature's algorithms are not always easy to figure out and not always susceptible to interpretation. Two aspects of the data led to construction of a niche based picture and an algorithm which yields a species-site population matrix in excellent agreement in all respects with the data set. The first is the complete lack of dependence of the number of species shared on the separation over the globe of the sites in a pair. The second aspect is the possibility of interpreting the exponential distributions \( S(n) \) and \( M(s) \) in terms of niche structures opening and closing.

If the system is dynamical (like the internal workings of a gas and suggested by the spread of alien species over the globe) the dynamics of such a system coming to an exponential equilibrium distribution can be modelled with a simple master equation
(see e.g. Volkov et al 2003, Bowler & Kelly 2012, Bowler 2014)

\[
\frac{dS(n)}{dt} = -(b+d)S(n) + bS(n-1) + dS(n+1)
\]  

(2)

Here, \( S(n) \) is the number of species found at \( n \) sites. The parameters \( d \) and \( b \) are rates at which a species vanishes from a site or appears at a site from which it was previously absent. The content of Eq.(2) is that if a species is present at \( n \) sites \( S(n) \) is reduced by 1 if that species vanishes from one site or if that species appears at one additional site. Similarly, \( S(n) \) is augmented by 1 by adding to a new site a species present at \( n-1 \) sites, or by losing from a site a species present at \( n+1 \) sites. The equation evolves to a steady state at which

\[
S(n+1) = \frac{b}{d} S(n)
\]

Iterating this equation the solution is, apart from a normalising constant

\[
S(n) = \left(\frac{b}{d}\right)^n
\]

(3)

which is an exponential in \( n \), a negative exponential if \( b < d \). If a species is at \( n \) sites, it is removed from one of those sites a little more often than it is added to a new site. More generally, an exponential results if the ratio of \( b \) for \( n \) sites to \( d \) for \( n+1 \) sites is independent of \( n \). The most economical (and possibly least implausible) way of achieving this is to have the parameters \( b \) and \( d \) not depend on \( n \); they are not per capita rates. It seems obvious that the chance of adding a species to a new site is independent of at how many sites it is already found. Suitable niches open and close; filled niches do not give birth to offspring and do not of themselves fill other sites. In contrast, \( d \) independent of \( n \) seems counter intuitive. For the moment, we ask the reader to live with this successful assumption. The matter is discussed in detail in the Appendix.

The content of Eqs. (2) and (3) can be simulated very simply for the purpose of obtaining the distribution of species over sites represented by the matrix \( N(J,K) \) introduced in section 2. If species \( J \) is present at site \( K \) then that element of the matrix is 1, otherwise 0. Apply the following operations to this matrix- it is perfectly reasonable to start with all elements zero. Choose at random a species \( J \). Then choose at random whether to open a new site for this species or remove one of the sites already filled, the ratio of choices being \( \frac{b}{d} \). If the choice is for putting \( J \) into a site it does not already occupy, choose one of the empty sites and change that element of the matrix from 0 to 1. If on the other hand the lot was cast for emptying a site, choose one of the elements for \( J \) with occupancy 1 and change it to 0. It does not matter by what recipe the empty site to be filled is chosen, nor the full site to be emptied. Repeat this operation a sufficiently large number of times for the
equilibrium configuration to emerge and then for each species count the number of filled \( K \) elements (sites). Each species can then be counted into the appropriate bin to yield the distribution \( S(n) \). It is an exponential and the exponent parameter \( \beta \) is given by

\[
\beta = -\ln \left( \frac{b}{d} \right)
\]  

Thus this picture of alien species available for suitable niches, which open and close at rates independent of the total niche space already filled, straightforwardly accounts for the exponential distribution of the number of species as a function of the number of sites (as does breaking MacArthur's stick). The results of a model calculation (in fact, from the first run of the final code) are shown in Fig.1, which may be compared with Fig.1 of Kelly et al (2011). Simulation was not needed to obtain this result, but this is only the \( S(n) \) part of the problem; there is the second dimension concerning \( M(s) \) and for this simulation is required to generate \( N(J, K) \).

![Figure 1](image_url)  

Fig.1 The model distribution of species over the number of sites at which they are found. The distribution is exponential, generated by the weighted species algorithm.
3. Complications in two dimensions

Table 1 shows a portion (a small portion) of the matrix $N(J, K)$ from which the model results in Figs. 1 – 3 of this paper were drawn.

$$
\begin{array}{cccccccccccccccc}
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}
$$

Table 1 A portion of the model calculation of $N(J,K)$. Each row gives the status of a single species at each of 16 sites. In the full matrix there are thousands of species and hence thousands of rows.
The number of sites at which a species is present is found by counting horizontally along the appropriate row; thus the first species in the list is found at one site only but the second is found at two. The number of species found at any given site is obtained by counting vertically. It is immediately clear that although an exponential distribution \( S(n) \) results from opening or closing as appropriate any of the sites for a given species (a species algorithm), regardless of how selected, the distribution of species over sites is determined by how that selection is made. Thus if a choice is made at random among the filled sites as to which to delete and similarly a random choice is made of which empty site to fill, then all sites are being treated in the same way and the distribution of occupancy will be approximately normal about the mean (which is approximately 650 species per site). This is very unlike the data where the singlets are consistent with being drawn from an exponential distribution, extending out to over 1600 species at a site. Thus in a species algorithm, selection of the next site to open or close must be made according to a recipe that will yield a singlet distribution consistent with exponential.

4. Distribution of site species populations

If the matrix \( N(J, K) \) is addressed differently, there is an obvious way of generating an exponential singlet distribution. Choose at random any site (out of only 16) and then either add a species not already present or, slightly more often, delete a species at that site. Regardless of how a species is chosen, an exponential distribution (more accurately, a set consistent with having been drawn from an exponential distribution) will result (a site algorithm). However, in looking for both an exponential distribution of \( S(n) \) and for an exponential distribution of singlets \( M(s) \) with occupation number \( s \), elements of \( N(J, K) \) are being changed and the question is how to do this consistently, working both horizontally (species algorithm) and vertically (site algorithm). For the horizontal approach yielding an exponential \( S(n) \) the ratio \( b/d \) needed for an exponential matching the data is \( \sim 0.6 \), yet to generate a singlet distribution with a mean of \( \sim 650 \) (as observed) requires a ratio \( b/d \sim 0.9985 \). The length of the stick to be broken, that is, resource to be partitioned, is in both cases the sum of elements in the matrix, the alien footprint. Nonetheless, the exponential singlets distribution can be made consistent with the mean number of sites per species and yet be attributed to opening and closing of niche structures in essentially the same way as \( S(n) \).

We envisage a site as having a degree of receptivity to alien species (rather than specific niches) and that receptivity fluctuating with time. It might correspond to space for a certain number of alien species, that number increasing or decreasing by amounts independent of the number itself. This could be described by a master equation for \( M(s) \) of the same type as (2) and with a suitable ratio of the frequency of increasing space to decreasing space yields outputs consistent with being drawn from an exponential with a mean \( \sim 650 \). Site populations from a single run of the site algorithm can only be displayed as an ideogram with structure consistent with being
drawn from an exponential, but the sum over many separate runs turns into a
histogram of clearly exponential nature, realising the underlying probability
distribution. We suppose, then, that alien species are available to colonize sites whose
receptivity has evolved to an exponential probability distribution. The recipe for
picking sites from which to remove a species or add a new species must reflect this
underlying receptivity, in the form of weights.

One way is to generate a set of individual occupations with a single run of a site
algorithm. Examples of two independent runs are in Table 2,

|   |    |    |    |
|---|----|----|----|
|1  |  97 | 15 |  256 | 10 |
|2  | 782 |  6 | 1165 |  3 |
|3  | 468 |  8 |  435 |  5 |
|4  |1526 |  3 |  349 |  7 |
|5  | 133 | 13 |  873 |  4 |
|6  | 859 |  5 |  427 |  6 |
|7  |1636 |  1 |  163 | 13 |
|8  | 187 |12 | 2874 |  1 |
|9  | 767 |  7 | 1781 |  2 |
|10 |1136 |  4 |  168 |12 |
|11 |  75 |16 |  220 |11 |
|12 |1601 |  2 |  307 |  9 |
|13 | 106 |14 |   16 |16 |
|14 | 467 |  9 |  328 |  8 |
|15 | 242 |11 |   40 |14 |
|16 | 357 |10 |   35 |15 |

Table 2. Results from two independent runs of the site algorithm. \( \{R\} \) denotes site rank.

where the left hand column specifies the site \( K \) and the right hand columns the
number of potential species accommodated. These sets of numbers give the relative
receptivity of each site. The label \( K \) has no significance, but ordering the sites by
receptivity creates a rank \( \{R\} \). Separate runs give sets of numbers drawn from the
same underlying exponential distribution. (If all that is wanted from the site algorithm
is a set of numbers of species at each site, the individual species can be ignored and
the site algorithm collapsed to the one dimension \( K \), computationally much more
efficient.) Table 2 indicates how much scatter there is between runs; we have chosen
rather to present results obtained using averaged weights determined by the
underlying exponential probability distribution.

5. A method of averaged weighting

Fig. 2 shows ideograms of the singlet populations for both the data and the output of
our final algorithm. They are obviously consistent with each other. Since a tick mark
is made for each individual population number, the exponential nature of the
underlying distributions is manifest in ticks being denser at the low end of the occupancy axis. Similar distributions are obtained for separate runs of the singlet algorithm (for example, ideograms drawn from Table 2) and many runs of the algorithm for singlet populations can be summed into histogram bins and accumulated define an exponential.

![Figure 2](image_url)

Fig. 2 The arrows indicate the number of species at each of 16 sites, covering a range from ~ 100 to ~ 1600. The upper panel is the model simulation, the lower panel the data of Kelly et al (2011).

If the site algorithm numbers are ordered by rank, the averages of \( s_R \) over many runs can be calculated for each \( R \). Thus for the most receptive sites \( s_1 \sim 2000 \) but for the least \( s_{16} \sim 50 \) (see Table 2). In fact, as samples accumulate it becomes clear that the means cluster ever closer to a straight line with \( s_R \) proportional to \( \ln R_0 - \ln R \), where the constant \( R_0 \) is, for 16 sites, close to 17. An exact calculation of these averages can be made analytically from the underlying exponential probability distribution and shows that the above relation is quite sufficiently accurate for our purposes. The relative weight (or receptivity) of a site of rank \( R \) is therefore taken as \( \ln R_0 - \ln R \). These averaged rank ordered receptivities were used in weighting the rank ordered sites when applying the species algorithm to produce our final results shown in the figures.

To implement the species algorithm with the sites appropriately weighted, the procedure is straightforward. If a chosen species is to be removed, choose one of its occupied sites at random and delete the species. If on the other hand a species is to be
added to a site, choose among the (ordered) sites without this species according to their relative weights (receptivities) \( \ln R_0 - \ln R \). (Thus the site with \( R = 15 \) seldom gains a new species.) This implements the notion of some sites having wandered over aeons to a state more receptive of alien species than others.

When the species algorithm is run according to this weighted recipe the output of a single run is in admirable agreement with the data, from which the value of the alien footprint for \( n > 1 \) and the 2049 such species determines the ratio \( b/d \) (0.6) for the species algorithm. The simulation then yields a distribution \( S(n) \) with a mean of about 2 sites per species. The same number also determines for the 16 sites the value of \( b/d \) for the site algorithm, a mean of \(~ 650\) species per site (Fig. 2). The full matrix \( N(J, K) \) thus generated was interrogated to yield the number of site pairs sharing species as a function of the number of species shared and the number of triplets as a function of the number of species common to three sites at which they are alien. The simulation reveals that these shapes are indeed exponential; Fig. 3 cannot easily be distinguished from the pairs and triplets shown in Fig. 3 of Kelly et al (2011).

The results from our simulation shown are for the very first run after the code for making the multiplet interrogations was added. The only parameter that had been fine tuned was the ratio \( b/d \) for adding or subtracting a site to a species. Nonetheless, this first run makes all the important points. The distribution \( S(n) \) is exponential (unavoidable) and the pair and triplet distributions are also exponential. The

![Fig. 3 Left panel: The number of site pairs as a function of the number of shared species. Right panel: The number of triplets as a function of the number of species shared. Both are exponential, generated by the weighted species algorithm. To be compared with Fig.3 of Kelly et al (2011).](image-url)
parameters of these exponentials are then given; the exponential curves shown in Fig. 3 are not fits to the data, but were calculated from the distribution \( S(n) \), just as in Kelly et al (2011). In this first run, the number of alien species on the loose was set rather arbitrarily to 7000. This resulted in approximately 4200 successfully naturalised species, while the data of Kelly et al (2011) contain approximately 3400 species under the exponential in Fig. 1 of that paper. This difference has the obvious effect on Fig.1 and the pair and triplet exponentials in Fig. 3 are marginally less steep than those in Kelly et al (2011). It would take a hawk eyed reader to perceive these differences, fixed if desired by changing to somewhere between 5000 and 6000 alien species looking for homes.

6. Variations on the algorithm

Forming the weights from a single run of the site algorithm shows the same features as described in section 5, but the output is noisier and anyway it is not known what single run of the site algorithm is most representative of the real world. Nonetheless, if relative weights are taken from, for example, either of the runs listed in Table 2, the resulting multiplet distributions are in good accord with Fig. 3 and with the observations. The alien footprint for the successfully naturalised species and 16 sites is the single parameter in the model, unifying the various exponential distributions found by Kelly et al (2011). We finally note that in the above discussion we have tacitly assumed that the distribution of site receptivities has settled down before the alien species are unleashed. While this seems a natural assumption to make, it is not necessary and our conceptual structure is more general. The site algorithm (operating only in the \( K \) dimension) can be embedded in the weighted species algorithm and (as an example) a site capacity updated each time a species has its complement of sites updated. The distributions nonetheless emerge in agreement with Figs. 1-3 (but noisier).

7. Conclusions

The lowest level conclusion is that there is indeed a simple algorithm - the weighted species algorithm - that generates an exponential distribution of species over the number of sites at which they are present and simultaneously species populations of sites drawn from an exponential distribution. This same algorithm yields exponential distributions for pairs and triplets of sites, as shown in Fig. 3 above (c.f. Fig. 3 of Kelly et al (2011)). This algorithm might be regarded merely as a recipe for generating distributions of alien species according to a twofold example of MacArthur's broken stick, and so it is. A resource (represented by the alien footprint) is partitioned among 16 sites and then partitioned again among over ~5000 species alien to those sites. MacArthur's broken stick failed in its original application to the very different problem of species abundance, but it is realised in abundance of species - the global distribution of alien species. The algorithm also bears a biological interpretation in a dynamical picture. Here the mathematics is that of stochastic processes, master equations for birth/death problems but with the analogues of birth and death rates
absolute and not *per capita*. A species can gain or lose a site at rates independent of the number of sites at which it is present (see Appendix). The underlying picture is of niches either potential or specific opening and closing over time and thus sorting out the distribution of these alien species. More detailed discussion of the relevant mathematics may be found in Bowler & Kelly (2012) and Bowler (2014).

The real world is undoubtedly more complex but this work has established a simple theoretical foundation unifying the separate observations to be found in Kelly et al (2011), much as conjectured therein. The inferences about the biology of alien species and sites in that work are thus strongly supported. Perhaps the most important inference comes from the mere applicability of elementary statistical mechanics to this problem. In the model, no site is special and it is as a result of random processes that it reaches a particular capacity for alien species. Similarly, no species is special and it reaches a particular number of sites through random processes. These equivalences among sites and among species might come about through genuine identity, known as neutrality in the context of species abundance distributions. This is not credible, given the variety of sites and species. Rather, the equivalence must arise through extreme individuality of both sites and species – the idiosyncracy of Pueyo et al (2007). This idiosyncracy may be highly relevant to the general problem of community assembly – Kelly et al (2011) describe two documented examples of species within guilds distributed exponentially over sites.

We have an indication of the machinery underlying the statistical mechanics of alien species - a global population of alien species, each awaiting the opening or rejected by the closing of suitable geographic sites. Underlying the ebb and flow of species is a conserved quantity, the number of alien establishments – the alien footprint of Kelly et al (2011). This alien footprint, for a given number of sites and of species, is (mathematically) the only free parameter in the model. Biologically, it represents some resource that may be related to global net primary productivity and it is an instance of the idea of biotic resistance.

**Appendix Can d be independent of n ?**

The algorithm described above is based on the master equation (2) and generates exponentials provided that both the opening (or filling) of an unoccupied site and the closure of a site already occupied by the target species are independent of the number of sites that species occupies. If the algorithm is regarded as merely a convenient way to generate the matrix $N(J, K)$ as a two fold broken stick, then any question about the independence of the $b$ and $d$ parameters is irrelevant. If however the algorithm is to be taken as representing biological process, the parameters $b$ and $d$ have biological significance. The first parameter $b$ represents the rate at which a new site becomes available to a species and the second parameter $d$ the rate at which an occupied site becomes inhospitable. These parameters are analogous to the birth and death rate parameters encountered when applying the master equation to species abundance problems; in such problems both birth and death rates are taken to be proportional to
the number of individuals – each individual has a probability of giving birth and of itself dying. For a population of \( n \) individuals the birth and death rates are per capita, proportional to \( n \).

The problem of suitable sites opening and closing is wholly different and there is no \textit{a priori} reason to expect rates proportional to the number of occupied sites. Indeed, it seems intuitively obvious that the rate at which a new site opens to an alien species will not depend on the number of sites at which it is already naturalised – one site does not give birth to another somewhere else in the world. (This might not be case if each occupied site sent out propagules that settle in some site already suitable but not yet occupied. This can be discounted because of the evidence that species are to be found throughout the world most of the time, namely the absence of any correlation between the number of species shared and geographical separation.) There is however an intuitive expectation that the more sites that are occupied by a species, the sooner one will close – just as the more individuals there are in a community the sooner one will die – an expectation that \( d \) must be proportional to \( n \).

There is an assumption, not always explicit, underlying this expectation. It is that all occupied sites have the same probability of closing in some interval of time and that there are no correlations involved. This assumption is not necessarily true and there are simple scenarios (of varying degrees of plausibility) under which it is not true. The simplest is that at any one time one site is much more vulnerable than the others and it is this one that gets hit. A much more vivid picture can be expressed in cartoon terms. Suppose every so often Nature despatches two kinds of emissary, which self destruct after completing their missions. The first is programmed to find a site not occupied by species \( x \) and settle it there. The second is programmed to find a site (any site) that is occupied by species \( x \) and extinguish it at that site. In this cartoon scenario both \( b \) and \( d \) have the same dependence on the number of sites occupied - they are independent of that number. (Although there is no logical necessity, the idea that the opening rates and time-reversed closing rates have the same dependence is attractive). The notions embodied above can in fact be expressed in biological terms. If a suitable site for species \( x \) opens, it constitutes an increase in some kind of available niche space for species \( x \). If an occupied site closes then that niche space decreases. For \( b \) and \( d \) to be independent of the number of occupied sites, all that is required is that the niche space fluctuates by absolute amounts rather than by fractional amounts of the space occupied. This is the kind of process we envisage as generating the exponential distributions reported in Kelly et al (2011). We note that if niche space fluctuations were fractional rather than absolute, both \( b \) and \( d \) would be proportional to \( n \) (again with the same \( n \) dependence) and the master equation would then generate not an exponential distribution but a log series (Bowler 2014).

References

Bowler, M. G., \textit{Species abundance distributions, statistical mechanics and the priors of maxent}. Theoretical Population Biology, 2014 \textbf{92C} p.69-77
Bowler, M. G. & Kelly, C. K., *On the statistical mechanics of species abundance distributions*. Theoretical Population Biology, 2012. 82: p.85-91

Kelly, C. K., et al., *The statistical mechanics of community assembly and species distribution*. New Phytologist 2011 191: p. 819

MacArthur, R. H., *On the relative abundance of species*. American Naturalist 1960 94: p. 25

Pueyo, S., et al, *The maximum entropy formalism and the idiosyncratic theory of biodiversity*. Ecology Letters 2007 10 p. 1017

Volkov, I., et al., *Neutral theory and relative species abundance in ecology*. Nature, 2003. 424: p. 1035-1037