TISCIA 28, 67-72

LINEAR ALGORITHM TO CALCULATE INDIRECT SPATIAL STATISTICS FOR COMPLETELY RANDOM MULTI-SPECIES COMMUNITIES

Zs. Erdei, B. Tóthmérész and A. Erdei

Erdei, Zs., Tóthmérész, B. and Erdei, A. (1994): Linear algorithm to calculate indirect spatial statistics for completely random multi-species communities. - Tiscia 28, 67-72.

Abstract. We present a linear algorithm to calculate the diversity of species combinations (or "species list - number of plots" diversity) for completely random communities in an indirect spatial series analysis. It serves as a null model to compare completely random multispecies patterns to real ones which are observed on the field. An efficient algorithm to derive explicitly all the possible species combinations and their frequencies is also proposed. Turbo Pascal algorithms to IBM-compatible PC's and results about the running time of the algorithms are also presented.

Keywords: spatial statistics, indirect spatial series analysis, null model.

Zs. Erdei, B. Tóthmérész, Ecological Institute, Kossuth L. University, Debrecen, POBox 71, H-4010 Hungary; A. Erdei, Student of Computing Science, Eötvös L. University, Budapest

Introduction

A spatial point pattern is a set of locations within a region of interest, which have been generated by some unknown mechanisms (Diggle, 1983). Communities of sedentary organisms, like plants, typically can be viewed as multispecies point patterns. Sometimes the "multidimensional point pattern" terminology is used which is rather confusing. Nowadays the multi-type point pattern terminology also tends to be more popular. In mathematics and biomathematics the period of the last 20 years was the golden age of pattern analysis of one- and two-species patterns (Greig-Smith, 1983; Kershaw, 1964). However, hardly any attention was paid to multispecies point patterns.

Juhász-Nagy developed a brand new way of analyzing multispecies point patterns from the 60-s onward (Juhász-Nagy, 1963, 1967, 1976; Juhász-Nagy and Podani, 1983; Podani et al., 1993). He was especially interested in the partial and multiple association of species in a community. The methods developed by him needs large sample size and a lot of computations (Bartha, 1990; Szollát and Bartha, 1991).

Analyzing one-species spatial point patterns the hypothesis of complete spatial randomness (CSR) has crucial importance. This asserts that (i) the number of individuals in any finite region follows a Poisson distribution and (ii) given n individuals x_i (i=1,..,n) in a region, the x_i's are independent random sample from a uniform distribution on a region. CSR acts as a dividing hypothesis between patterns which are classifiable as regular or aggregated.

Using a multi-species CSR hypothesis we can derive the spatial characteristics of a random multi-species community and we can use this one just as in the case of the one-dimensional pattern; i.e. we can compare the characteristics of a random community to the actual one studied on the field.

In this paper we propose an effective algorithm to calculate the multi-species random characteristics, especially the "species list - number of plots" diversity. We also studied the computing time for an algorithm which explicitly calculate all the possible species lists and we proved that without an efficient algorithm it is very easy to waste inordinate computing time even for very small communities. Throughout he paper we use synonymously the terms "species lists", "floristic composition" or "species combinations".

General description of the algorithm

The abundance vector of a community is denoted by $n=(n_1,n_2,...,n_i,...,n_S)$, where n_i is the abundance of the i-th species of the community. N= n_i is the total number of individuals. When the CSR hypothesis is valid then the distribution of individuals in the sampling plots can be described by a Poisson distribution. Thus the probability that we find zero individual of the i-th species in a plot is

$$p(n_i = 0) = q_i = \exp(-\lambda_i), \quad \lambda_i = n_i \frac{t}{A} \qquad (1)$$

where t is the plot size and A is the total studied area. This was recognized very early by the botanists (Stevens, 1935). Clearly,

$$p_i = 1 - q_i$$

is the probability that we find at least one individual of the species i in a randomly chosen sample plot of size A.

For a multi-species community the probability of the floristic composition vectors can be calculated as a multiplication of the probability of presence and/or absence of the species:

$$\prod_v = p_1 p_2 \cdots p_{i-1} p_i (1-p_{i+1})(1-p_{i+2}) \ldots (1-p_s) \qquad (2)$$

where the species 1,...,i are present and species i+1,...,S are absent in the plot. There are 2^S possible species list vectors and evidently

$$\sum_{v \in 2^s} \prod_v = 1$$

The "species list - number of plots" diversity, $H(2^S)$, for a community of S species is defined as

$$H(2^s) = \sum_{v \in 2^s} \left(\prod_v \log \prod_v\right) \qquad (3)$$

where the summation is taken from 1 to 2^S (Czárán 1992).

It is evident that a direct calculation of (3) is very time-consuming because the computing time is increased by 2^S as S increases. We prove, however, that there is a linear algorithm to calculate (3). For a community having species S+1, the "species list - number of plots" diversity can be calculated in the following way when the diversity H(S) of a community having S species is known:

$$H(2^{s+1}) = H(2^s) + p_{s+1} \log p_{s+1} + (1-p_{s+1}) \log(1-p_{s+1})$$

$$= -\sum_{i=1}^{s} \{p_i \log p_i + (1-p_i)\log(1-p_i)\} \qquad (4)$$

Evidently

$$H(2^1) = p_1 \log p_1 + (1-p_1)\log(1-p_1) \qquad (5)$$

Proof of (4) is in the appendix. The computing time of our algorithm to calculate (3) increases linearly with S.

Algorithm implementations

Two functions are presented. CSR_Lin_Div and CSR_Diversity. The source code is written in Borland's Turbo Pascal 7.0.

function CSR_Lin_Div calculates and returns the Shannon diversity of the species combinations for CSR pattern without the calculation of the species combinations and it has a linear growth of running time as S increases. All the indirect spatial statistics can be calculated using this procedure which are related to the "species list - number of plots" diversity; e.g., florula evenness, distinctiveness, etc. It is defined as data_type which is an extended variable in the presented subroutine. We propose to use extended variables because the rare species combinations have very small contribution to the overall "species list - number of plots" diversity.

function CSR_Diversity presents all the species combinations and their relative frequencies, thus the computing time grows with 2^S as S increases. This procedure may be useful when the species combinations are directly utilized; e.g. in the case of global space series analysis (Tóthmérész, 1994b).

Both functions return the "species list - number of plots" diversity in logarithm of 2; it can be modified easily in the source code. The program can be run in most 286, 386, and 486 IBM compatible PCs with or without built-in mathematical coprocessor. Although input data information provided through a keyboard in an interactive fashion is nice, it is more convenient and efficient to read all input information from a file. In the driver to the procedures we did not present any special data input-output procedures.

Major scalars and vectors

The major scalars and vectors are summarized next.

message = A label to remember what data set is used during the calculations.

Species = Total number of species of the studied community.

plot_size = Area of the sample plots in standard units.

total_size = Area of the whole study area in standard units.

n = The abundance vector of the studied

community; i.e. n[i] is the number of individuals of the i-th species.

function CSR_Lin_Div:

lambda = Parameter of the Poisson distribution; average number of the species in the plots which size is plot_size.

Rel_Fr = A matrix with two rows. In the first row (Rel_Fr[i,0]) are the relative frequencies of the plots where the species i is missing and in the second one (Rel_Fr[i,1]) are the relative frequencies of the plots where the species i is present.

function CSR_Diversity:

SC_Frequency = Relative frequency of the species lists or species combinations.

ListStr = The species combinations are in this string variable in a "0/1" form, where "0" means that the species where missing from the plot while "1" means that the species where present.

Case demonstration

3-species community
plot size = 0.2000000000

relative frequency	Species Combination
6.69044737400155E-0008	000
5.87218293072251E-0004	100
3.27824225584282E-0006	010
2.87730209078975E-0002	110
2.21158009230573E-0006	001
1.94109633362172E-0002	101
1.08364880634898E-0004	011
9.51114875855356E-0001	111

Diversity of Species Combinations :
H1 = 0.3342899617
H2 = 0.3342899617

4-species community
plot size = 0.2000000000

relative frequency	Species Combination
3.98324036839300E-0007	0000
3.27931257830673E-0003	1000
5.60018268143305E-0006	0100
4.61050496820707E-0002	1100
2.12988347416320E-0006	0010
1.75348535894881E-0002	1010
2.99448073486243E-0005	0110
2.46528891835199E-0001	1110
8.72303062951313E-0007	0001
7.18147573802028E-0003	1001
1.22640264063998E-0005	0101
1.00966982495757E-0001	1101
4.66430269431993E-0006	0011
3.84001593674454E-0002	1011
6.55771300596408E-0005	0111
5.39881823753949E-0001	1111

Diversity of Species Combinations :
H1 = 1.8796904272
H2 = 1.8796904272

Fig. 1. Example runs for the 3-species and 4-species communities; plot size is 0.2 unit.

The run of the program is demonstrated on a small data set to make it possible to recalculate the result by tedious work using a pocket calculator. Actually, the three species community is identical with three dominant species of the shrub community of the NE-slope of the "Rejtek Project" Research Area while the four species community is from the plateau area (Tóthmérész, 1994a). The aim of the case demonstration is to illustrate the performance and output of the program, and not to thoroughly solve a biological problem. We tried to keep the format of the output as simple as possible. The

3-species community
plot size = 0.1000000000

relative frequency	Species Combination
2.58658991222063E-0004	000
2.39753088544691E-0002	100
1.57031523756926E-0003	010
1.45553775810097E-0001	110
1.25080600209370E-0003	001
1.15938209128305E-0001	101
7.59362632263790E-0003	011
7.03859299653606E-0001	111

Diversity of Species Combinations :
H1 = 1.3339804687
H2 = 1.3339804687

4-species community
plot size = 0.1000000000

relative frequency	Species Combination
6.31129176032372E-0004	0000
5.66376310894350E-0002	1000
1.81805573292660E-0003	0100
1.63152606173046E-0001	1100
9.58904632099549E-0004	0010
8.60522518452213E-0002	1010
2.76225871013929E-0003	0110
2.47885529206502E-0001	1110
4.96091787129072E-0004	0001
4.45193546629410E-0002	1001
1.42906167532580E-0003	0101
1.28244218529046E-0001	1101
7.53735892254506E-0004	0011
6.76403770029278E-0002	1011
2.17124150184361E-0003	0111
1.94847552383130E-0001	1111

Diversity of Species Combinations :
H1 = 2.8694726582
H2 = 2.8694726582

Fig. 2. Example runs for the 3-species and 4-species communities; plot size is 0.1 unit.

output is arranged into two blocks or columns. The first block contains the relative frequency of the species combinations. The second one contains the species list of the plots. "0" means that the species were absent and "1" means that the species were present. Thus "0000" means that the plot was empty; i.e. there were no one species present in the sample plot. "1000" means that the first species was present and the others were absent. "0100" means that the second species was present while the others were absent, etc.

The output is presented in the Fig. 1 for the plot size of 0.2. In the case of three-species community the relative frequency of the "full" plots, where all the species are present is 0.9511. In the case of four-species community it is 0.5399. The diversity of species combinations is especially low for that plot size in the case of three-species community.

For the plot size of 0.1 the output is presented in the Fig. 2. The diversity of species combinations is much higher this case, although the plots where all the species of the community were present are still dominant.

Source Code of the Algorithms

```
{$N+,E+}
Program Linear_Algorithm;
Const Ln2=0.69314718056;
  MaxSpeciesN = 256; {max num of species }

Type
  StringL   = String[159];
  data_type = extended;
  IntRowVector = array[1..MaxSpeciesN] of integer;

var
  message        : StringL;
  species        : integer;
  H1, H2, plot_size, total_size: data_type;
  n              : IntRowVector;
{————————————————————————————————}
function CSR_Lin_Div(
    plot_size, total_size : data_type;
    Species : integer;
   var n : intRowVector) : data_type;

var i      : integer;
   lambda, sum_piLOGpi : data_type;
   Rel_Fr  : array[1..MaxSpeciesN,'0'..'1'] of extended;

begin
{ Calculation of the relative frequency of the plots where species
'i' where absent
(Rel_Fr[i,'0']) and present (Rel_Fr[i,'1']) }
  for i:=1 to species do begin
   lambda:=n[i]*(plot_size/total_size);
   Rel_Fr[i,'0']:=exp(-lambda);
   Rel_Fr[i,'1']:=1.0-Rel_Fr[i,'0'];
  end;
{--- End of calculation of relative frequencies                    }
{ Calculation of the "Species Combination - Number of Plots" }
{ diversity by the linear algorithm.                                 }
  sum_piLOGpi:=0.0;
  for i:=1 to species do begin
  sum_pilogpi:=sum_pilogpi
   +Rel_Fr[i,'0']*ln(Rel_Fr[i,'0'])/ln2
   +Rel_Fr[i,'1']*ln(Rel_Fr[i,'1'])/ln2;
  end;
  CSR_Lin_Div:=-sum_piLOGpi;
end;
{————————————————————————————————}

function CSR_Diversity(plot_size,
      total_size : data_type;
      Species : integer;
      var n : intRowVector) : data_type;

var
  ListStr          : String;
  i, i_flt         : integer;
  lambda, SC_Frequency,
  sum_piLOGpi          : data_type;
  Rel_Fr  : array[1..MaxSpeciesN,'0'..'1'] of extended;

begin
{ Calculation of the relative frequency of the plots where }
{ species 'i' where absent (Rel_Fr[i,'0']) and             }
{ present (Rel_Fr[i,'1'])                                  }
  for i:=1 to species do begin
    lambda:=n[i]*(plot_size/total_size);
    Rel_Fr[i,'0']:=exp(-lambda);
    Rel_Fr[i,'1']:=1.0-Rel_Fr[i,'0'];
  end;
{   ------ End of calculation of relative frequencies. ---------  }
{ Generation of all the species combinations and their          }
{ relative frequencies.The species combinations are in          }
{ the "ListaStr" string in a "0/1" form, where "0" means that  }
{ the species where missing from the plot while "1" means      }
{ that the species where present.                               }
{ The relative frequency of a species combination is           }
{ in the "SC_Frequency" variable.                               }
  sum_piLOGpi:=0.0;
  ListStr:='';
  for i:=1 to species do
    ListStr:=ListStr+'0';
  i_flt:=0;
  repeat
    SC_Frequency:=1.0;
    for i:=1 to species do
      SC_Frequency:=SC_Frequency
        *Rel_Fr[i,ListStr[i]];
  sum_piLOGpi:=sum_piLOGpi
  +SC_Frequency*LN(SC_Frequency);
{--------------------------------------------------------------------}
{ Print a species combination and its relative frequency.        }
    writeLn(SC_Frequency,'   ',ListStr);
{--------------------------------------------------------------------}
    inc(i_flt);
    inc(ListStr[1]); { increment the first binary digit }
    i:=1;
    while (ListStr[i]='2') AND (i<species) do begin
      dec(ListStr[i],2); { set up '0' because of overflow }
      inc(i);            { increment the next binary digit }
      inc(ListStr[i]);
    end;
    until (ListStr[species]='2'); { all species combinations are
counted }
  CSR_Diversity:=-sum_piLOGpi/ln2;
end;
{————————————————————————————————}
procedure DataInput;
begin
{Number of individuals of the dominant shrubs on the NE-slope:}
 message:='3-species community'; species:=3;
 n[1]:=1135; n[2]:= 489; n[3]:= 441;
{Number of individuals of the dominant shrubs on the plateau:}
{
 message:='4-species community'; species:=4;
 n[1]:=1127; n[2]:= 339;
 n[3]:= 231; n[4]:= 145;
}
```

```
 plot_size := 0.2;
{ plot_size := 0.1; }
 total_size := 25;

end;

BEGIN
  DataInput;
  writeLn; writeLn;
  writeLn(message);
  writeLn('   plot size = ', plot_size :20:10);
  writeLn('relative frequency    Species Combination');
  writeLn('----------------------------------------');

  H1:=CSR_Lin_Div(plot_size, total_size, Species, n);
  H2:=CSR_Diversity(Plot_Size, total_size, Species, n);

  writeLn;
  writeLn('Diversity of Species Combinations :');
  writeLn('    H1 = ', H1 :20:10);
  writeLn('    H2 = ', H2 :20:10);
  readLn;
END.
```

Conclusions

The algorithms presented in the paper are appropriate for determining the diversity of species combinations in the case of completely spatially random multispecies communities. The function CSR_Lin_Div may be used when the species combinations themselves are irrelevant. This is a very fast linear algorithm. The function CSR_Diversity may be used to identify all the species combinations. This algorithm is, however, very time-consuming because the run-time grows exponentially as the number of species of the community grows. The code presented here is a Turbo Pascal implementation of the procedures.

Acknowledgements

We are indebted to Dr. Tamás Czárán who focused our attention on the calculation of the "theoretical" value of "species list - number of plots" diversity (i.e. the value for an infinitely large community). The research was supported by the Hungarian Research Fund (OTKA) F 006082 to the first author and T 5066 to the second author.

References

Bartha, S. (1990): Spatial processes in developing plant communities: pattern formation detected using information theory. - In: Krahulec, F., Agnew, A.D.Q., Agnew, S. and Willems, J. (eds): Spatial Processes in Plant Communities, Akademia, Praha, Czechoslovakia, pp. 31-47.

Bartha, S. (1992): Spatial pattern development in primarily succession on dumps from strip coal mining. - PhD thesis, Vácrátót, Hungary.

Czárán, T. (1992): Calculation of the "theoretical" value of florula diversity. - Personal communication to the second author.

Greigh-Smith, P. (1983): Quantitative Plant Ecology. - Blackwell, Oxford.

Juhász-Nagy, P. (1963): Investigations on the Bulgarian vegetation. II. Study of interspecific correlations in the "Ravnako complex" (Pirin Mountains). - Acta Biologica Debrecina 2, 58-62.

Juhász-Nagy, P. (1967): On some "characteristic areas" of plant community stands. - Proceedings of the Colloquim on Information Theory, Bolyai Mathematical Society (ed. A. Rényi). Akadémiai Kiadó.

Juhász-Nagy, P. (1976): Spatial dependence of plant populations. Part 1. - Acta Botanica Hungarica 22, 61-78.

Juhász-Nagy, P. and Podani, J. (1983): Information theory methods for the study of spatial processes and succession. - Vegetatio 51, 129-140.

Kershaw, K.A. (1964): Quantitative and Dynamic Ecology. - Edward Arnold, London.

Podani, J., Czárán, T. and Bartha, S. (1993): Pattern, area and diversity: the importance of spatial scale in species assemblages. - Abstracta Botanica 17, 37-51.

Szollát, Gy. and Bartha, S. (1991): Pattern analysis of dolomite grassland communities using information theory models. - Abstracta Botanica 15, 47-60.

Stewens, W.L. (1935): The relation between plant density and number of empty quadrats. - Annals of Botany 49, 798-802.

Tóthmérész, B. (1994a): Statistical analysis of spatial pattern in plant communities. - Coenoses 9, 33-41.

Tóthmérész, B. (1994b): Diversity orderings and spatial series analyses. - DSc Dissertation, Debrecen (in Hungarian).

Appendix

For a 1-species community the "species list - number of plots" diversity is defined as

$$H(2^1) = p_1 \log p_1 + (1 - p_1)\log(1 - p_1)$$

Evidently, for a 2-species community it can be calculated

$$H(2^2) = H(2^1) + p_2 \log p_2 + (1 - p_2)\log(1 - p_2)$$

because

$$H(2^2) = p_1 \log p_1 + p_2 \log p_2 + (1 - p_1)\log(1 - p_1) + (1 - p_2)\log(1 - p_2)$$

Now we prove for a (S+1)-species community that the "species list - number of plots" diversity can be calculated as it was indicated by (4); i.e. we prove that

$$H(2^{s+1}) = H(2^s) + p_{s+1} \log p_{s+1} + (1 - p_{s+1})\log(1 - p_{s+1})$$

We know from (3) that

$$H(2^s) = \sum_{v=1}^{2^s} \left(\prod_v \log \prod_v\right)$$

To get $H(2^{S+1})$ we have to multiply all the Π_V's (we have 2^S of them) by both p_{S+1} and $(1-p_{S+1})$; i.e.

$$H(2^{s+1}) = \sum_{v=1}^{2^s} \left(\prod_v p_{s+1} \log\left(\prod_v p_{s+1}\right) + \left(\prod_v (1 - p_{s+1})\log\left(\prod_v (1 - p_{s+1})\right)\right)\right)$$

Using the basic identities of logarithm we get

$$= \sum_{v=1}^{2^s} \left(\prod_v p_{s+1} \log \prod_v + \prod_v p_{s+1} \log p_{s+1} + \prod_v (1 - p_{s+1})\log \prod_v + \prod_v (1 - p_{s+1})\log(1 - p_{s+1})\right)$$

$$= \sum_{v=1}^{2^s} \left(\prod_v p_{s+1} \log \prod_v + \prod_v p_{s+1} \log p_{s+1} + \prod_v \log \prod_v - \prod_v p_{s+1} \log \prod_v + \prod_v (1 - p_{s+1})\log(1 - p_{s+1})\right)$$

Rearranging the expression we get

$$= \sum_{v=1}^{2^s} \left(\prod_v \log \prod_v\right) + \sum_{v=1}^{2^s} \left(\prod_v p_{s+1} \log p_{s+1} + \prod_v (1 - p_{s+1})\log(1 - p_{s+1})\right)$$

$$= H(2^s) + p_{s+1} \log p_{s+1} \sum_{v=1}^{2^s} \prod_v + (1 - p_{s+1})\log(1 - p_{s+1}) \sum_{v=1}^{2^s} \prod_v$$

Because

$$\sum_{v=1}^{2^s} \prod_v = 1$$

thus

$$= H(2^s) + p_{s+1} \log p_{s+1} + (1 - p_{s+1})\log(1 - p_{s+1})$$

and we have proved the proposition.

Table 1. Example run-times of the calculation of the diversity of species combination by direct calculation of all the species combinations

Number of species	IBM PC/XT	AT/20 MHz	386DX/33 MHz + 80387	486DX2/66 MHz
10	42 minutes	4 minutes	10 seconds	1.6 second
20	6 weeks + 2 days	4 days + 8 hours	3 hours + 41 minutes	2 minutes
30	183.5 years	18 years	30 weeks + 1 day	1 day + 22 hours
40	278'432 years	27'269 years	789 years	7 years + 16 weeks
50	422'000'000 years	41'300'000 years	1'100'000 years	10'073 years