The Index–of–Dispersion Test Revisited

D. Pfeifer, H. Ortleb, U. Schleier–Langer[1], H.–P. Bäumer[2]

[1] Fachbereich Mathematik, Carl von Ossietzky Universität Oldenburg,
Postfach 2503, D–26111 Oldenburg

[2] Hochschulrechenzentrum, Carl von Ossietzky Universität Oldenburg,
Postfach 2503, D–26111 Oldenburg

Arbeitsgruppe Statistik in der
Ökosystemforschung Niedersächsisches Wattenmeer

Summary: In quantitative ecology the classical index–of–dispersion is widely used for testing the hypothesis of spatial randomness. However, spatial aggregation of individuals is often observed in field experiments, so that the test will frequently reject the hypothesis without indicating any alternatives. In this paper we consider a modified index–of–dispersion test which allows for testing the hypothesis of a spatial Poisson point process with intensity measure having a density $\lambda(\mathbf{x})$ of the form

$$\lambda(\mathbf{x}) = \sum_{j=1}^{n} a_j f_j(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^2$$

with known non–negative regression functions $f_j(\mathbf{x})$ and unknown non–negative parameters a_j which are to be estimated by the observed data. This model includes the classical case for $n = 1$ and $f_1(\mathbf{x}) \equiv 1$. Further applications to testing local geographical influences on health data are also pointed out.

1. Introduction

In quantitative ecology the classical index–of–dispersion (normalized variance–to–mean ratio of quadrat counts) is widely used for testing the hypothesis of spatial randomness (see e.g. Greig–Smith (1983), p. 61ff, or Diggle (1983), p. 23ff). One reason for this is the fact that the experimental estimation of spatial distribution and abundance of animals living on small scales is usually faced with taking samples by means of physical devices (quadrat sampling; cf. eg. Krebs (1985), p. 160ff). Due to individual aggregation effects, patchiness of the spatial distribution can be observed even on those scales: the following figure represents two examples of 5 × 5 multicorer samples of benthic fauna *(Harpacticus obscurus)* in the Wadden Sea, taken from Pfeifer, Schleier–Langer, and Bäumer (1994).

95	1	3	0	42
1	1	1	4	2
0	5	8	81	24
11	1	6	71	116
1	5	116	2	10

165	22	1	94	68
11	82	111	97	153
0	0	24	13	15
31	1	46	22	11
2	0	5	8	6

Fig. 1

W. Gaul et al. (eds.), *From Data to Knowledge*
© Springer-Verlag Berlin · Heidelberg 1996

The aggregation effect observed in these samples can e.g. be explained by oxidation and fertilization of the surrounding sediment by lugworms *(Arenicola marina)* or the polychaete *Lanice conchilega*, attracting meiofauna in high numbers around their burrows and funnels (cf. Reise (1985), p.126ff; Ekschmitt (1993) contains a more general discussion of aggregation effects). It is therefore clear that the hypothesis of a random spatial distribution of meio– and microfauna must in general be rejected, although the spatial dispersion of lugworms and polychaetes – as the "parent" point pattern – is in good coincidence with the Poisson process assumption (Pfeifer, Bäumer and Albrecht (1993)). However, it might be that the spatial distributional pattern of individuals within each cluster (represented by a parent polychaete) may possibly stem from a Poisson process itself and is hence "locally random". In this paper we therefore suggest a modification of the classical index–of–dispersion test which allows for testing the hypothesis of a spatial Poisson point process with intensity measure possessing a density $\lambda(\mathbf{x})$ of the form

$$\lambda(\mathbf{x}) = \sum_{j=1}^{n} a_j f_j(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^2 \tag{1}$$

with known non–negative regression functions $f_j(\mathbf{x})$ and unknown non–negative parameters a_j. As usual in linear statistical models, these parameters are to be estimated by the observed data via suitable least–squares–methods. This model includes the classical case for $n = 1$ and $f_1(\mathbf{x}) \equiv 1$. (Note that this procedure is not restricted to the two–dimensional Euclidean space, but is generally applicable in any dimension.)

2. The Model

Consider a spatial (in general non–homogeneous) Poisson point process ξ with an intensity (= Lebesgue–density for its intensity measure) of type (1). Suppose further that observations $Z_1, \ldots, Z_m$ are taken with $m > n$, representing aggregate point counts within pairwise disjoint subsets $B_1, \ldots, B_m$ of $\mathbb{R}^2$, i.e.

$$Z_i = \xi(B_i), \quad i = 1, \ldots, m. \tag{2}$$

Then

$$\mu_i := E(Z_i) = \int_{B_i} \lambda(\mathbf{x})\, d\mathbf{x} = \sum_{j=1}^{n} a_j \int_{B_i} f_j(\mathbf{x})\, d\mathbf{x} = (\mathbf{Wa})_i, \; i = 1, \ldots, m \tag{3}$$

where $\mathbf{W}$ is the weight matrix $\mathbf{W} = (w_{ij})$ with

$$w_{ij} = \int_{B_i} f_j(\mathbf{x})\, d\mathbf{x}, \quad i = 1, \ldots, m, \; j = 1, \ldots, n \tag{4}$$

and $\mathbf{a}$ is the column vector of parameters a_j. By the central limit theorem, if the μ_i are sufficiently large, the column vector $\mathbf{Z}$ consisting of the components Z_i is approximately normally distributed as $\mathcal{N}(\mathbf{Wa}, \Sigma)$ with a variance–covariance matrix $\Sigma = \Delta(\mathbf{Wa})$ of diagonal form where

$$\Delta(\mathbf{b})_{ij} = \begin{cases} b_i, & \text{if } i = j \\ 0, & \text{otherwise,} \end{cases} \qquad 1 \leq i, j \leq m \tag{5}$$

for any column vector $\mathbf{b} = (b_1, \ldots, b_m)^{\text{tr}} \in \mathbb{R}^m$. Hence

$$\hat{\mathbf{a}} = (\mathbf{W}^{\text{tr}}\mathbf{W})^{-1}\mathbf{W}^{\text{tr}}\mathbf{Z}, \quad \hat{\mu} = \mathbf{W}\hat{\mathbf{a}} \tag{6}$$

are appropriate least–squares–estimates for the parameter vectors $\mathbf{a}$ and μ, resp., and the *generalized index–of–dispersion* D is just the sum of squares of residuals:

$$D = \sum_{i=1}^{m} \frac{(Z_i - \hat{\mu}_i)^2}{\hat{\mu}_i} = (\mathbf{Z} - \mathbf{W}\hat{\mathbf{a}})^{\text{tr}}\hat{\Sigma}^{-1}(\mathbf{Z} - \mathbf{W}\hat{\mathbf{a}}) \tag{7}$$

where $\hat{\Sigma} = \Delta(\hat{\mu}) = \Delta(\mathbf{W}\hat{\mathbf{a}})$ is the estimated diagonal matrix with entries $\hat{\mu}_i$, $i = 1, \ldots, m$. Hence D is approximately χ^2–distributed with $m - n$ degrees of freedom if $\mathbf{W}$ has full rank n and can thus be used as a test statistic for the null–hypothesis that the underlying point process is Poisson with an intensity of the form (1) (for technical details, see e.g. Fahrmeir and Hamerle (1984), p.84ff).

However, the estimate $\hat{\mathbf{a}}$ given in (6) is not optimal; a better estimate would be

$$\mathbf{a}^* = (\mathbf{W}^{\text{tr}}\Sigma^{-1}\mathbf{W})^{-1}\mathbf{W}^{\text{tr}}\Sigma^{-1}\mathbf{Z} \tag{8}$$

if Σ was known. The problem here is that Σ is *unknown*, so that we should use an iterative procedure in order to get a better estimate of $\mathbf{a}$, inserting consecutively the updated estimate $\hat{\Sigma}$ in (8):

$$\mathbf{a}_1^* := \hat{\mathbf{a}}, \quad \mathbf{a}_{k+1}^* = (\mathbf{W}^{\text{tr}}\Delta(\mathbf{W}\mathbf{a}_k^*)^{-1}\mathbf{W})^{-1}\mathbf{W}^{\text{tr}}\Delta(\mathbf{W}\mathbf{a}_k^*)^{-1}\mathbf{Z}, \; k \in \mathbb{N}. \tag{9}$$

It is not always ensured that this iterative procedure will give meaningful results. However, as long as the estimates of the parameters a_j remain non–negative, it seems that in most cases the corresponding dispersion indices according to (7) will decrease and thus by boundedness converge to a lower limit which would then also imply convergence of the estimator sequence $\{\mathbf{a}_k^*\}$. [If by chance some of the a_{kj}^* become negative, they should be set to zero, omitting thus the corresponding regression functions in the basic setup.]

3. Examples

For simplicity we shall here assume $m = 4$, $n = 2$ and regression functions f_j of the following type:

$$f_j(\mathbf{x}) = \sum_{k=1}^{4} c_{jk} I_{B_k}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^2, \; j = 1, 2 \tag{10}$$

with disjoint sets B_k of unit area each (for example, a 2×2 square subdivided into 4 equally large subsquares), indicator functions I_B defined by

$$I_B(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in B \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

for any Borel subset $B \subseteq \mathbb{R}^2$, and coefficient matrix $\mathbf{C} = (c_{jk})$ given by

$$\mathbf{C} = \begin{pmatrix} 7 & 4 & 3 & 1 \\ 1 & 2 & 5 & 9 \end{pmatrix}. \tag{12}$$

If we think of the B_k forming a 2×2 gridded square, these coefficients could also be thought of being arranged in matrix form:

$$\mathbf{C}_1 = \begin{bmatrix} 7 & 4 \\ 3 & 1 \end{bmatrix} \qquad \mathbf{C}_2 = \begin{bmatrix} 1 & 2 \\ 5 & 9 \end{bmatrix} \tag{13}$$

Poisson point processes with intensities f_1 and f_2, resp. would hence place the majority of points in the north–west and the south–east corner of the square, resp. By our assumptions, it follows that the weight matrix $\mathbf{W}$ coincides with the coefficient matrix $\mathbf{C}^{\mathrm{tr}}$ since all B_j have equal area of unit measure. By (6), we obtain

$$\mathbf{W}^{\mathrm{tr}}\mathbf{W} = \begin{pmatrix} 75 & 39 \\ 39 & 111 \end{pmatrix}$$

and hence

$$(\mathbf{W}^{\mathrm{tr}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{tr}} = \frac{1}{1134}\begin{pmatrix} 123 & 61 & 23 & -40 \\ -33 & -1 & 43 & 106 \end{pmatrix} \tag{14}$$

Suppose we have obtained the sample $\mathbf{Z} = (41\ 36\ 57\ 95)^{\mathrm{tr}}$ from quadrat counts, or, in matrix form (corresponding to the gridded square above):

$$\mathbf{Z} \simeq \begin{bmatrix} 41 & 36 \\ 57 & 95 \end{bmatrix}. \tag{15}$$

Then (14) yields the estimate

$$\hat{\mathbf{a}} = (2375/567 \quad 5566/567) = (4.188 \quad 9.816) \tag{16}$$

with an index–of–dispersion of

$$D = 0.508. \tag{17}$$

Iteration of the procedure gives the following table for $\mathbf{a}_k^*$:

k	1	2	3	4
$\mathbf{a}_k^*$	(4.188 9.816)	(4.290 9.684)	(4.287 9.687)	(4.287 9.687)
D	0.508	0.488	0.488	0.488

Fig. 2

A stabilisation of the recursive estimates of $\mathbf{a}$ is visible here already for values of k smaller than 5; the corresponding χ^2–test would further not reject the null–hypothesis at any reasonable level of significance while the ordinary index–of–dispersion gives a value of 233.706 which is far beyond the 0.999–quantile of the χ^2_3–distribution.

For the sample $\mathbf{Z} = (75\ 71\ 79\ 68)$ or, in matrix form,

$$\mathbf{Z} \simeq \begin{bmatrix} 75 & 71 \\ 79 & 68 \end{bmatrix} \tag{18}$$

the situation is different: here (14) gives the initial estimate

$$\hat{\mathbf{a}} = (12653/1134 \quad 8059/1134) = (11.157 \quad 7.106) \tag{19}$$

with $D = 5.856$ and iterations

k	1	2	3	4
$\mathbf{a}^*_k$	(11.175 7.106)	(11.496 7.091)	(11.510 7.079)	(11.511 7.078)
D	5.856	5.6363	5.6346	5.6344

Fig. 3

such that the χ^2–test would reject the null–hypothesis at a significance level of 0.10 with a critical value of 4.61 (from the χ^2_2–distribution). Note that the ordinary index–of–dispersion is 5.866 here which is below the 0.90–quantile of the χ^2_3–distribution, hence the hypothesis of a purely random spatial distribution would not be rejected at this level.

An application of this method for instance to benthos data seems reasonable if clustering of meio– and microfauna is caused by several superposed parent macrofauna patterns with different intensities of species attraction. For such a kind of analysis, it is therefore necessary to register macrofauna data from the same samples as well.

Distributional patterns as in Fig. 1 are in good coincidence with simulated patterns from Thomas–processes with radially symmetric bivariate normal distributions for the daughter points (see e.g. Diggle (1983), p. 54ff.). Due to the possibly quadrat–overlapping variability of such a normal distribution, quadrats should, however, be grouped again before further investigation. For example, the left hand sample in Fig. 1 could be grouped as follows:

95	1	3	0	42
1	1	1	4	2
0	5	8	81	24
11	1	6	71	116
1	5	116	2	10

or, in matrix form :

$$\begin{bmatrix} 115 & 153 \\ 140 & 199 \end{bmatrix} \tag{20}$$

A closer look at several of the obtained benthos data sets suggests that clustering of meio– and microfauna occurs typically around the high peaks within the quadrats, hence the following assumptions could be justified:

$$\mathbf{C}_1 = \begin{bmatrix} 4 & 6 \\ 5 & 7 \end{bmatrix} \qquad \mathbf{C}_2 = \begin{bmatrix} 9 & 6 \\ 6 & 4 \end{bmatrix} \tag{21}$$

Here $\mathbf{C}_1$ corresponds to the local intensities for the parent point pattern of polychaetes attracting meiofauna [not actually counted in this sample], while $\mathbf{C}_2$ corresponds to a pure background noise component of homogeneous Poisson type per quadrat. The method described above gives here

k	1	2	3	4
$\mathbf{a}_k^*$	(27.316 0.223)	(27.024 0.498)	(27.033 0.491)	(27.022 0.488)
D	1.309	1.290	1.290	1.290

Fig. 4

which means that the average cluster size (within a quadrat) is about 27 individuals with a superimposed Poisson noise of about one individual per two quadrats on average. The modified index–of–dispersion test would not reject the null–hypothesis at any reasonable level here, while the ordinary index–of–dispersion for these four data points gives a value of 24.53.

A similar procedure for the right hand sample in Fig. 1 yields:

165	22	1	94	68
11	82	111	97	153
0	0	24	13	15
31	1	46	22	11
2	0	5	8	6

or, in matrix form :

$$\begin{bmatrix} 392 & 412 \\ 109 & 75 \end{bmatrix} \tag{22}$$

Choosing

$$\mathbf{C}_1 = \begin{bmatrix} 12 & 13 \\ 3 & 2 \end{bmatrix} \qquad \mathbf{C}_2 = \begin{bmatrix} 6 & 4 \\ 9 & 6 \end{bmatrix} \tag{23}$$

we obtain

k	1	2	3	4
$\mathbf{a}_k^*$	(31.393 1.857)	(31.420 1.815)	(31.420 1.815)	(31.420 1.815)
D	0.122	0.120	0.120	0.120

Fig. 5

The average cluster size (within a quadrat) is slightly larger as before, while the Poisson background noise here is about 2 individuals per quadrat. The modified index–of–dispersion test would again not reject the null hypothesis, while the simple index–of–dispersion gives a value of 392.21.

4. Further Applications

A particular case of interest is $n = 1$, i.e. there is only one parent macrofauna type causing patchiness in the spatial distribution. In this case, the calculations simplify to a great extent: let f denote the regression density and a the parameter to be estimated. The design matrix $\mathbf{W}$ then reduces to a column vector $\mathbf{W} = (w_1, \ldots, w_m)^{\text{tr}}$, with $w_i = \int_{B_i} f(\mathbf{x})\,d\mathbf{x}$, hence by (6),

$$\hat{a} = \frac{\sum_{i=1}^{m} w_i Z_i}{\sum_{i=1}^{m} w_i^2} \tag{24}$$

and the index–of–dispersion is given by

$$D = \sum_{i=1}^{m} \frac{(Z_i - \hat{a}w_i)^2}{\hat{a}w_i}. \tag{25}$$

Note that since here **a** is one–dimensional only, there is no difference between the estimate $\hat{a}$ and the estimate a^* according to (8).

Another more natural application of the latter model to geographic health data was incorporated into the CARLOS–project recently (**Ca**ncer **R**egistry **Lo**wer **S**axony; see Appelrath et al. (1993)). The establishment of a country cancer registry, based on a homogeneous, area–wide, population–based registration of cancer cases, is pushed by the social ministry of Lower Saxony (FRG) since 1992. CARLOS is the respective project for the Weser–Ems region, providing also statistical tests besides regional information and the calculation of epidemiological indices. Questions whether for instance the presence of nuclear power plants rises the risk of childhood Leukaemia are usually discussed controversely (see Gardner (1993) for a recent U.K. study). The simple index–of–dispersion test cannot directly be applied here due to the different population structures in the various geographic regions so that adjustments have to be made in advance (see Kafadar and Tukey (1993) for an alternative approach). For instance, the weights w_i could be chosen proportional to the local population densities n_i, $i = 1, \ldots, m$, for the m local districts in the country. Another problem is that not all cancer cases are registered with the same registration probability p_i, $i = 1, \ldots, m$. This means that in the medical data actually a *thinned* (Poisson) process is observed, so that the weights w_i should be chosen proportional to $p_i n_i$. The problem of a reliable estimation of the p_i is, however, an unsolved problem until now, since it requires further considerations other than mere statistical investigations.

The multi–parameter regression approach as in (1) might likewise be fruitful for the analysis of cancer data, for instance, if it is reasonable to consider age classes or other personal or social factors.

On the other hand, local influences on health data could be tested by this approach as well, by increasing the values of the regression functions over the corresponding areas. Experience with such kind of statistical procedures is presently in progress.

Acknowledgements:

Financial support was given by the Ministry of Research and Technology, FRG, within the research project "Ökosystemforschung Niedersächsisches Wattenmeer, Hauptphase B (ELAWAT)". The contribution of H. Ortleb was, in particular, financed by Terramare e.V., Wilhelmshaven.

References:

APPELRATH, H.–J., BEHRENDS, H., JASPER, H., ORTLEB, H. ET AL. (1993): Endbericht der Projektgruppe "Aktive Informationssysteme", Bericht IS 15, Teil A. Fachbereich Informatik, Universität Oldenburg.

DIGGLE, P.J. (1983): *Statistical Analysis of Spatial Point Patterns.* Mathematics in Biology. Ac. Press, N.Y.

EKSCHMITT, K. (1993): Über die räumliche Verteilung von Bodentieren. Zur ökologischen Interpretation der Aggregation und zur Probenstatistik. Dissertation, Universität Bremen.

FAHRMEIR, L., AND HAMERLE, A. (Eds.) (1984): *Multivariate statistische Verfahren.* W. de Gruyter, Berlin.

GARDNER, M.J. (1993): Investigating childhood leukaemia rates around the Sellafield nuclear plant. *Int. Stat. Rev.* 61, 231 – 244.

GREIG–SMITH, P. (1983): *Quantitative Plant Ecology.* Studies in Ecology, Vol. 9. 3rd edition, Blackwell Scientific Publ., Oxford.

KAFADAR, K., AND TUKEY, J.W. (1993): U.S. cancer death rates: a simple adjustment for urbanization. *Int. Stat. Rev.* 61, 257 – 281.

KREBS, C.J. (1985): *Ecology.* The Experimental Analysis of Distribution and Abundance. 3rd edition, Harper & Row, N.Y.

PFEIF.ER, D., BÄUMER, H.–P., AND ALBRECHT, M. (1992): Spatial point processes and their applications to biology and ecology. *Modeling of Geo–Biosphere Processes* 1, 145 – 161.

PFEIFER, D., SCHLEIER–LANGER, U., AND BÄUMER, H.–P. (1994): The analysis of spatial data from marine ecosystems. To appear in: H.–H. Bock, W. Lenski, and M.M. Richter (eds.): *Information Systems and Data Analysis.* Prospects – Foundations – Applications. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, N.Y.

REISE, K. (1985): *Tidal Flat Ecology.* An experimental approach to species interactions. Ecological Studies 54, Springer, N.Y.