The Analysis of Spatial Data from Marine Ecosystems

Dietmar Pfeifer, Ulrike Schleier–Langer[1], Hans–Peter Bäumer[2]

[1] Fachbereich Mathematik, Carl von Ossietzky Universität Oldenburg,
Postfach 25 03, D–26015 Oldenburg, Germany

[2] HRZ – Angewandte Statistik, Carl von Ossietzky Universität Oldenburg,
Postfach 25 03, D–26015 Oldenburg, Germany

Summary: In this paper we show how established methods from the fields of point process theory, stochastic geometry, and geostatistics can be applied to analyze spatial data from marine ecosystems. In particular, examples and data sets from the research project "Ökosystemforschung Niedersächsisches Wattenmeer" are presented.

1. Introduction

The use of statistical methods has a very good tradition in applied sciences such as medicine or biology, and seems to play a more and more important role also in related fields like ecology (cf. Richter and Söndgerath (1990).) While the "classical" statistical theory is well–established and largely adapted to the particular problems arising here (cf. Krebs (1985) or Begon, Harper, and Townsend (1990)), more modern advances in stochastic modelling have seemingly not yet found their final way into the field. The recent papers of Rothschild (1992) and Pfeifer, Bäumer, and Albrecht (1992) as well as the forementioned monograph by Richter and Söndgerath (1990) show, however, that applications of point process theory and stochastic geometry may be fruitful to obtain a deeper insight into the highly complicated stochastic structures which are the basis for the analysis of many of the spatial data obtained in biological or ecological field experiments (cf. also Diggle (1983) and Cressie (1991)).

Besides these two disciplines, we want to show in this paper that also geostatistical aspects could well be included in an analysis of such data, especially in marine ecology (cf. also the recent monograph by Haining (1990)).

2. Point patterns and processes: small–scale communities

Among the many species which form typical communities in marine ecosystems there are several whose spatial distribution may by physical reasons supposed to be more or less uniform over certain areas. This assumption can be justified if e.g. the offspring is in larval form, being drifted by hydrodynamical forces and distributed "at random". Since the physical extension of the larger part of meio– and microfauna species is typically in the millimeter range or even less (e.g. the copepode species *Harpacticus obscurus*) we may neglect this aspect of statistical modelling, such that point process theory becomes an appropriate tool for the

H.-H. Bock et al. (eds.), *Information Systems and Data Analysis*
© Springer-Verlag Berlin · Heidelberg 1994

description and analysis of distributional patterns created by such species.
A first basic model is the so–called spatial *Poisson process*, denoted by $\xi(\,.\,)$. Finite versions of them are characterized by the following two properties ($d \in \mathbb{N}$ denoting the dimension):

i) For any measurable (i.e. Borel–)set $A \subseteq \mathbb{R}^d$, the random variable $\xi(A)$ is Poisson distributed with mean $\lambda(A) \in [0,\infty)$.

ii) For any countable collection of disjoint measurable sets $A, B, C, \ldots$, the random variables $\xi(A), \xi(B), \xi(C), \ldots$ are independent.

Here $\lambda(\,.\,)$ (also denoted as $E\xi$) is a finite measure (called *intensity measure*) on the Borel σ–field over $\mathbb{R}^d$. A canonical representation of ξ is

$$\xi(A) = \sum_{k=1}^{N} \varepsilon_{X_k}(A) = \sum_{k=1}^{N} \mathbb{1}_A(X_k) \tag{1}$$

where the random variables N and $X_1, X_2, X_3, \ldots$ are independent, N follows a Poisson distribution with mean $E(N) = \lambda(\mathbb{R}^d) = E\xi(\mathbb{R}^d)$, and $X_1, X_2, X_3, \ldots$ follow the distribution Q given by

$$Q(A) = P(X_k \in A) = \frac{E\xi(A)}{E\xi(\mathbb{R}^d)} \tag{2}$$

for all measurable sets $A \subseteq \mathbb{R}^d$ and $k \in \mathbb{N}$, provided $E\xi(\mathbb{R}^d) > 0$ (otherwise there are no points realized by the process). Here $\mathbb{1}_B$ denotes the indicator random variable of the event B, which shows that $\xi(A)$ counts the number of points in the set A. If, in particular, the intensity measure $E\xi$ has the form

$$E\xi(A) = c \cdot m(A \cap \mathcal{X}), \tag{3}$$

for some bounded measurable region $\mathcal{X} \subset \mathbb{R}^d$ and $c > 0$, where $m(\,.\,)$ denotes the Lebesgue measure (i.e. area or volume in case $d = 2$ or $d = 3$, resp.), then ξ is called *homogeneous Poisson process over* $\mathcal{X}$. In this case, the positions of points $\{X_n\}$ are uniformly distributed over $\mathcal{X}$ since here

$$Q(A) = P(X_k \in A) = \frac{E\xi(A)}{E\xi(\mathbb{R}^d)} = \frac{m(A \cap \mathcal{X})}{m(\mathcal{X})} \tag{4}$$

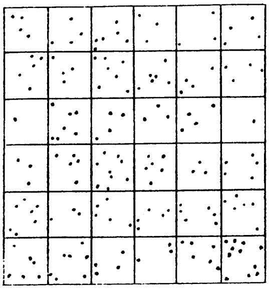
for all Borel sets A, provided $m(\mathcal{X}) > 0$.
In order to test the hypothesis of homogeneity it is sometimes convenient to use the *index–of–dispersion test*, especially if only spatially aggregated data are available. For this purpose, the region $\mathcal{X}$ has to be subdivided into n, say, disjoint observation windows $A_1, \ldots, A_n$ of equal size (i.e. $m(A_k) = m(\mathcal{X})/n$ for all k). Under the assumption of homogeneity the random variables $\xi(A_1), \ldots, \xi(A_n)$ are independent and follow a Poisson distribution with mean $(c/n)m(\mathcal{X})$, which is identical to their variance. Hence the distribution of the test statistic

$$D_n = (n-1)\frac{\sigma_n^2}{\overline{\xi}_n} \tag{5}$$

with

$$\bar{\xi}_n := \frac{1}{n}\sum_{k=1}^{n}\xi(A_k), \quad \sigma_n^2 := \frac{1}{n-1}\sum_{k=1}^{n}\left(\xi(A_k)-\bar{\xi}_n\right)^2$$

is asymptotically independent of c and asymptotically χ^2_{n-1}, by the central limit theorem for Poisson distributed random variables. D_n is called normalized index–of–dispersion. The following figure shows the positions of individuals of the polychaeta species *Arenicola marina* in a sample of size 174, together with the corresponding 6 × 6 abundance matrix of spatially aggregated data:



4	4	6	3	3	4
5	4	7	6	4	4
1	7	4	3	4	1
3	5	9	6	3	5
7	3	5	5	6	6
6	7	4	3	6	11

Fig. 1

spatial distribution *aggregated data*

$\bar{\xi}_n = 4.8333 \qquad \sigma_n^2 = 4.0857 \qquad D_n = 29.59$

The corresponding lower and upper 10%–quantiles of the χ^2_{35}–distribution here are $\chi^2_{35;0.10} = 24.80$, $\chi^2_{35;0.90} = 46.05$, hence the test will not reject the homogeneity hypothesis at a significance level of $\alpha = 0.2$ or less. (A similar result would be obtained with the data set of the centipede species *Lithobius crassipes* in Krebs (1985), Figure 10.2, p. 160.)

The following two tables contain the aggregated data of the abundance of the copepode species *Harpacticus obscurus*, taken at two different sites (labeled sites 6 and 8):

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. 2

site 6 *site 8*

$D_n = 1464.20 \qquad D_n = 1490.14$

The index–of–dispersion test here clearly rejects the homogeneity hypothesis for both sites, at all reasonable significance levels α. In this particular case, alternative models for the spatial distribution of individuals should be taken into account. For a discussion of non–homogeneous Poisson processes and, more generally, cluster processes, we refer the reader to Richter and Söndgerath (1990), Stoyan, Kendall and Mecke (1989) or Pfeifer, Bäumer and Albrecht (1992). Geostatistical aspects of such distributional patterns will be treated in more detail in Chapter 5 below.

3. Dynamic point patterns: revitalization & relative stability

One of the most important factors for relative stability in marine ecosystems seems to be the enormous dynamics due to water and wind forces, which guarantee to a certain extent the revitalization potential of the system. It is therefore desirable to introduce a time–dependent dynamic component in stochastic point patterns of the above type, which enables a study of the long–time behaviour of (at least a part of) the entire system. In the recent paper of Pfeifer, Bäumer, and Albrecht (1993) such an attempt is made by considering a family of Poisson point processes $\{\xi_t \mid t \geq 0\}$ which allows for birth, death and movements of points over time. In particular, the counting process $M_t = \xi_t(\mathsf{R}^d)$ of particles at time $t > 0$ forms a Markovian birth–death process with time– and state–dependent birth and death rates. Depending on the choice of system parameters, extinction, explosion and stabilization of the system over time can be modelled. Simulation studies show that the revitalization of (artificially) depopulated areas in the wadden sea by e.g. the gastropode species *Hydrobia ulvae* can very well be described through such models. However, a general framework for dynamic point processes is not yet fully developed, such that only particular models are available at present (cf. also Chapter 5.5.5 in Stoyan, Kendall and Mecke (1989)).

4. A Boolean model: mussel banks

Point processes can of course also serve as a basis for more complicated geometric structures, such as random sets. *Boolean models* are obtained just in this way: the points of a homogeneous Poisson process, say, are the centers of geometric objects such as discs or balls with fixed or random radius. The random sets Ξ, which are created in this manner, are suitable models for patchiness or spatial clustering, such as mussel bank structures. The picture below shows a part of a juvenile *Mytilus edulis* bank.

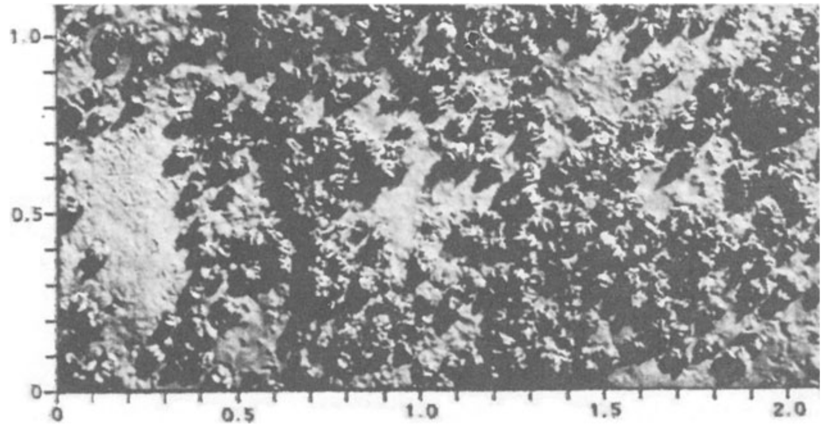


Fig. 3
mussel bank of Mytilus edulis

In applications one of the important parameters of the model is the so–called *area* or *volume fraction* p, i.e. the average area or volume per unit square or unit cube, resp., covered by the image. In the model outlined above it is given by

$$p = 1 - \exp\{-c \cdot E(m(B))\} = \begin{cases} 1 - \exp(-c\pi E(R^2)), & d = 2 \\ 1 - \exp(-\frac{4}{3}c\pi E(R^3)), & d = 3 \end{cases} \tag{6}$$

where R is the random variable describing the (random) radius of the "typical" disc or ball B in the model. This quantity can also be interpreted as the probability that a point of the Poisson process hits the "average" disc or ball with radius $E(R^2)$ or $E(R^3)$, resp. In marine ecology, the knowledge of p for mussel banks is basic to estimate the biomass in the bank, for instance. For stationary and isotropic random sets (which means shift and rotation invariance of the distribution of Ξ, as in the above Boolean model) there are simple and efficient estimators for p, for instance

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\Xi}(x_i) \tag{7}$$

where $\{x_i\}$ is a grid of n points in R^d, typically larger in extent than the given image. Thus $\hat{p}$ counts the corresponding number of grid points $\{x_i\}$ that hit the image. This estimator is unbiased; under appropriate conditions, its variance is of order $\mathcal{O}(1/n)$ for large n (cf. Stoyan, Kendall and Mecke (1989), Chapter 6.3). For the (complete) mussel bank above, the grid estimate gives values of $\hat{p}$ about 0.4 which is in good coincidence with the value that was obtained by a deterministic technique measuring the covered area by exhaustion with "small" rectangles.

5. Geostatistical models for spatial dependence: benthos data

If we interpret the spatial data as being realizations of a *random field* $\{Z(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathsf{R}^d\}$, then geostatistical approaches could be tried for the data analysis as well. Such techniques have gained increasing importance in environmental sciences over the years (see Haining (1990); in particular in connection with GIS's (Geographical Information Systems)). For instance, for the Harpacticus data (Fig. 2 above), the assumption of a homogeneous spatial distribution is surely inappropriate, as can be seen even with the bare eye. Corresponding observations can be made throughout for a great deal of other benthic species; it seems that such species have a tendency to spatial aggregation in general (cf. also the recent Ph.D. Thesis by Ekschmitt (1993)). A simple but nevertheless efficient way to model such spatial dependence structures is the assumption of a weakly stationary random field with constant mean μ and variance σ^2 in each point $\boldsymbol{x}$, such that information on the spatial dependence is given through the *variogram* function

$$2\gamma(\boldsymbol{h}) = V(Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})) = 2(\sigma^2 - C(\boldsymbol{h})) = 2(C(0) - \mathrm{C}(\boldsymbol{h})) \tag{8}$$

for vectors $\boldsymbol{x}, \boldsymbol{h} \in \mathsf{R}^d$. Here

$$C(\boldsymbol{h}) = Cov(Z(\boldsymbol{x}), Z(\boldsymbol{x}+\boldsymbol{h})) \tag{9}$$

denotes the *covariance* function of the random field. If stationarity and isotropy can be assumed, $C(\boldsymbol{h})$ depends only on the length $\|\boldsymbol{h}\|$. The function $\gamma(\cdot)$ is also called *semi–variogram* function. (Note that in the literature, variograms and semi–variograms are sometimes identified, and that only in the case of normal distributions, this function uniquely determines the distribution of the random field $\{Z(\boldsymbol{x})\}$.) The behaviour of the variogram or semi–variogram function in the neighbourhood of the origin determines the degree of "smootheness" of the random field. For variograms of Gaussian type, i.e.

$$\gamma(\boldsymbol{h}) \sim c \cdot \|\boldsymbol{h}\|^2 \qquad (\boldsymbol{h} \to 0) \tag{10}$$

with some positive constant c there exist realizations of a random field with normal marginal distributions that have "smooth", i.e. differentiable paths; in the case

$$\gamma(\boldsymbol{h}) \sim c \cdot \|\boldsymbol{h}\| \qquad (\boldsymbol{h} \to 0) \tag{11}$$

the corresponding paths will be continuous, but non–differentiable. If $\gamma(\cdot)$ is not continuous in the origin (so–called *nugget effect*) then the paths of the random field will also not be continuous. Although such an effect cannot be discovered by finite sampling it is sometimes convenient to incorporate it into the model in order to describe *micro–scale variation*, which can be considered as being caused by some white noise process superposed to the underlying "continuous" random field (cf. also the discussion in Cressie (1991), p. 59). Another effect that is occasionally considered is the so–called *hole effect*, which produces some kind of periodic oscillations in the (semi–)variogram, corresponding to spatial correlations at fixed distances.

In a geostatistical analysis, a first goal is to estimate the (semi–)variogram function from data taken at measurement points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathsf{R}^d$. The "classical" estimator here is

$$\hat{\gamma}(\boldsymbol{h}) = \frac{1}{2m} \sum_{j=1}^{m} \left\{ \hat{Z}(\boldsymbol{x}_{i_j}) - \hat{Z}(\boldsymbol{x}_{i_j} + \boldsymbol{h}) \right\}^2 \tag{12}$$

where $\hat{Z}(\boldsymbol{x})$ denotes the observed value of the random field at the point $\boldsymbol{x} \in \mathsf{R}^d$, and $\boldsymbol{x}_{i_j} - \boldsymbol{x}_{i_{j-1}} = \boldsymbol{h}$ for all j (transsect sampling). In a second step, a curve fitting procedure is applied in order to estimate $\hat{\gamma}(\cdot)$ also at intermediate distances.

A second goal in the geostatistical analysis is a prediction of values of the random field at arbitrary points $\boldsymbol{x} \in \mathsf{R}^d$, on the basis of the complete set of data and the empirical (semi-)variogram observed. With the aid of this *Kriging*[1] *procedure*, contour maps of the underlying random field can be established. In *block Kriging*, the prediction of an averaged value

$$Z_V = \frac{1}{m(V)} \int_V Z(\boldsymbol{x})\, d\boldsymbol{x} \tag{13}$$

over some bounded and measurable region $V \subset \mathsf{R}^d$ is required. *Simple Kriging*, i.e. a pointwise prediction for $\boldsymbol{x}_0 \in \mathsf{R}^d$, is obtained from this by taking the limit $m(V) \to 0$ over regions V which contain $\boldsymbol{x}_0$; then also $Z_V \to Z(\boldsymbol{x}_0)$ if the random field has continuous paths. Usual statistical requirements for a "good" prediction $\hat{Z}_V$ are:

[1] named after the south–african statistician D.G. Krige

a) *Linearity:* $\hat{Z}_V = \sum_{i=1}^n \lambda_i Z(\boldsymbol{x}_i)$ with weights $\lambda_1, \dots, \lambda_n \in \mathbf{R}$;

b) *Unbiasedness:* $E[\hat{Z}_V] = E[Z_V] = \mu$, i.e. $\sum_{i=1}^n \lambda_i = 1$;

c) *Minimum Variance:* $E\big[(\hat{Z}_V - Z_V)^2\big] \to \min!$ subject to $\sum_{i=1}^n \lambda_i = 1$.

A solution of this Lagrangian problem can be obtained from the linear system

$$\boldsymbol{K\lambda} = \boldsymbol{\gamma} \qquad \text{or} \qquad \boldsymbol{\lambda} = \boldsymbol{K}^{-1}\boldsymbol{\gamma}, \tag{14}$$

where

$$\boldsymbol{K} = \left(\begin{array}{c|c} & 1 \\ [\gamma_{ij}] & \vdots \\ & 1 \\ \hline 1 \cdots 1 & 0 \end{array}\right) \qquad \boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \nu \end{pmatrix} \qquad \boldsymbol{\gamma} = \begin{pmatrix} \overline{\gamma}_V(\boldsymbol{x}_1) \\ \vdots \\ \overline{\gamma}_V(\boldsymbol{x}_n) \\ 1 \end{pmatrix}. \tag{15}$$

Here ν is a Lagrangian multiplier, and the matrix $[\gamma_{ij}]$ and the entries $\overline{\gamma}_V(\boldsymbol{x}_i)$ are given by

$$\begin{aligned} \gamma_{ij} &= \gamma\big(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|\big) \\ &\qquad [\text{or variogram estimate } \hat{\gamma}(\cdot) \text{ for } \gamma(\cdot), \text{ resp.}] \\ \overline{\gamma}_V(\boldsymbol{x}_i) &= \frac{1}{m(V)} \int_V \gamma\big(\|\boldsymbol{x}_i - \boldsymbol{y}\|\big)\, d\boldsymbol{y} \end{aligned} \tag{16}$$

For simple Kriging, i.e. $V = \{\boldsymbol{x}_0\}$, the last terms reduce to

$$\overline{\gamma}_V(\boldsymbol{x}_i) = \gamma\big(\|\boldsymbol{x}_i - \boldsymbol{x}_0\|\big). \tag{17}$$

The minimum variance of c) above can be expressed in terms of $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ as follows:

$$E\big[(\hat{Z}_V - Z_V)^2\big] = \boldsymbol{\gamma}^{tr}\boldsymbol{\lambda} = \boldsymbol{\gamma}^{tr}\boldsymbol{K}^{-1}\boldsymbol{\gamma}. \tag{18}$$

(For a more thorough discussion of the foundations of geostatistics, we refer the reader to the monographs of Cressie (1991), especially Chapter 3.2, and Journel and Huijbregts (1978) or Haining (1990).)

The following figures show variogram estimates for the Harpacticus data from Fig. 2 (sites 6 and 8), as well as contour maps for the corresponding random fields with simple and block Kriging. In the latter case, blocks V of four neighboured quadrats each were considered. All calculations were performed with the program GEO–EAS (Geostatistical Environmental Assessment Software). For the curve fitting procedure in the semi–variogram, a Gaussian model with nugget effect was used.

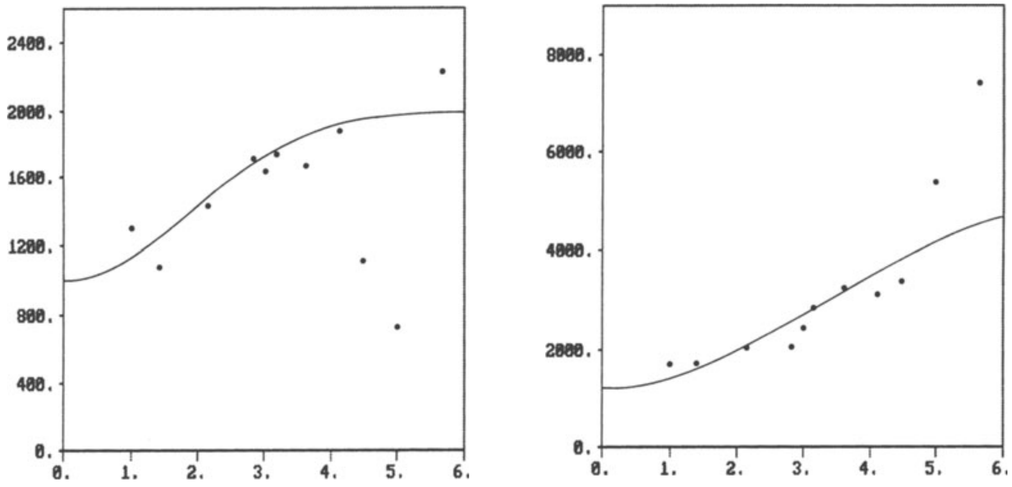


Fig. 4
semi–variogram estimates for Harpacticus obscurus, sites 6 and 8

Seemingly, in the semi–variogram for site 6 a hole–effect can be detected. This is due to the fact that in opposite corners of the data set (north–west and south–east), high count values are observed.

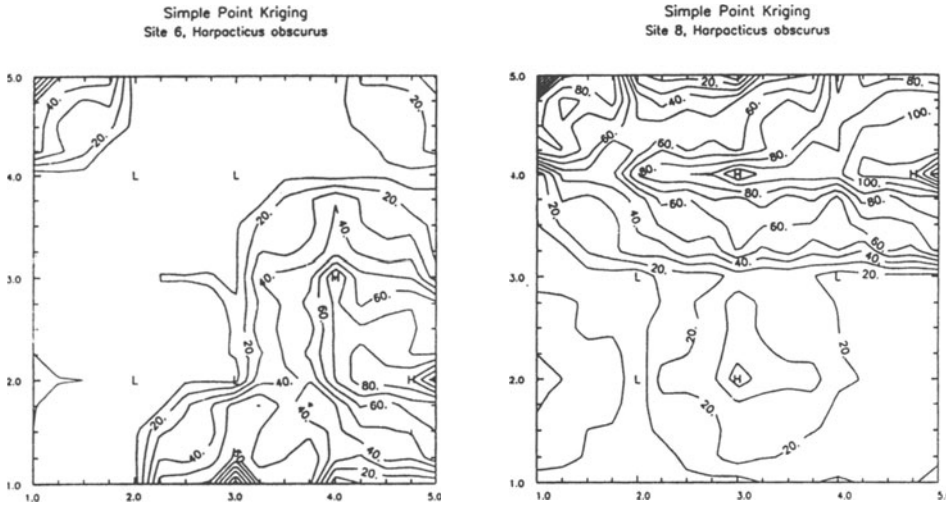


Fig. 5
contour maps for Harpacticus data: simple Kriging, sites 6 and 8

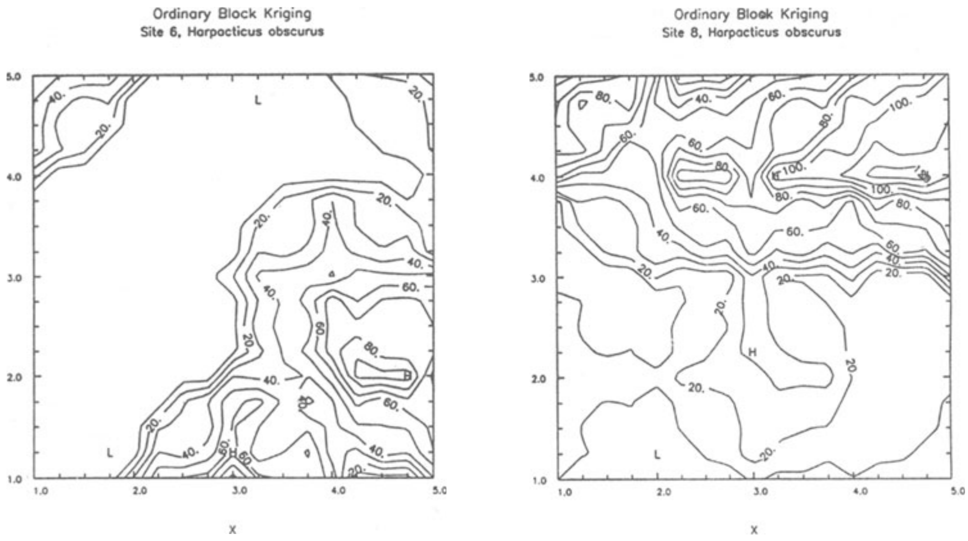


Fig. 6
contour maps for Harpacticus data: ordinary Kriging, sites 6 and 8

The last two figures show 3D–plots produced by a numerical interpolation of the Harpacticus data of the two sites (routine "inverse" of the program SYSTAT). A comparison with the corresponding contour plot above shows that the estimated distribution map is in good coincidence with this numerical data representation.

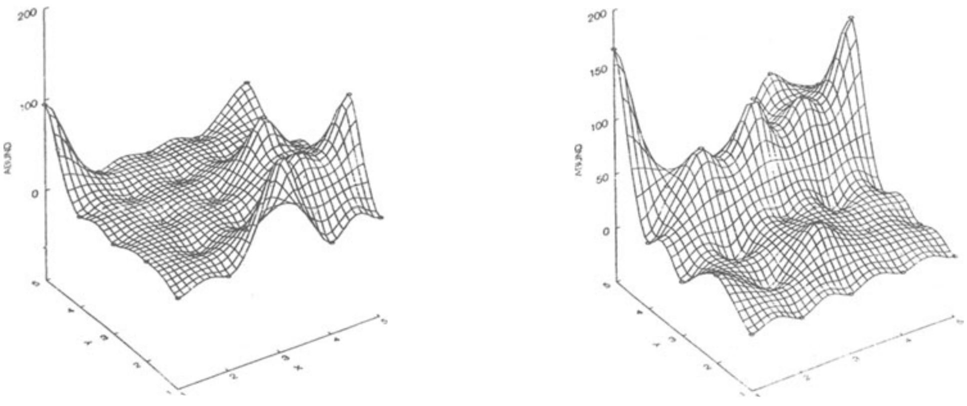


Fig. 7
3D–plots for Harpacticus data, sites 6 and 8

As a final comment it should be pointed out that in the case of count data (like above) the application of geostatistical methods might mathematically not be fully justified since the observed values are realizations of discrete distributions. However, at least in some approximative way, the analysis makes sense in order to obtain information about spatial dependence structures. Such information is e.g. necessary in order to determine *minimal areas* for probe schemes in field experiments, for instance in order to reduce the costs for necessary laboratory analyses. In the above examples, the structure of the semi-variograms shows that a reduction to a four by four probe scheme could be recommended for site 6, while at site 8, a comparable reduction might be inappropriate.

Acknowledgements. A part of the mathematical modelling presented in this paper has been developed and applied in the research project *Ökosystemforschung Niedersächsisches Wattenmeer (Pilotphase): Beiträge der Angewandten Statistik zur Bearbeitung von Maßstabsfragen und zur Versuchsplanung für die Analyse räumlicher Strukturen und dynamischer Vorgänge im Watt*, granted by the *Bundesminister für Forschung und Technologie*, Förderkennzeichen FZ 03F0023F. Special thanks are given to the following colleagues: W. Ebenhöh, ICBM, Universität Oldenburg, who supplied us with photographic material for Fig. 1; B. Hertweck, Forschungsinstitut Senckenberg, Wilhelmshaven, who provided the photograph for Fig. 3; and K.H. van Bernem, Forschungszentrum Geesthacht (GKSS), who contributed lots of data, among them the Harpacticus obscurus data from Fig. 2.

6. References

BEGON, M., HARPER, J.L., and TOWNSEND, C.R. (1990): *Ecology - Individuals, Populations, Communities.* Blackwell Sci. Publ., Oxford.

CRESSIE, N. (1991): *Statistics for Spatial Data.* Wiley, N.Y.

DIGGLE, P.J. (1983): *Statistical Analysis of Spatial Point Processes.* Academic Press, N.Y.

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

HAINING, R. (1990): *Spatial Data Analysis in the Social and Environmental Sciences.* Camb. Univ. Press, Cambridge.

JOURNEL, A.G. and HUIJBREGTS, C.J. (1978): *Mining Geostatistics.* Academic Press, London.

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

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

PFEIFER,D., BÄUMER, H.-P., and ALBRECHT, M. (1993): Moving point patterns - the Poisson case. In: O. Opitz and B. Lausen (eds.): *Information and Classification: Concepts, Methods and Applications.* Springer, N.Y., 248 - 257.

RICHTER, O. and SÖNDGERATH, D. (1990): *Parameter Estimation in Ecology. The Link between Data and Models.* VCH, Weinheim.

ROTHSCHILD, B.J. (1992): Application of stochastic geometry to problems in plankton ecology. *Phil.Trans.R.Soc.Lond. B, 336*, 225 - 237.

STOYAN, D., KENDALL, W.S. and MECKE, J. (1989): *Stochastic Geometry and Its Applications.* Wiley, N.Y.