Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

M.J. Rottschäfer, L.G. Barendregt

A statistical analysis of spatial point patterns
A case study

Department of Mathematical Statistics Report MS-R8813 September

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

A Statistical Analysis of Spatial Point Patterns

A case study

M.J. Rottschäfer & L.G. Barendregt
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Within the context of the project "statistical image analysis" of CWI we have studied some spatial point patterns that originated from biological observations. These observations were the positions of so called EGF-receptors on the surface of human carcinoma cells.

We have put forward a stochastic model for these point patterns. Since the EGF-receptors appear in clusters on the cell surface, we have opted for the Poisson-cluster-process as the model. We estimated the three parameters in this process by means of a method described by Diggle. We did also some work in assessing the statistical reliabiblity of our estimates.

1980 Mathematics Subject Classification: 62M07, 62M09, 62P10, 60G55.
Key Words & Phrases: spatial point process, complete spatial randomness, Poisson cluster process, point to nearest event distances, nearest neighbour distances, *K*-function.

1. Introduction.

The objects of our study are the spatial point patterns, formed by immunogold-labeled epidermal growth factor (EGF) receptors on the surface of A431 human epidermoid carcinoma cells. The cells have been treated with EGF for varying periods of time. Electron micrographs of immunogold-labeled cells were digitized, and the coordinates of goldlabeled receptors were interactively determined by computer. We have used these data, which originated from a joint project of the Department of Molecular Cell Biology of the University of Utrecht and the Netherlands Institute for Developmental Biology (Hubrecht Laboratory), for our statistical analysis.

However, it must be noted that the method, known as immunogoldlabeling only labels 50-80% of the receptors. Even worse is that one does not know how these labeled receptors have been selected. For this reason one has to be careful with the interpretation, especially with the biological meaning, of the results.

For a description of the knowledge that is currently available on the rate of EGF-receptors we refer to VAN BELZEN e.a. (1988).

For an understanding of the significance of our statistical analysis it is of help to know the following of the biological background.

Growth factor receptors play an important role in the regulation of cell growth and differentiation. The electron microscopical images give the impression that, after treating cells with EGF, these receptors have a tendency to aggregate. One suspects that there is a connection between this aggregation and the cellular response to EGF. Previously these effects were investigated using a modified Poisson variance test. We, however, have chosen for a different method, by means of which it was tested whether the observed point pattern might be generated by a 2-dimensional homogeneous Poisson-process (or, as it is formulated in the biomedical literature, whether the point-pattern was "random" or "non-random"). We have investigated the data mentioned above by means of a different method but with the same objective (homogeneous Poisson vs. unspecified alternative). We found that none of the observed point-patterns could be thought of as being generated by a Poisson process; they all showed a tendency towards clustering.

The question arises then which random process might give a good description of the observations, or in other words, which random process might generate point patterns that are very much alike to

the observed point patterns.

In the construction of such a process one can follow two lines of thought:

1. generate clusters of points
2. generate individual points and let them move as if there were a pairwise mutual attraction between them.

We have followed the first line of thought and opted for the Poisson cluster process, because this process has already been described in the literature, and because it is mathematically tractable.

2. Models and methods

We have started our analysis with a test on complete spatial randomness (CSR), also called a homogeneous Poisson point process, by using a graph based on the empirical distribution function of the 'point to nearest event distances' and 'nearest neighbour distance' respectively. In case of rejection of CSR these graphs could give us some information about a possible alterative point process. A homogeneous Poisson point process has the following properties:

(1) For any finite region, the number of events in that region has a Poisson distribution with mean $\lambda|A|$, for some constant $\lambda>0$ and $|A|=$ the area of region A.

(2) Conditional on the number of events in A, the events are distributed according to a uniform distribution on A and independently w.r.t. each other.

We will give a short introduction to the distance-based methods which have been used.

2.1. First method:

Choose a regular grid of m points and calculate in each of those points the distance x_i, $i=1,\ldots,m$, to the nearest event. Diggle (1983) makes a distriction between a point of the grid and a realization of the point process by calling them 'point' and 'event' respectively. We will do the same.

Define

$$\hat{F}_j(x)=m^{-1}\#(x_i\leqslant x), \tag{2.1.1}$$

the empirical distribution function (EDF) of the point to nearest event distance of our dataset ($j=1$) and of $s-1$ independent simulations ($j=2,\ldots,s$) of CSR, conditional on the number of events. We also define

$$\bar{F}_j(x)=(s-1)^{-1}\sum_{k\neq j}\hat{F}_k(x), \tag{2.1.2}$$

for $j=1,\ldots,s$. Now we can compare $\hat{F}_1(x)$, the EDF of our dataset, with the mean of simulation-EDF's, by plotting $\bar{F}_1(x)$ against $\hat{F}_1(x)$. In case of CSR the plot should be roughly linear. To create some sort of confidence band, we draw

$$L(x)=\min_{j=2,\ldots,s}\hat{F}_j(x), \tag{2.1.3}$$

the lower simulation envelope, and

$$U(x)=\max_{j=2,\ldots,s}\hat{F}_j(x), \tag{2.1.4}$$

the upper simulation envelope, in the same plot. Finally, for a formal test of CSR we use the rank of the test statistic within the sequence U_1, with U_j defined as

$$U_j=\int\{\hat{F}_j(x)-\bar{F}_j(x)\}^2dx,\ j=1,\ldots,s, \tag{2.1.5}$$

which can be seen as a measure for the discrepancy between $\hat{F}_j(x)$ and $\bar{F}_j$— the mean of all the other EDF's — through the whole range of x.

2.2. Second method:
An other way of investigating a spatial point pattern is by looking at the nearest neigbour distances: Suppose that there are n events. For every event we calculate the distance y_i, $i=1,\ldots,n$ to the nearest other event. We now consider

$$\bar{G}_j(y)=n^{-1}\#(y_i\leqslant y),\ j=1,\ldots,s, \tag{2.2.1}$$

the EDF's of the nearest neighbour distance, and

$$\bar{G}_j(y)=(s-1)^{-1}\sum_{k\neq j}\hat{G}_k(y),\ j=1,\ldots,s. \tag{2.2.2}$$

with, just as in § 2.1, $j=1$ refering to the set of observed events and $j=2,\ldots,s$ refering to simulations. Now we can plot $\bar{G}_1(y)$ against $\hat{G}_1(y)$ with their lower- and upper simulation envelopes. Again a formal test has been done, this time with the rank of w_1, where w_j is defined as

$$w_j=\int\{\hat{G}_j(y)-\bar{G}_j(y)\}^2dy,\ j=1,\ldots,s, \tag{2.2.3}$$

completely analogously to the previous case.

The main purpose of the graphs is to compare our point process with the chosen model, which, in this case, is a Poisson point process. However, the plots might provide us additional information. In case of rejection of *CSR* the deviation in the plots from a straight line could give us some idea about an alternative model.

2.3. In the previous analysis the plots indicated clustering, as will be explained when we discuss the results, so we decided that a Poisson cluster process could be a possible alternative. We have chosen for a simple version of this model with the following properties:

(1) Parent events form a Poisson process with intensity ρ, so $\rho|A|$ is the expectation of the number of parents in region A.
(2) Each parent creates a stochastic number S of offspring, distributed independently and identically according to a Poisson distribution with parameter μ, so the intensity of the Poisson cluster process is $\lambda=\rho\mu$.
(3) The location of the offspring with respect to the related parent is determined by a bivariate normal distribution with density function

$$h(x,y)=(2\pi\sigma^2)^{-1}\exp(-(x^2+y^2)/(2\sigma^2)). \tag{2.3.1}$$

REMARK. The parents do not take part in the final pattern. They only serve to determine the location of the offspring. We will try to estimate the parameters ρ and σ by using the K-function

$$K(t)=\lambda^{-1}E\{\#\text{ events within distance } t \text{ of some other event}\}, \tag{2.3.2}$$

for several values of t.
The theoretical expression of $K(t)$ in our model is known to be

$$K(t)=\pi t^2+\rho^{-1}(1-\exp(-t^2/(4\sigma^2))). \tag{2.3.3}$$

Let us define t_{ij} as the distance between the events z_i and z_j, and let

$$I_t(t_{ij})=\begin{cases}1, & \text{if } t_{ij}\leqslant t\\ 0, & \text{otherwise,}\end{cases} \tag{2.3.4}$$

then

$$\sum_{i\neq j}\sum I_t(t_{ij}) \tag{2.3.5}$$

is the number of inter-event distances $\leqslant t$. A (not completely unbiased) estimator for $K(t)$ is

$$\hat{K}(t) = n^{-2}|A|\sum_{i\neq j}\sum w_{ij}^{-1} I_t(t_{ij}), \tag{2.3.6}$$

where w_{ij}^{-1} is meant as a correction for edge-effects: $w_{ij}(t)$ is the conditional probability that an event z_j is observed, given that it is at a distance t from a certain other event z_i. Expressions for $w_{ij}(t)$ for a rectangular region can be found in DIGGLE (1983).

The idea is to estimate the unknown parameters ρ and σ of the model by minimizing

$$D(\theta) = \int_0^{t_0} \{\hat{K}(t)^c - (K(t;\theta))^c\}^2 dt, \ \theta=(\rho,\sigma), \tag{2.3.7}$$

for some t_0 and some c. An initial estimate for θ can be found by assuming that $\hat{K}(t)-\pi t^2$ has its maximum $1/\rho$ for $t=2.\sigma$. By choosing several values for t_0 one can concentrate on certain scales of clustering. The constant c only serves to "dampen" the effect of the statistical variation of $\hat{K}(t)$, which becomes more pronounced as t increases. For minimizing $D(\theta)$ we have used the NAG-routine E04CGF.

2.4. The question arises what sort of reliability could be attached to these estimates. This question can be put more precisely as follows:

If one generates a point pattern according to a Poisson-cluster-process, with known parameter-vector $\theta_0=(\mu_0,\rho_0,\sigma_0)$, and one estimates the parameter-values by means of the method described in section 2.3, what can be said about the deviation of the estimated $\hat{\theta}$ from the original θ_0?

One might answer this question by generating, say, 100 point-patterns where the unknown θ_0 is substituted by $\hat{\theta}_d$, the estimate for θ from the dataset and re-estimating θ each time. This yields 100 re-estimates $\hat{\theta}_{(r)i}$ ($i=1,\ldots,100$; r indicating re-estimation), which one might compare with $\hat{\theta}_d$, and out of which one might compute the variance-covariance matrix. The estimation-method for θ is however too expensive to let it run a 100 times. To be able to say at least something about the distribution of $\hat{\theta}_{(r)}-\hat{\theta}_d$ (which should be close to the distribution of $\hat{\theta}-\theta_0$). R.D. Gill came up with the following idea:

Let X_d be our dataset, and let $X\sim P_\theta$ be a short notation for "X is distributed according to a Poisson cluster process with parameter θ". Let $\hat{\theta}$ be the solution of minimizing $D(X;\theta)$ over θ for a simulation $X\sim P(\theta_0)$. Furthermore, let $\hat{\theta}_{(r)}$ be the solution of minimizing $D(X^*,\theta)$ for a simulation $X^*\sim P_{\hat{\theta}_d}$. Now define

$$g(X;\theta) = \frac{\partial}{\partial\theta} D(X;\theta) \tag{2.4.1}$$

and suppose that $g(X;\theta)$ is a smooth function of θ. Tayloring $g(X;\hat{\theta})$ around θ_0 gives

$$0 = g(X;\hat{\theta}) \approx g(X;\theta_0) + \left[\frac{\partial}{\partial\theta} g(X;\theta)\right]_{\theta_0} (\hat{\theta}-\theta_0) \tag{2.4.2}$$

$$\hat{\theta}-\theta_0 \approx -\left[\frac{\partial}{\partial\theta} g(X^*;\theta)\right]^{-1}_{\hat{\theta}_0} \cdot g(X;\theta_0) \tag{2.4.3}$$

Now estimate $\frac{\partial}{\partial\theta} g(x;\theta)|_{\theta_0}$ by $\frac{\partial}{\partial\theta} g(x;\theta)|_{\hat{\theta}_d}$ and hence the distribution of $\hat{\theta}-\theta_0$ by the bootstrap distribution (i.e. under $X^*\sim P_{\hat{\theta}}$) of

$$-\left[\frac{\partial}{\partial\theta} g(X;\theta)\right]^{-1}_{\hat{\theta}_d} g(X^*;\theta_0) \tag{2.4.4}$$

The advantage of this estimation-procedure is, that one avoids unnecessary minimization of $D(\theta)$. This only has to be done once.

3. Description of the data

We have applied the statistical methods described in section 2 to five photographical pictures, each representing the surfaces of a different human cell. Each of the cells had been "treated" in a different way. The experimental treatment of a human cell consisted of an exposure to the epidermal growth factor (EGF) during a certain amount of time.

The scheme of the pictures was as follows:

indication	duration of exposure to EGF
B041	0
B014	30 sec
B015	60 sec
B019	90 sec
B012	240 sec

Table 1.

Each picture was a square of 1024×1024 pixels, out of which we used only a square of 1000×1000 pixels.

On each picture there were a few hundred EGF-receptors visible, ranging from 170 for B015 to 415 for B014. We have represented each picture by a plot of the position of the EGF-receptors. These plots have been included in this report in Appendix I.

4. Results of the tests for complete spatial randomness.

For each of the five pictures described in section 3, we applied two tests for CSR, namely

a. the one based on point-to-nearest-event-distances (see section 2.1)
b. the one based on event-to-nearest-event-distances (see section 2.2).

The first method did not always yield a clear-cut-result, but the test based on event-to-nearest-event-distances yielded a decisive result for each of the five pictures, namely rejection of CSR in favour of clustering.

In order to illustrate this we have given in figure 1 a graphical representation of both tests for one picture, namely B014 (30 seconds exposure). The plot of $\bar{F}_1(X)$ against $\hat{F}_1(X)$ is nearly diagonal, $\bar{F}_1(X)$ being outside its (0.01 - 0.99) limits only for large values of $F_1(X)$. The plot of $\bar{G}_1(X)$ against $\hat{G}_1(X)$ on the other hand runs far beyond its 99%-limit for the greater part of its range.

This means that CSR is rejected in favour of clustering. Near the origin however $\bar{G}_1(x)$ behaves in the opposite way: its graph runs initially horizontally and touches the 1%-limit. This means that the minimum of the event-event-distances is lower than could be expected from CSR, and suggests therefore that there is a repulsion between the EGF-receptors at very short distance.

We think that the test results point to clustering in each of the five figures, even though the results of the point-to-nearest-event-method is not always significant. The latter test is powerful for alternative processes with a preference for large void areas, a process that doesn't seem appropriate for the EGF-receptors.

5. Estimation results

We applied the estimation method, described in section 2.3 to the 5 point-patterns of the pictures B041, B014, B015, B019 and B012.

The estimates for the parameters μ, ρ and σ have been tabulated in table 2.

Identification	$\hat{\mu}$	$\hat{\sigma}$	Duration of exposure
B041	0.49	0.0093	0 sec.
B014	1.07	0.0165	30 sec.
B015	1.85	0.0136	60 sec.
B019	1.44	0.0120	90 sec.
B012	0.98	0.0104	240 sec.

TABLE 2
Estimates for μ and σ for the 5 pictures of series B

For the coefficient c in the formula for $D(\theta)$ we choose 0.25 (thereby following the suggestion by Diggle) and for t_0 0.05 $\times$ the side of the square within which the points lie.

In the remaining part of this report we shall discuss the following subjects:
- the biological meaning of the estimated parameter-values, more especially those of μ and σ.
- the choices for c and t_0 in the formula for $D(\theta)$.
- the examination of the minima $D(\theta)$.

In a separate section we shall take up the matter of the statistical reliability of estimates, both from the biometrical point of view (differences between different cells with the same treatment) and from the mathematical statistical point of view (variance due to the estimation method).

5.1. The biological meaning of the parameter-estimates
The estimates for σ (the dispersion) are

0.0093 for B041 (no exposure)
about 0.015 for the other four cells (exposed to EGF)

Since a distance of 0.001 on the picture corresponds to a distance of 1.9 nm on the cell, a σ of 0.015 on the picture corresponds to 28 Nm on the cell. The mean distance of a daughterpoint to its parent-point is about $\sigma\sqrt{2}$, which is 0.013 for B041 and 0.021 for the other 4 cells.

The range of one cluster is therefore small as compared to the size of one picture.

The estimates for μ (the expected number of daughters per parent) were smaller than might be expected from the illustrations in the literature on cell-biology that have come to our attention. Their order of size is 1, ranging from

0.49 for B041 (no exposure)
to 1.85 for B015 (exposure for 1 minute)

Such values give the impression that the mean cluster-size is 1 and that there is therefore no clustering at all! This would be in contradiction with the results of the tests on Complete Spatial Randomness, which clearly pointed towards clustering. Two comments can be made here.

In the first place it is likely that a part of the EGF-receptors has not been observed, for instance because they were not labeled. If we assume that the positions of the EGF-receptors are generated by a Poisson-cluster-process, and that the observability of an EGF-receptor is independent of its position, then the positions of the observed EGF-receptors are situated according to a Poisson-cluster-process with the same ρ and σ as the original one, but with a reduced μ. If a fraction p of the EGF-receptors is actually observed, then

$$\mu \text{ (observed receptors)} = p \cdot \mu \text{ (receptors present)}$$

Our values for μ might therefore be under-estimates.

In the second place μ is not the mean clustersize, but the mean number of daughters per parent. A parent with 0 daughters does not yield a cluster. If somebody would perform a cluster-analysis on the

data, and calculate the mean number of points per cluster, he would only take account of daughter-points of those parents who have generated at least one daughter.

The following example may clarify this. For picture B014 (30 second exposure) the estimates are: $\hat{\mu}=1.07$ $\hat{\rho}=388$ and $\hat{\sigma}=0.0165$. For a Poisson-cluster-process with $\mu=1.07$ $\rho=388$ and $\sigma=0.0165$ the mean number of parents with 0 daughters is $\rho e^{-\mu}=388*0.343=133$. The mean number of observable clusters is $388-133=255$ rather than 388.

The mean number of offspring of these 255 parents is $\mu[1-e^{-\mu}]=1.63$ rather than $\mu=1.07$. Out of these 255 parents, 56% has one daughter and 44% more than one daughter.

The expected number of daughters without sister is $0.56*255=142$ and the total number of daughters is $388*1.07=415$. The fraction of daughters without sister is $0.343=e^{-\mu}$.

These figures are not in contradiction with the impression that one gets while looking at B014. There are indeed quite a few isolated points.

We generated a point-pattern with a Poisson-cluster-process with $\mu=1.07$, $\rho=388$ and $\sigma=0.0165$. See figure 2.

The circle in the lower right corner has a radius of 0.0165 and indicates therefore the size of a cluster. To be more precise: within a cluster 40% of the points (on the average) lie within a circle of that size around the center of the cluster.

5.2. The comparison of the 5 pictures

As has already been mentioned in section 5.1, the estimate for σ was 0.0093 for B041 (no exposure) and about 0.015 for the other 4 cells (exposed to EGF).

In figure 3 the 5 estimates for μ (expected number of daughters per parent) have been plotted against the duration of the exposure. This figure suggests a curvilinear relation between duration of exposure and μ, with μ being maximal for an exposure of about 1 minute.

5.3. The choices for c and t_0

The variance of $\hat{K}(t)$ is not constant for t between 0 and t_0; it increases for increasing t as will be discussed in section 7.

For the minimalization of an integral of squares such as $D(\theta)$ it is of importance that the integrand has its variance as constant as possible.

This goal can be approximated by applying a power transformation to $\hat{K}(t)$ and $K(t)$, with a power-coefficient between 0 and 1. Diggle suggests taking a coefficient $c=0.5$ if there is a moderate clustering in the pattern, and $c=0.25$ if there is a strong clustering. We have opted for $c=0.25$, because the testresults for CSR suggested a strong clustering.

We repeated all our computations with the choice $c=1$ instead of $c=0.25$, just in order to see how this would influence the outcomes.

In the computations with $c=0.25$ no difficulties of numerical nature were encountered. With $c=1$, on the contrary, sometimes minima were found with values for ρ and σ that were not credible. In other computations with $c=1$, no minimum was found at all.

For t_0, the right-hand-bound of the integration-interval, we have chosen the value of 0.05 $\times$ the side of the square, within which the points lie. This is smaller than what Diggle suggests (0.25), but larger than each of the estimates for σ ($\sim$ 0.015). The reason for our choice is our experience, that for each of the 5 pictures $\hat{K}(t)-\pi t^2$ fluctuates considerably for $t>0.05$, and even becomes <0 for some of them, whereas its expectation is positive.

Phenomena of this kind are also reported by Diggle in his fig. 3.1, and also in his fig. 3.9.

5.4 Examination of the minima of $D(\theta)$

Whenever a function of several parameters is minimized by means of a numerical algorithm, the possibility exists that the attained minimum is not the global minimum but a local minimum or only a saddle point. One should examine the behaviour of the function by graphical means, if one wants to be certain that the global minimum has indeed been obtained.

We have examined the behaviour of $D(\theta)$ for one of the five point patterns, namely B015.

In the first place we have examined two perspective plots of the function $D(\rho,\sigma)$: one on a relatively large area of the parameter-space, namely

$$13 \leqslant \rho \leqslant 190$$

$$0.0015 \leqslant \sigma \leqslant 0.031$$

(fig 4.)
and one on a relatively small area in the parameter-space:

$$90 \leqslant \rho \leqslant 140$$

$$0.006 \leqslant \sigma \leqslant 0.016$$

(fig 5)
This last plot shows clearly that $D(\rho,\sigma)$ has only one local minimum (which is therefore the global minimum) and that this minimum is equal to the one that was calculated ($\rho=92$, $\sigma=0.0136$).

The contourlines are oblong and banana-shaped. This means that there are parameter-combinations (ρ,σ) that are rather different from (92, 0.0136) but that nevertheless yield almost the same function-value $D(\rho,\sigma)$. An example is $(\rho=120,\sigma=0.12)$ for which $D(\rho,\sigma)$ is only 5% larger than the minimal D.

6. The differences between cells with equal treatments

In section 5.2 the estimates for μ of the 5 pictures have been compared to each other. In order to see whether these differences mean anything, one should compare them with the differences of μ's of a number of cells who all have been treated in the same way. The cells are namely the units of treatment.

There were some pictures available to us of cells that were not exposed to EGF. We have analyzed 12 of them in the same way as the 5 pictures described before.

The pictures differed from B041 (also without EGF-exposure) in that they were prepared by a different investigator.

The tests for CSR, using the event-to-nearest-event-distances gave a significant result for all three pictures.

The parameter-estimates were:

	n	μ	σ	ρ
A011	174	0.40	$10.3 \star 10^{-3}$	435
A028	112	0.40	$7.84 \star 10^{-3}$	277
A029	107	0.27	$7.48 \star 10^{-3}$	398
A030	71	0.33	$9.13 \star 10^{-3}$	213
A034	275	0.80	$11.6 \star 10^{-3}$	345
A038	284	0.57	$11.9 \star 10^{-3}$	497
A043	360	0.36	$10.9 \star 10^{-3}$	1010
A052	329	1.09	$11.5 \star 10^{-3}$	329
A072	77	0.45	$9.2 \star 10^{-3}$	172
A076	260	0.53	$12.5 \star 10^{-3}$	487
A083	197	0.50	$12.6 \star 10^{-3}$	394
A084	212	0.34	$10.5 \star 10^{-3}$	620

Table 3

The variation in the values for σ is not very large. The coefficient of variation of the 12 estimates for σ is 16%.

If this coefficient of variation is representative for what will be found in any other series of replications, this has favourable consequences for the design of future comparative experiments. If two treatments have to be compared with each other with respect to σ, and both treatments are replicated 5 times (on different cells), a relative difference of 23% between the the mean of the σ-values of the first treatment and the mean of the σ-values of the second treatment is already significant (two-sided Student-test; $p<0.05$).

The variation coefficient of the 12 estimates for μ is much larger than the one for σ, namely 40%. A reason for this large spread might be that the detection-rate for the EGF-receptors varies from one cell to another, even if those cells have been prepared by the same investigator.

7. The variability of $\hat{K}(t)$.

We have, for each of the pictures analyzed by us, made a plot of $\hat{K}(t)-\pi t^2$ against t.

Our primary aim was model-validation: that is: comparison of $\hat{K}(t)-\pi t^2$ with its expectation

$$K(t)-\pi t^2 = \frac{1}{\rho}(1-e^{-\frac{t^2}{4\sigma^2}})$$

This function is increasing, and almost $=1/\rho$ beyond $t=3.5\sigma$.

$\hat{K}(t)-\pi t^2$ is however far from constant for larger t. It fluctuates wildly and becomes sometimes negative. We have included one example of such a plot (fig. 6).

Diggle does report this kind of behaviour as well; see his fig. 31 and also his fig. 39. This means that $\hat{K}(t)-\pi t^2$ can't be of much value for estimating θ, if $t>2\sigma$.

The large deviations of $\hat{K}(t)$ from its expectation $K(t)$ for larger values of t can be explained by the rapidly increasing variance of $\hat{K}(t)$ for increasing t.

The formula (2.3.6) for $\hat{K}(t)$ is a sum over $n(n-1)$ terms, with n being stochastic, having at least the Poisson-variance. For large t $var(\hat{K}(t))$ becomes proportional to $K^2(t)$ rather than proportional to $K(t)$.

8. Variability due to the estimation method

We might put forward the question how large the variance is of the estimators for μ, ρ and σ, obtained by applying Diggle's method, described in section 2.3. This variance is of a purely mathematical statistical origin, and is therefore to be distinguished from the "biometrical" variance discussed in section 6.

We have estimated the variance of the parameter-estimators for one data-set, namely the point-pattern of A028 (see section 6).

We did this according to the method described in section 2.4. We generated 100 point-patterns according to a Poisson-cluster-process with parameters $\mu=0.40$, $\sigma=0.00784$, $\rho=0.000377$ — the values that were obtained as estimates for this picture. For each of the generated point patterns we calculated:

$$\hat{K}(t),\ D(\theta),\ g_\rho \equiv \frac{\partial D}{\partial \rho} \text{ and } g_\sigma \equiv \frac{\partial D}{\partial \sigma},$$

where the derivates of $D(\rho,\sigma)$ were taken at the point ($\rho=0.000277, \sigma=0.00784$).

From those 100 values for g_ρ and g_σ we calculated their variance-covariance-matrix, and from that matrix we could calculate approximative values for the variances of $\hat{\rho}$ and $\hat{\sigma}$, using formula (2.4.4).

For the standard deviations of $\hat{\rho}$ and $\hat{\sigma}$ we found

$$st(\hat{\rho}) = 0.000095 \quad (\hat{\rho}=0.000277)$$

$$st(\hat{\sigma}) = 0.00163 \quad (\hat{\sigma}=0.00784)$$

We approximated the variance of $\hat{\mu}$ by means of:

$$Var(\hat{\mu}) = var(n/\tilde{\rho}|A|) =$$

$$\frac{1}{|A|^2}\{\frac{var(n)}{\hat{\rho}^2} - \frac{2cov(n,\hat{\rho}).En}{\hat{\rho}^3} + \frac{var(\hat{\rho}).(En)^2}{\hat{\rho}^4}\}$$

For $var(n)$ we took the variance of the numbers n of points in the 100 simulated point-patterns.

$cov(n,\hat{\rho})$ can be estimated by means of

$$cov(n,\hat{\rho}) = b_{\rho\rho}.cov(n,g_\rho)+b_{\rho\sigma}cov(n,g_\sigma)$$

with $b_{\rho\rho}$ and $b_{\rho\sigma}$ elements from the 2×2 matrix $B=-(\frac{\partial^2 D}{\partial\theta})^{-1}$, and $cov(n,g_\rho)$ and $cov(n,g_\sigma)$ esitmated from the simulations.

For the standard deviation of $\hat{\mu}$ we found:

$$st(\hat{\mu}) = 0.148 \quad (\hat{\mu}=0.40)$$

It is natural to compare the variability due to the estimation method with the biometrical variability, which has been described in section 6. One might expect the biometrical variability to be the bigger one as is the case on so many occasions but for σ the opposite is true.

The coefficient of variation for σ due to the estimation-method is 0.00163/0.00784=21%, whereas the biometrical coefficients of variation is 16%.

The case for μ is about 40% for both sources of variation.

9. Discussion

The method for the estimation of the parameters μ, ρ and σ of the Poisson-cluster-process appears to work well in practice; that is: the convergence to the minimum of $D(\theta)$ is straightforward in most examples that we have encountered, and the variance of the estimators is not that large, that it would inhibit application of the method in practice.

A condition for success is however that the preliminary assumptions are credible for the point pattern that has to be analyzed. The density of the points must be homogeneous and there must be strong evidence for clustering, preferably by means of a statistical test of Complete Spatial Randomness vs. clustering.

The length t of the integration-interval in the formula for $D(\theta)$ should be chosen rather small, because of the huge fluctations of $\hat{K}(t)$ for large t. If one has any idea of how large σ might be, t_0 should not be chosen larger than 3σ.

The Poisson-cluster-process is a process with secondary clustering. Some cell-biologists have argued that EGF-receptors exercise some direct pairwise attraction onto each other.

A mathematical model for the positions of EGF-receptors should, in that view, have a primary clustering-mechanism built in rather than a secondary clustering mechanism.

Such a model is the Gibbs-point-process with pairwise potential. In this model it is assumed that between each pair of points (x_i,x_j) there exists a potential $V(x_i,x_j)$. $V(x,y)$ depends only on the distance $|x-y|$. The probability density for a point-pattern with n points at positions $x_1,x_2,\cdots,x_n$ is proportional to

$$\exp[-\sum_{i<j} V(x_i,x_j)]$$

Point-patterns with a low total potential have a preference in comparison with point patterns with a large total potential.

A popular choice for $V(x,y)$, which seems very suitable for for EGF-receptors, is

$$V(x,y) = +\infty \text{ if } 0<|x-y|<h$$
$$V(x,y) = -b \text{ if } h<|x-y|\leq R$$
$$V(x,y) = 0 \text{ if } |x-y|>R$$

There is a preference for point-patterns with many pairs of points less than R apart. It is however not allowed that any pair of points comes closer than h to each other.

There have been some statistical methods proposed for the estimation of the parameters h,b and R (see STOYAN e.a. (1987)). It seems appropriate to apply these methods to the point-patterns that have been analyzed in this study.

REFERENCES

N. VAN BELZEN (1988). Direct Visualization and Quantitative Analysis of Epidermal Growth Factor-Induced Receptor Clustering Journal of Cellular Physiology *134,* pgg. 413-420.

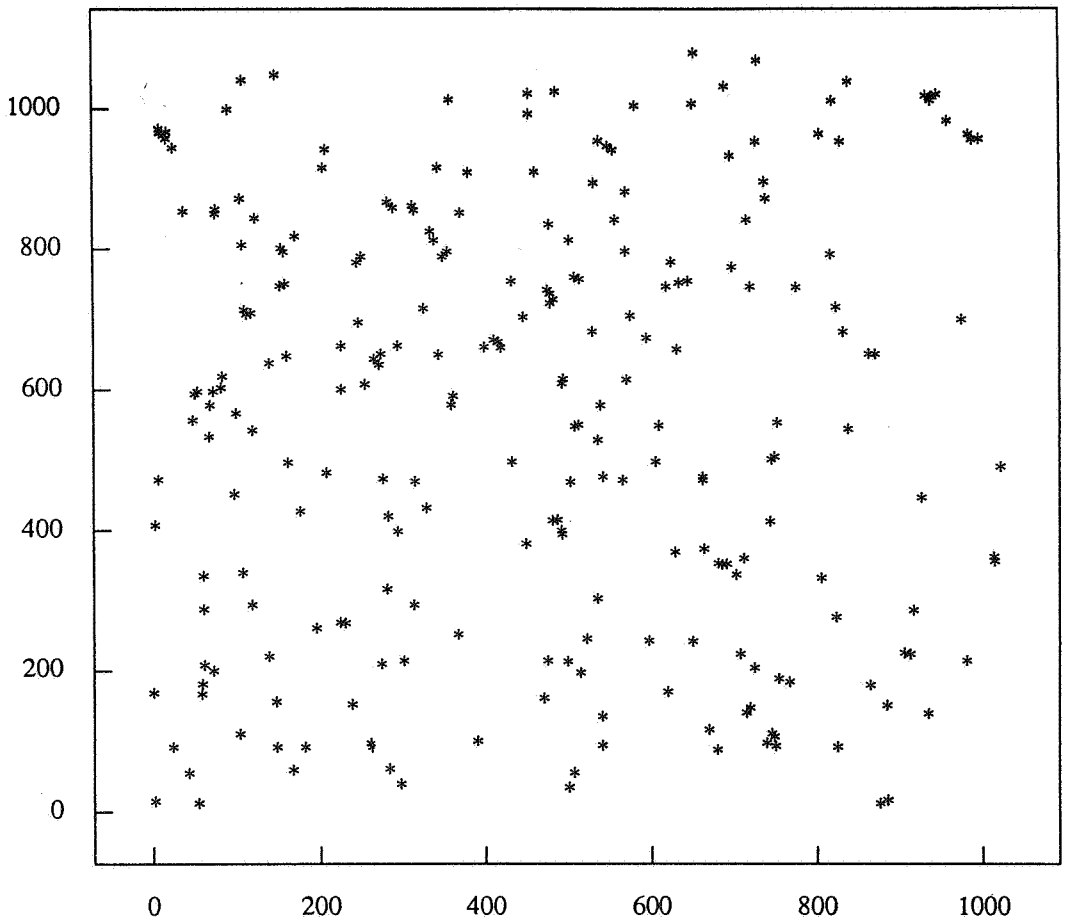
P.J. DIGGLE, (1983), Statistical Analysis of Spatial Point Patterns, Academic Press.

D. STOYAN, W.S. KENDALL, J. MECKE, (1987), Stochastic Geometry and its Applications, John Wiley & Sons.

Appendix I

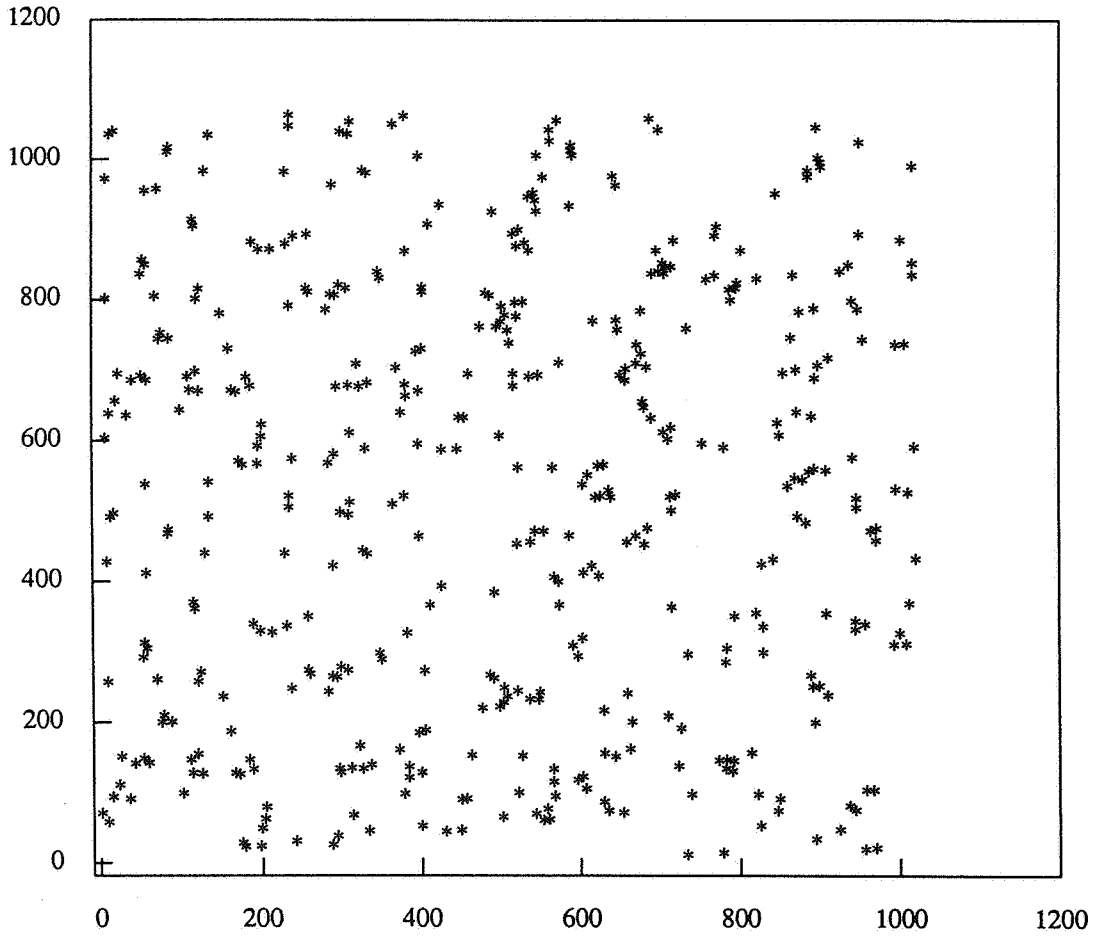
12

B041



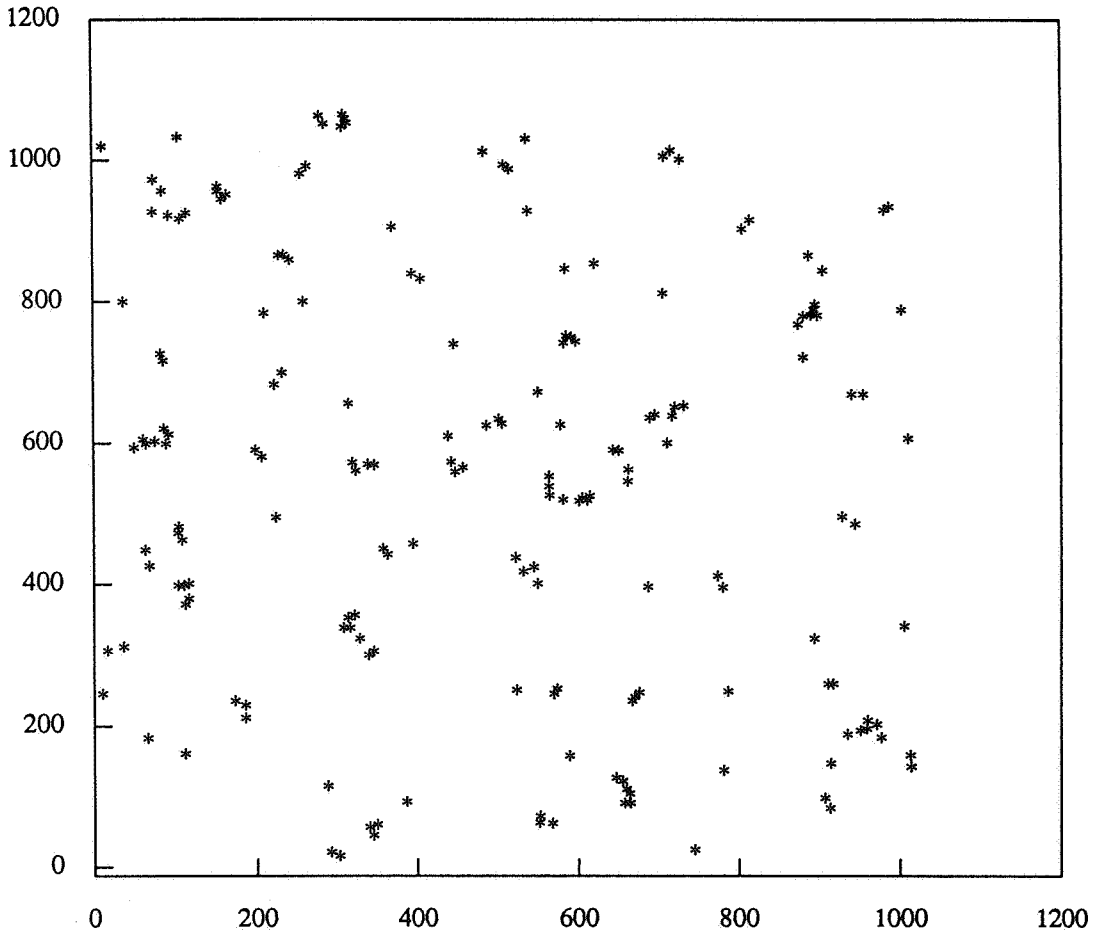
no exposure to EGF

B014



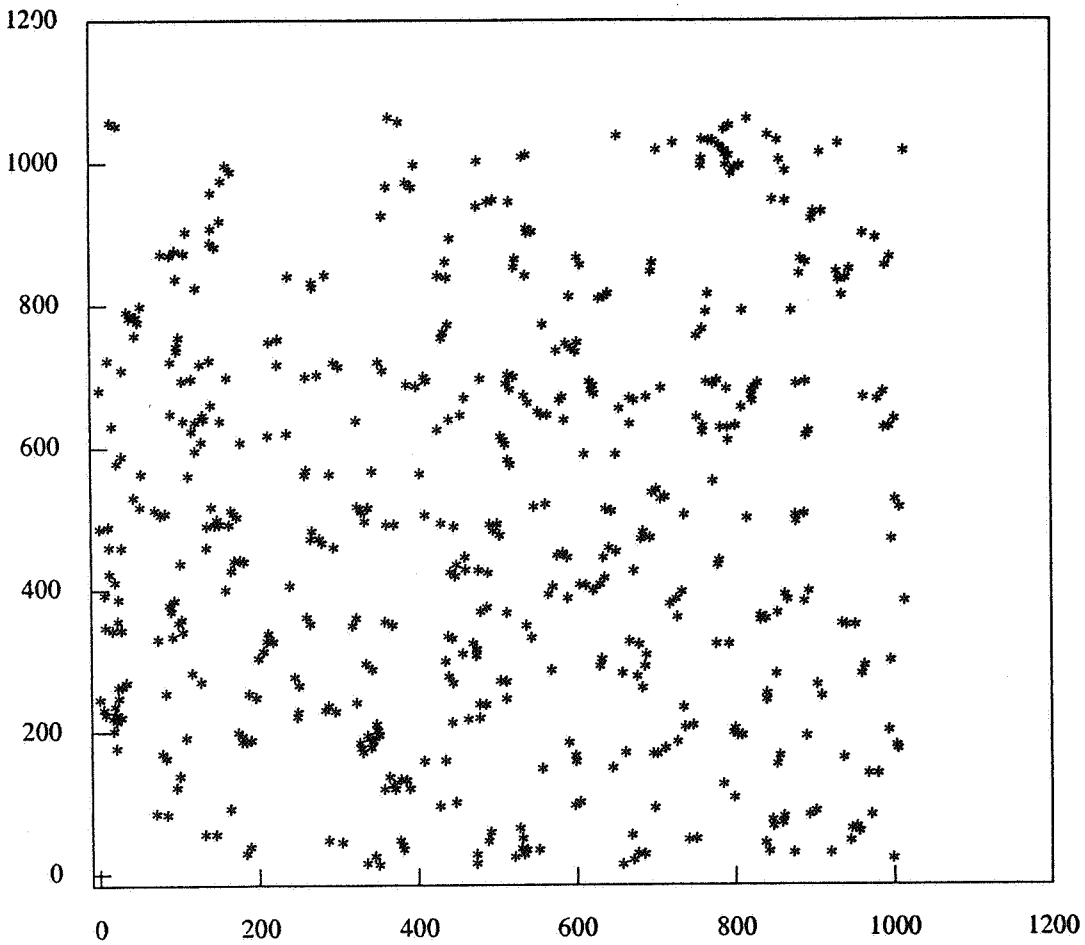
exposure for 30 seconds

B015



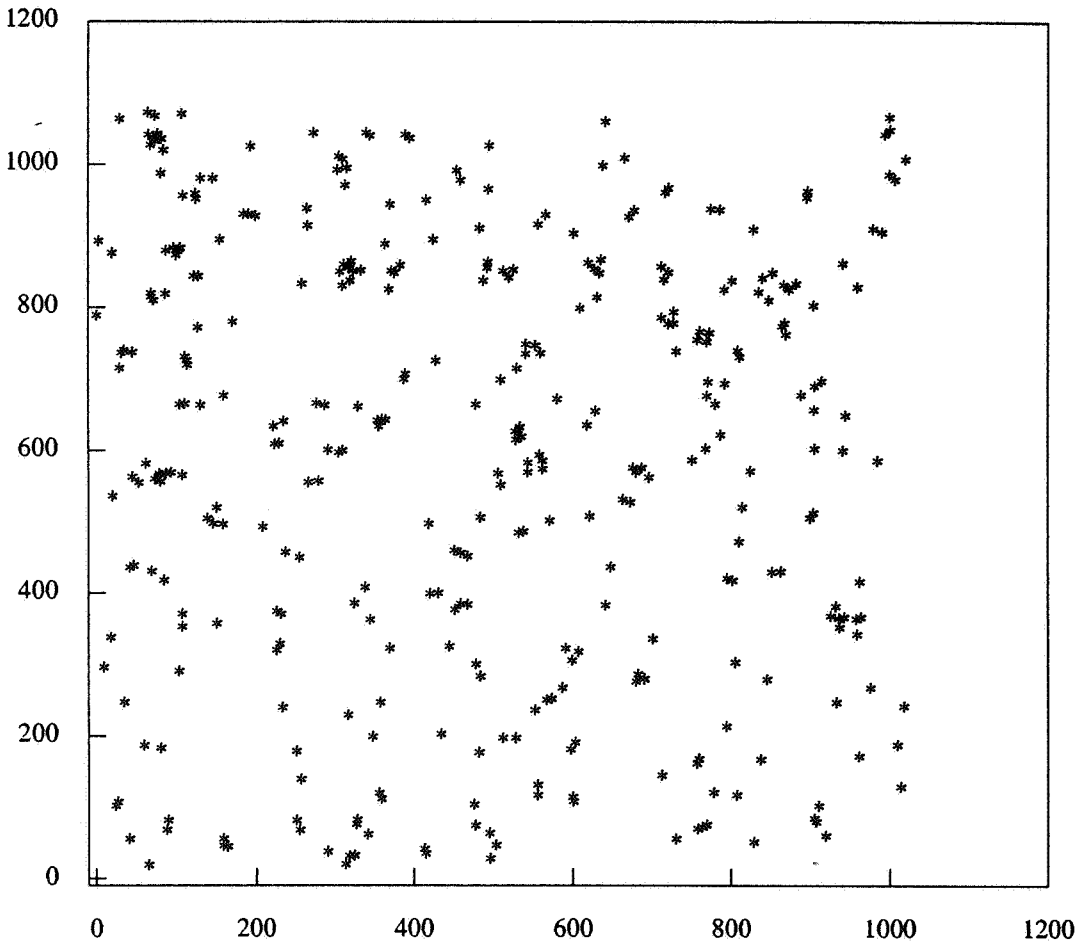
exposure for 60 seconds

B019



exposure for 90 seconds

B012



exposure for 240 seconds

FIGURE 1a. EDF-plot voor B014 van de event-event distances

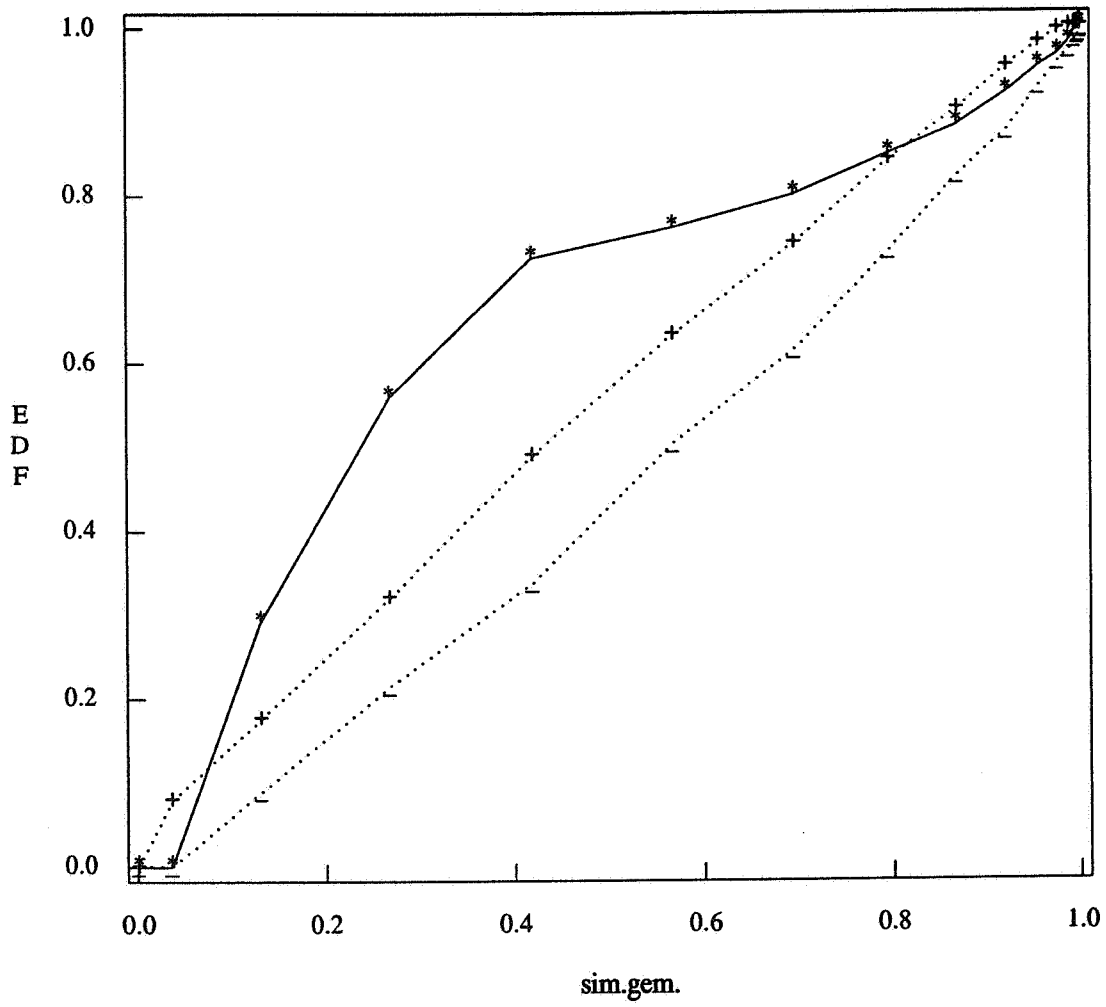
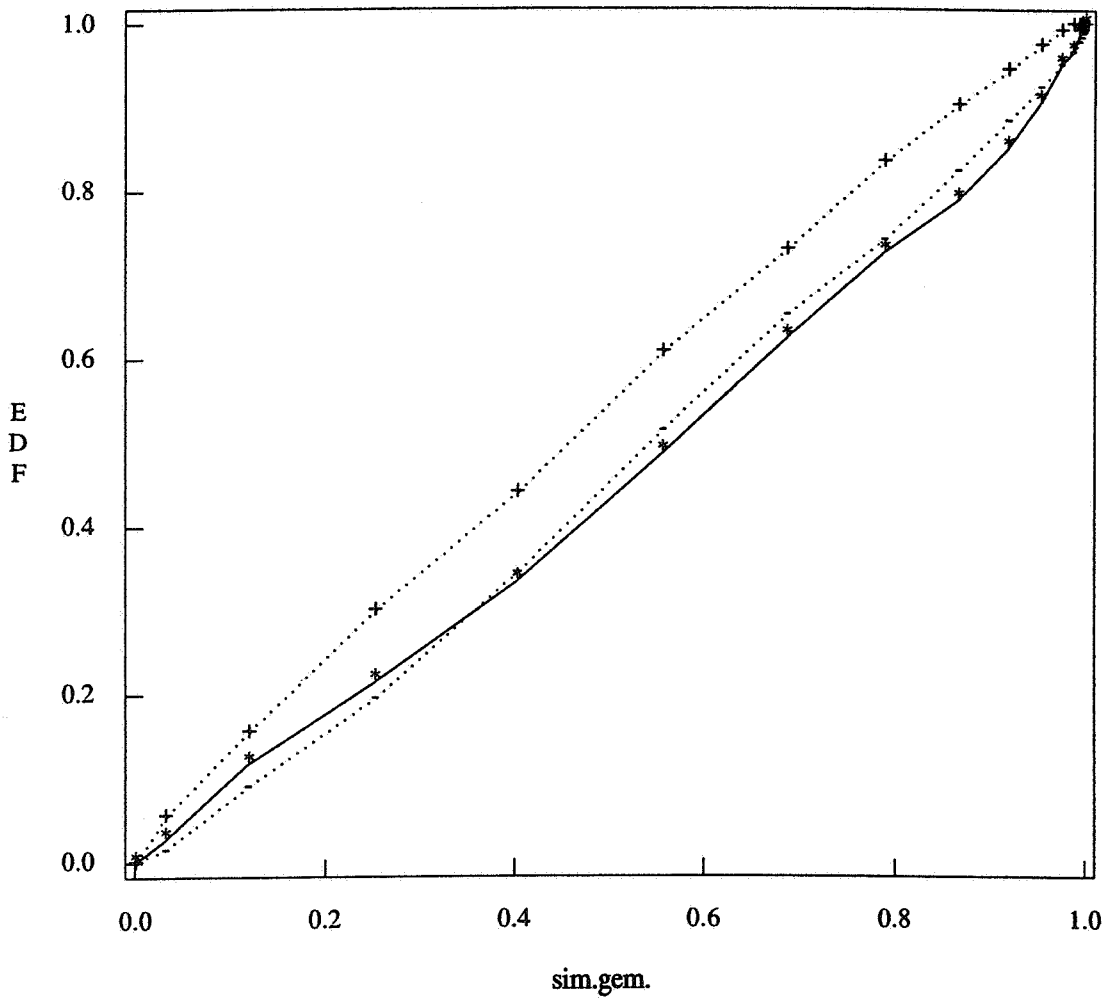


FIGURE 1b. EDF-plot voor B014 van de point-event distances



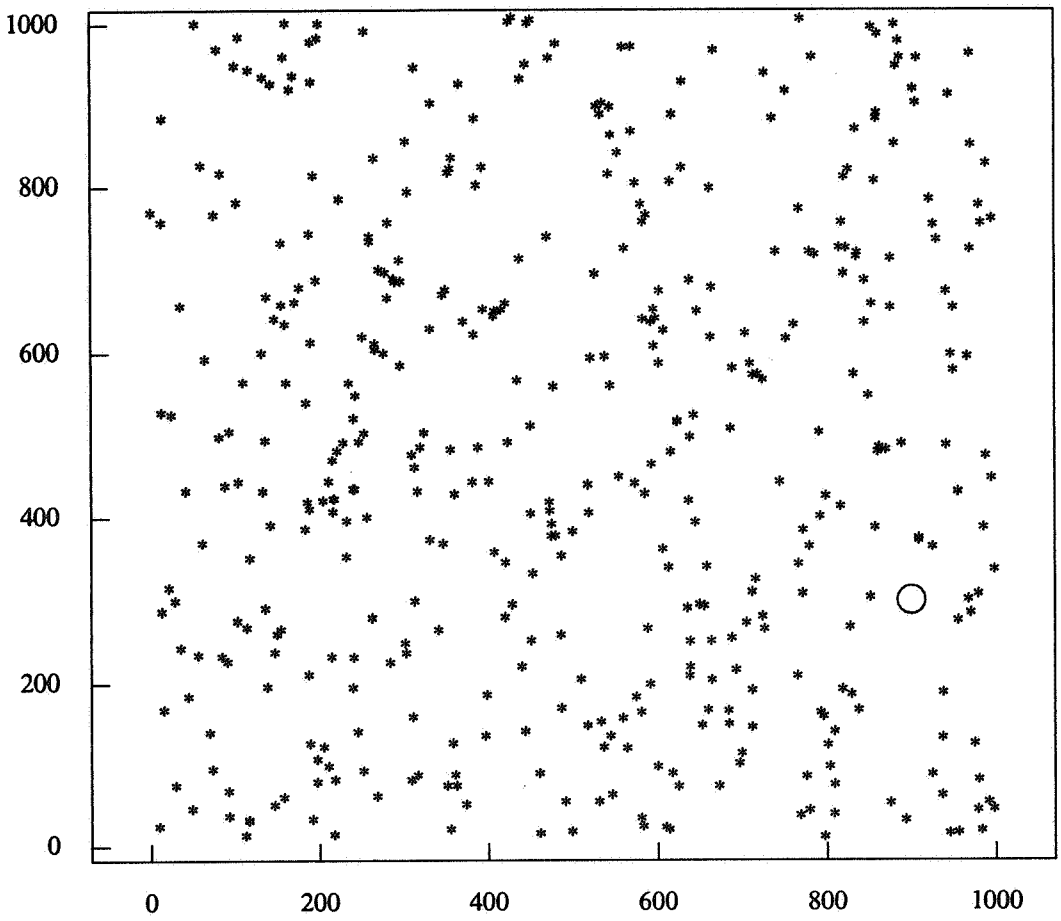


FIGURE 2. Simulated point pattern with mu = 1.07, sigma = 0.0165

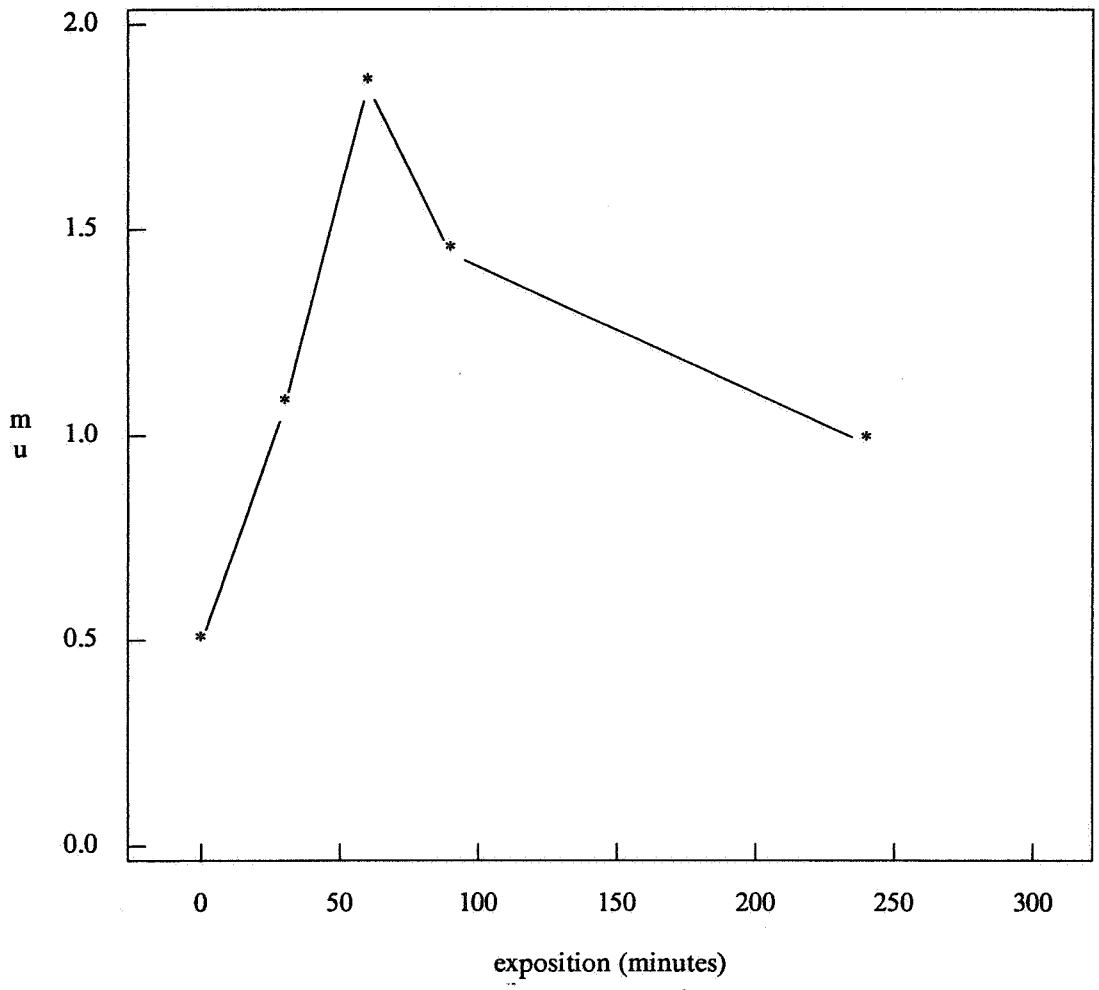


FIGURE 3. Estimated number of offspring per parent vs. exposition duration

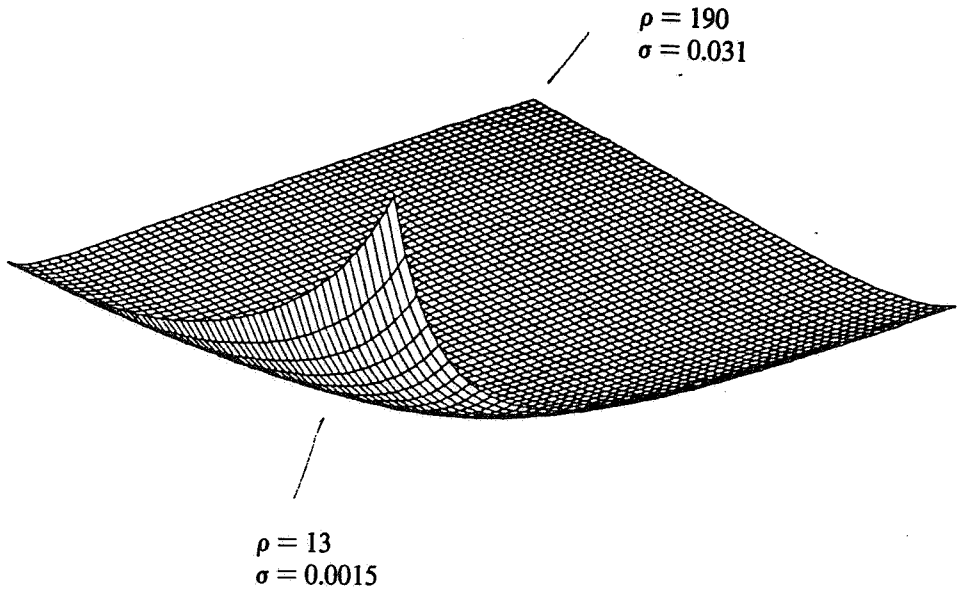


FIGURE 4: 3-dimensional plot for the function $D(\rho, \sigma)$ (B 015)
ρ : 13 − 190
σ : 0.0015 − 0.031

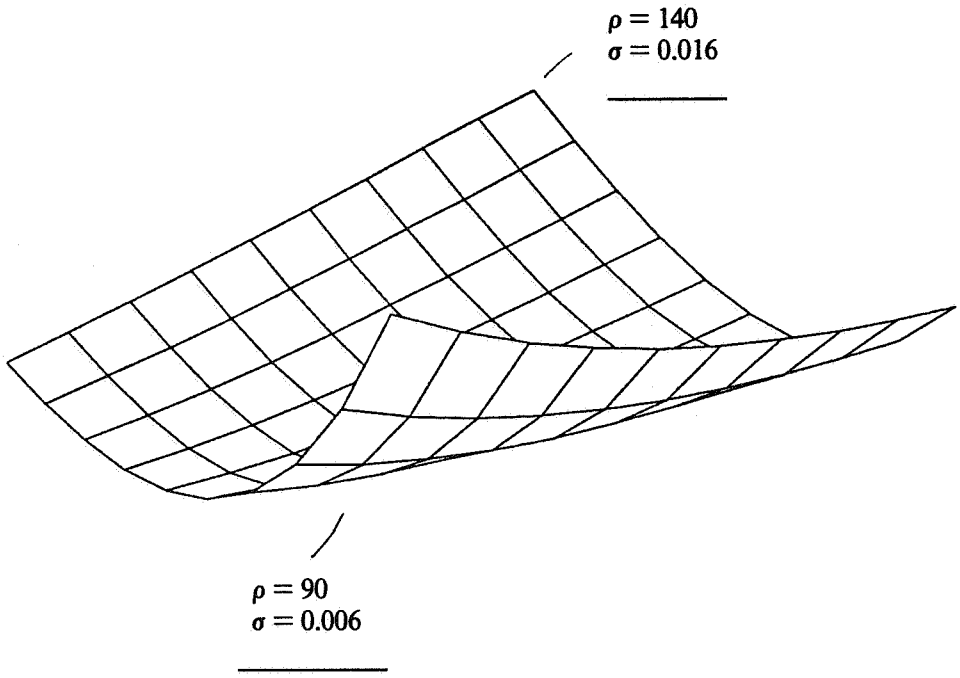


FIGURE 5: 3-dimensional plot for the function $D(\rho, \sigma)$ (B 015)
ρ : 90 − 140
σ : 0.006 − 0.016

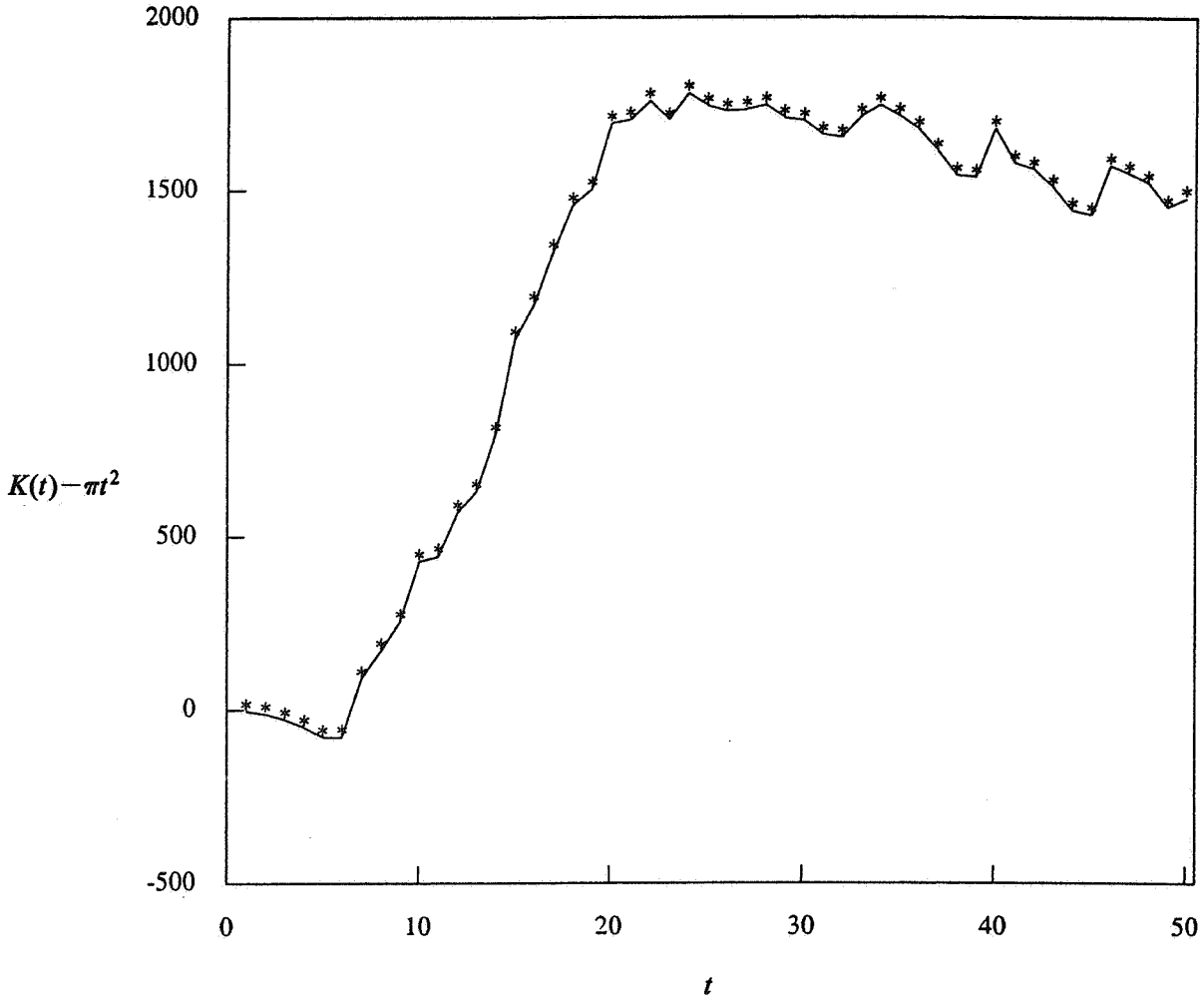


FIGURE 6. K-functie voor B014