SLAC – PUB – 3452
September 1984
(M)

ON SOME PROPERTIES OF THE KNN AND KMST*

Vito Di Gesu'
Stanford Linear Accelerator Center
Stanford, California 94305

ABSTRACT

This report studies the statistical behavior of some parameters defined in a graph, such as are used in pattern recognition and classification applications. It analyzes the changing structure, with increasing K, of the K-nearest neighbor (KNN) and the K-minimum spanning tree (KMST) of a graph. The results show that the connectivity of the KMST grows more rapidly and that the KNN is a subgraph of the KMST for each K. A conjecture is made on the value of the parameters K for which the KMST becomes a complete graph. Some experimental results on simulated data are reported.

Author's Permanent Address:
Dip. di Matematica - Univ. di Palermo - ITALY

Submitted to *Journal of Pattern Recognition*

* Work supported by the Department of Energy, contract DE – AC03 – 76SF00515

1. INTRODUCTION

The statistical and combinatorial properties of the K-nearest neighbor and the K-minimal spanning tree of an undirected, weighted and complete graph G, as denoted by KNN and KMST respectively, have been extensively studied. Friedman and Rafsky [1] have shown how to apply the KNN and KMST to define a generalized correlation coefficient, which allows one to do predictive data analysis about a sample of ordered pairs $(x_i, y_i), i = 1, \ldots n$. A comparison between the Bayesian classification function and the classification via KNN is given by T.H.Cover [3]. Within this framework, we provide a characterization of some useful quantities, such as the outdegree of a node OD and the number of components Nc of the KNN in order to use them as estimators of the degree of uniformity of a distribution of a set of points in a multidimensional space X. We ntroduce a technique that allows detection of the most significant volume in a set of points. This problem is of great importance with two-dimensional images for which the pixel-values belong to the set 0,1), and they appear as sparse matrices (mostly zero values). Maps of the sky in $\gamma - ray$ astronomy are an example of such data [4], [5].

Section 2 provides the definitions and the notation used during the exposition. Section 3 is dedicated to the study of the statistical behavior of the parameter Nc of the KNN of a graph G. Section 4 provides some limits about the dynamics of the growing of the KNN and KMST with increasing K. A conjecture on the value K=Kmax, for which the KMST becomes a complete graph, is formulated. In Section 5 we report the results of the application of the methodology to classify the presence of signal in a sparse image or to detect the presence of significant clusters in a multidimensional space.

In the present phase of the work, simulated data have been used in order to quantify the value of the results. Applications on real data will be made after the testing and calibration of the method. Section 5 also includes some implementation information. Section 6 is dedicated to final remarks.

2. DEFINITIONS AND NOTATIONS

We consider undirected complete graph $G =< N, W >$ where N is the set of nodes whose cardinality is denoted by $|N|$, and W is a weight function $W : N \times N \rightarrow R^+$. Here R is the set of positive real numbers.

Note that in this context we assume that given a node x, the set $N_x = N - \{x\}$ is ordered as follows:

$$\exists\, y, z \epsilon N_x \quad y \prec z \Leftrightarrow W(x,y) \leq W(x,z).$$

The K nearest nodes to x are the first K nodes in N_x.

DEF 1. The KNN of G is a subgraph of G such that each node is linked with the K nearest nodes.
Fig.1a, b, c shows a graph G and their 1NN and 2NN.

DEF 2. The MST is a spanning tree of G, such that the sum of the weights of its arcs is minimal.
Fig.2b shows the MST of G.

DEF 3. The KMST of G is the spanning subgraph of G such that:

$$KMST = MST(G) \qquad K = 1$$
$$KMST = [K-1]MST \bigcup MST(G - [K-1]MST) K > 1$$

where "$G - [K-1]MST$" denotes the set difference between graphs.

Fig. 2c shows the 2MST of G.

DEF 4. The outdegree (OD) of a node x of G is the number of arcs linked to x.

DEF 5. A component of a graph G is a connected subgraph $G' =< N', W' >$ such that $\forall x \epsilon N'$ and $\forall y \epsilon N$ there does not exist a path between x and y.

DEF 6. A complete connected component on a graph G is a component which is a complete subgraph of G.

The number of components and complete components of G are denoted by Nc and Ncc respectively.

3. THE PARAMETER Nc FOR THE STUDY OF "CLUSTERING TENDENCY"

The study of the clustering tendency is an important problem in exploratory data analysis; it becomes more difficult as the number d of measured parameters is increased or the sample size becomes small so that and the central limit theorem is no longer valid. Considerable work has been done to develop various methods based on the cooperation of mathematical, statistical and heuristic reasoning. In [6], W.G.S. Hines and R.J.O. Hines analyze spatial data by using the SSI (single sequential inhibition) process. The packing density is calculated for the case of uniform density and then a tendency test is performed.

Many methods have been proposed to formalize the problem. In [7], E. Panayirci and R.C. Dubes provides a good review of the subject and propose a d-dimensional version of the Cox-Lewis statistical test. So far, the results have provided only partial answers and most of them are suggested by the particular application.

One approach to the problem is to define parameters that characterize the distribution of the components in G. For example, V.Di Gesu' and B.Sacco [4] studied the parameter Nc as a function of the cutting threshold, ϕ, of the edges of the MST of a random graph G. The application of such criterion gave nice results when applied to the study of the $\gamma-ray$ sky maps as seen from the COS-B satellite.

In this report, the parameter Nc is considered as a function of K in the KNN of G, under the hypothesis of Poisson process. This is then applied to define a uniformity test (UT).

3.1 STUDY OF REGULAR TOPOLOGICAL CONFIGURATIONS

The expected number of $Nc(K)$ is, of course, a decreasing function of K, ranging from 1 to $|N|$:

$$Nc(0) = |N|, \qquad Nc(N-1) = 1$$

The first value K for which $Nc = 1$ will be denoted by K_1 .

The value of Nc also depends upon the topological configuration of the nodes in G. Consider, first, the case on which the points are arranged in a regular square lattice, see Fig.3a. It is easy to see that, for such configuration, the following results hold:

LEMMA 1. Let G be a square regular lattice in a d-dimensional space X, with number of edges equal to 2^n and $|N| > d$, then the maximum number of components of the KNN of G versus K is:

$$Nc(K) = \begin{cases} |N|/2^K & for\ 0 \leq K \leq d \\ 1 & for\ K > d \end{cases}$$

This is obvious for $d = 1$. the proof is given by construction for the case of $d = 2$; it is easy to generalize the results for $d > 2$. The strategy to maximize Nc is to increase by 1 the OD of each node at each step of the computation of the KNN. Of course for $K = 0$, $Nc = |N|$; for $K = 1$, by connecting each node to its right neighbor in the same row starting from the left (see Fig.3b), $Nc = |N|/2$; for $K = 2$, by linking each node to its lower neighbor of the same column starting from the top, $Nc = |N|/4$ (see Fig.4c); for $K = 3$, the third nearest node of the corner nodes are still in the same component, but in order to compute the 3NN, the algorithm must link in both directions with other nodes and $Nc = 1$ (see Fig.4d).

— In order to generalize the demonstration of LEMMA 1 for $d > 2$ it must be noted that the corner-nodes of each hyperplane, J, will be connected to the corresponding corner-nodes of a nearest hyperplane, J', for $K > 2$.

If $|N|$ is not a power of two, the problem cannot be stated in a simple way because the dependence of Nc on K is more complicated, however one can show that

$$Nc(K) = \begin{cases} \leq \lfloor |N|/2^K \rfloor & for\ 0 \leq K \leq d \\ 1 & for\ K > d \end{cases}$$

Let us now consider a second limit case.

LEMMA 2. Under the hypothesis that the nodes of a graph G, with $|N| = 2^n$, are grouped in hierarchical clusters forming a balanced binary tree (i.e, at level "0" there is G, at level "i" there are 2^i clusters of size 2^{n-i}), the value of Nc versus K is:

$$Nc(k) = - \begin{cases} 2^n & for\ K = 0 \\ 2^{n-i} & for\ 2^{i-1} \leq K < 2^i \quad i = 1, 2, \ldots, n-1 \\ 1 & for\ k = 2^{n-1} \end{cases}$$

In fact, the number Nc decreases by a factor two whenever the components of KNN are completed.

LEMMA 2 may be proved by induction:

$$For\ 2^0 \leq K < 2^1.$$

It is easy to see that $Nc = 2^{n-1}$. Suppose now that we are at a certain step $K = 2^{i-1}$ of the computation and let the number of the component be:

$$Nc = 2^{n-i} \qquad i < n$$

Before we reduce the number of components to 2^{n-i-1}, the actual computation must be completed in $2^i - 1$ steps, therefore:

$$Nc = 2^{n-i}\ for\ 2^{i-1} \leq K < 2^i.$$

For $K = 2^{n-1} - 1$, $Nc = 2$ and the components are completed, therefore for $K = 2^{n-1}$ the KNN is a connected subgraph of G. This last result may be

generalized for any number of nodes and with hierarical clusters arranged in a non-balanced tree.

3.2 Case Of Poisson Process

In the following, a rule is derived to compute the Nc in the KNN under the hypothesis that the nodes, $x = (x_1, x_2, \ldots, x_d)$, belong to a d-dimensional normed space X and the weight function W is the Euclidean distance. The nodes are distributed following a Poisson process. Moreover the following assumption are considered:

ASSUMPTION 1. The $OD(K)$ of each node x does not depend on x (uniformity of X).

ASSUMPTION 2. The $OD(K)$ of each node x is uniformly distributed in all directions (symmetry of X).

Deriving an exact law of $Nc(K)$ is difficult because of the following constraints:

(a) The process of growing the KNN does not depend only on the distribution of the points. Given a value for K, each node in the KNN must have at least $OD = K$.

(b) The maximum and minimum number of edges in the KNN is an increasing function of $|N|$ and K, and it is not easy to derive its analytical form.

Nevertheless, in the present report an empirical and simplified solution is proposed, which fits very well with experimental data. It seems, therefore, that it could be a good starting point for further formal studies. This rule is suggested by the theoretical results in 3.1. The main idea is that a Poisson process must have some regularity united to a hierarchical behavior.

CLAIM 1. The mean number of components Nc in a KNN of a random graph

G is given by:

$$N_c(K) = 1 + (|N| - K - 1) * 2^{\overline{OD}(K)} \tag{1}$$

Where $\overline{OD}(K)$ is the mean OD of KNN, empirically computed by:

$$\overline{OD}(K) = \left[k \sum_{i=1}^{|N|} OD(K)_i \right] / |N|.$$

For $K = 0 : \overline{OD}(K) = 0$ and $Nc(K) = |N|$; for $K = |N| - 1 : \overline{OD}(K) = |N| - 1$ $Nc(K) = 1$.

In the following, an estimation is given of $OD(K)$ by handling the problem without constraints and assuming a binomial model.

Let $p(K)$ be the probability of increasing by one the OD of a node in the KNN. Therefore the probability of having $OD(K) = K + L$, $0 < L < |N| - K - 1$ is:

$$P(OD = K + L|K) = \binom{|N| - K - 1}{L} * p(K)^L * [1 - p(K)]^{|N|-K-L-1} \quad 0 \leq K \leq |N| - 1 \tag{2}$$

CLAIM 2. If the nodes, x, are uniform Poisson distributed:

$$p(K) = K/(|N| - 1)$$

and $P(OD = K + L|K)$ may be rewritten:

$$P(OD = K + L|K) = \binom{|N| - K - 1}{L} * [K/(|N| - 1)]^L * [1 - K/(|N| - 1)]^{|N|-K-L-1} \tag{3}$$

and the $\overline{OD}(K)$ is:

$$\overline{OD}(K) = K + (|N| - K - 1) * K/(|N| - 1). \tag{4}$$

The expression claimed for $p(K)$ may be explained by assuming that it is pro-

portional to the number of nodes included in a hypersphere surrounding each of the nodes in the KNN, if the density of the points is assumed constant, the number is proportional to K, as a matter of fact that the radius of the hypersphere is proportional to K. The $P(K)$ is equal to 0 for $K = 0$ and 1 for $K = |N| - 1$, as expected, therefore $P(OD = K + L|K)$ becomes a delta function for $K = 0$ and $K = |N| - 1$, with value 1 for $L = 0$. The last property is in accordance with the theoretical limit values for L.

In Fig.4, we show the results of fitting the experimental distributions obtained in a run of 1000 points, uniformly distributed in a 2d and 4d space with the claimed distribution function. The results seem to verify the assumption that has been made.

Figures 5 and 6 display the experimental values of $\overline{OD}(K)$ for 2d and 4d respectively. Also, in this case there is good agreement of the experimental data with the rule (4).

Note that the fitting has been performed assuming:

$$P(K) = [K/(|N| - 1)]^{\alpha}$$

The value of the parameter α is greater than one and in the experiments performed its value is of the order of 1.5 and does not depend strongly on the dimension of X.

The parameter takes partially into account the constraints on the sum of the degrees in the KNN at each K. The $\overline{OD}(K)$ is of the order of K, as expected, and its maximum deviation from K is for:

$$K_M = \frac{\alpha}{\alpha + 1}(|N| - 1)$$

3.3 THE UNIFORMITY TEST UT

The results above suggest that a uniformity test (UT) is valid even when classical UT testing cannot be done ($\chi^2 - test$, Kolmogorof-Smirnov, least square fitting, t-student,...). In fact, the probability of having the Nc(K) component may be drawn from (1) and (2), in the general case, and from (1) and (3) for Poisson process. From eqn.(1):

$$\log_2(Nc(k) - 1) = -\overline{OD}(K) + \log_2(|N| - K - 1)$$

But for a given K $\log_2(|N| - K - 1)$ is a constant and therefore:

$$P[-\log_2(Nc(K) - 1)] = P[\overline{OD}(K)]$$

Now let $f_1(x)$ and $f_2(g(x))$ be the probability density function of x and $g(x)$ a continuous and derivable function of x. There exists the following between f_1 and f_2 relation:

$$f_1(x) = f_2(g(x)) * g'(x)$$

In our case:

$$P[Nc(K)] = P[\overline{OD}(K)]/(Nc - 1) \tag{5}$$

Where the value of $P[\overline{OD}(K)]$ may be derived from (2), (3) and (4).

The UT is now defined by the following procedure:

Given a set of experimental components $\{Nc^*(K)\}, K = 1, 2, ...K_1$, the probability of this set being consistent with the expected $\{Nc(K)\}$ is (see [8]):

$$Q_1 = P_1$$

$$Q_K = P_K * Q_{K-1} \quad * [1 - ln(P_K * Q_{K-1})] \qquad K = 2, ...K_1$$

where P_K (which may be computed from (5)) is the probability of $Nc(K)$ clusters for a given K.

4. DYNAMICS OF THE GROWING OF THE KNN AND KMST

This section shows some results concerning the quantity Kmax defined as:

DEF 7. The Kmax is the value of K such that KNN and KMST are equal to G.

If the nodes of G are points of a normed space X, the following lemma holds:

LEMMA 3. In the process of the computation of the KNN of G, let T be the subgraph added to KNN to build the $[K+1]NN$, then T is a forest (i.e. does not have cycles).

In fact, suppose that there exists in T a cycle of order L:

$$(x_1, x_2, \ldots, x_{L-1}, x_1)$$

It is easy to see that the following order must hold between the weights of the arcs

$$W(x_i, x_{i+1}) > W(x_{i+1}, x_{i+2}) \qquad for\ 1 = 1, 2, \ldots, L-3$$

$$W(x_{L-2}, x_{L-1}) > W(x_{L-1}, x_1)$$

$$W(x_{L-1}, x_1) > W(x_1, x_2)$$

which cannot hold if the space X is normed.

By using the LEMMA 3 it is easy to see:

LEMMA 4. For the KNN, Kmax $= |N| - 1$.

Suppose $Kmax < |N| - 1$, then:

$$\exists K < |N| - 1\ such\ that\ \forall x \epsilon N\ OD_x > K$$

i.e., the graph T added to $[K-1]NN$ to compute the KNN has a cycle which cannot hold by LEMMA 3.

Note that the lemma is valid only if it is assumed that X is a normed space.

The evaluation of Kmax for the KMST is more difficult and it depends on the topology of G. In this present report, only limit values are given and a conjecture is formulated for its mean value, in the case of a random graph.

The two limit configurations considered are:

DEF 7. A graph G is said to be in a "Pure Linear Configuration," PLC, if its nodes are topologically configurated as follows:

$K = 0 \quad G_{(0)} = G$ And there exists an ordering of the nodes $I_1, I_2, \ldots I_{|N|}$ such that

$$W(I_i, I_{i+1}) = \min_{e \epsilon N} \{W(I_i, e)\}$$

$K > 0 \quad G_{(K)} = G - KMST$ And there exists an ordering of the nodes $I_1, I_2, \ldots I_{|N|}$ such that

$$W(I_i, I_{i+1}) = \min_{e \epsilon G^{(K)}} \{W(I_i, e)\}$$

DEF 8. A graph G is said to be in a "Pure Star Configuration," PSC, if its nodes are topologically configurated as follows:
There exist an ordered sequence of nodes $I_1, I_2, \ldots, I_{|N|}$ such that:

$$\forall J \epsilon [1, |N|], \forall x \epsilon N - \bigcup_{r=1}^{J} I_r \Rightarrow W(I_J, x) = \min_{y \epsilon N - \bigcup_{r=1}^{J} I_r} \{W(y, x)\}$$

NOTE. The two configurations are unlikely, nevertheless they are the two extreme configurations which allow the establishment of the following limits for Kmax.

LEMMA 5. The value of Kmax for G=PLC ℓ s $N/2$ and it is the minimum possible value. In fact, for the topology of PLC, at each step of the computation are added $|N|-1$ edges to the KMST and therefore:

$$Kmax * (|N|-1) = \frac{|N| * (|N|-1)}{2} \Rightarrow Kmax = |N|/2$$

The value is the minimum because at each step we add the maximum number of edges.

LEMMA 6. The value of Kmax for G=PSC is $|N|-1$ and it is the maximum possible value. If G is PSC at each step I, there are $|N|-I$ added edges, therefore:

$$\sum_{I=1}^{Kmax} (|N|-I) = \frac{|N| * (|N|-1)}{2} \Rightarrow Kmax = |N|-1$$

The value is the largest possible because the maximum degree of a node in G is $|N|-1$.

From LEMMAs 5 and 6, it follows:

THEOREM 1. For a general topological configuration of G:

$$|N|/2 < Kmax < |N|-1$$

CONJECTURE. The mean value of Kmax for the Kmst of G depends linearly on $|N|$:

$$Kmax = \alpha \cdot |N|$$

with $\alpha < 1$.

Figure 7 shows the experimental value of Kmax versus the number $|N|$ found by computing the KMST of a random graph, the nodes of which follow the Poisson distribution ($\alpha = .58$). The predicted law fits the experimental data quite well.

Note that given the nature of the algorithm for constructing the KMST, the results do not depend on the dimensionality of space X.

This section is ended by stating a conclusion relation between the KNN and the KNMST.

LEMMA 7. The 1NN $\subset$ 1MST.

From the definition of KMST each node is linked with the nearest neighbor.

THEOREM 2. The KNN is a subgraph of the KMST for each K.

By induction:

$K = 1 \quad 1NN \subset 1MST$ from LEMMA 7.

suppose now that it is true that:

$$K > 1 \quad KNN \subset KMST$$

Therefore the arcs of the KNN are included in KMST, however by LEMMA 3 the arcs added to the KNN, to compute the $[K+1]$NN are subtree of the $[K+1]$MST and therefore:

$$[K+1]NN \subset [K+1]MST$$

These results could be useful in evaluating the sum of the OD of the nodes of a KNN in order to define the constraints in the computation of the probability distribution function of OD(K), because in this case the combinatoric problem seems to be easier. For example, for $K = 1$ the value of such sum is $2 * (|N| - 1)$.

5. APPLICATIONS AND IMPLEMENTATION-NOTES

The experiments have been performed on simulated data, generated by Monte Carlo procedures in accordance with typical images of the sky as detected from γ and X-ray astronomy. Multidimensional data are also considered to show the behavior of the UT proposed method by varying the dimensionality of the space. The graph G is still considered as a set of points in a normed space X and only Euclidean distance is considered at the present.

5.1 EXPERIMENTS FOR THE UT

The UT has been tested on samples with increasing point density (ρ=.5, 1., 2., 3.counts/unit volume) and for $d = 2$ and 4. In Fig.8, a sample for $\rho = 1.$ and $d = 2$ is shown, Fig.9 displays Nc versus K and the agreement with the predicted law seems good ($Q = .49$). Fig.10 shows the application of the UT for $d = 4$ and $\rho = 10^{-4}$, the agreement with the predicted law is still good ($Q = .42$) and it does not seem to hold strong dependence from the dimension of the space X, because of the costraints a) and b) of section 3.1.

In order to test the performance of the UT, data has also been generated with increasing number of clusters (Ncl=2,3,4,5,6,30) of equal size, imbedded in uniform background for $d = 2$ and $d = 4$. Figure 11 shows a map with Ncl=5 and, in Fig.12, a more complicated situation with Nc=30. Figure 13 displays Nc vs. K for the analyzed cases on which $Q < 10^{-4}$.

Table 1 reports the results of the application of the UT under different experimental conditions. The results have been carried out after a run of 200 samples of size $|N| = 100$ and $\rho = 1$. The values of Q computed by the formula (6) are lower for $d = 2$, as expected. In both cases, Q is of the order of 10^{-4} for $S/N > .66$. The result seems encouraging for the application of the method in real situations.

TABLE 1

$d=2$		$d=4$	
S/N	Q	S/N	Q
.0	.24	.0	.29
.25	.004	.25	.01
.66	.0002	.66	.0005
1.	.00001	1.	.0001
1.5	$< 10^{-6}$	1.5	.00001
2.0	"	2.0	$< 10^{-4}$
3.0	"	3.0	"
4.	"	4.0	10^{-5}

To study the gestaltical nature of the KNN, experiments have been performed where a set of maps with an increasing number of clusters have been generated, as explained before. In this framework, a cluster is defined as relevant if it is a complete component of the KNN and, for the given K, $Q_K < \beta$ (in our case, $\beta = 10^{-3}$). Table 2 shows the mean number of clusters detected, Ncc versus the number of clusters generated, Ncl. The results show the degradation of the performance in the detection with the increasing of Ncl. The average has been performed in a run of 1000 events.

TABLE 2

d=2		d=4	
Ncl	Ncc	Ncl	Ncc
5	5	5	5
6	6	6	4
7	7	7	5
8	7	8	6
10	8	10	8
15	10	15	12

Although number of clusters is underestimated, the position and intensity (number of points contained in) of those that were detected is correct. Figure 14 shows the results of the analysis for a two-dimentional map. Five of the eight clusters generated were detected, the three that were missed are overlapping and very spread out in the picture.

5.2 Implementation Notes

At present, the work has not optimized algorithms that have been used to compute the KNN of G. Many of such algorithms are available; in [9] J.H.Friedman, J.L.Bentley and R.A.Finkel proposed an algorithm based on the storing of the data in a K-d tree, which required a computation time proportional to $K|N|\log|N|$. To compute the components of the KNN, a quick deep first search algorithm has been used [10] which requires a computation time of the order of $|N|$.

The experiments and the growing limits of the KNN obtained in Section 4 show that to perform the complete analysis of an input sample of size $|N|$, it requires computation of the KNN for $K = 1, 2, \ldots, \log|N|$; therefore the average computation time is:

$$Tcomp = 0\left(\sum_{K=1}^{\log|N|} K * |N| * \log|N| + |N|\right) =$$

$$0\left(\frac{|N|}{2} * \log^3|N| - \frac{|N|}{2}\log^2|N| + |N| * \log|N|\right)$$

6. FINAL REMARKS

In this report, the dynamic behavior of the KNN of a graph G has been analyzed. It seems that good results may be obtained whenever exploratory data analysis, as clustering tendency, is performed on very sparse data. The gestaltical power of graph theoretical methods, based on the MST computation, has been dramatically shown by C.T.Zahn [11], and the KNN has also shown adaptive properties to the topology of the data. Nevertheless, it is less sensible to line structure and gives better results whenever the morphology of the signal is compact, where the compactness may be defined roughly as the ratio between the volume and the surface.

The exact computation of $P(Nc = K + L|K)$ seems to be very difficult as it also is for standard p.d.f of the points in X, and much work remains to be done in that direction.

Another interesting topic would be to explore the case in which X is not normed; in this case the combinatorics will be more complicated.

ACKNOWLEDGMENTS

I wish to express my thanks to Mrs. Harriet L. Canfield for the review of my English and for the beautiful typing. I would also like to thank Professor Jerome H. Friedman for the fruitful discussions.

FIGURE CAPTIONS

Figure 1. Example of KNN of a graph G.

Figure 2. Example of KMST of a graph G.

Figure 3. Maximum $Nc(K)$ for $d = 2$.

Figure 4. Distribution of OD for $K = 6$ and $d = 2$.

Figure 5. $\overline{OD}(K)$ versus K for $d = 2$.

Figure 6. $\overline{OD}(K)$ versus K for $d = 4$.

Figure 7. Value of Kmax versus $|N|$.

Figure 8. Example of input data, $\rho = 1$.

Figure 9. Nc versus K for $d = 2$.

Figure 10. Nc versus K for $d = 4$.

Figure 11. Input with $Ncl = 5, d = 2$ and $\rho = 1$.

Figure 12. Input with $Ncl = 30, d = 2$ and $\rho = 1$.

Figure 13. Results of the application of the UT, $Q = 10^{-4}$.

Figure 14. Example of clustering detection.

REFERENCES

[1] J.H.Friedman, L.C.Rafsky, "Graph-Theoretic Measures of Multivariate Association and Prediction," The Annals of Statistics, Vol.11, No.2, pp.377-391, 1983.

[2] J.C.Friedman, W.Stuetzle, A.Schroeder,"Projection Pursuit Density Estimation," SLAC-PUB-3215, 1983.

[3] T.M.Cover, "Estimation by Nearest Neighbor Rule," IEEE-Trans. Inf. Theory, 1966.

[4] V.DiGesu', B.Sacco, "Some Statistical Properties of the Minimum Spanning Forest," Journal of Pattern Recognition, Vol.16, No.5, pp.525-531, 1983.

[5] G.De Biase, V.Di Gesu, B.Sacco, "Detection and Classification of extended sources in X and astronomy," submitted to Signal Processing, 1984.

[6] W.G.S.Hines, R.J.O.Hines, "The Eberhard Statistics and the Detection of Nonrandomness of Spatial Point Distribution," Biometrika, Vol.66, pp.73-79, 1979.

[7] E.Panayirci, R.C.Dubes, "A New Statistic for Assessing Gross Structure of Multidimentional Patterns," Tech.Rep. TR No. 81-04, Dep.of Computer Science, Michigan State University, 1982.

[8] W.T.Eadie, D.Drijard, F.E.James, M.Roos, B.Sadoulet: "Statistical Method in Experimental Physics," North Holland Pub.Comp., pp.282-284, 1971.

[9] J.H.Friedman, J.L.Bentley, R.A.Finkel, "An Algorithm for Finding Best Matches in Logarithmic Expected Time," ACM Trans. on Math. Software, Vol.3, No.3, pp.209-226, 1977.

[10] E.M.Reingold, J.Nievergelt, M.Deo, "Combinatorial Algorithms Theory and Practice," Prentice Hall, pp. 327-330, 1977.

[11] C.T.Zahn, "Graph-Theoretical Method for Detecting and Describing Gestalt Clusters," IEEE Trans.Comp. C-20, pp.68-86, 1971.

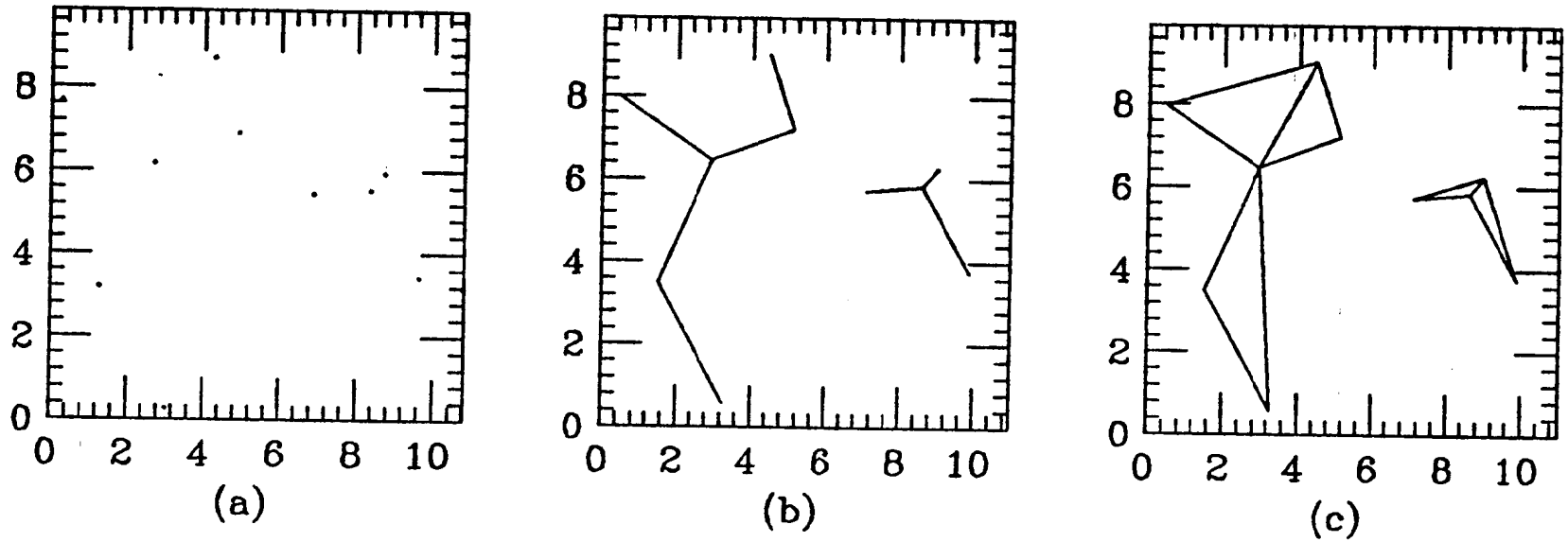


Fig.1 (a) Graph G;
(b) 1NN Of G;
(c) 2NN Of G.

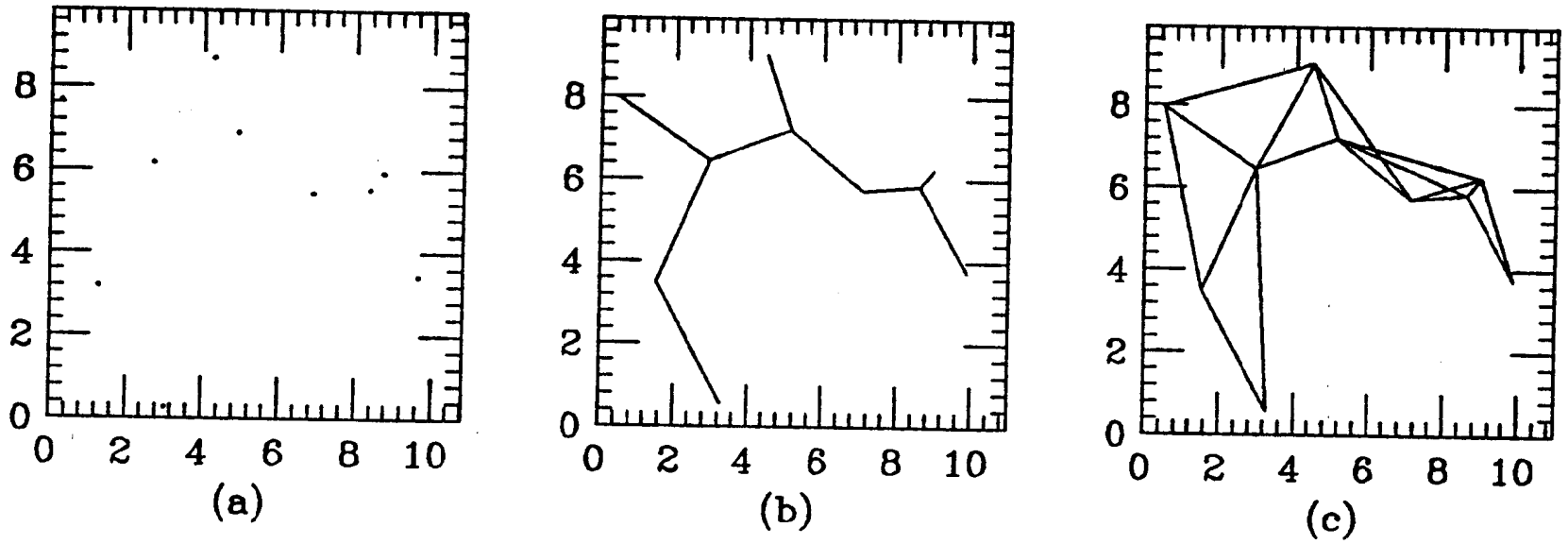


Fig.2 (a) Graph G;
(b) 1MST Of G;
(c) 2MST Of G.

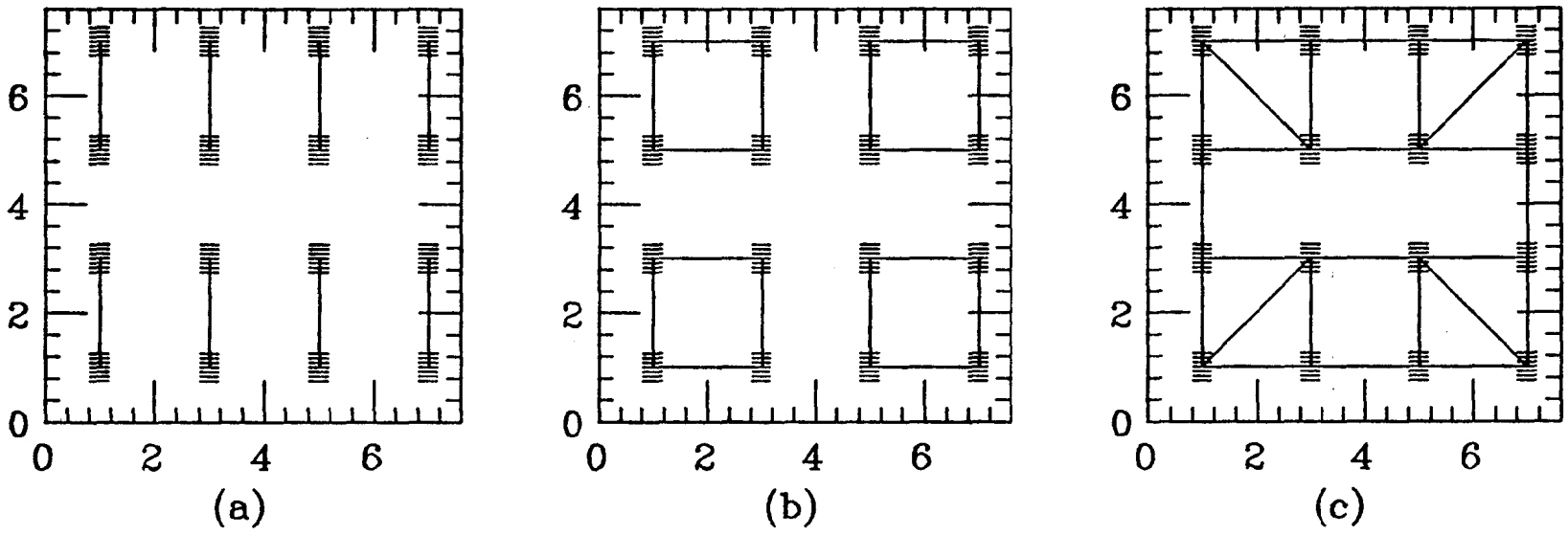


Fig.3 (a) 1NN of G;
(b) 2NN of G;
(C) 3NN OF G.

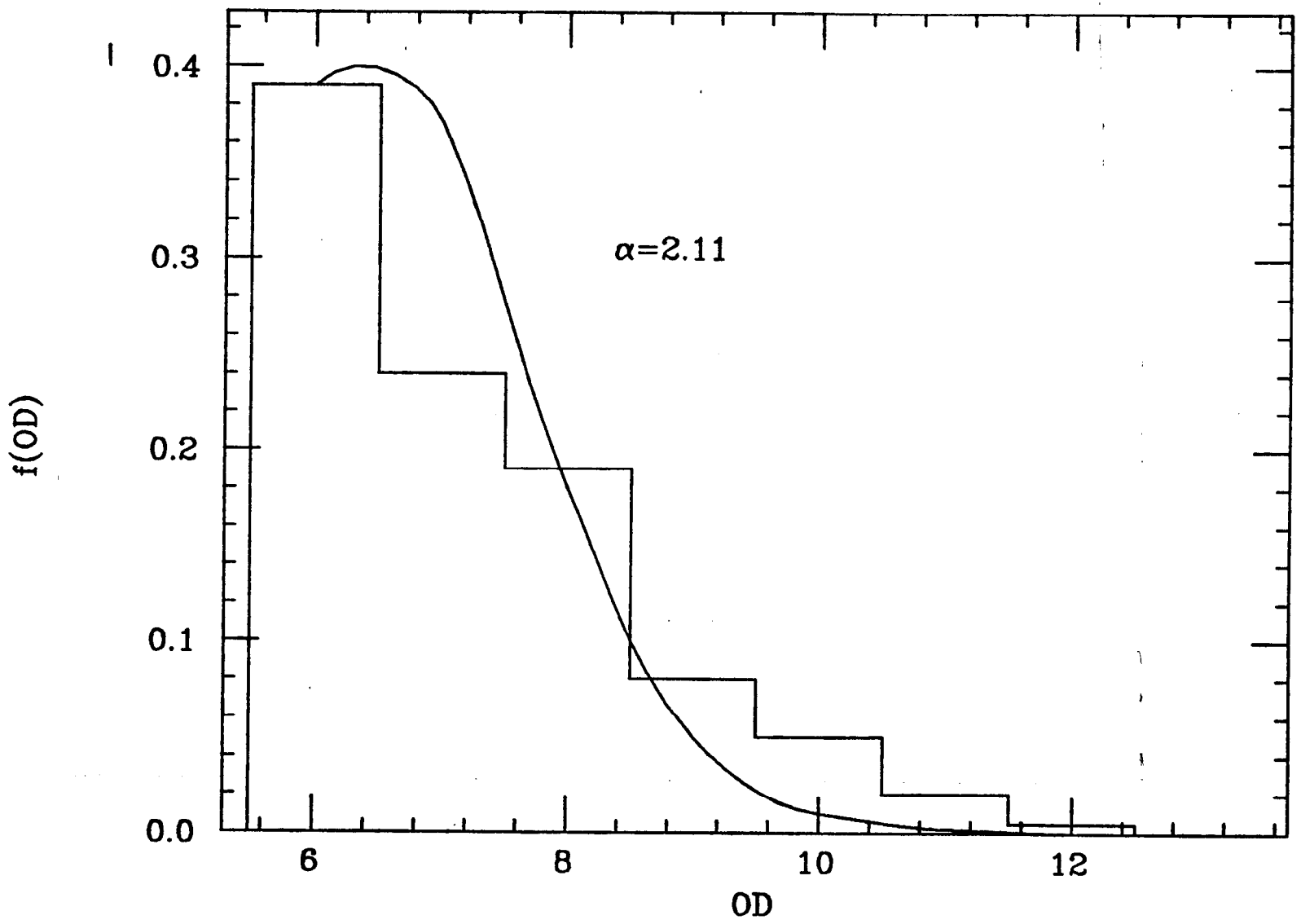


FIG.4

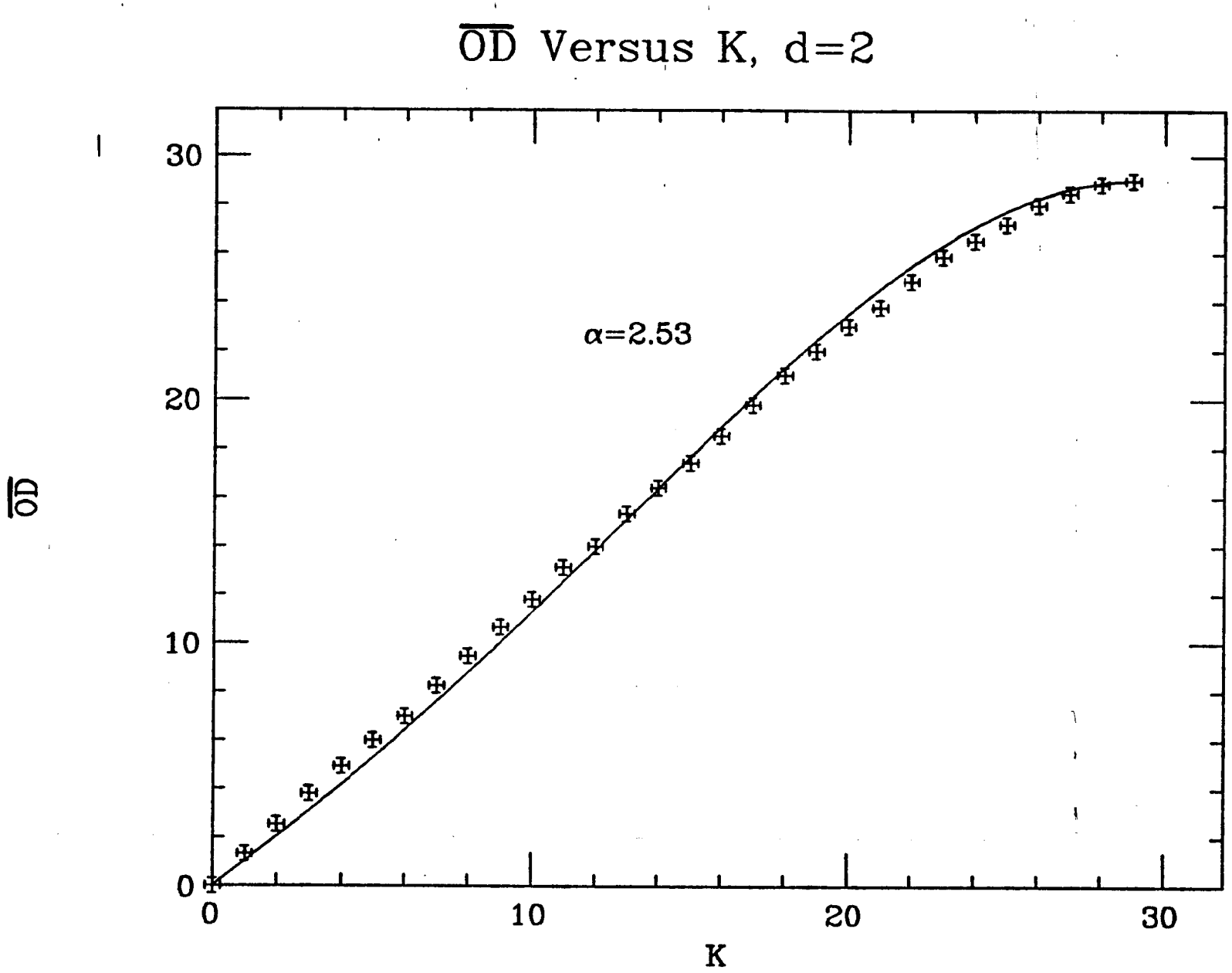


Fig.5

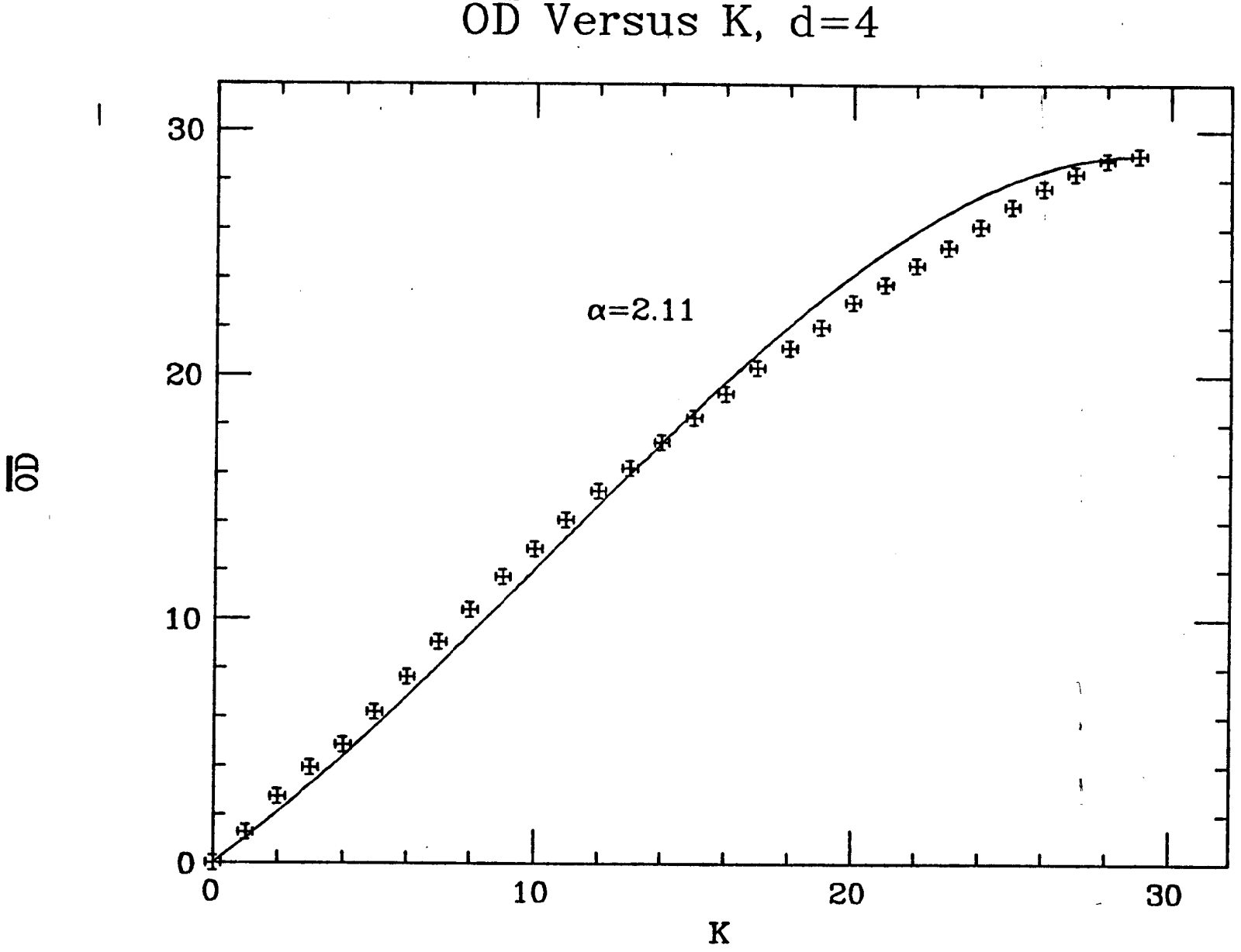


Fig.6

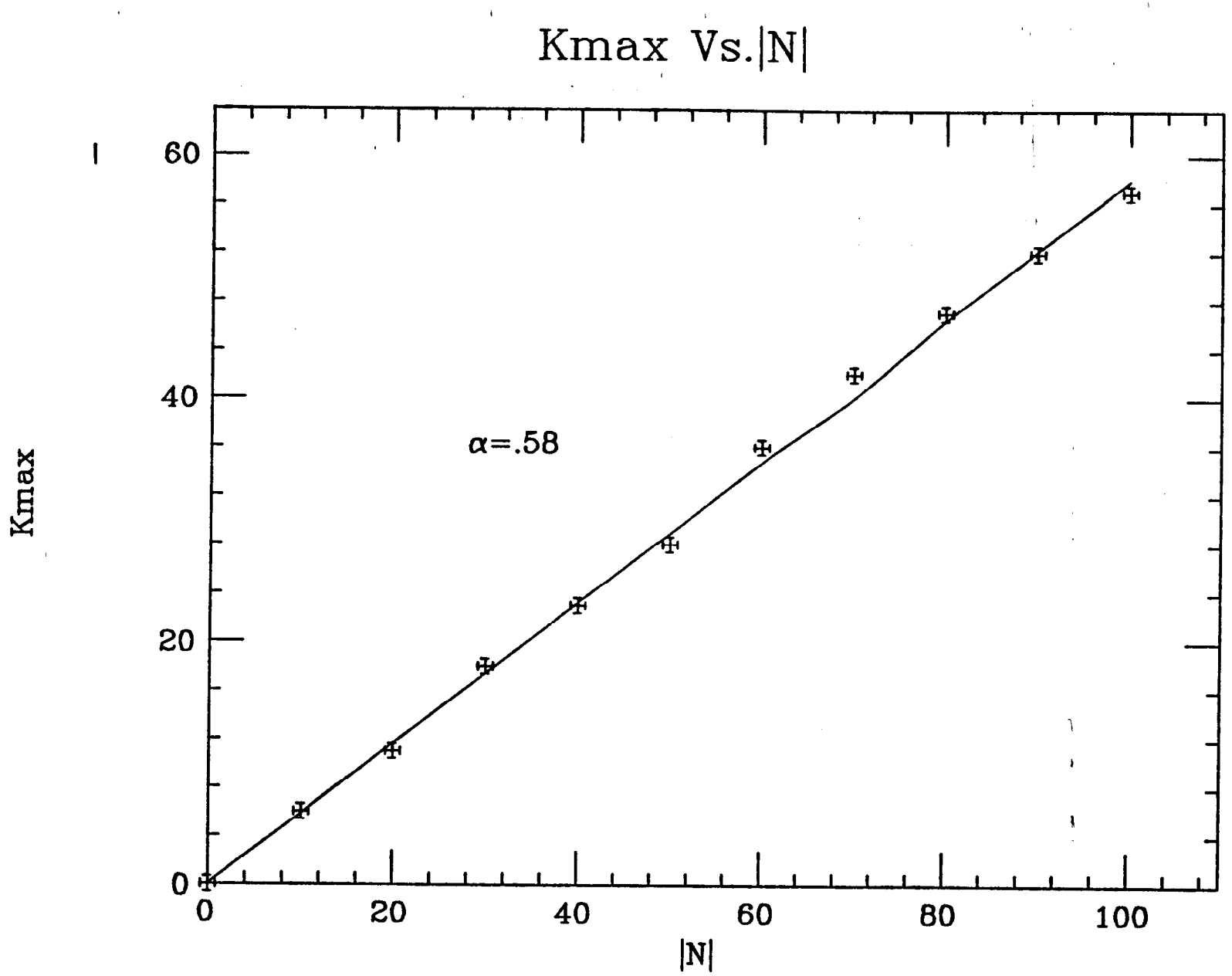


Fig.7

Input data ρ=1.

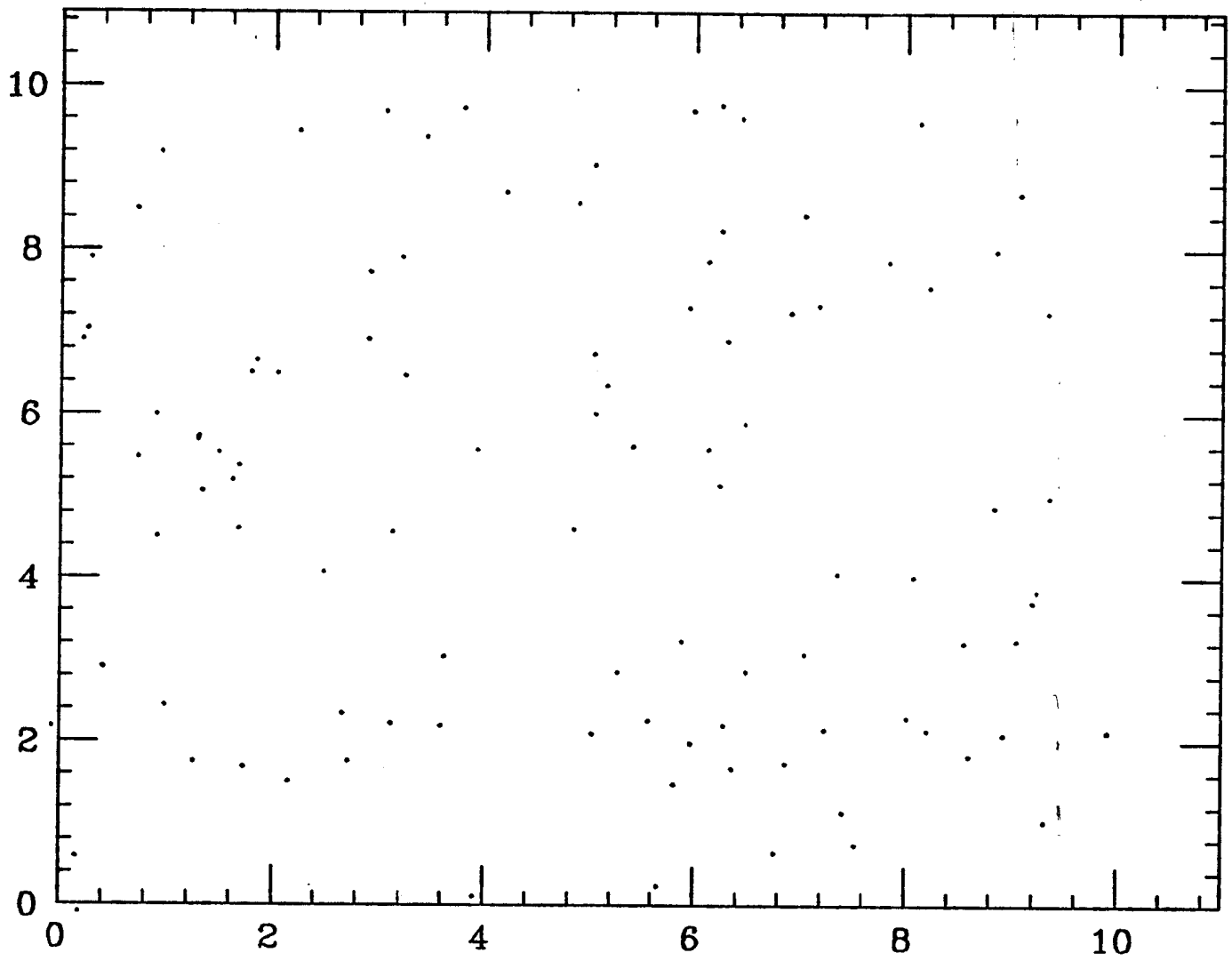


Fig.8

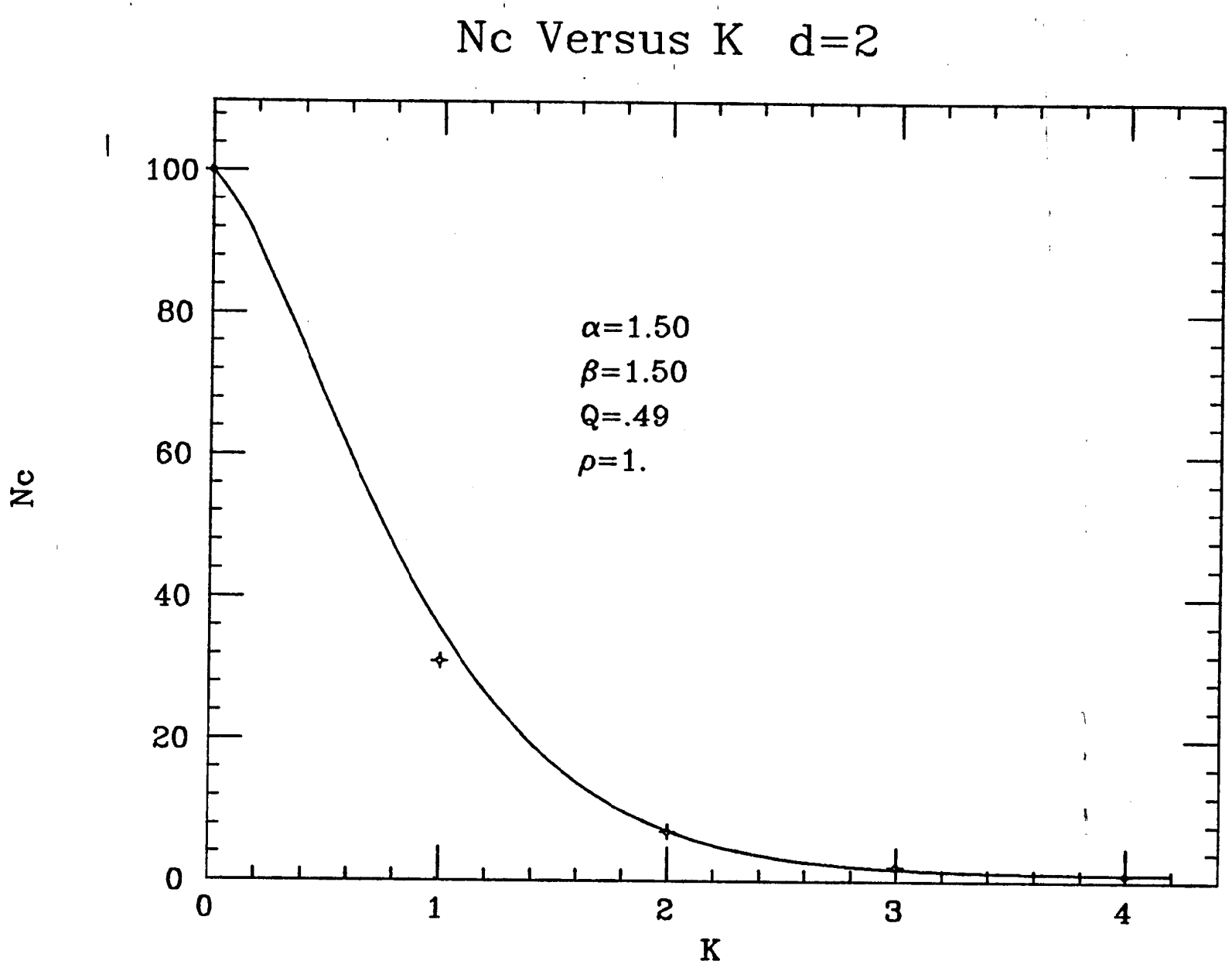


Fig.9

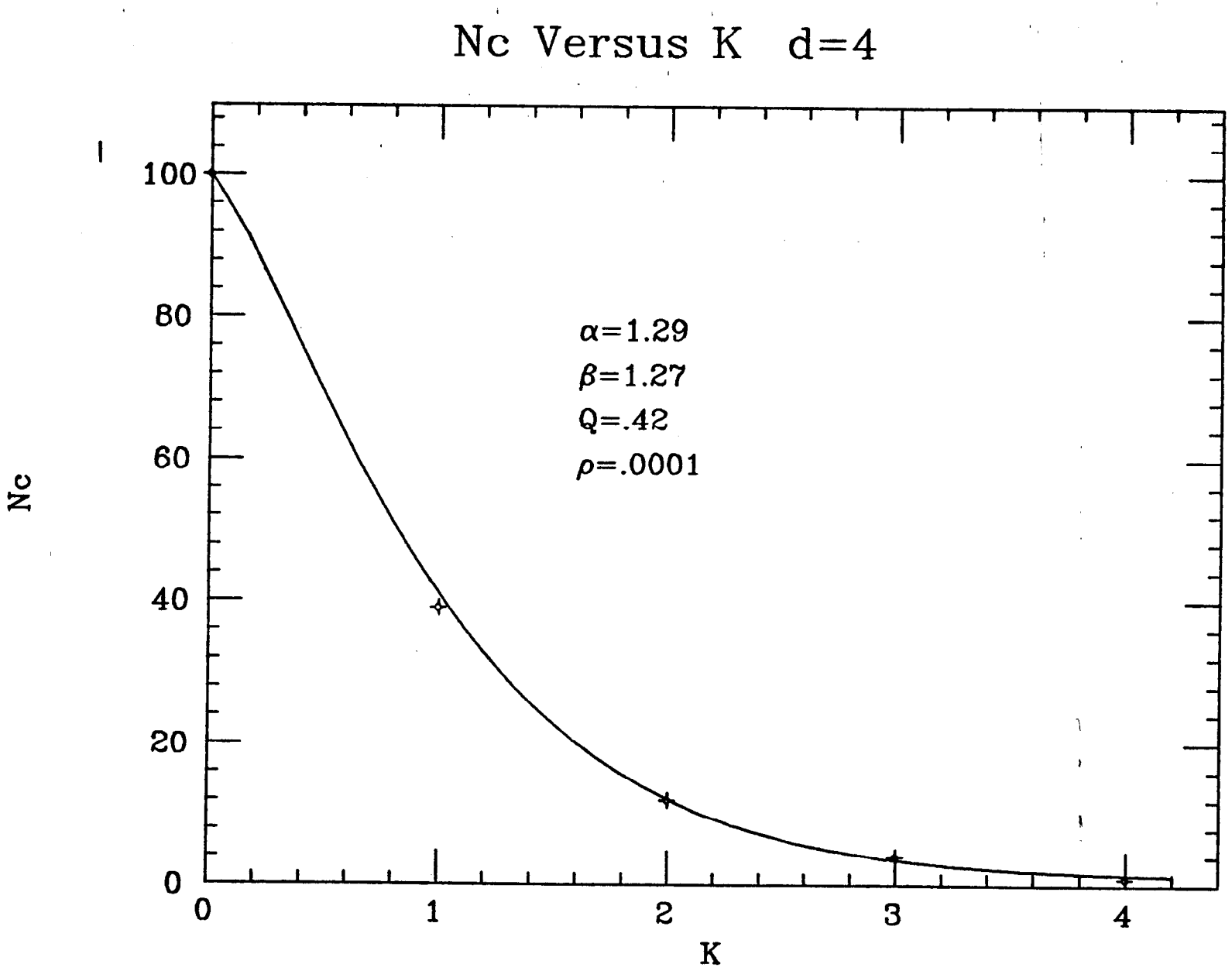


Fig.10

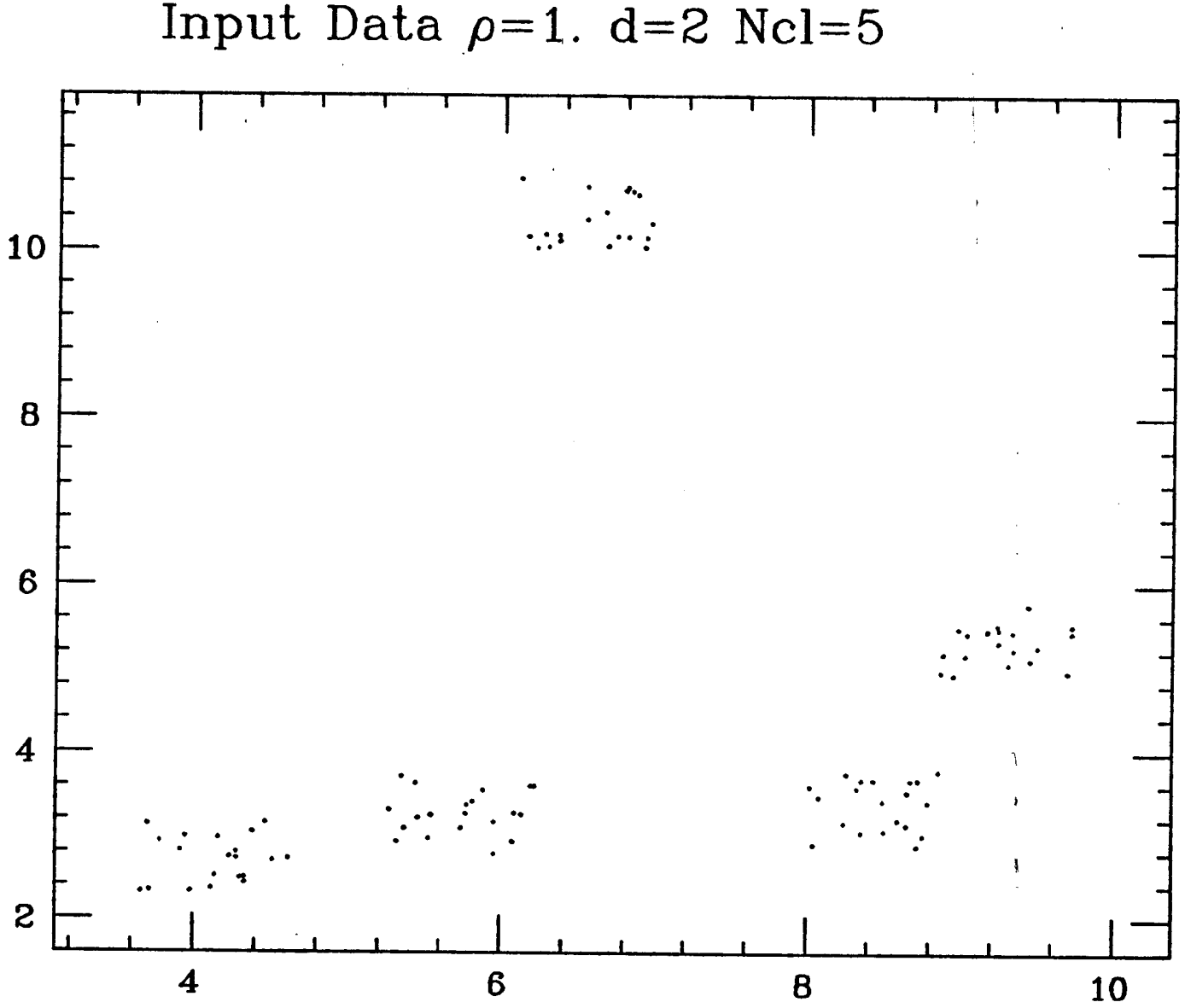


Fig.11

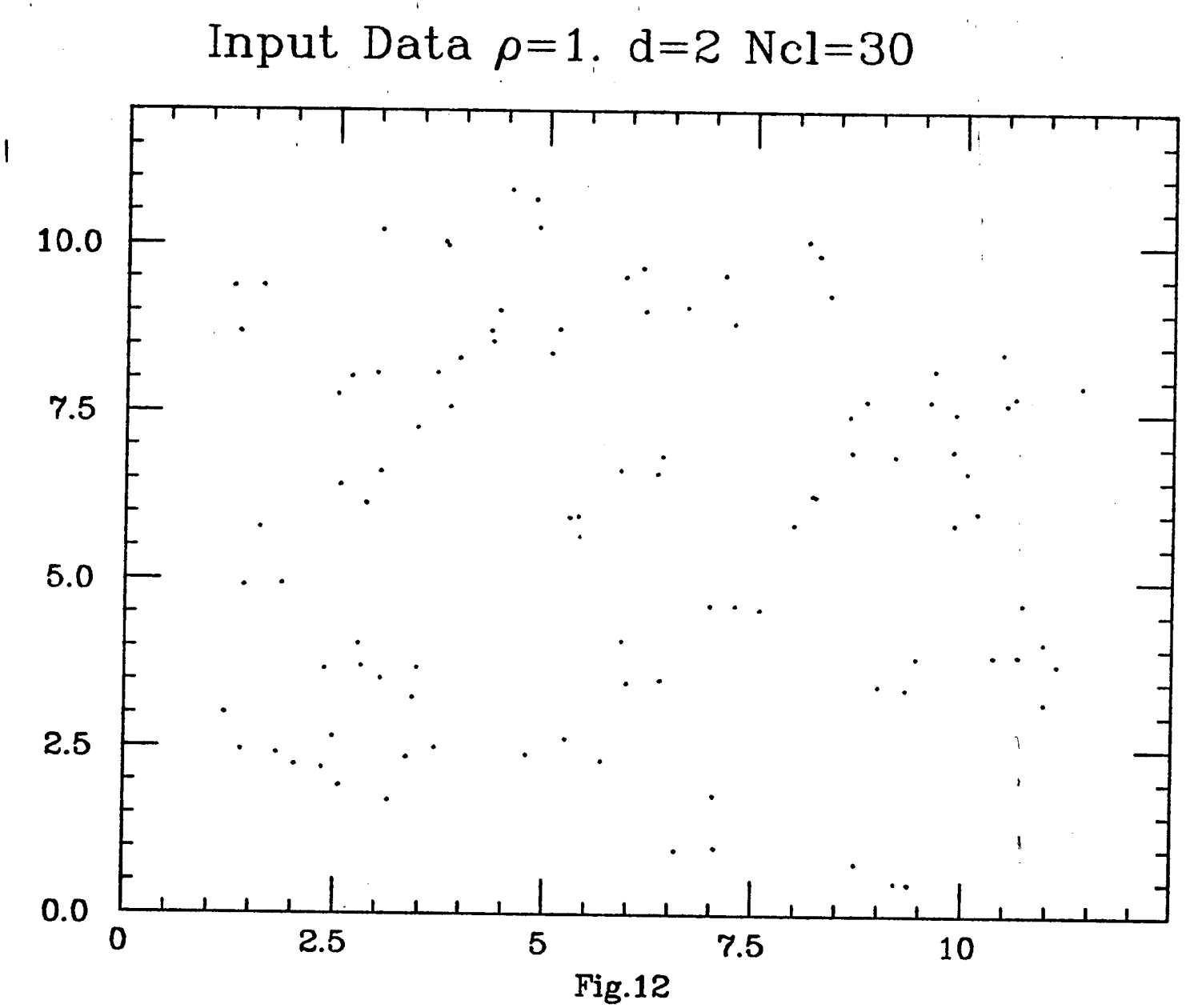
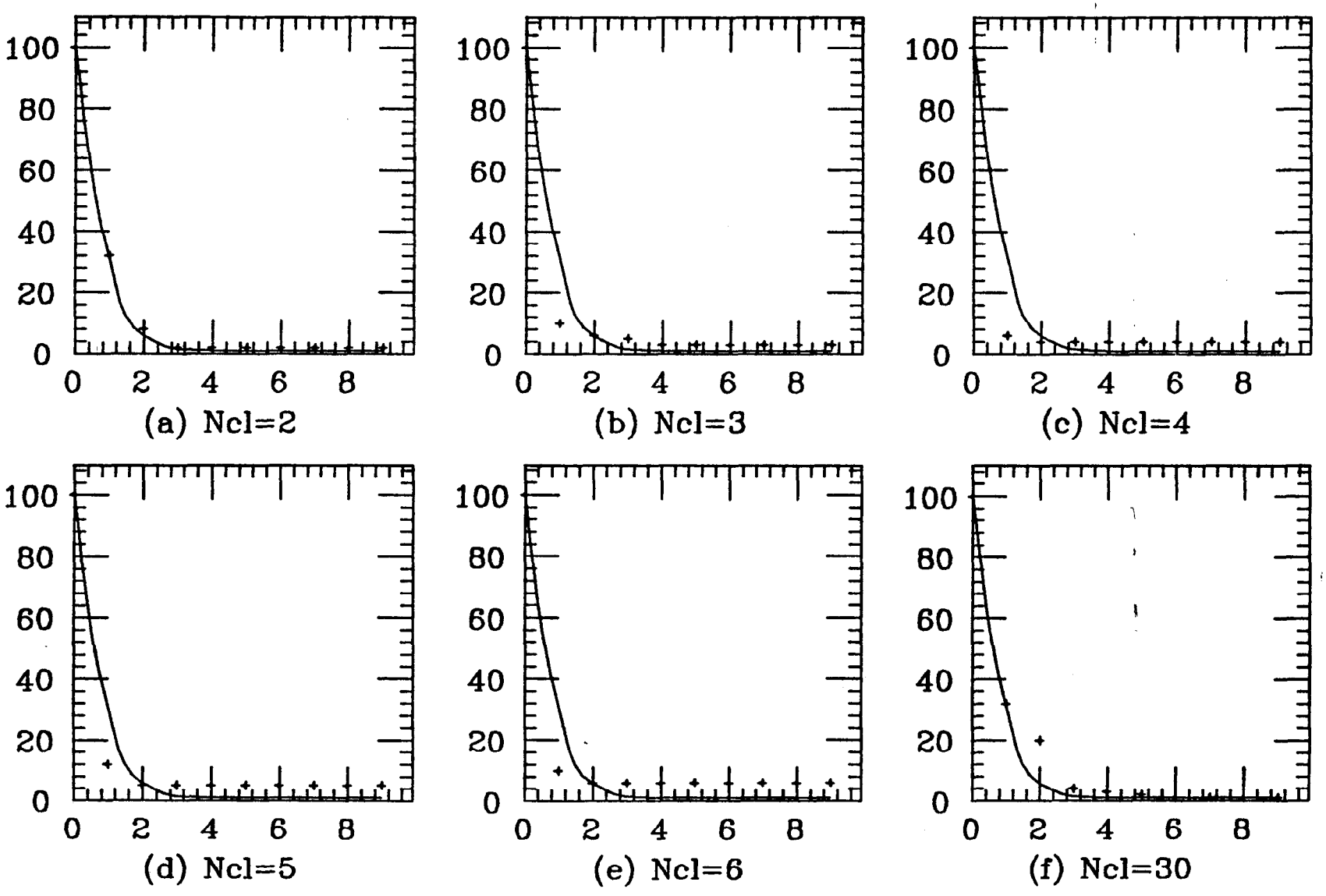


Fig.12

Fig.13 Results of The UT, Q<10.E−4



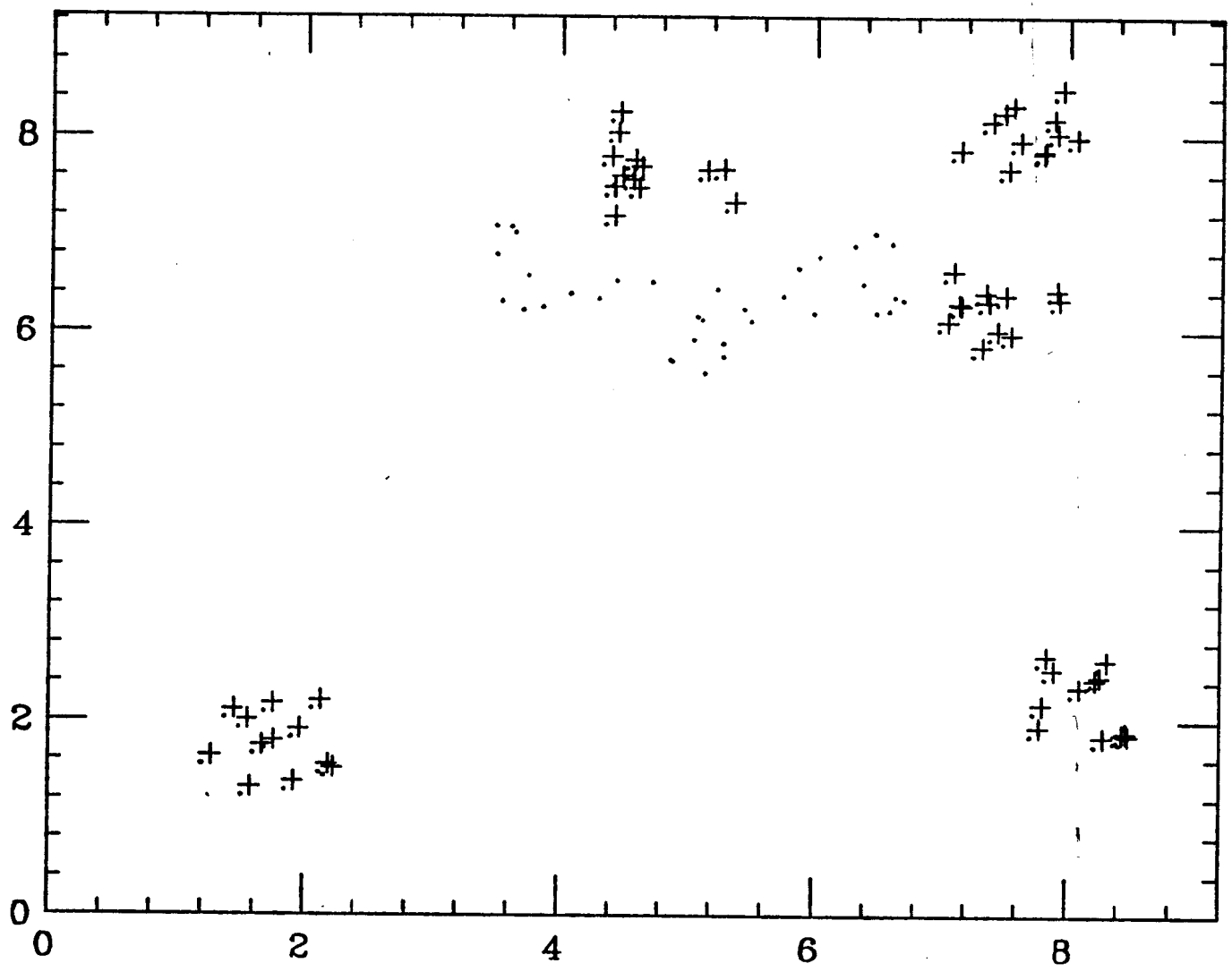


Fig.14