Disclaimer: unless otherwise agreed by the Council of UPOV, only documents that have been adopted by the Council of UPOV and that have not been superseded can represent UPOV policies or guidance.

This document has been scanned from a paper copy and may have some discrepancies from the original document.

Avertissement: sauf si le Conseil de l'UPOV en décide autrement, seuls les documents adoptés par le Conseil de l'UPOV n'ayant pas été remplacés peuvent représenter les principes ou les orientations de l'UPOV.

Ce document a été numérisé à partir d'une copie papier et peut contenir des différences avec le document original.

Allgemeiner Haftungsausschluß: Sofern nicht anders vom Rat der UPOV vereinbart, geben nur Dokumente, die vom Rat der UPOV angenommen und nicht ersetzt wurden, Grundsätze oder eine Anleitung der UPOV wieder.

Dieses Dokument wurde von einer Papierkopie gescannt und könnte Abweichungen vom Originaldokument aufweisen.

Descargo de responsabilidad: salvo que el Consejo de la UPOV decida de otro modo, solo se considerarán documentos de políticas u orientaciones de la UPOV los que hayan sido aprobados por el Consejo de la UPOV y no hayan sido reemplazados.

Este documento ha sido escaneado a partir de una copia en papel y puede que existan divergencias en relación con el documento original.

E

BMT/5/7
ORIGINAL: English
DATE: August 24, 1998

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

WORKING GROUP ON BIOCHEMICAL AND MOLECULAR TECHNIQUES AND DNA-PROFILING IN PARTICULAR

Fifth Session
Beltsville, United States of America, September 28 to 30, 1998

PHENOTYPIC DISTANCES PREDICTION ACCORDING TO MOLECULAR DATA

Document prepared by experts from France

Phenotypic Distances Prediction According to Molecular Data

C.Baril, G.Nuel and S.Robin

August 14, 1998

Abstract

In order to study the relationship between genetic and phenotypic distances. we propose a linear model linking phenotypic variables to molecular markers. Assuming that parameters are known, and conditionally to the markers. this model provides a confidence interval for phenotypic distances. Preserving a validation population, we have applied this model to maize data. Results are presented and discussed, as well as possible perspectives.

Contents

1 Introduction

The phenotypic distance between two lines calculated from measures of phenotypic characteristics depends on their sensitivity to the environments. The Mahalanobis distance involves the correction of the dependance structure between the quantitative variables.

The RFLP molecular markers have the advantage of being obtained quickly in laboratory with a good reproducibility. Using genetic markers linked to quantitative trait loci (QTL) , even partially, for phenotypic distances calculations could be of a great benefit for varietal description.

Many genetic distances have been proposed, elaborated from genetic data on the basis of similarity indexes. To analyse the relationship between genetic and phenotypic distances, Burstin and Charcosset (1997) have given some evidence of a triangular linking structure.

This linking structure founds some biological explainations:

- Compensation of such elementary characteristics in complex characteristics expression,
- the low part of genom involved in phenotypic caracteristics expression,
- the different histories in co-selection of characteristics not physically linked but which combination give an adaptative or selective advantage (variation of linkage desequilibrium),
- the lack of environmental conditions necessary for the expression of all characteristics.

We could also explain this phenomenon by the internal dependance structure of the data. Indeed if 1,2 and 3 are three individuals, D(1,2) and D(2,3) are (strongly) linked so it is possible that the triangular shape observed is a data artefact.

Starting from Burstin and Charcosset work, we have tried to develop a model linking genetic data to phenotypic variables.

2 Material

145 lines of maize, representative of the material released in France, were characterised by using both RFLP markers and morphological traits used in current distinctness studies (Dillmann et al, 1997).

Modified Rogers Distance (MRD) was calculated on RFLP data obtained from 80 monolocus probes, with one enzyme per probe.

Mahalanobis distance was calculated on morphological data for 10 quantitative traits , collected at three locations in France, during four years. Each location was planted with two replications in a block design.

Figure 1 shows the triangular shape of the relation between phenotypic and genetic distances (the weighting matrix beeing the variance-covariance matrix of the residual of the ANOVA model considering all the known factors). Only 100 lines were used in order to keep 45 lines for the validation of the model.

3 Method

The purpose is to link quantative phenotypic variables to genetic data by a classical linear model relation. For each variable, we consider a subset of specific markers (subset of Quantitative Trait Loci - QTL - linked to this variable) that will have a linear effect on this variable. The residuals are supposed to have a gaussian distribution.

Notations: n is the number of individuals ($i \in \{1, \ldots, n\}$)
m is the number of markers ($k \in \{1, \ldots, m\}$)
s_k is the number of alleles of the kth marker
$M = \sum_{k=1}^{m} s_k$
X_{ik} is the value of the kth marker for the ith individual ($X_{ik} \in \{1, \ldots, s_k\}$)
p is the number of morphological variables ($j \in \{1, \ldots, p\}$)
Yij is the value of the jth variable for the ith individual
The model is:

$$\begin{array}{ccccccccc} Y_{ij} & = & \mu_j & + & \mathbf{X}_i & \times & \Theta_j & + & E_{ij} \\ (1,1) & = & (1,1) & + & (1,M) & \times & (M,1) & + & (1,1) \end{array}$$

or:

$$\begin{array}{ccccccccc} \mathbf{Y} & = & \mu & + & \mathbf{X} & \times & \Theta & + & \mathbf{E} \\ (n,p) & = & (n,p) & + & (n,M) & \times & (M,p) & + & (n,p) \end{array}$$

and phenotypic distance between two individuals i and i' can be expressed in terms of molecular markers as follow:

$$d_P^2(i,i') = (\mathbf{Y}_i - \mathbf{Y}_{i'})\mathbf{M}(\mathbf{Y}_i - \mathbf{Y}_{i'})'$$

with:

$$(Y_{ij} - Y_{i'j}) = (\mathbf{X}_i\Theta_j - \mathbf{X}_{i'}\Theta_j) + (E_{ij} - E_{i'j})\,.$$

The parameter estimation, once chosen the QTL subset, needs a large number of lines. In practice, even if our model does not forbide interaction between markers, the low number of data forces us to ignore it.

4 Results

4.1 Triangular Linking Structure

Once estimated the parameters, the triangular distribution appared in simulation. Figure 2 shows the relationship between phenotypic and molecular distances. The weighting matrix used to workout mahalanobis distance was given by the variance-covariance matrix of the residual of the ANOVA model considering all the known factors (the same than the one of figure 1).

Moreover, we proved we could reject the assumption that such a structure would be a data artefact.

If we use the weighting matrix equal to the variance-covariance matrix of residuals considering the linear model between the two distances, we still obtain similar results with real data and simulated data (results not shown). Thus, we can consider that this model reflect the reality.

4.2 Phenotypics Distances Predictions

We decided to proceed conditionally to the markers rather than to the genetic distances. This gives the advantage to get rid of the choice of a genetic distance, which is anyway an abstract of the genetic informations we have, and of the lack of models for marker expression

In this condition and assuming that the estimated parameters are the true ones, the phenotypic distance between two individuals displays a non central χ^2 distribution, in which the degrees of freedom number is the phenotypic variables number, and the non central parameter is the distance between two lines predicted by the model. This predicted distance is both computed with markers, and thus is a kind of genetic distance, and corrected by model parameters which depend on phenotypic data.

Based on this distribution, we can predict for two lines a confidence interval which will contain the phenotypic distance.

As we can see in the table below, the predictions are not very good.

Prediction and 95% confidence interval

Type	Observed confidence
2	86 %
1	48 %
0	36 %

Type 2 stands for distances between two lines that has been both used for estimation (100 lines), type 1 stands for distances between a line used for estimation (100 lines) and a line not used for estimation (45 lines) and type 0 for distances between two lines of the validation population (45 lines). Figures 3 and 4 present the relationship between predicted data and real data for type 2 and for type 0, respectively.

Two explanations can be given to the lack of robustess of these predictions:

- the quality of the parameters estimation is poor, due to the choice of QTL and to the low number of data.
- parameters are only estimated ones (if they were true ones, the type 2 observed confidence would be 95%).

Some simulations indicated that the first hypothesis is the most important one. However we think the study of parameter estimation handling has to be pursued, as the question of lines number in relation to markers number will probably be important.

5 Conclusion

In order to be usefull in varietal management our prediction have to be more efficient. It will be obtain by several way:

- a more suitable choice of QTL.
- a larger number of lines for the same number of markers,
- the development of prediction that take care of the statistic of parameter estimation.

However, a prediction allow to make the distinction only if its lower bound is upside a threshold value, so our interval predictions have to be as shorter as possible.

6 References

Burstin J and Charcosset A (1997) Relationship between phenotypic and marker distances: theoretical and experimental investigations. Heredity 79:477-483

Dillmann C, Bar-Hen A, Guerin D, Murigneux A, Charcosset A (1997). Comparison of RFLP and morphological distances between maize *Zea mays L.* inbred linrd. Consequences for germplasm protection purposes.Theor Appl Genet

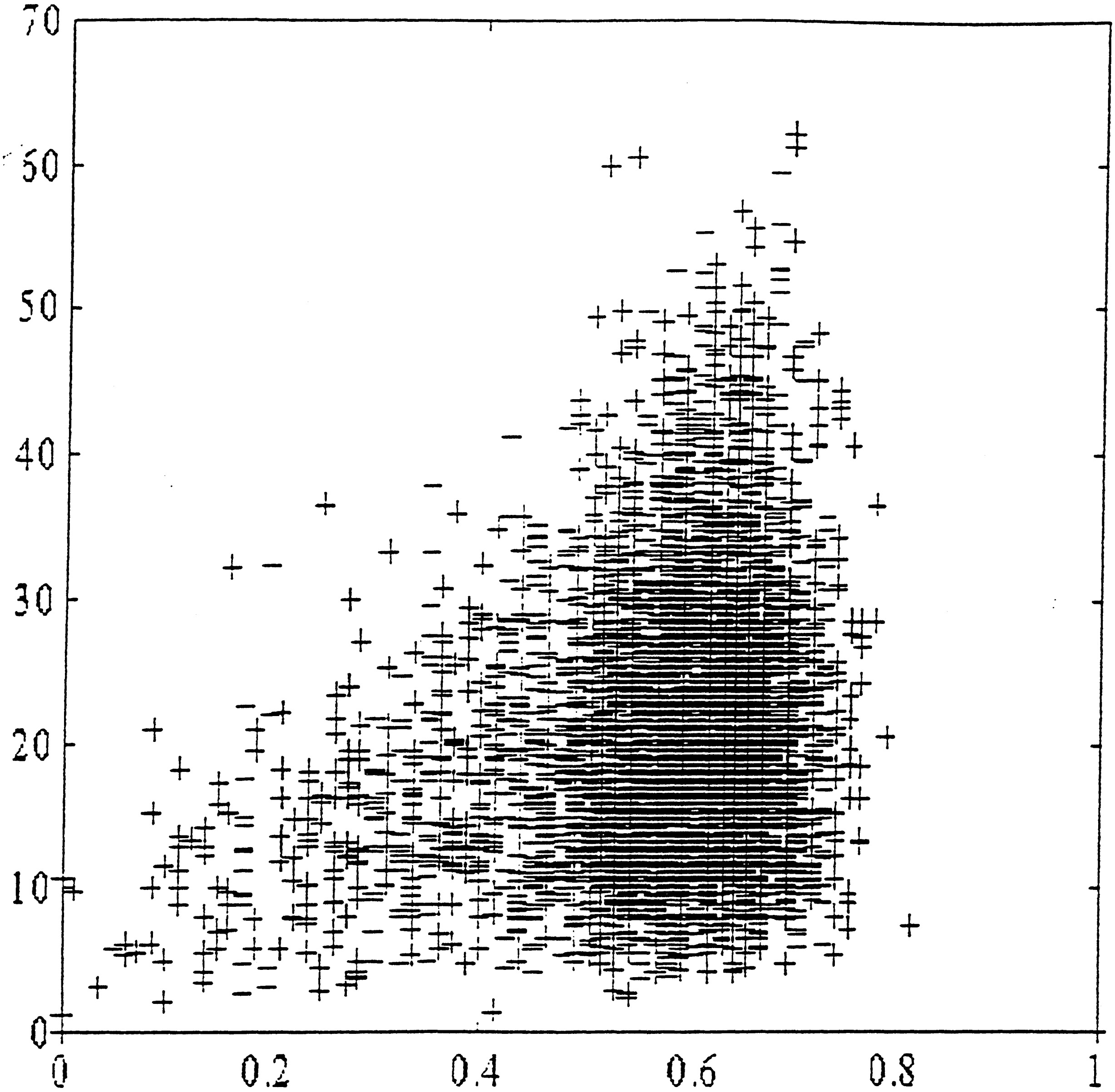

Figure 1: Relationship between phenotypic and genetic distances (real data)

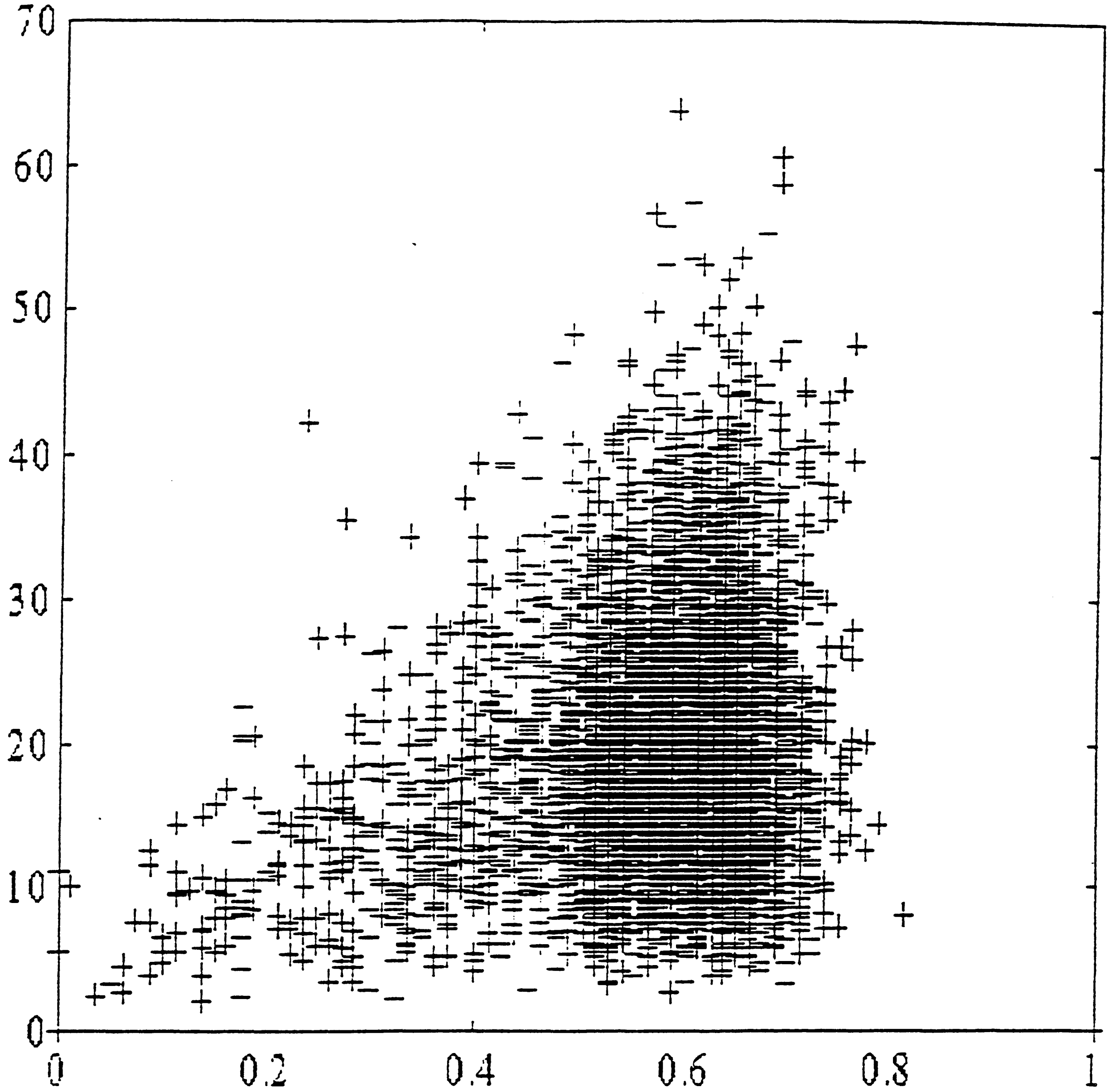

Figure 2: Relationship between phenotypic and genetic distances (simulated data)

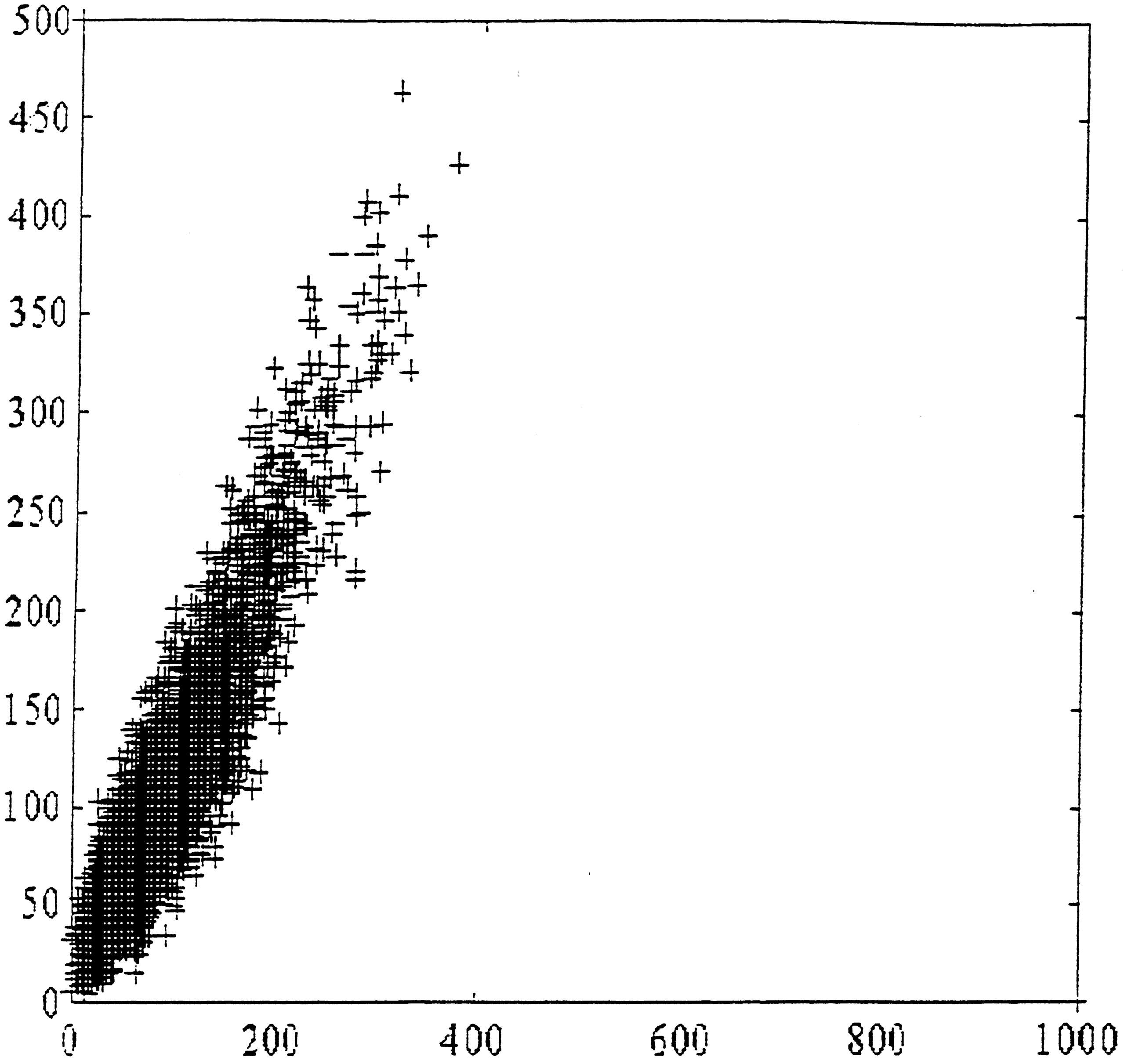

Figure 3: Relationship between predicted and real data of type 2

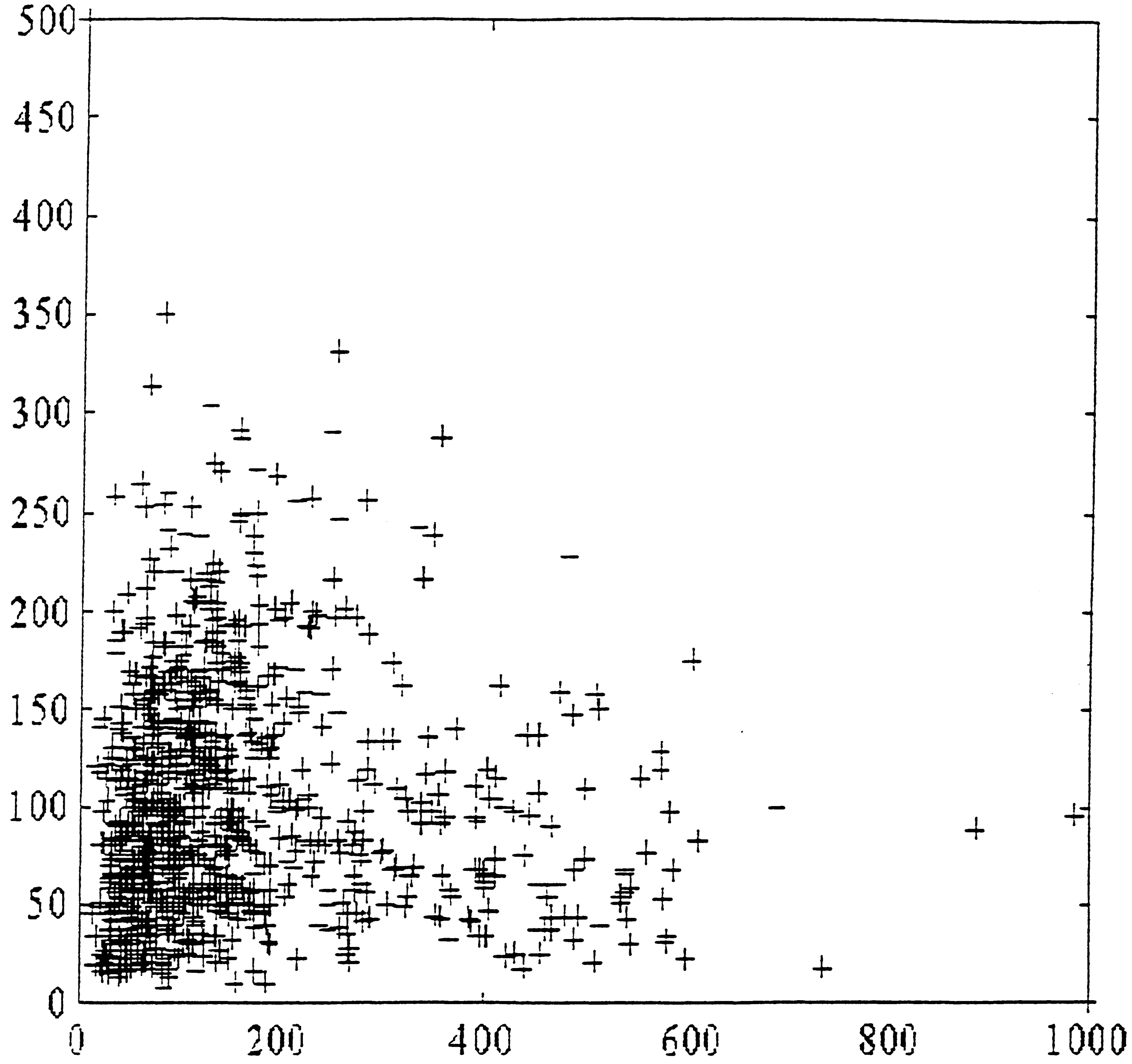

Figure 4: Relationship between predicted and real data of type 0

[End of document]