Biometrical Letters
Vol. 46 (2009), No. 1, 55-70

Using Multidimensional Unfolding in Plant Breeding Programs

Ana Alexandra A.F. Martins[1], Iola M.S. Pinto[1], Margarida G.M.S. Cardoso[2]

[1]Instituto Superior de Engenharia de Lisboa, Mathematics Cientific Area, Portugal,
anamartins@deea.isel.ipl.pt, ipinto@deetc.isel.ipl.pt
[2]ISCTE Business School, Departament of Quantitative Methods, Portugal,
margarida.cardoso@iscte.pt

SUMMARY

In the present work Multidimensional Unfolding (MDU) is suggested as a new approach to support decision-making in plant breeding programs. It is an exploratory data analysis technique that yields the construction of a map, picturing the attractiveness of cultivars towards planting environments. The proposed approach is illustrated using data from a wheat plant breeding program in Portugal (1986–1999). MDU precedes the use of Linear Joint Regression Analysis (LJRA) (e.g. Pinto 2006) in the study of relationships between cultivars and environments, taking into account the cultivars' performance evaluation. As regards the selection of the best cultivars, both approaches agree. Furthermore, MDU provides an additional advantage related to its easily interpreted results. In fact simplicity and interpretability may be considered the main advantages of the proposed approach.

Key words: Multidimensional Unfolding, Linear Joint Regression Analysis, Plant Breeding Programs, Cultivar Selection.

1. Introduction

In a plant breeding program, a set of experiments occur during a certain number of years and locations (environments) with an open set of cultivars. This means that during the breeding program, cultivars are discarded at the same time as new ones are admitted into the plan. The use of different locations provides information concerning the responses of the different varieties to environmental conditions characterized by a set of factors (e.g. climatic conditions and types of soil). The selection process is based on the cultivars' performance measured by the corresponding yields. In addition, the genotype × environment interactions are taken into account for evaluation of the cultivars' performance.

Multidimensional Unfolding (MDU) is a spatial distance model for the proximities between elements from two distinct groups of entities. The MDU output is a map where the elements of both sets of entities are represented by points and the distances between them represent the input proximity data. Common applications of MDU refer to non-metric data: typically subjects' preferences for a set of objects or stimulus, marketing and psychometry being the best-known fields of application (Desarbo et al. 1997, Carrol, Green 1997, Heiser, Busing 2004).

The present work suggests the use of metric MDU to construct a map representing data on cultivars' performance. In the (two-dimensional Euclidean) derived map, cultivars and environments are represented by points, and the yield of each cultivar in a given environment is associated with the corresponding distance. Besides this introduction, the paper contains three sections. In the following section the metric MDU model is presented, as well as the PREFSCAL algorithm which is used to perform MDU analysis. In the third section, an application using the Portuguese wheat plant breeding program data (1986–1999) is presented, illustrating the use of MDU. This analysis is complemented by the use of an inferential tool – Linear Joint Regression Analysis (LJRA) – which is commonly used in plant breeding program management. In the final section the application results are discussed and some future research topics are suggested.

2. Multidimensional Unfolding

MDU input refers to two-mode, two-way data, corresponding to the proximities between the elements from two distinct sets of entities. The goal of MDU is to obtain a configuration (commonly bi-dimensional and Euclidean) where the elements of both sets are represented by points. In the output map, the distances between the elements of one set of entities relative to the other represent the initial proximity data. The transformation process of proximities into distances depends on the metric nature of the input data: a distinction is made between metric and non-metric approaches. The non-metric models refer to ordinal data (e.g. subjects' preferences for a set of objects). In the metric case the input data refer to distance measures.

The unfolding model introduced by Coombs (1950) was generalized to the multidimensional case by Bennet, Hays (1960). These original techniques are non-metric.

Schönemann (1970) found an algebraic solution to the problem of locating two sets of points in a joint space, given the Euclidean distances between the elements of both sets. The metric MDU technique makes it possible to represent H-dimensional data distances with a map (2 dimensions) providing the fit between the original distances and the final configuration.

Further developments for multidimensional unfolding can be found in the context of multidimensional scaling, where unfolding is seen as a multidimensional scaling problem of off-diagonal matrices (e.g. Borg, Groenen 2005).

2.1 The Metric MDU Model

Metric MDU input data can either be a matrix of dissimilarities or a matrix of similarities between the elements of two sets of entities. Let δ_{ij} be the dissimilarity between i $(i=1,...,I)$ and j $(j=1,...,J)$ elements of the first and second sets respectively. Then the input dissimilarities matrix is $\Delta_{(I\times J)} = \lfloor \delta_{ij} \rfloor$. The MDS output is a configuration in a bi-dimensional Euclidean space, where the elements of both sets are represented by points with coordinates $\mathbf{u}_i$ and $\mathbf{v}_j$, and the Euclidean distance between them is denoted by d_{ij}.

The relationship between δ_{ij} and d_{ij} is given by

$$d_{ij}=f\left(\delta_{ij}\right)+e_{ij}, \tag{2.1}$$

f being a parametric function, and e_{ij} the random part corresponding to the measurement errors and the deviations associated with the obtained configuration. In metric models f is linear:

$$d_{ij} = a + b\delta_{ij} + e_{ij}. \tag{2.2}$$

The values $f\left(\delta_{ij}\right)$ are called disparities, and are usually represented by $\hat{d}_{ij}$. As minimization objectives, different alternative loss functions can be considered based on the error $e_{ij} = \hat{d}_{ij} - d_{ij}$. The simplest form for the loss function is the sum of the squared errors known as raw STRESS (Standardized Residuals Sum of Squares)

$$\mathrm{raw-STRESS} = \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \left(\hat{d}_{ij} - d_{ij} \right)^2 \right)^{1/2} . \tag{2.3}$$

Since this function is not invariant under uniform stretching and shrinking of the resulting configuration, a normalization factor is needed. $\mathrm{STRESS}-1$ and $\mathrm{STRESS}-2$ (Kruskal 1964, Kruskal, Carrol 1969) are the best-known formulae:

$$\mathrm{STRESS}-1 = \left(\frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \left(\hat{d}_{ij} - d_{ij} \right)^2}{\sum_{i=1}^{I} \sum_{j=1}^{J} d_{ij}^2} \right)^{1/2} , \tag{2.4}$$

$$\mathrm{STRESS}-2 = \left(\frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \left(\hat{d}_{ij} - d_{ij} \right)^2}{\sum_{i=1}^{I} \sum_{j=1}^{J} \left(d_{ij} - \bar{d} \right)^2} \right)^{1/2} . \tag{2.5}$$

Since there is no analytical solution to the STRESS minimization problems, an iterative optimization procedure involving the approximation of disparities by distances is used.

2.2 The PREFSCAL Algorithm

To the authors’ knowledge, PREFSCAL (Busing et al. 2005) is the most recent algorithm specifically developed for MDU. It considers a new objective (loss) function: the penalized STRESS (P-STRESS)[1]

[1] The Penalized STRESS formula admits several variants depending on the adopted model and on two parameters’ values (Busing et al. 2005). The present formula corresponds to the model adopted in this work.

$$P-STRESS=\left[\left(\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\hat{d}_{sc}-d_{sc}\right)^{2}}{\sum_{i=1}^{I}\sum_{j=1}^{J}d_{sc}^{2}}\right)^{1/2}\left(1+\left(\frac{cv(\delta)}{cv\left(\hat{d}\right)}\right)^{2}\right)\right]^{1/2} \qquad (2.6)$$

where $\mathrm{cv}(\delta)$ and $\mathrm{cv}(\hat{\mathrm{d}})$ are the variation coefficients of input data and disparities respectively. The first factor of P-STRESS corresponds to STRESS-1. The second factor was an innovation proposed in order to penalize solutions with "small" variation coefficients for disparities, trying to avoid degenerate solutions with equal inter-set distances.

P-STRESS is minimized by an alternating iterative procedure. It alternates between updating the configuration given a current estimate of the disparities, and updating the disparities given a current estimate of the configuration. Both steps are carried out using an iterative majorization procedure for minimizing P-STRESS (Busing et al. 2005, Borg, Groenen 2005).

3. Wheat Plant Breeding Data Analysis

3.1 The Data

The present work suggests MDU analysis as an exploratory tool to support a preliminary analysis of plant breeding data. The results provide information concerning the relationships between cultivars and environments based on the cultivars' yields for each environment.

The data relate to a wheat plant breeding program in Portugal (1986–1999), kindly forwarded by the Portuguese Plant Breeding Station. The data used correspond to nine years. For each year, a series of trials on cultivars are conducted at several locations, allowing study of the responses of the same set of varieties to different environmental conditions. For each location and year there are four replicates of the yield per cultivar. The locations and cultivars used in each year are presented in Table 1. This plant breeding program was managed without taking account of objective criteria, selection of cultivars being based on experts' knowledge and practical experience.

Table 1. Cultivars and locations used in each year

	Years								
	1986	1987	1988	1989	1991	1992	1995	1997	1999
	anza	anza	anza	almansor	almansor	almansor	almansor	almansor	almansor
C	flycatcher	flycatcher	flycatcher	alva	anza	anza	anza	anza	anza
U	hahn-s	hahn-s	lima1	anza	lima1	mondego	te9111	te9113	te9203
L	lima1	lima1	te8501	lima1	milan	te9002	te9112	te9114	te9406
T	miwivet-s	miwivet-s	te8502	liz1	te8802	te9101	te9113	te9203	te9503
I	neelkant-s	te8501	te8504	liz2	te8901	te9102	te9114	te9301	te9504
V	sunbird-s	te8502	te8601	te8603	te8902	te9111	te9203	te9302	te9712
A	te8401	te8504	te8602	te8701	te8906	te9112	te9301	te9303	te9713
R	te8501	te8601	te8603	te8702	te9001	te9113	te9302	te9406	te9714
S	te8502	te8602	te8701	te8801	te9002	te9114	te9303	te9503	te9715
	te8504	te8603	te8702	te8802	te9003		te9406	te9504	te9716
	Almeirim	Almeirim	Beja	Beja	Abrantes	Benavila	Beja	Beja	Comenda
Lo-	Comenda	Coruche	E.N.M.P.	E.N.M.P.	Beja	E.N.M.P.	Comenda	Comenda	Revilheira
ca-	Coruche	Fundão	Évora	Évora	E.N.M.P.	Mirandela	Revilheira	Revilheira	V.F. Xira
tions	Évora		Fundão	Fundão	Santarém	V.F. Xira			
	Mirandela		Lamaçais	Mirandela					

3.2. MDU Analysis

Let $y_{i,t,l,r}$ represent the data on the yield observed for the ith cultivar in year t, location l and repetition r, with $i = 1,...,I_t$, $t = 1,...,9$, $l = 1,...,L_t$ and $r = 1,...,4$, where I_t and L_t refer to the number of cultivars and locations respectively, considered in year t.

Since the selection process is typically analyzed yearly and the set of cultivars used changes accordingly, an MDU analysis is performed for each year. For a given year[2], a dissimilarity measure is considered

$$\delta_{ij} = \underset{S}{Max}\left(y_{ij}\right) - y_{ij}\,, \qquad (3.1)$$

where y_{ij} is the yield of the ith cultivar in environment j, S the set of all y_{ij} and each environment j corresponds to a pair (location, repetition), with $i = 1,...,I$ and $j = 1,...,J$. In the proposed approach each repetition in the same location is characterized by different environmental conditions. Thus the MDU input data matrix is given by $\Delta_{(I \times J)} = \lfloor \delta_{ij} \rfloor$.

[2] For notational simplicity, the index *t* (referring to the year) is dropped.

Metric MDU analysis is performed using the PREFSCAL algorithm[3]. For the starting configuration, 200 random starts are considered. In addition, the classical scaling configuration, which considers the triangular inequality for computing the distance values between rows and between columns, is used.

The MDU output provides the coordinates of the points representing cultivars and environments in the resulting bi-dimensional Euclidean map, illustrating their relationships. In this configuration, lower distances between the points correspond to higher yields. These distances can be used for selection of the best cultivar in a given environment.

3.3 MDU Results

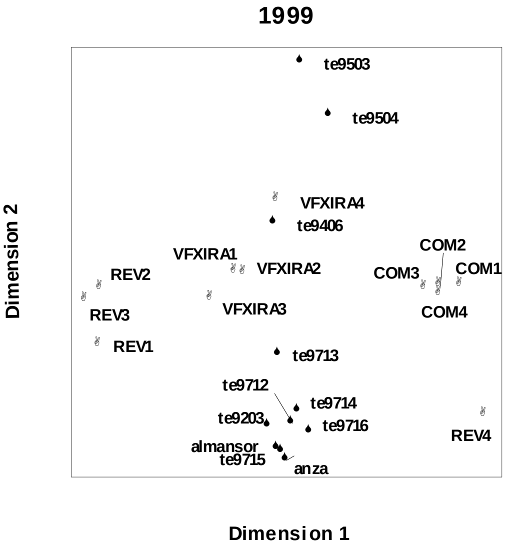
In order to evaluate the quality of MDU results, the R^2 measure is used (e.g. Busing et al. 2005)

$$R^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(d_{ij} - \bar{d}\right)^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\hat{d}_{ij} - \bar{\hat{d}}\right)^2}, \qquad (3.2)$$

where $\bar{\hat{d}}$ and $\bar{d}$ represent the average of disparities and distances respectively. It is the proportion of disparity variance which is accounted for by the distances in the map.

The derived MDU maps regarding plant breeding data from 1986 to 1988 are presented in the Appendix (Figures A.1 and A.2). In order to fully illustrate the maps' interpretation, the year 1999 is considered (Figure 1). Results from MDU analysis correspond to a good model fit: R^2 ranges from 0.785 to 0.966. This fact enables interpretation of the maps, which helps support the selection of cultivars.

[3] SPSS implementation.



$R^2 = 0.902$

Figure 1. MDU result maps for 1999. The circles represent the locations and the stars the cultivars. Legend for locations: Comenda – COM; Revilheira – REV. The number after the location indicates the repetition. Below each map the R^2 result is presented.

In the maps, the lower is the distance between a cultivar and an environment, the higher is the yield. Thus for a given environment, the best cultivar is the closest one. For example, in the 1999 map (see Figure 1), the cultivar te9406 is the closest to the environment VFXIRA4, which means that this cultivar is the one with the highest yield for this environment.

Furthermore, the map provides a global overview of the relationships between cultivars and environments. Some environments suggest particular productivity conditions. For example (see Figure 1), the environment REV4 is isolated, which means that it offers particular productivity conditions. The map also enables the identification of groups of environments characterized by similar productivity conditions, for example, the environments COM1, COM2, COM3 and COM4 constitute a separate group. Furthermore, the map also suggests the existence of cultivar groups characterized by similar yield behaviour for the considered environments. The cultivars are distributed into three distinct groups: 1) te9503 and te9504; 2) te9406; 3) the remaining cultivars.

3.4 LJRA in Plant Breeding Programs

Linear Joint Regression Analysis (LJRA) integrates a set of techniques used to study the genotype × environment interactions (Aastveit, Mejza 1992). It has been widely used for comparison of cultivars' performance. Specifically, it is based on a set of linear regressions, one per cultivar, which consider the yield as the dependent variable. The independent variable is the environmental index that measures, for each location, the corresponding productivity.

Initially, randomized blocks were used and their mean yields were taken as the corresponding environmental indices. Later on, following Patterson and Williams (1976), α-designs, which have incomplete blocks, superseded randomized blocks. In this context, the environmental indices for each block could not be measured by the respective mean yield, which would lead to biased estimates. This problem was solved by the introduction of L_2 environmental indices (Mexia et al 1999). The authors proposed approach relies on the application of a zigzag algorithm, which simultaneously estimates the regression coefficients and the environmental indices. Its estimation is based on the following quadratic goal function:

$$S(\alpha_I, \beta_I, x_L) = \sum_{i=1}^{I} \sum_{l=1}^{L} w_{il} (y_{il} - \alpha_i - \beta_i x_l)^2$$

In order to deal with different sets of cultivars, as well as of locations for each of the different years, the weights w_{il} are considered (Dias 2000) which equal 1[0] when the ith cultivar is present [absent] in the lth location. I and L refer to the number of cultivars and locations, respectively, considered in the year, and y_{il} is the mean yield corresponding to the four replicates of the ith cultivar in the lth location.

The goal is to minimize S, obtaining estimates of the intercept α_i $(i = 1,..., I)$, of the β_i $(l = 1,..., L)$, the slope of the regression line to be adjusted for the i^{th} cultivar and of the x_l $(l = 1,..., L)$ the environmental indices. To perform the minimization, we may use the zigzag algorithm (Mexia et al. 1999) in which the minimization is carried out alternately in the regression coefficients and in the environmental indices. Once the adjustment is complete, a joint representation of the regression estimated lines called the upper contour (introduced by Mexia et al. 1997) is obtained. The information supplied by the upper contour allows easy identification of cultivars with maximum yields, for certain ranges of the environmental indices. Geometrically, the upper contour is a convex polygonal. The cultivars that appear on the upper contour are the dominant ones. For each one of these, there is a dominance range constituted by the environmental indices for which they have highest yield. Non-dominant cultivars should be compared with the dominant ones using the appropriate parametric tests (Pinto 2006). As a result of these comparisons, cultivars that are significantly dominated should be discarded.

3.5 LJRA Results

For each of the nine years presented in the data in section 3.1, the LJRA technique was applied. As a result of application of the zigzag algorithm, by year, the vector of estimated environmental indices is obtained, as well as the regression estimated lines for each one of the cultivars. Once the adjustment is complete, a joint representation of the regression estimated lines – the upper contour – is obtained (for more details see Pinto 2006). The derived results enable one to select the best cultivar for each environment.

In order to fully illustrate such a process, the year 1999 is considered. The results of the zigzag algorithm, estimated environmental indices and coefficients of the estimated regression lines are presented in Tables 2 and 3 respectively. The corresponding determination coefficients range from 0.709 to 0.909. In Figure 2 the estimated regression lines are represented, where the bold line defines the upper contour. The estimated lines that integrate the upper contour correspond to the dominant cultivars. They have higher estimated yields for certain ranges of environmental indices.

Table 2. The estimated environmental indices for 1999

Environments	Comenda	Revilheira	V.F.Xira
$\hat{x}_1$	3.828	3.054	5.271

Table 3. Coefficients of the estimated regression lines for the cultivars in 1999

Culti-vars	alman-sor	anza	te9203	te9406	te9503	te9504	te9712	te9713	te9714	te9716
$\hat{\alpha}_i$	0.573	0.347	0.278	-1.318	-0.505	-1.152	0.182	-0.511	1.080	0.756
$\hat{\beta}_i$	0.843	0.872	0.936	1.440	1.019	1.235	0.961	1.209	0.785	0.819

Table 4 shows the dominant cultivars and the corresponding dominance ranges for the year 1999.

Finally, given an environment, the selection of the best cultivar proceeds as follows. For example, . considering the environment VFXIRA, the corresponding estimated environmental index is 5.271 (Table 2). This value belongs to the second dominance range (Table 4), leading to identification of te9406 as the dominant cultivar.

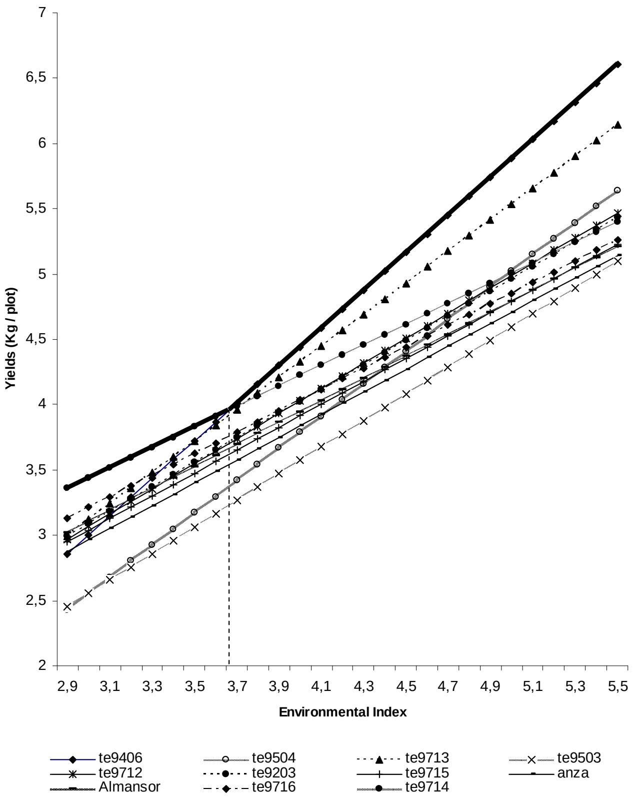


Figure 2. Upper contour obtained for the year 1999

Table 4. Derived dominance ranges and the corresponding dominant cultivars for 1999

Dominance Range	Dominant Cultivar
[2.901, 3.66]	te9714
[3.66, 5.422]	te9406

4. Discussion and Perspectives

In present work we first propose MDU analysis as an exploratory tool making it possible to obtain preliminary conclusions regarding the relationships between cultivars and the environments. Observing the MDU output map, we easily obtain information concerning the cultivars' performance. Specifically, for each environment, the best cultivar is the closest one.

In addition, the LJRA is suggested in order to quantify the obtained preliminary results, providing them with inferential support.

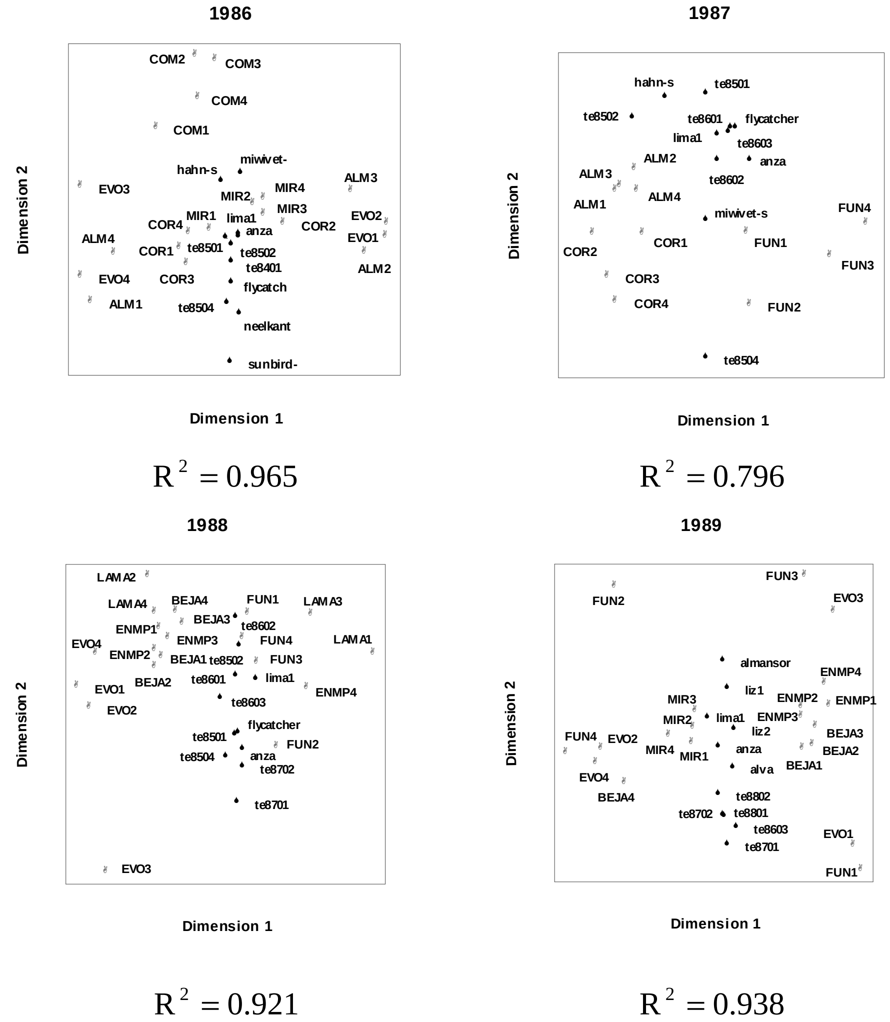
For testing the association between LJRA and MDU results, concerning the selection of the best cultivar, the chi-square test is used. For each environment (location, year), the cultivars are dichotomically classified as being or not being the best ones according to the classification criteria based on the results from both techniques. The test indicates rejection of the null hypotheses at a significance level of 0.001 (p-value close to zero). Thus there is significant agreement between the MDU and LJRA results.

Furthermore, MDU provides an additional advantage related to its easily interpreted results. In fact, simplicity and interpretability may be considered the main advantages of the proposed approach.

The research presented in this paper suggests a methodology for the analysis of data from plant breeding programs. It uses MDU as an exploratory tool and LJRA for conducting an inferential study. Future work could use more complex models to reach a more detailed inferential analysis (e.g. Calinski et al. 1997, Calinski et al. 2005).

Finally, additional information concerning the characterization of environments could improve the analysis of the responses of the different varieties to environmental conditions.

5. Appendix



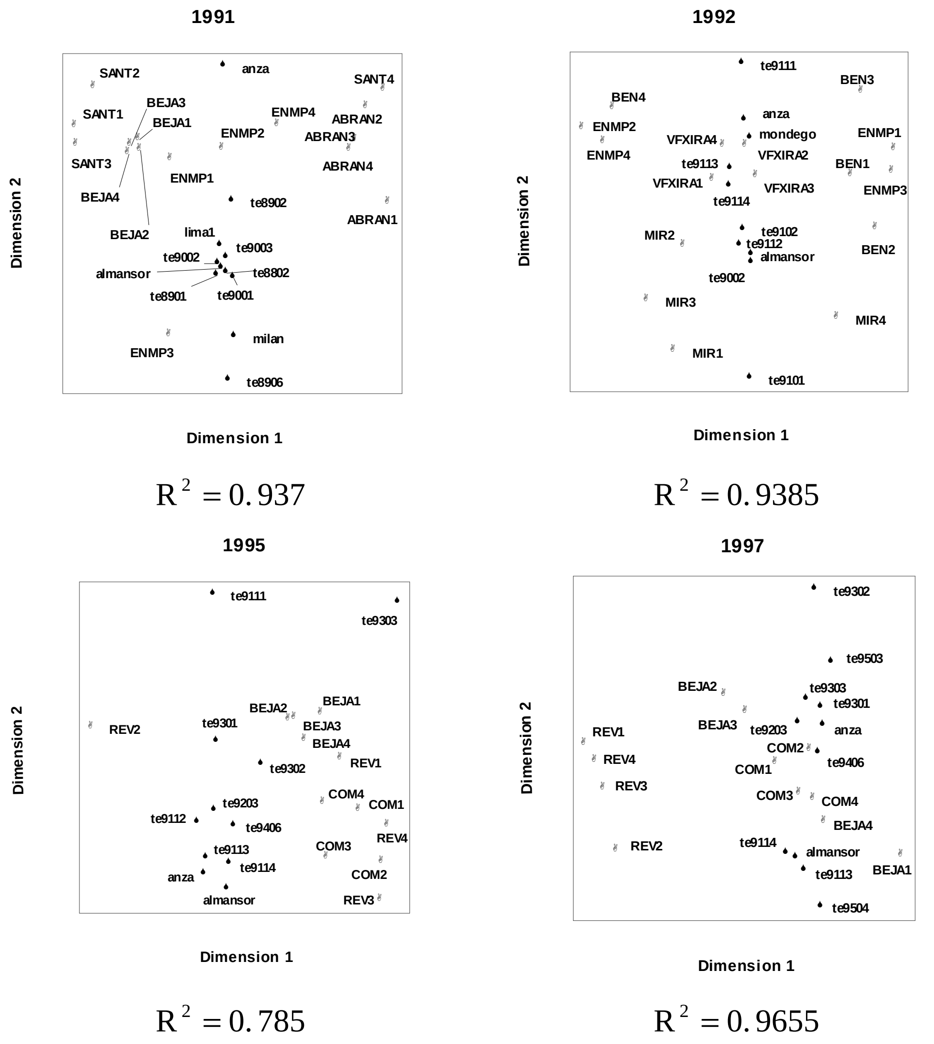
$R^2 = 0.965$

$R^2 = 0.796$

$R^2 = 0.921$

$R^2 = 0.938$

Figure A.1. MDU result maps for 1986, 1987, 1988 and 1989. The circles represent the locations and the stars the cultivars. Legend for locations: Almeirim – ALM; Comenda – COM; Coruche – COR; Évora – EVO; Fundão – FUN; Lamaçais – LAMA. The number after the location indicates the repetition. Below each map the R^2 result is presented



$R^2 = 0.937$ $R^2 = 0.9385$

$R^2 = 0.785$ $R^2 = 0.9655$

Figure A.2. MDU result maps for 1991, 1992, 1995 and 1997. The circles represent the locations and the stars the cultivars. Legend for locations: Abrantes – ABRAN; Santarém – SANT; Mirandela – MIR; Benavila – BEN; Comenda – COM; Revilheira- – VER. The number after the location indicates the repetition. Below each map the R^2 result is presented.

REFERENCES

Aastveit A.H., Mejza S. (1992): A selected bibliography on statistical methods for the analysis of genotype x environment interaction. Biuletyn Oceny Odmian 24–25: 83–97.

Bennet J., Hays W. (1960): Multidimensional unfolding: Determining the dimensionality of ranked preference data. Psychometrika 25: 27–43.

Borg I., Groenen P. (2005). Modern Multidimensional Scaling: Theory and Applications. Second Edition, Springer.

Busing F., Groenen P., Heiser W. (2005): Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. Psychometrika 70(1): 71–98.

Calinski T., Czajka S., Kaczmarek Z. (1997): A multivariate approach to analyzing genotype-environment interactions. Advances in Biometrical Genetics, Proceedings of the Tenth Meeting of the EUCARPIA section Biometrics in Plant Breeding Poznan, 14–16 May 1997. P. Krajewski and Z. Kaczmarek (eds.): 3–14.

Calinski T., Czajka S., Kaczmarek Z., Krajewski P., Pilarczyk W. (2005): Analysing Multi-environment Variety Trials Using Randomization-Derived Mixed Models. Biometrics 61: 448–455.

Carrol J.D., Green P. (1997): Psychometric methods in marketing research: Part II, Multidimensional Scaling. Journal of Marketing Research XXXIV: 193-204.

Coombs C. (1950): Psychological scaling without a unit of measurement. Psychological Review 57: 148–158.

Coombs C., Kao R. (1960): On a connection between factor analysis and multidimensional unfolding. Psychometrika 25: 219–231.

DeSarbo W., Young M., Rangaswamy A. (1997): A parametric multidimensional unfolding procedure for incomplete nonmetric preference/choice set data in market research. Journal of Marketing Research, XXXIV: 499–516.

Dias C.P. (2000): Análise Conjunta Pesada de Regressões. Master's Thesis. Évora.University.

Heiser W., Busing F. (2004): Multidimensional Scaling and Unfolding of symmetric and asymmetric proximity relations. In The Sage Handbook of Quantitative Methodology for the Social Sciences. David Kaplan (eds), Sage Publications: 26–48

Kruskal J. (1964): Multidimensional scaling by optimization goodness-of-fit to a nonmetric hypothesis. Psychometrika 29: 1–27.

Kruskal J., Carroll J. (1969): Geometrical models and badness-of-fit functions. In Multivariate Analysis, vol. 2. Krishnaiah (eds.), Academic Press, New York: 639–671.

Mexia J.T., Amaro A.P., Gusmão L., Baeta J. (1997): Upper Contour of a Joint Regression Analysis. J. Genet. and Breed (eds.) 51: 253–255.

Mexia J.T., Pereira D.G., Baeta J. (1999): L2 environmental indices. Biometrical Letters 36: 137–143.

Patterson H.D., Williams E.R. (1976): A new class of resolvable incomplete block designs. Biometrika 63: 83–92.

Pinto I. (2006): Joint Regression Analysis and Plant Breeding Programs. PhD Thesis, Faculty of Sciences and Technology of New University of Lisbon.

Schönemann P. (1970): On metric multidimensional unfolding. Psychometrika 35: 349-366.