*Statistical Science*
1990, Vol. 5, No. 2, 175–226

# OMEGA (Online Multivariate Exploratory Graphical Analysis): Routine Searching for Structure

**Claus Weihs and Heinz Schmidli**

*Abstract.* A strategy for online multivariate exploratory graphical analysis is presented and illustrated, motivated by the need for a routine procedure for searching for structure in multivariate data sets arising in the context of a major pharmaceutical, dyestuffs and agrochemical company.

*Key words and phrases:* Multivariate exploratory data analysis, interactive graphics, variable selection, resampling, Procrustes.

## 1. INTRODUCTION

### 1.1 Background and Motivation

Multivariate exploratory data analysis is an area of user-oriented applied statistical research of considerable interest to the Mathematical Applications group of CIBA-GEIGY in Basel, whose members, serving as internal statistical consultants, are faced with developing statistical methodologies routinely applicable in a wide range of industrial settings. In this context, where sizeable data bases are now being routinely generated as a consequence of technological advance in information systems and data collection techniques, statistical investigations are aimed both at initially uncovering structure and eventually proposing models in a variety of application contexts. Examples of such applications include:

- identification of structure-activity relationships, e.g., relating molecular structure with biological properties;
- relating product quality, e.g., of dyestuffs and pigments, to the mix or characteristics of the chemical compounds in different stages of the production process;
- understanding the relationships among different possible measures of the same phenomenon; e.g., can spectroscopic techniques replace the human eye in assessing color quality of dyestuffs.

*Claus Weihs and Heinz Schmidli are members of the Mathematical Applications unit at CIBA-GEIGY, Switzerland. The mailing address of the first author is CIBA-GEIGY Ltd., Mathematical Applications, R-1045.2.09, CH-4002 Basel, Switzerland.*

The industrial and scientific motivations in these and numerous other applications include the desire to improve design and control of production processes, to identify uninformative variables in order to reduce the scale of data recording and monitoring, and to obtain greater scientific insight into relationships between physicochemical and biological or other quality characteristics of a product at and between different production stages.

*Algebraic dimension reducing* Multivariate Exploratory Data Analysis (MEDA) methods and *interactive, dynamical graphical* analysis of their outcome by means of Visual Exploratory Data Analysis (VEDA) methods provide the two starting points for the analysis of the kinds of multivariate data sets encountered over and over again in CIBA-GEIGY applications. Young, Kent and Kuhfeld (1988) are primarily responsible for the terms MEDA and VEDA as they are used here, and also for the proposal of combining these two kinds of methods. The paper by Young, Kent and Kuhfeld is part of a recent overview by Cleveland and McGill (1988) of the history and the state of the art of VEDA techniques. Unfortunately, most of the exposition in this book and elsewhere is largely technical, presenting just one more method or software tool, illustrated by examples particularly fitted to demonstrate their usefulness. From the perspective of an industrial applied statistics group, such presentations proceed as if, given an individual data set, the investigator has virtually unlimited time and creative energy to try out a wide variety of methods. In contrast, in industrial practice the investigator is faced with an ongoing stream of many data sets, limited time and the need for a fairly general single routine strategy. It is this latter "gap" between methodological exposition and applied exigencies which

motivates this paper, in which we describe and illustrate a strategy (OMEGA), combining MEDA and VEDA techniques, that has been found to be an effective routine research tool in our own industrial context.

### 1.2 Outline of the Paper

In Section 2, we give an overview of the OMEGA strategy. Sections 3 and 4 describe the dimension-reducing MEDA tools and the VEDA tools for dynamic graphics which are the building blocks of the strategy. In Section 5, the practical use of the strategy is demonstrated with an extended analysis of a CIBA-GEIGY application. Section 6 concludes the paper with a short discussion of the generality of the strategy and of possible extensions.

## 2. THE OMEGA STRATEGY

### 2.1 The Concept

The strategy of Online Multivariate Exploratory Graphical Analysis (*OMEGA* strategy) combines dimension reduction methods (*MEDA* techniques) with dynamic graphics methods (*VEDA* techniques) within the framework of an overall data analysis perspective. OMEGA is an atttempt to meet the need implicitly identified by Young, Kent and Kuhfeld (1988, page 422): "A highly interactive, highly integrated MEDA/VEDA system that incorporates all of the MEDA models and all of the ... VEDA notions ... would indeed be a ... very useful tool for exploring, understanding, and forming hypotheses about the structure of multivariate data." The organization of an OMEGA is very much motivated by the concept of a *viewing pipeline* as introduced by Buja, Asimov, Hurley and McDonald (1988) for graphical analyses. Indeed, insofar as it proposes a sensible ordering for using the involved MEDA/VEDA techniques, an OMEGA can also be looked at as a data analysis pipeline corresponding to a sensible organization for exploratory data analysis.

Let us now introduce the *OMEGA concept* and the *OMEGA pipeline* in more detail. The term *Online Multivariate Exploratory Graphical Analysis* (*OMEGA*) describes an online study of relationships in a multivariate data set, where, rather than testing one specific property, as many clues as possible for interesting structures are searched for by different dimension reductions and succeeding graphical analysis. At least three characteristic aspects of such an OMEGA should be emphasized:

1. the *experimental character* of the search, in which the data are analyzed by interactively (online) switching between different approaches;
2. the *dimension reduction*, i.e., the transition from high-dimensional data to a low-dimensional transformation which represents the properties of interest of the data as well as possible;
3. the interactive, dynamical *graphical analysis* of the low-dimensional data, exploiting the great perspective power of the human eye for the identification of structure.

Interactive, dynamic computer graphics is a relatively young field of research mainly due to previously restricted access to suitable hardware. However, since about 1984 a generation of low-cost computer graphics systems has become available, giving the user the capability to realize virtually instantaneous graphical changes on a display. This is sometimes called "real-time-graphics: plots are recomputed and redrawn so rapidly that the visual effect of smooth motion can be achieved" (Buja, Asimov, Hurley and McDonald, 1988, page 278). If this is combined with a user-friendly control device like a mouse or well arranged keys, dynamic high-interaction graphical methods are at hand, the importance of which for exploratory data analysis can hardly be overestimated.

### 2.2 The OMEGA Pipeline

At the time of writing, our OMEGA pipeline at CIBA-GEIGY includes only a selection of the classical linear MEDA techniques, but most of the new interactive VEDA techniques. The available *MEDA techniques* are: Principal Components Analysis on COVariances (PCA-COV), Principal Components Analysis on CORrelations (PCA-COR), Canonical Discriminant Analysis (CDA), Canonical Correlation Analysis (CCA), and Successive OrthoGonalization (SOG). Possible extensions to the MEDA-tool box will be discussed in Section 6. The available *VEDA techniques* may be classified in broad terms as: one-window techniques, such as rotation, interpolation, and viewport transformations such as scaling, translating, subsetting, marking and identifying; multiwindow techniques, such as brushing in scatterplot matrices, as well as parallel analyses of nonlinear transformations and of pseudo-samples generated for graphical testing or for studying variation of the outcomes of the MEDA techniques with resampling. Since the MEDA techniques are well-discussed in the literature, we will give here only a short unified description, sufficient to form the basis for the exposition of some extensions and for the discussion of properties relevant for graphical representation. In comparison, the VEDA techniques are much less discussed in the literature and therefore need much more consideration concerning their merits and drawbacks. One of the advantages of an integrated look at

algebraic and graphical methods is that in this way the relation of some VEDA techniques to practical problems will immediately become clear. This will also be illustrated by the discussion of our OMEGA *pipeline* as it is presented in Figure 1.

Let us start the discussion of the pipeline by means of a very simple application: a scatterplot of two variables from a multivariate data set. For this, we need no more than a capability to select two variables (dimension reduction), and to specify the so-called window-to-viewport transformation, which affinely maps a rectangle in the two variable plane to a rectangle in the screen plane (viewporting) (Buja, Asimov, Hurley and McDonald, 1988, pages 279–280). This illustrates the general use of the pipeline: no step between "raw data" and "plot" is indispensable except "viewporting," and "dimension reduction" if more than two dimensions are involved. All the other steps are optional. Moreover, the ordering of the building blocks in the pipeline is just one of several reasonable forms of organization; others may be more adequate for other kinds of problems. For the rest of this section, the building blocks will be surveyed one by one, starting from the blocks which are the most important and most central in the pipeline, and elaborating more on new ideas than on ideas already well-rehearsed in the literature.

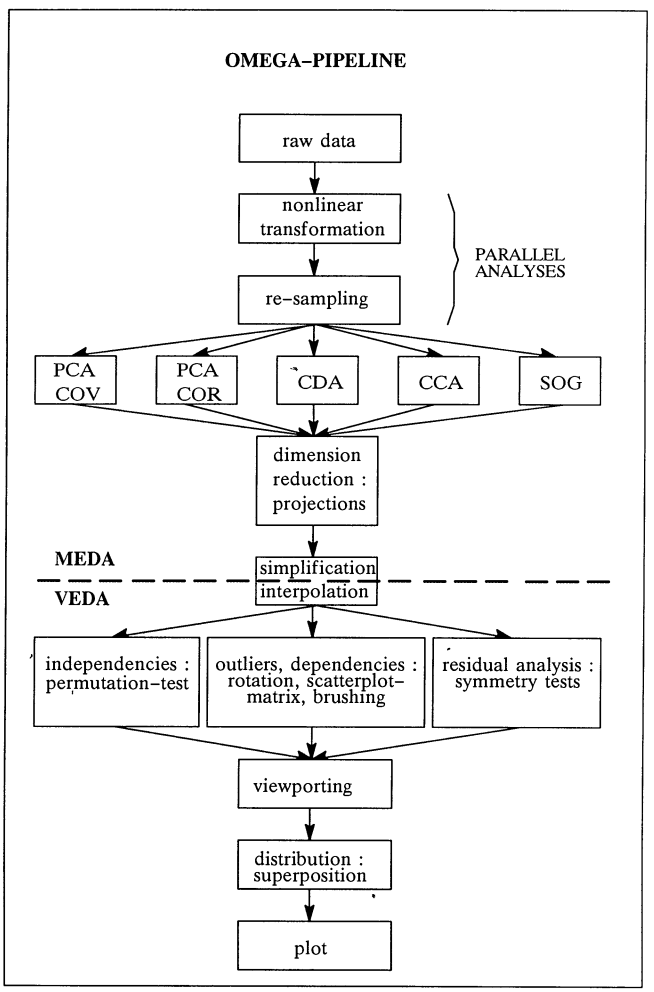


FIG. 1. *OMEGA pipeline.*

For *dimension reduction* (from an initial $h$D space, say) parallel *projections* are used. Thus the projections have to be orthogonal with regard to some reasonable scalar product in the full $h$D space. In some cases, one would like to declare the original raw variables to be orthogonal (see Buja, Asimov, Hurley and McDonald, 1988, page 280). However, this typically implies a scalar product different from the Euclidean one, which may cause troubles with distance judgments. Fortunately, use of the MEDA techniques PCA, CDA and SOG leads to orthogonal directions corresponding to the Euclidean scalar product. But this is not true for CCA, since the corresponding canonical variables of the two groups are, on the contrary, constructed to be as highly correlated as possible. See Sections 3.2 to 3.5 for a more detailed discussion.

All MEDA techniques suffer from at least one drawback: the resulting variables are usually linear combinations of many original variables, the interpretation of which may be difficult, if not impossible, if the original variables do not all measure the same phenomenon. Suitably defined "simplified" representations of these resulting variables would therefore be helpful for interpretation. One possible interpretation aid is the so-called biplot (Gabriel, 1971), in which not only the observation scores are plotted, but also the projections of the original variables are indicated (see, e.g., Young, Kent and Kuhfeld, 1988, page 339). In this way, variables important for the two directions projected upon can be identified. In our OMEGA we use a different technique for *simplification*. We have developed an importance criterion for the original variables in the linear combinations resulting from the different MEDA techniques, in order to eliminate the effect of unimportant variables. The relevance of such a simplification is then judged by a graphical *interpolation* of the projection on the full linear combinations and on the simplifications, respectively. Indeed, this combination of simplification and graphical interpolation very nicely illustrates how graphical techniques may support the understanding of algebraic techniques. Since interpolation moves each observation in a smooth motion from one projection to the other, with an individual but constant speed depending only upon the distance of the observations in the two projections, interpolation quite obviously helps to judge the similarity of the two projections. See Section 3.6 for the importance criteria and Section 4.3 for definition and discussion of interpolation.

Sometimes, graphical methods are seen as suspect because of the potential problem of overexploring data and finding spurious structure. However, in our case this is only rarely likely to be a serious issue, particularly because the insights of exploratory data analysis are mainly more or less qualitative (see Buja, Asimov, Hurley and McDonald, 1988, page 292). In any case, tools for judging stability under strategic and random fluctuations in the data may sharpen one's ideas about properties and relevance of structure. Following Buja, Asimov, Hurley and McDonald (1988, pages 293–295), in an OMEGA strategic fluctuations are used to graphically "test" for some very crude nonparametric hypotheses such as *independence* of variables and some types of distributional *symmetry*. Testing for independence appears to be very useful in an exploratory setting, e.g., for identification of nonpredictors. Distributional symmetry is particularly important in *residual analysis*. Generating random fluctuations in the original sample, e.g., by *resampling* using cross validation or bootstrapping, can allow one to study the stability of the projection resulting from the MEDA techniques. This may help to distinguish relevant from irrelevant structure. For a graphical study of fluctuations, there are at least two different possible approaches: either the different samples are studied in parallel, or the distribution of (the outcome of the analysis of) all samples together is displayed by some suitable form of *superposition*. *Parallel analyses* require a multiwindow facility; superposition does not. But adequate superposition requires careful reasoning about equivalence of projections and identification of the distributions corresponding to different original observations. See Section 3.7 for resampling, Section 4.6.2 for graphical tests, and Section 4.6.3 for resampling.

Another kind of parallel analysis is the study of the linear MEDA techniques after a set of different *nonlinear transformations*. This may support the identification of "optimal" transformations for different purposes (see Section 4.6.1).

A multiwindow technique used for purposes other than parallel analyses is that of *scatterplot matrices*. This consists of looking at more than one 2D projection of the $h$D space simultaneously, in order to discover multidimensional *outliers* and surfaces. Here a linked analysis is particularly useful, marking corresponding observations in the different scatterplots by *brushing* (see Section 4.5 for both scatterplot matrices and brushing). Another means of identifying 3D outliers, clusters and surfaces is 3D *rotation* (see Section 4.3, particularly for the relation between rotation and interpolation).

As we have indicated before, the one block in the OMEGA pipeline which is always used is *viewporting*, or more specifically, the "window-to-viewport transformation," although viewporting will also be understood to include viewport transformations such as rescaling, translating, etc. (see Section 4.4 for a discussion).

This concludes the survey of the pipeline. We will now discuss the building blocks in detail.

## 3. ALGEBRAIC TECHNIQUES

### 3.1 An Overview

The main purpose of this paper is to demonstrate the usefulness of the various techniques for practical purposes. Let us begin therefore with a classification of the linear techniques involved according to the preinformation available to us about the objects or variables involved in the problem. An overview can be found in Table 1.

If no preinformation is available, particularly if one is not willing to ignore the effect of the different scaling of the variables, then PCA-COV is a candidate method for representing the variation in the data by as few as possible linear combinations of the variables. Obviously, if the scaling of the variables is very different, one should expect variables with large variation to be important in determining at least the first principal component. This effect can be avoided by declaring the scaling of variables to be unimportant and using PCA-COR. If there is one variable available which classifies the objects into a small number of classes, and the ability of the other variables to predict this classification is of interest, CDA can be used for class discrimination by as few as possible linear combinations of the variables. If the variables can be naturally attached to more than one group and the predictability of one group by another is of interest, CCA can be used for identification of corresponding linear combinations of the variables within the groups which maximize between-group correlation. If there exists a natural ordering of the variables, and the predictability of a variable by its predecessors is of interest, SOG can be used to generate successive residuals. Obviously, PCA, CDA and SOG provide natural dimension reduction criteria, in that it is their common intention to represent as much information

TABLE 1
*Classification of linear multivariate techniques*

| Analysis | Pre-information | Representation of |
|---|---|---|
| PCA-COV | | data variation |
| PCA-COR | scaling unimportant | data variation |
| CDA | object classification | class discrimination |
| CCA | variables grouping | between-group correlation |
| SOG | variables ordering | successive residuals |

as possible in as few dimensions as possible. For CCA, dimension reduction may not be that obvious, since it is the intention of CCA to provide as many as possible "independent" corresponding combinations with high correlation. However, one natural dimension reduction is to concentrate on just one pair of corresponding linear combinations to demonstrate predictability.

The rest of this section will be organized as follows. In Sections 3.2–3.5 a unified representation of PCA, CDA, CCA and SOG will be given, including dimension reduction criteria. In Section 3.6 simplification procedures of the full linear combinations will be described, and in Section 3.7 equivalent projections will be characterized. The presentation here is heuristic and mathematically informal. Full technical details are provided in Appendices 1.2–1.7.

### 3.2 Principal Components Analysis

The basic idea of *principal components analysis* is to describe the dispersion of an array of $m$ observations in $h$-dimensional space by introducing a new set of coordinates, the principal components, which are (in the Euclidean sense) orthogonal linear combinations of the original coordinates. These principal components are ordered so that the sample variances of the observations with respect to these derived coordinates are maximal regarding the orthonormality restrictions and in decreasing order of magnitude (see, e.g., Gnanadesikan, 1977, page 7).

By considering the first $k$ such principal components which explain 95%, say, of the variation, we achieve a dimension reduction from an $h$- to a $k$-dimensional space. Both the new coordinates and the projections of all the observations on the corresponding subspace are orthogonal (in the Euclidean sense).

Principal components analyses are typically based on either the covariance matrix (*PCA-COV*) or the correlation matrix (*PCA-COR*).

### 3.3 Canonical Discriminant Analysis

Given a set of $m$ observations in $h$D space and a classifier for these observations, the basic idea of *canonical discriminant analysis* (*CDA*) is to introduce a new ordered set of coordinates, the discriminant coordinates, which are linear combinations of the original variables and orthogonal corresponding to the within-class sample covariance matrix, so that class separation is maximal regarding the orthonormality restrictions and decreasing, in the sense that the sample variances of the discriminant coordinates are successively decreasing (see, e.g., Gnanadesikan, 1977, pages 84–85).

Dimension reduction can be performed as for PCA. Note, however, that the discriminant coordinates are orthonormal corresponding to the within-class sample covariance matrix, but the corresponding projections are orthogonal in the Euclidean sense.

### 3.4 Canonical Correlation Analysis

Given two sets of variables and $m$ observations corresponding to the same $m$ objects for all variables, the basic idea of *canonical correlation analysis* (*CCA*) is to introduce two ordered sets of pairwise corresponding new coordinates, the canonical variables, which are linear combinations of the original coordinates of the corresponding set and orthonormal corresponding to within-group sample covariance matrix, respectively, so that correlation between corresponding coordinates is maximal regarding the orthonormality restrictions and decreasing. For an extension of this concept to more than two groups, see, e.g., Gnanadesikan (1977); for the applications in this paper two groups are sufficient.

Interest in high predictability may lead to a dimension reduction criterion of keeping the $k$ first pairs of canonical variables with a correlation of at least 0.9, say. Note that the two sets of canonical variables are orthonormal corresponding to the two sample within-group covariance matrices, respectively, but that the projections with respect to them are orthogonal in the Euclidean sense, not only inside the groups but also with respect to the noncorresponding canonical variables in different groups. Indeed, only projections on corresponding pairs of canonical variables are correlated.

### 3.5 Successive Orthogonalization

Given $m$ observations of a set of $h$ variables for which a natural ordering exists, the basic idea of *successive orthogonalization* (*SOG*) is to introduce new coordinates which are free of all (linear) information "explained" by their predecessors. The dimension reduction works as for PCA. Note that the new coordinates are not orthogonal in the Euclidean sense, but that the corresponding projections are.

### 3.6 Simplification

All linear multivariate techniques suffer from the drawbacks that the resulting new coordinates are typically linear combinations of *all* the original variables, the interpretation of which may be difficult, if not impossible, if the original variables are not all measuring the same phenomenon. What would be helpful to a user is a "slight" change of the new coordinates, which results in a more interpretable transformation. What we shall look for, therefore, for each of the techniques we have discussed is a simplification criterion, indicating the "importance" of the original variable for the linear combination, in order to eliminate the effect of unimportant variables. In practice,

simplification will be achieved by setting as many "nonimportant" loadings as possible to zero, and by rounding. After this has been done for, say, the first new coordinate, the whole orthogonal system has to be rotated so that the first axis coincides with the "simplified" first new coordinate. This is carried out by a procedure analogous to SOG. For PCA the importance criterion is identical with the absolute values of the elements of the new coordinates themselves.

The *PCA simplification* is implemented as follows: Let $0 < p < 100$, and $0 < s$ be given integers. Set to zero those elements of the new coordinate with absolute value $< p\%$ of the maximal absolute value of the elements. Round the unchanged elements to $s$ significant digits, and re-normalize.

Note that the adequacy of the resulting simplification, particularly of the chosen parameters $p$, $s$, needs to be checked by a graphical technique, e.g., by interpolation of the original and "simplified" projections (see Section 4.3). In order to more or less automate the selection of $p$, the following procedure appears to be helpful. Increase $p$ stepwise from 10 to 90 by 10. For each $p$, regress the observations of the principal component $v_i$ upon the observations of the selected variables. Select that $p$ after which the corresponding goodness of fit substantially drops for the first time.

After having selected in this way the important variables, a new analysis is performed on the selected variables. This then takes care of the fact that not all members of a highly correlated subset of variables may have been selected, and that the loadings are highly correlated if the corresponding original variables are highly correlated.

For CDA and CCA the above simplification procedure is also used in principle. Only the importance criteria will be different. At first glance, the problem seems to be easy to solve. Indeed, after certain transformations, for both methods the PCA situation applies in that the first $r$ new coordinates deliver the $r$D best approximation to the data. Unfortunately, those transformations already correspond to linear combinations of the original variables. Thus, simplifications of the transformed variables do not tend to solve the interpretation problem. What is also needed, is a simplification of the transformations (see Appendix 1.6 for details). Last, but not least, note that for SOG simplification makes no sense, since the interpretation of the new coordinates is obvious. They are just the residuals orthogonal to their predecessors. Other simplification procedures are indicated at the end of Section 3.7.

### 3.7 Resampling and Procrustes Transformation

Most of the time, the outcomes of a multivariate analysis are interpreted to be valid not only for the objects for which data are available, but for all objects "of a similar type." In statistical terms, this relates to thinking of the observed objects as a representative sample of a relevant population. The question naturally arises: How sensitive are the outcomes to resampling; i.e., how much would the outcomes change for different sets of objects? Unfortunately, most of the time the actual generation of a reasonable number of samples with reasonable sample size is too time consuming and costly. Thus, sample simulation is the only possibility. For this, the cross-validation idea is used in the following way.

*$p\%$ resampling.* Let $0 < p < 100$ be a given integer. Randomly choose $(100 - p)\%$ of the objects in the observed sample. Apply the multivariate method to this subsample. Project the not included objects into the $k$D space defined by dimension reduction with subsample analysis, where $k$ is defined by dimension reduction with the whole sample (prediction step). Repeat this step $N$ times. Display the "distribution" of the predictions object by object. Note that for an appropriate uniform random number generator $N$ replications will produce approximately $pN/100$ predictions per object for large enough $N$.

Unfortunately, the resampled projections cannot be compared directly. Indeed, experienced PCA users will doubtless be aware that even small changes in the data can lead to a change in orientation of the principal components. But how much do the projections really differ? Having fixed the dimension of the space to be projected on, we shall assume that translation, rotation/reflection, and global scale change do not change a projection qualitatively. The process of constructing an optimal matching of "configurations" under such operations is called *Procrustes analysis*, and the optimal transformation in our case is excellently reviewed by Sibson (1978).

After all resampled projections have been optimally matched with the original projections, the distributions of the projections corresponding to one object are approximated by normal distributions with the simulated sample means and variances. This provides the basis for constructing confidence ellipsoids (see Section 4.6.3).

Note that the same method can be applied to projections from PCA-COR, CDA, the two groups of variables in CCA individually, and SOG, since in every case projections are represented corresponding to an orthogonal basis. In the case of a 2D projection in CCA, the projection corresponding to an individual group will be 1D and normalized, so that matching consists of translation and reflection only.

Let us finish with some comments. First, the size of resampling-ellipsoids should depend upon the chosen projection dimension $k$. The less information is lost by reduction, the smaller the size to be expected. This

may also give a dimension reduction criterion analogous to the Predictive Residual Error Sum of Squares (PRESS) suggested by Wold (1978), and Eastment and Krzanowski (1982). Secondly, instead of studying distributions by simulation, it may be worth looking for analytic expressions derived from data disturbances analogously to Sibson (1979). Thirdly, Procustes analysis may even be used for simplification of the outcome of linear multivariate techniques. Indeed, Krzanowski (1987) proposed using it to identify that subset of $p \geq k$ original variables which generate $k$D projections best-matching to the corresponding $k$D projections of all original variables, where $k$ was fixed beforehand by dimension reduction. Note that this, in a way, relates to "biplot simplification" (see Section 2.2). Fourthly, types of transformations other than those considered in "standard" Procrustes analysis appear to be more appropriate if the variables should form "smooth" curves when plotted against an appropriate other variable, e.g., time. For such cases, Arbuckle and Friendly (1977) proposed simplification by rotation of the new coordinate vectors such that they are as smooth as possible in a least-squares sense, subject to orthogonality. Note that here it is not the projections which are transformed (as with the other methods in this section) but the coordinate vectors (as with the simplification techniques in Section 3.6).

This completes the discussion of the linear multivariate analyses which are currently part of our OMEGA pipeline. In the following sections the VEDA techniques will be discussed.

## 4. GRAPHICAL TECHNIQUES

### 4.1 An Overview

Before discussing the different graphical methods in detail, we will give a rough classification of the techniques, basically distinguishing static and dynamic techniques (see Young, Kent and Kuhfeld, 1988, pages 393–394).

*Static techniques* generate 2D snapshots of the $h$D space.

*Dynamic techniques* generate 2D movies. At best, these are perceived by a user as smooth motions of point clouds. One can trace premarked objects across changing views. The speed of moving points can give information in addition to location for the identification of clusters and outliers (see Section 4.3). Dynamic techniques can be further subclassified as active and passive.

*Active dynamic techniques* require the user to create the movie interactively, e.g., by stepwise controlling direction and speed of rotation (see Section 4.2).

*Passive dynamic techniques* only require the user passively to watch the movie that is created automatically. Again, there are at least two kinds of passive techniques. The first where only the algorithm for generating the movie is stored before execution and one observes the algorithm working on a given data set (as with iterative dimension reduction algorithms, not discussed here, like projection pursuit of Huber, 1985; or the grand tour of Asimov, 1985); the second where the sequence of views itself is precomputed, stored and only replayed at the time of viewing. For both kinds of techniques the term *animation* is used.

The classification used for presentation in this paper is different. Here, we subdivide the methods into *one-window and multiwindow techniques*. Obviously, techniques available for one-window can, at least in principle, also be used simultaneously in a multiwindow environment.

### 4.2 Interactive User Control

Unfortunately, writing about dynamic graphics is a very frustrating business, since writing is static and nongraphical. While some "intellectual" understanding can be gained from the words, equations and static graphics presented in what follows; true insight about dynamic graphics can only be obtained by seeing the movie (see Young, Kent and Kuhfeld, 1988, page 394). This is particularly true for the understanding of the various active dynamic techniques for controlling a movie.

A valuable technique for controlling dynamic graphics is the good old menu. Typical situations for using it are the selection of colors, variables, subsets, observations or transformations. In all these cases, the cursor is moved by a key, or the mouse, or whatever, to a field representing the choice. Then key strokes or mouse clicks will start the action, e.g., coloring a subset of observations by two mouse clicks. Imagine now that you want to rotate the (projected) 3D point cloud tracing that subset across changing views. For such rotations there are at least three types of control in use.

With *stop control* one initiates a rotation of a prefixed angle in a prefixed direction around a prefixed screen axis by one key stroke or mouse click. Then, motion stops immediately. By changing the direction one obtains a rocking motion.

With *animated control* repeated steps are initiated by holding down the key or mouse button. This may be implemented with constant prefixed speed or as an increasing-step-and-reversal control, where rotation starts slowly and then accelerates to some maximum speed. Releasing the key or mouse button stops the motion, and pressing it down again resumes rotation in the other direction.

Finally, with *lasting control* the motion is initiated by one key stroke or mouse click and is continued

until deactivation or the choice of another action. Again, increasing-step-and-reversal control may be implemented, where change of direction may be activated by another key stroke or mouse click. For more details on user control, see, e.g., Cleveland and McGill (1988).

The rest of this section will be organized as follows. In Sections 4.3–4.4, one-window techniques will be discussed, i.e., rotation, interpolation and viewporting. In Sections 4.5–4.6, multiwindow techniques will be discussed, i.e., linked analyses by means of brushing in scatterplot matrices, and parallel analyses corresponding to nonlinear transformations, graphical tests or resampling.

### 4.3 Rotation and Interpolation

The most widely available feature of existing interactive dynamic graphics packages is *3D rotation.* As an alternative, Young, Kent and Kuhfeld (1988) offer *"3D" interpolation.* Let us start with a comparison of rotation and interpolation in the special case of transformation of the $(x, y)$ plot into the $(x, z)$ plot.

Let us assume that we initially see the projection of objects onto the first two coordinates of a 3D space, the $z$-axis being vertical to the screen. Then, rotation around the $x$-axis corresponds to using a linear combination of the $y$- and $z$-axes as the new $y$-axis. The new 2D subspace to be projected on is then defined by the old $x$-axis and the new $y$-axis. Again, with stop control one obtains a rotation by a prespecified angle, and with animated or lasting control one generates a scatterplot smoothly spinning around the $x$-axis. Interpolation and rotation only differ in the way this spinning proceeds.

Let $x, y, z$ be observation vectors corresponding to the 3 spatial coordinates. Then

$$y_R(t) := y\cos(t) + z\sin(t), \quad 0 \le t \le \pi/2,$$

are the observations corresponding to the intermediate vertical axes with *rotation*, and

$$y_I(t) := y(1 - t) + zt, \quad 0 \le t \le 1,$$

are the observations corresponding to the intermediate vertical axes with *interpolation.* The different effect of rotation and interpolation then becomes clear, e.g., by discussing speed of motion and norm preservation, elaborating on ideas of Buja, Asimov, Hurley and McDonald (1988, pages 288–290).

Let us start with *motion speed*, for which

$$(d/dt)(y\cos(t) + z\sin(t)) = z\cos(t) - y\sin(t).$$

With rotation the speed is changing with the rotation angle and depends upon the size of the coordinate orthogonal to the screen provided $y, z$ are orthonormal. In this way, the invisible coordinate is made "visible" by rotation speed. Note that it is this "parallax effect" (see Fisherkeller, Friedman and Tukey, 1988, page 98) that delivers information additional to location during 3D rotation. This cannot be delivered by interpolation, since

$$(d/dt)(y(1 - t) + zt) = z - y,$$

so that with interpolation the speed is uniformly equal to the distance between the values in the $z$- and $y$-direction. Thus, the spatial information delivered by speed is restricted to the situation at the beginning of the motion ($t = 0$).

Let us now consider in what senses the axes of the rotated or interpolated 2D subspaces are normalized. First, let us again imagine that the variables $y, z$ are orthonormal. Then for $y_R(t), y_I(t)$ one obtains

$$y_R(t)'y_R(t) = \cos(t)^2 + \sin(t)^2 = 1,$$

and

$$y_I(t)'y_I(t) = (1 - t)^2 + t^2 = 1 - 2t + 2t^2 \le 1,$$

which is $= 1$ if and only if $t = 0$ or $t = 1$.

Thus, rotation is *norm-preserving*, whereas interpolation leads to an intermediate shrinking of the point cloud in the interpolated coordinate, which is not caused by the data structure, but only by the transformation method. Thus, again interpolation does not seem to be appropriate if the axes interpolated are orthogonal.

However, let us also consider the other extreme $y = z$. In such a case

$$y_R(t)'y_R(t) = 1 + 2\sin(t)\cos(t) \ge 1,$$

and

$$y_I(t)'y_I(t) = (1 - t + t)^2 = 1.$$

Thus, if two very similar variables were rotated into each other, an intermediate bulging in the rotated coordinate would be obtained, whereas interpolation would preserve the norm.

Rotation therefore appears to be appropriate if one is interested in spatial information in an orthogonal coordinate system, e.g., looking for nonlinear structure, clusters or outliers, whereas interpolation appears particularly suitable for illustrating the relation between two very similar coordinates. Note, e.g., that in the extreme case $y = z$ the view is not changing with interpolation for $0 \le t \le 1$, but that this is *not* true for rotation.

Generalizations beyond 3D are also discussed in the literature. Buja, Asimov, Hurley and McDonald (1988, pages 288–289), e.g., propose a special, but genuine, 4D rotation; here, however, we will only use the *"4D" interpolation* technique proposed by Young, Kent and Kuhfeld (1988, pages 410–411).

Let $(x^{(1)}, y^{(1)})$ and $(x^{(2)}, y^{(2)})$ be the observation vectors corresponding to 4 spatial coordinates. Then, by interpolation of the $(x^{(1)}, y^{(1)})$ plot and the $(x^{(2)}, y^{(2)})$ plot one obtains the following observation vectors corresponding to intermediate coordinates:

$$x_I(t_1) := x^{(1)}(1 - t_1) + x^{(2)}t_1,$$

$$y_I(t_2) := y^{(1)}(1 - t_2) + y^{(2)}t_2, \quad 0 \le t_i \le 1, \quad i = 1, 2.$$

Among possible applications of 4D interpolation are the comparison of projections on two coordinates generated by a linear multivariate method and on their simplifications (see Section 3.6), or the comparison of projections on two similar nonlinear transformations, or of two samples, respectively (see Section 4.6).

Note that 4D interpolation corresponds to moving each point in the $(x^{(1)}, y^{(1)})$ plot on an individual straight line to the corresponding point in the $(x^{(2)}, y^{(2)})$ plot (see Young, Kent and Kuhfeld, 1988, page 405). Thus, the dynamic interpolation might be replaced by static disconnected arrows in the superimposed $(x^{(1)}, y^{(1)})$ and $(x^{(2)}, y^{(2)})$ plots. Another static substitute for dynamic interpolation is the parallel view at intermediate phases of the movement (see Figure 4). For further discussion of the merits and drawbacks of rotation and interpolation see Goodall and Thoma (1987).

### 4.4 Viewporting

Regardless of which data transformation is applied to high-dimensional data to portray them in a 2D plane, one has to address the decision as to what part of the data is to be displayed on the screen. That is, one has to decide which rectangle in the 2D data plane should be mapped onto the screen rectangle assigned to the plot. This corresponds to an affine transformation from the data scale to the screen scale and clipping at the boundaries of the screen rectangle (see Buja, Asimov, Hurley and McDonald, 1988, page 282).

First, the center of the data rectangle to be displayed has to be identified. This center may well be different from the data center, which often is the sample mean (see Appendix 1.1). Indeed, for graphics the median or the midrange may be appropriate. The graphical center is then identified with the screen center. Centering has to be followed by choosing the initial scale of the displayed data rectangle. This can be done, e.g., by normalization with the standard deviation, the median absolute deviation or the half-range. Also, one may wish to use a common scale for several variables. Graphical centering and normalization can be adequately supported by a menu.

After having had a look at the plot produced by this window-to-viewport transformation, however, one might change one's mind about appropriate centering and scaling. Then, flexibility in viewport transformations would be helpful. Buja, Asimov, Hurley and McDonald (1988, page 283) proposed the most complete class of transformations:

- simultaneous expanding and shrinking in both directions with the same factor;
- inverse expanding and shrinking: expanding in one direction, shrinking in the other;
- expanding and shrinking in one direction only;
- shifting (translating) in one direction (slicing);
- simultaneous shifting in both directions.

These may be implemented by means of a menu and animated control. Inverse expanding and shrinking would be particularly convenient for scaling time series to obtain quickly a reasonable shape parameter, which is the physical length of the vertical screen axis divided by the length of the horizontal axis (see Cleveland, McGill and McGill, 1986). The shape parameter appears to be very important in the detection of such simple structures as trends, periodicities and exotic points of time series. This is impressively demonstrated by an example of Buja, Asimov, Hurley and McDonald (1988, page 285).

Other important viewport transformations are:

- marking an observation on the screen (clicking the observation);
- subsetting, i.e., actively choosing observations forming a subset (clicking or roping in the observations, joining or intersecting already built subsets);
- identifying observations on the screen corresponding to a variable or a subset (clicking the variable or subset name in a list).

These operations may result in highlighted, colored, named or invisible points. In this way, particularly interesting points may be tracked during transformations. Last, but not least, a facility for recovering from accidents is an absolute necessity.

This completes the discussion on one-window techniques. Note that asking for a common scale for several variables makes particular sense in a multiwindow environment.

### 4.5 Scatterplot Matrices

Often, one is interested in looking at more than one view at a time. One may, e.g., hope to understand the high-dimensional space better by inspecting different 2D or 3D projections simultaneously. The basic implementation idea is that of a generalized draftsman's plot (front, top and side views). Indeed, adding coordinate-reflected views leads to the concept of *scatterplot matrices*, which seems to be first discussed by

Chambers, Cleveland, Kleiner and Tukey (1983) (see Figure 5). A scatterplot matrix of $k$ variables is nothing other than a "matrix" of graphs with a scatterplot of the $i$th variable against the $j$th as the $(i, j)$-entry of the matrix, $i \neq j$, with the diagonal free for additional information. Since this results in $k^2$ plots, such an approach is only practical if $k$ is not too large. The most important feature of the scatterplot matrix is that one can visually scan a row or a column of the matrix and see one variable graphed against all others. The diagonal may be utilized to display names and scale limits of variables (see Becker and Cleveland, 1988), or all the different values of the variables (see Wang and Gugel, 1988), or histograms of the variables (see also Stuetzle, 1988). Obviously, there is no need for the variables represented in a scatterplot matrix to be the original variables. Indeed, principal components may be included or regression residuals (see also Becker and Cleveland, 1988, page 216). Obviously, these concepts may be generalized for 3D scatterplots (see also Stuetzle, 1988).

Now that we have a matrix of views, we may want visually to link corresponding points on the different scatterplots. A very convenient linking method is *brushing* (see Becker and Cleveland, 1988), in which the central tool, the brush, is a rectangle superimposed on the screen. This brush is moved to different positions on the scatterplot matrix by moving a mouse or pressing direction keys. There are four basic brushing operations: highlighting (or coloring), shadow highlighting (or shadow coloring), deleting and labeling. Each operation can be carried out in one of three paint modes: transient, lasting and undo. Moreover, the shape of the brush is changeable (see Becker and Cleveland, 1988, page 203).

With highlighting, the points underneath the brush are highlighted, at the same time highlighting the corresponding observations in the other scatterplots. With shadow highlighting, only highlighted points appear, in all scatterplots except that one where the brush is acting. This reduces overlapping and may help in identifying highlighted observations. With deletion and labeling, the points underneath the brush are deleted or labeled, respectively.

In transient mode, the brushing operation is undone automatically after the brush has left a point. In lasting mode, the result of a brushing operation lasts until it is undone in undo mode. Also, added histograms for individual variables (constructed in the free diagonal) may be linked to a brushed scatterplot, forming each bar of the histogram from the icons corresponding to observations falling into the displayed class, and highlighting (etc.) those icons in the histograms corresponding to highlighted points in the scatterplots (see Stuetzle, 1988).

Becker, Cleveland and Weil (1988) compared brushing a scatterplot matrix with rotation analysis of multivariate data. Their conclusions were that rotation is most revealing in those cases where higher than two-dimensional structure is present, whereas brushing is particularly appropriate for studying dependencies of two of the chosen variables for different values of a third variable.

Scatterplot matrices formed by including both original and simplified coordinates (e.g., principal components; see Section 3.6) can be used to check the adequacy of the proposed simplification. For adequate simplifications, the scatterplot of the original and corresponding simplified coordinates should be essentially diagonal.

Up to now, only "square" scatterplot matrices have been considered. But, obviously, the concept can be easily extended to nonsquare matrices, even with different sets of variables representing rows and columns. Here, the $(i, j)$-entry of the matrix is the scatterplot of the $i$th row variable against the $j$th column variable. In this case, there is no free diagonal for additional information (see ISP, 1988; Wang and Gugel, 1988), but marginal histograms, e.g., could be included, using the margins of the matrix (see S-PLUS, 1988).

### 4.6 Parallel Analyses

Unfortunately, there are applications of multiple views for which a more general structure than "rectangular scatterplot matrices" is needed. Examples include parallel analyses of nonlinear transformations of the original data, and parallel analyses of pseudo-samples generated for judging stability under strategic and random fluctuations. For such analyses, linked scatterplots are needed, the axes of which do not have matrix structure, but are different for each scatterplot, e.g., the first two principal components corresponding to different nonlinear transformations or pseudo-samples of the original variables. Also, most of the time there is no natural ordering of the scatterplots on the screen. On the contrary, in a multiwindow environment one may want to place different scatterplots in individual windows and to be able to shuffle the windows in order that interesting features become clearer, e.g., by hiding an uninteresting part of a scatterplot behind another scatterplot (see also Stuetzle, 1988).

#### 4.6.1 Nonlinear Transformations

Nonlinear transformations can often be used for linearization, variance stabilization or symmetry enhancement. A popular class of such transformations due to Box and Cox (1964) takes the form

$$x_{BC}(a) := (x^a - 1)/a, \quad a \neq 0, \quad x_{BC}(0) := \log(x),$$

where $x$ is the original variable and the parameter "$a$" may be chosen to be different for different original variables.

In particular, the logarithmic transformation ($a = 0$) is one of the standard tools of model builders. An obvious way to implement such transformations as an active dynamic technique is to specify limits and increments for "$a$" and use stop control, activating a new "$a$" by one key, and reverse incrementing by another key. If, instead, animated control is implemented, it may be wise to precompute all possible transformations before plotting to guarantee a smooth motion from one transformation to the next (see Becker, Cleveland and Wilks, 1988, page 41).

Stop control could also be used to generate different views for different transformations of the original variables and subsequent dimension reduction, e.g., by canonical correlation analysis, in order to be able to compare the outcomes of different transformations by means of linked scatterplots (see Figure 14). Animated control can be used to interpolate between the outcomes of nearby transformations (see Section 4.3).

### 4.6.2 Graphical Tests

Buja, Asimov, Hurley and McDonald (1988) suggested the use of strategic random fluctuations to test, graphically, crude nonparametric distribution hypotheses, randomizing the data such that its distribution is preserved if an appropriate null hypothesis is true. Interesting hypotheses are: independence of variables; symmetry of the distribution of the $x$-variable (or the $y$-variable) about the center; symmetry of the joint $x$-$y$-distribution, so-called 45%-symmetry; point symmetry, which should be valid for every 2D normally distributed variable; rotation symmetry, which means that the distribution of $(x, y)$ and $(x \cos(t) + y \sin(t), -x \sin(t) + y \cos(t))$ are the same.

To clarify the idea behind all these tests, consider the realization of two of them. In order to check whether the variable plotted vertically is independent of the horizontal variable, the $y$-values are randomly permuted against the $x$-values. In order to check symmetry about the $x$-center, the $x$-values of 50% randomly chosen points are reflected about the center.

Stop control may be used for generating one new random plot at a time. In this way, different random plots may be placed into different views (see Figure 10), which may be further analyzed simultaneously. This may protect one against being trapped by spurious structure if the hypothesis is true. It should also produce views very different from the original if the hypothesis is not true. On the other hand, animated control can be implemented to flash random plots at the user at maximum speed. This may help to test the null hypothesis and may help the user in acquiring a sense of randomness under the null hypothesis. To understand this, let us imagine the independence test. Animated control produces a smooth density surface, where, e.g., on a monochrome screen, high brightness stands for high density. If the $x$, $y$-variables are really independent, the density surface should qualitatively coincide with the original data. In a static environment, some random plots superimposed in one view may be a substitute. For easy comparison, original and animated views may be displayed in parallel (see Figure 11), but obviously linking makes no sense here.

Testing for independence appears to be very useful, e.g., for the identification of nonpredictors. Distributional symmetry is particularly important in residual analysis.

### 4.6.3 Resampling

Random resampling can be used to study the stability of the projection resulting from dimension reduction by applying resampling together with optimally matching Procrustes transformations, as proposed in Section 3.7. Again, stop control may be used to produce one view per pseudo-sample for later comparison. But note that linking appears to be problematic here, also, since in the different views the represented objects will not be all the same. On the other hand, because of the optimally matching transformations, superposition, e.g., produced by animated control, generates a very informative impression of the (pointwise) distribution of the projections of all the involved objects. As a graphical aid for separating the replicates belonging to different objects, ellipsoids may be superimposed, surrounding, say, 95% of the replicates per object (having assumed normal distributions; see Figure 3).

This completes the discussion of graphical techniques. A discussion of the merits and drawbacks of the environment in which the OMEGA is actually implemented can be found in Appendix 2.

## 5. AN ILLUSTRATIVE APPLICATION

### 5.1 The Problem

The following data, collected as part of a major CIBA-GEIGY investigation into the quality of dyestuffs, will serve to typify the kind of problem for which a routine online multivariate exploratory graphical analysis strategy is required in an industrial applied data analysis context. The data consist of 93 observations of 29 variables. Each 29-dimensional observation corresponds to one produced batch of a particular dyestuff. The 29 variables fall into two

categories: 18 correspond to analytical properties, 11 relate to coloristic properties. The analytical properties correspond to portions of well-defined chemical compounds in the dyestuff (or derived quantities); the coloristic properties are visual judgments or technical measurements of the coloristic impression (strength, hue, brightness) of the produced dyestuff compared with a standard when applied to polyester. Details of the variables are provided in Table 2.

The analytical properties are mainly reported as percentages of the compounds in question relative to all compounds or to all organic compounds. Of the coloristic properties, five variables (19, 22, 23, 26, 27) are visual judgements translated to a numerical scale, and six variables (20, 21, 24, 25, 28, 29) are technical colorimetric measurements generated by an agreed, partly normed algorithm. The main questions of interest to the producer are the following.

*Can the coloristic properties of the dye be predicted from the analytical measurements, and, if so, how?* The motivation behind this question is the desire to control the production process in order to produce dyestuffs having the required coloristic properties. Batches which fail to come up to the standard require costly

TABLE 2
*Variables*

| Short title | No. | Title | Unit |
|---|---|---|---|
| ANALYTICAL PROPERTIES | | | |
| TERCUP | 1 | TERtiary CoUPling-component | %t[a] |
| DNANDI | 2 | DiNitro-ANinline-DIazo-component | %t |
| DNBZDI | 3 | DiNitro-BenZole-DIazo-component | %t |
| SECMC | 4 | SECondary Main Component | %t |
| TERTMC | 5 | TERTiary Main Component | %t |
| PRIMSC | 6 | PRIMary Side Component | %t |
| SECSC | 7 | SECondary Side Component | %t |
| DNSEC | 8 | DiNitro-SECondary side component | %t |
| TERTSC | 9 | TERTiary Side Component | %t |
| SUNKUV | 10 | Sum of UNKnown UV-absorbers | %t |
| SUMUV | 11 | SUM of known UV-absorbers | %t |
| SUMRED | 12 | SUM of RED dyestuffs | %t |
| SUMGRD | 13 | SUM of GReen Dyestuffs | %t |
| SUNKDY | 14 | Sum of UNKnown DYestuffs | %t |
| SUMDYE | 15 | SUM of DYEstuffs | %[b] |
| TOTORG | 16 | TOTal of ORGanic compounds | % |
| SEC/TE | 17 | SECondary/TErtiary main component | |
| LMBDAC | 18 | LaMBDA-Center-of-gravity-wavelength | nm[c] |
| COLORISTIC PROPERTIES | | | |
| STRVI | 19 | STRength, VIsually | %s[d] |
| STRREM | 20 | STRength, REMission | %s |
| STRTRA | 21 | STRength, TRAnsmission | %s |
| HUEVI | 22 | HUE, VIsually, daylight | codeH[e] |
| HUEVIAL | 23 | HUE, VIsually, Artificial Light | codeH |
| HUEREM | 24 | HUE, REMission, daylight | CIELAB[f] |
| HUEREMAL | 25 | HUE, REMission, Artificial Light | CIELAB |
| BRIVI | 26 | BRIghtness, VIsually, daylight | codeB[g] |
| BRIVIAL | 27 | BRIghtness, VIsually, Artificial Light | codeB |
| BRIREM | 28 | BRIghtness, REMission, daylight | CIELAB |
| BRIREMAL | 29 | BRIghtness, REMission, Artificial Light | CIELAB |

[a] %t: % of TOTal of ORGanic compounds (TOTORG).
[b] %: % of all compounds.
[c] nm: nanometer.
[d] %s: % relative to the standard.
[e] codeH: coloristic codes for hue differences from the standard translated into numerical equivalents: + = redder; − = greener.
[f] CIELAB: colorimetric international standard, hue and brightness differences, respectively.
[g] codeB: coloristic codes for brightness differences from the standard translated into numerical equivalents: + = brighter; − = duller.

COMMENTS on the chemical compounds. Coupling- and diazo-components are residues of preproducts. Main- and side-components are expected process outcomes. The sums and the "TOTal of ORGanic compounds" were constructed by the producer after having classified the compounds. SEC/TE and LMBDAC were constructed by the producer in order to predict hue differences. The center-of-gravity-wavelength was constructed from the spectra of all relevant compounds.

COMMENTS on the coloristic properties. Strength, hue and brightness are different characteristics of the coloristic impression of a dyestuff applied to a well-defined medium (here polyester). These characteristics are generated in two ways: by the dyer's visual judgment of differences between dyeings of the produced batch and of a standard, and by evaluating the remission spectra of these dyeings. The International Commission of Illumination (CIE) has recommended definitions for the hue and brightness differences (CIE, 1978). Here, CIELAB-formulas are used. Strength was defined internally, not only for remission, but also for transmission-spectra.

and time-consuming corrective treatment. On the basis of considerable experience with the process, the producer had already suggested two analytical variables, SEC/TE and LMBDAC, as possible predictors of one of the coloristic properties (namely, the hue).

*Can the visual judgements be predicted from the technical colorimetric measurements*? The motivation for this question is the desire to replace human judgements by more reproducible and faster technical colorimetric measurements. However, the human impression of a dyestuff color is, of course, the relevant one, and so a measured colorimetric variable must be able to mimic the human colorist.

The first question is investigated in Sections 5.2–5.5, the second in Section 5.6. In all the following sections, we try to give an honest, nonstylized account of how the OMEGA was actually carried out in practice in the course of this investigation.

### 5.2 Color Strength: Unexpected Nonpredictability

Obviously, Table 1 (see Section 3.1) could provide a means for choosing the linear multivariate analysis to be applied first to the data. Since we are interested in prediction between groups of variables, Canonical Correlation Analysis (CCA) would seem to be the method of choice. But is it really justified to impose variables grouping at the beginning of the analysis? Would we not miss something? Following this feeling, we postponed CCA and started with Principal Component Analysis on COVariances (PCA-COV), just as if we had no preinformation. And we were lucky; PCA-COV detected an inconsistency in the data.

PCA-COV identified two nearly equally large principal components, together explaining 93% of the data variation (see Table 3a).

TABLE 3
*Principal component analysis on covariances*

| Variables | (a) First 2 orthonormally transformed variables | | | (b) 50% resampling (PROCRUSTES) | | | | (c) Simplified variables | | Simplified orthonormalized variables | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | PC1C | | PC2C | | | | | |
| | PC1 | PC2 | PC3 | mean | sdv | mean | sdv | PC1S | PC2S | PC1SO | PC2SO |
| Explained variation | 55.9% | 92.8% | 95.6% | | | | | | | | |
| 1. TERCUP | 0.000 | −0.004 | | 0.000 | 0.002 | −0.004 | 0.002 | 0.0 | 0.0 | 0.000 | 0.000 |
| 2. DNANDI | −0.006 | −0.002 | | −0.006 | 0.004 | −0.002 | 0.004 | 0.0 | 0.0 | 0.000 | 0.000 |
| 3. DNBZDI | 0.002 | −0.007 | | 0.002 | 0.002 | −0.007 | 0.003 | 0.0 | 0.0 | 0.000 | 0.000 |
| 4. SECMC | −0.017 | −0.005 | | −0.018 | 0.009 | −0.004 | 0.015 | 0.0 | 0.0 | 0.000 | 0.000 |
| 5. TERTMC | 0.025 | 0.033 | | 0.025 | 0.012 | 0.031 | 0.016 | 0.0 | 0.0 | 0.000 | 0.000 |
| 6. PRIMSC | −0.002 | −0.003 | | −0.002 | 0.002 | −0.004 | 0.002 | 0.0 | 0.0 | 0.000 | 0.000 |
| 7. SECSC | −0.004 | −0.001 | | −0.005 | 0.002 | −0.001 | 0.002 | 0.0 | 0.0 | 0.000 | 0.000 |
| 8. DNSEC | 0.000 | 0.001 | | 0.000 | 0.001 | 0.000 | 0.001 | 0.0 | 0.0 | 0.000 | 0.000 |
| 9. TERTSC | −0.005 | −0.008 | | −0.004 | 0.005 | −0.008 | 0.007 | 0.0 | 0.0 | 0.000 | 0.000 |
| 10. SUNKUV | 0.010 | −0.001 | | 0.011 | 0.004 | −0.001 | 0.005 | 0.0 | 0.0 | 0.000 | 0.000 |
| 11. SUMUV | 0.001 | −0.016 | | 0.000 | 0.007 | −0.016 | 0.007 | 0.0 | 0.0 | 0.000 | 0.000 |
| 12. SUMRED | −0.023 | −0.008 | | −0.025 | 0.011 | −0.008 | 0.018 | 0.0 | 0.0 | 0.000 | 0.000 |
| 13. SUMGRD | 0.020 | 0.024 | | 0.021 | 0.014 | 0.024 | 0.019 | 0.0 | 0.0 | 0.000 | 0.000 |
| 14. SUNKDY | −0.008 | 0.000 | | −0.007 | 0.004 | 0.002 | 0.006 | 0.0 | 0.0 | 0.000 | 0.000 |
| 15. SUMDYE | −0.683 | −0.706 | | −0.682 | 0.005 | −0.704 | 0.006 | −0.7 | −0.7 | −0.700 | −0.700 |
| 16. TOTORG | −0.100 | −0.150 | | −0.104 | 0.023 | −0.154 | 0.026 | −0.1 | −0.1 | −0.100 | −0.100 |
| 17. SEC/TE | −0.002 | −0.002 | | −0.002 | 0.001 | −0.002 | 0.002 | 0.0 | 0.0 | 0.000 | 0.000 |
| 18. LMBDAC | 0.002 | 0.001 | | 0.002 | 0.001 | 0.001 | 0.002 | 0.0 | 0.0 | 0.000 | 0.000 |
| 19. STRVI | −0.431 | 0.411 | | −0.429 | 0.013 | 0.410 | 0.011 | −0.4 | 0.4 | −0.408 | 0.408 |
| 20. STRREM | −0.432 | 0.418 | | −0.430 | 0.013 | 0.417 | 0.012 | −0.4 | 0.4 | −0.408 | 0.408 |
| 21. STRTRA | −0.386 | 0.366 | | −0.387 | 0.029 | 0.365 | 0.026 | −0.4 | 0.4 | −0.408 | 0.408 |
| 22. HUEVI | −0.001 | 0.000 | | −0.001 | 0.001 | 0.000 | 0.002 | 0.0 | 0.0 | 0.000 | 0.000 |
| 23. HUEVIAL | −0.002 | −0.005 | | −0.002 | 0.002 | −0.005 | 0.004 | 0.0 | 0.0 | 0.000 | 0.000 |
| 24. HUEREM | 0.000 | 0.000 | | 0.000 | 0.000 | 0.000 | 0.001 | 0.0 | 0.0 | 0.000 | 0.000 |
| 25. HUEREMAL | −0.001 | −0.001 | | −0.001 | 0.001 | −0.001 | 0.001 | 0.0 | 0.0 | 0.000 | 0.000 |
| 26. BRIVI | 0.002 | 0.000 | | 0.002 | 0.001 | 0.000 | 0.002 | 0.0 | 0.0 | 0.000 | 0.000 |
| 27. BRIVIAL | 0.003 | 0.004 | | 0.003 | 0.002 | 0.004 | 0.003 | 0.0 | 0.0 | 0.000 | 0.000 |
| 28. BRIREM | 0.001 | 0.001 | | 0.001 | 0.000 | 0.001 | 0.000 | 0.0 | 0.0 | 0.000 | 0.000 |
| 29. BRIREMAL | 0.001 | 0.001 | | 0.001 | 0.001 | 0.001 | 0.001 | 0.0 | 0.0 | 0.000 | 0.000 |

PC1, PC2 are the loadings of a PCA on the covariance matrix of the dataset. The stability of this 2D projection is tested by resampling (including Procrustes transformations) giving mean loadings PC1C, PC2C. Simplified loadings PC1S, PC2S are also shown together with their orthonormalized versions PC1SO, PC2SO.

Since the third component represented less than 3% of variation, it was decided to reduce the 29-dimensional space to the 2D subspace generated by the first two principal components (see Figure 2).

At first glance, the resulting structure appeared to be three parallel lines in the direction of one of the diagonals. In order to check the stability of this structure against resampling, 50% resampling was applied together with optimal Procrustes transformation (see Section 3.7). Fortunately, the structure appeared to be extremely stable (see Table 3b and Figure 3).

Now that we have found some structure which appears to be relevant, easy interpretation would be helpful. Again, we were lucky, since a very rough simplification, equating to zero all loadings with absolute value $< 5\%$ of the maximum and then rounding to one significant digit (see Section 3.6), led to a very simple structure in the loadings (see Table 3c) and to a nearly identical projection. This can easily be demonstrated by 4D interpolation (see Section 4.3), or a scatterplot matrix (see Section 4.5) including both original and simplified components (see Figures 4 and 5).

Moreover, these simplifications are quite easily interpretable. Indeed, the diagonals of the projections can now be interpreted as representing "SUM of DYEstuffs" (SUMDYE) and "TOTal of ORGanic compounds" (TOTORG) (SW–NE-diagonal), and the three color strengths (NW–SE-diagonal) (see Table 3c). Rotation by (−45°) around the $z$-axis provides interpretable screen axes. In addition, the five original variables left influencing the projection can easily be represented in a scatterplot matrix (see Figure 6).

From this representation, we had the strong impression that the two classes of variables representing the above diagonals could be represented by one variable each, except for a cluster of unusually behaving batches ("outliers") with unexpectedly high "SUM of DYEstuffs" in relation to "TOTal of ORGanic compounds." By erasing these batches with brushing (delete operation) and by rescaling, the close connection between the variables within a class is clearly revealed (see Figure 7).

Moreover, now one has the impression that the remaining objects again divide into two subclusters, one SW and one NE in scatterplot (1, 2). Labeling the objects reveals that the clusters mainly correspond to successive batches (1–12, 13–46, 47–93) with the exception of batches 84 and 93 appearing in unexpected places (see Figure 8).

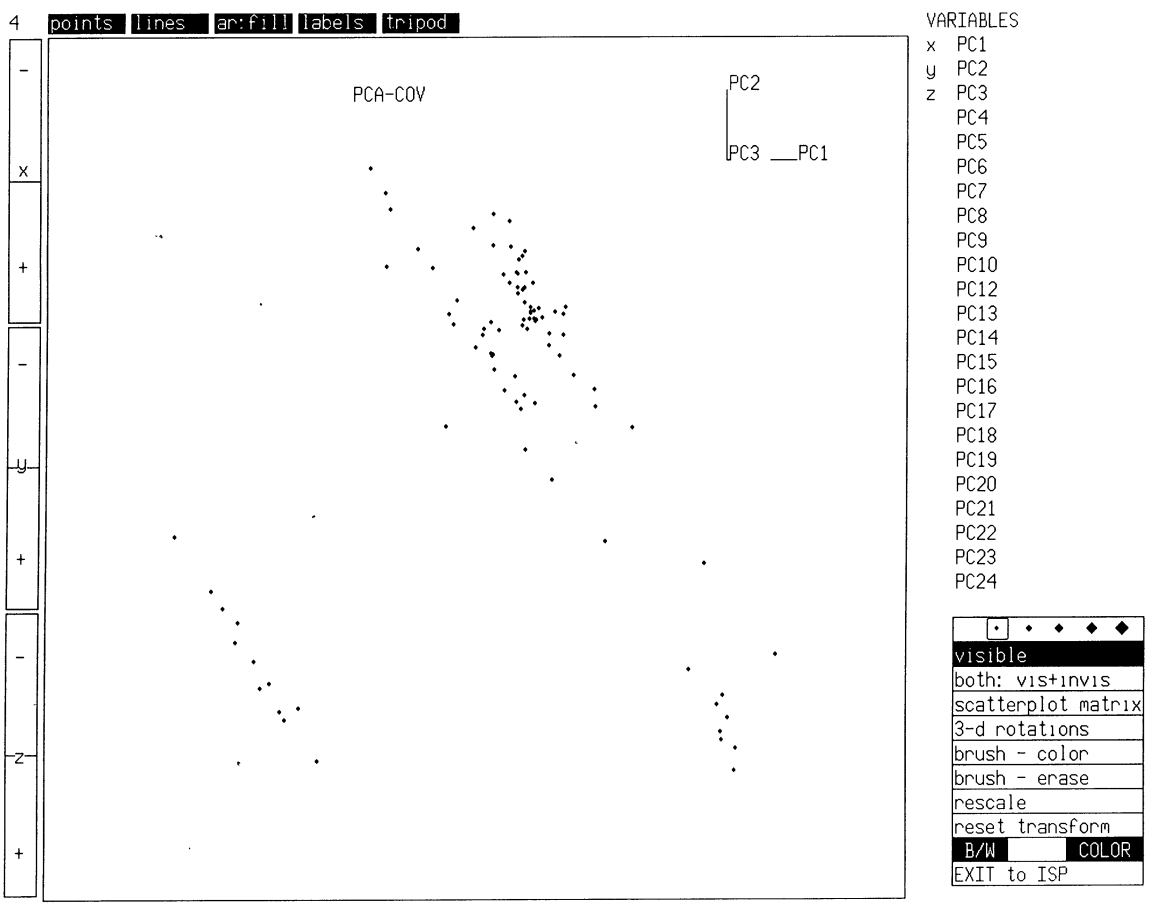


FIG. 2. *First two principal components,* PC1 *and* PC2.

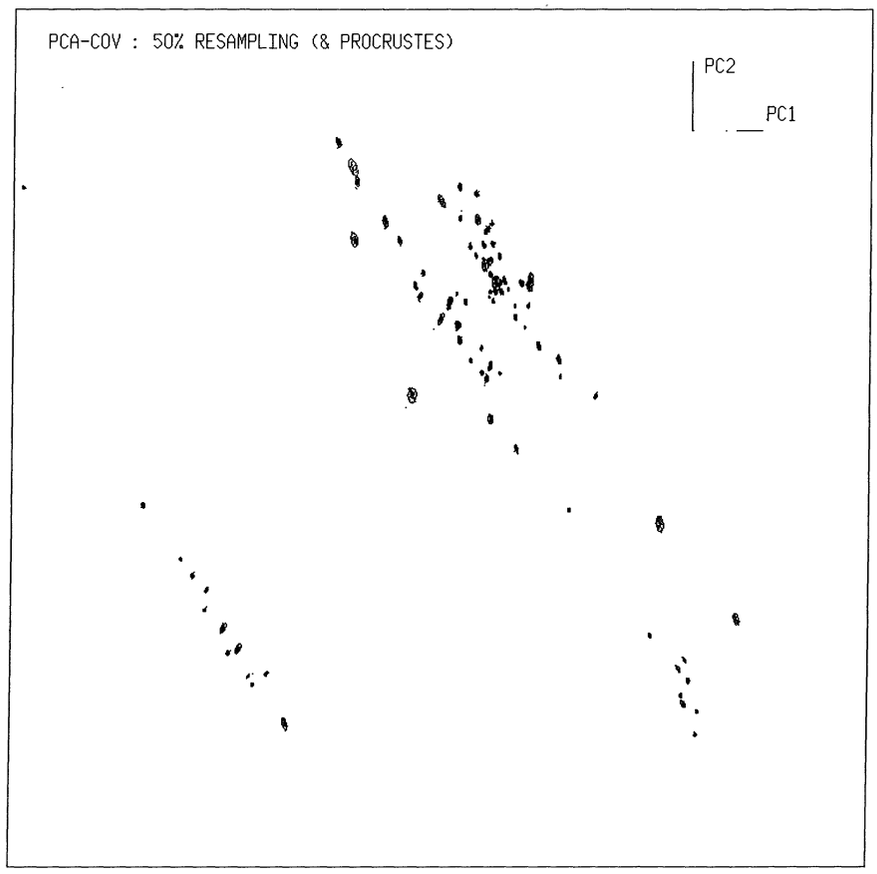


FIG. 3. *Stability test of the first two principal components by resampling.*

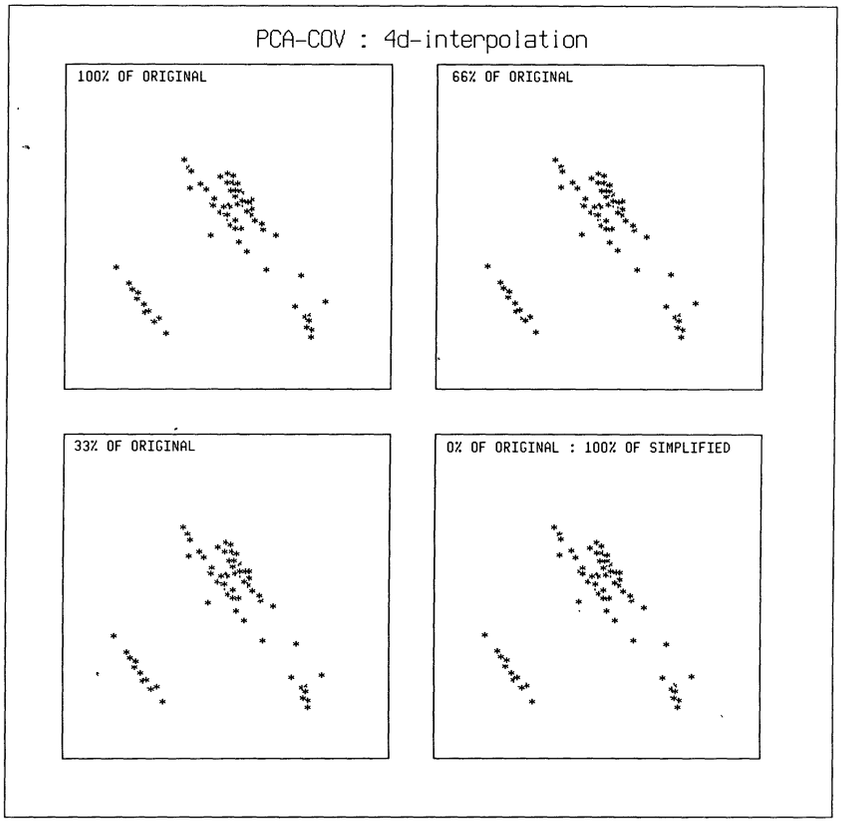


FIG. 4. *Comparing the first two principal components* PC1 *and* PC2 *with their simplifications* PC1SO *and* PC2SO *by* 4D *interpolation.*

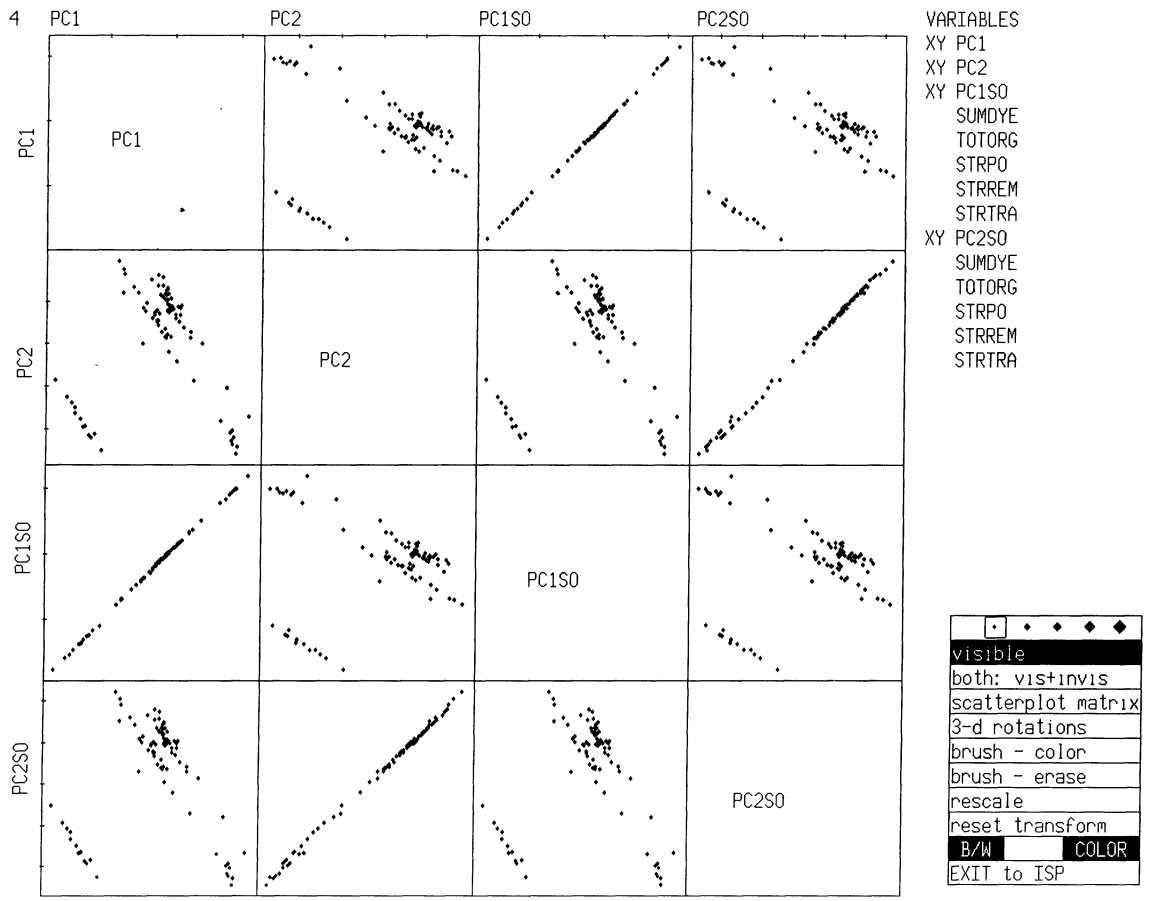


FIG. 5. *Scatterplot matrix of the first two principal components* PC1 *and* PC2 *and their simplifications* PC1SO *and* PC2SO.

In a final scatterplot matrix, the first two original principal components and one representative each of the two "diagonal classes" are included, displaying the three subsets of batches identified above (see Figure 9).

This reconfirms the adequacy of using the representatives (compare scatterplots (1, 2) and (3, 4)), but it also reveals again the structure in the principal component projections: the originally identified three parallel lines represent batches 1–12, 13–46, 47–93, respectively. Because of the time order of these clusters, it was natural to suspect that changes in the production process had taken place after the twelfth and after the forty-sixth batches. This was in fact confirmed by the producer and represented, we felt, a small success for the OMEGA.

But! Should one not be startled by the "orthogonality" of "SUM of DYEstuffs" and "color strength." Shouldn't there be a greater color strength if there is more dyestuff in the batch? Clearly, there was more work on color strength for OMEGA. At first, the producer refused to believe our finding and in order, therefore, to convince ourselves further, a permutation test (see Section 4.6.2) was applied to the two variables. As expected, this delivered qualitatively similar individual plots (see Figure 10) and superposition also indicated only weak dependence (see Figure 11).

Subsequently, after considerable discussion the producer realized that the unpredictability of color strength should have been clear from the beginning. In fact, what was happening was that between the measurement of the analytical properties and the strength, steps in the production process were carried out in such a way that the strength was being influenced randomly. This major finding has led to a revision of measurement procedures within the production process.

The outcomes of the OMEGA thus far can be summarized as follows. 1) The data set is not homogenous. Two production changes have caused a division of the data into three clusters, corresponding to different levels of the variable SUMDYE; this might make one wary of subsequent analyses based on the whole dataset. 2) The strength measurements are inconsistent with the analytical variables as a consequence of inappropriate measurement practice within the production process, and trying to predict this coloristic property makes no sense.

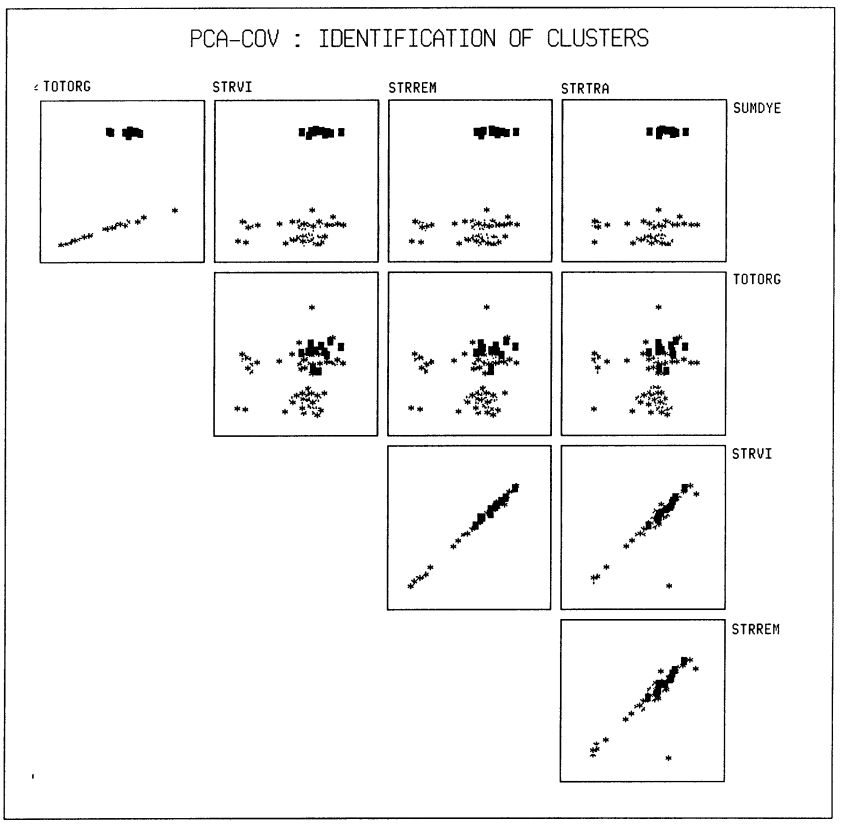


FIG. 6. *Upper half of the scatterplot matrix with the variables involved in the first two principal components.*

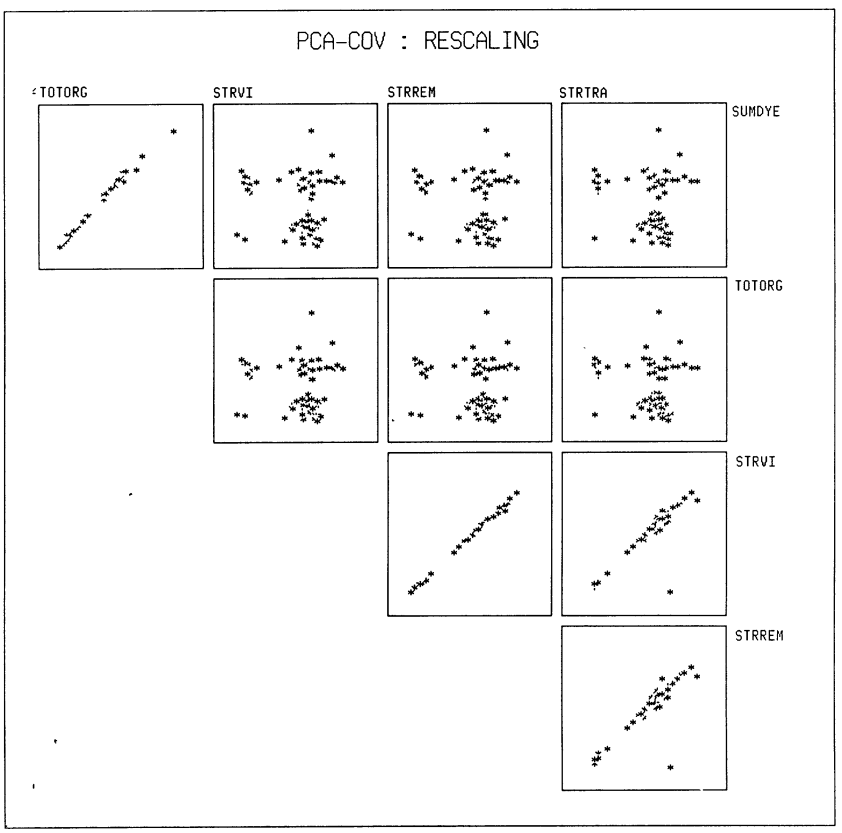


FIG. 7. *Upper half of the scatterplot matrix with the variables involved in the first two principal components, after elimination of outliers and rescaling.*

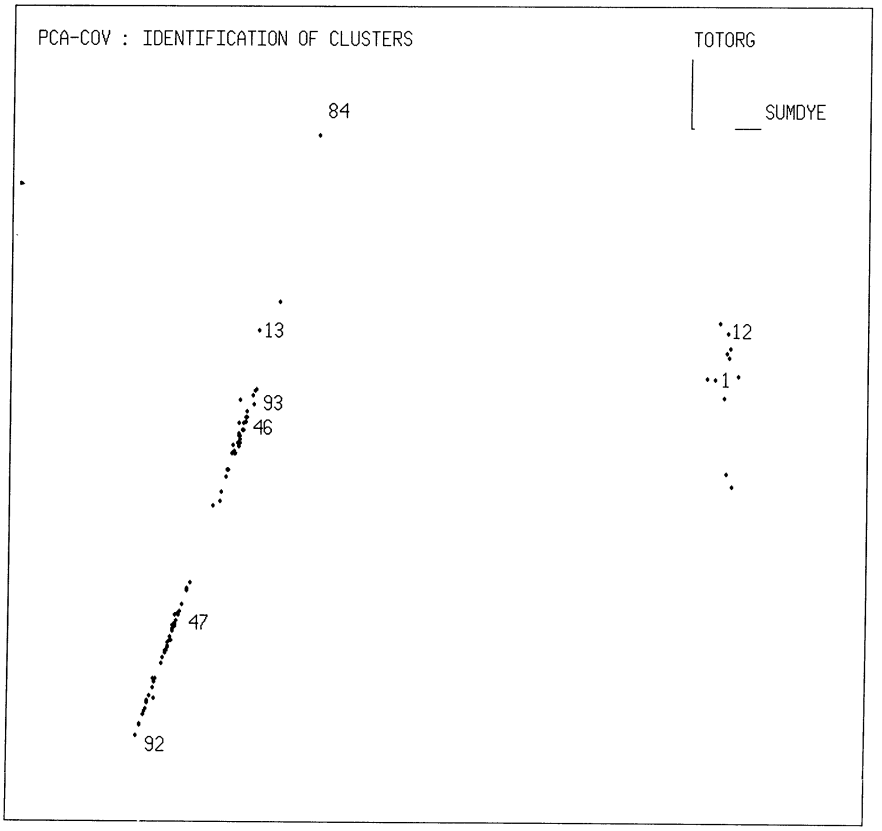


FIG. 8. *Clusters correspond to batches* 1–12, 13–46, 47–93, *but with batches* 84 *and* 93 *misplaced.*

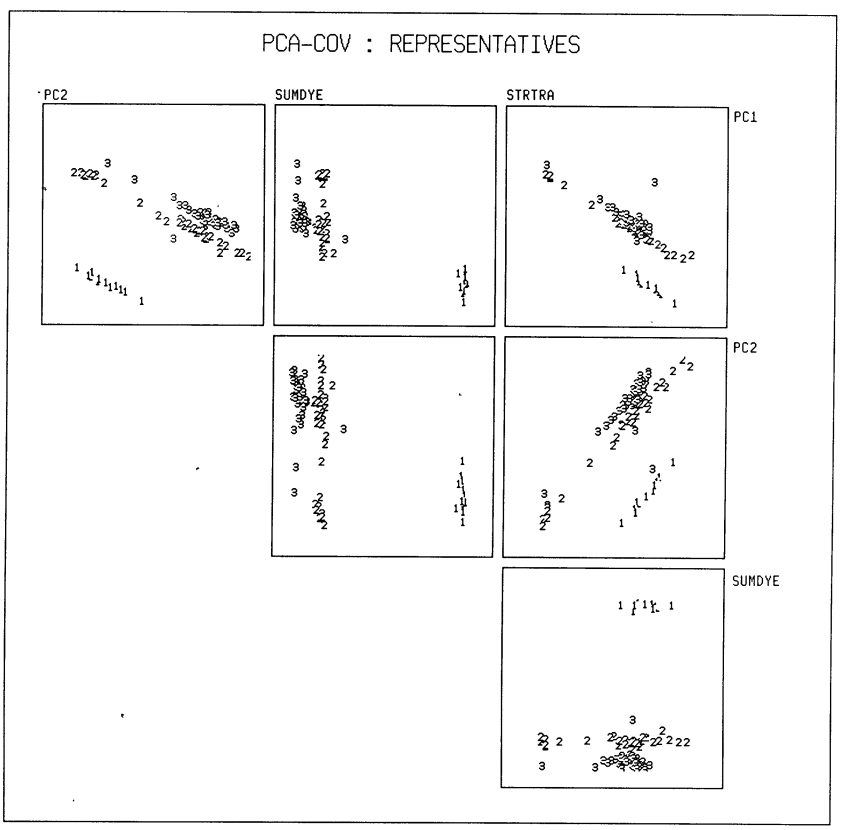


FIG. 9. *Upper half of the scatterplot matrix of the first two principal components and representatives with identified clusters. Clusters are* 1 = *batches* 1–12; 2 = *batches* 13–46; 3 = *batches* 47–93.

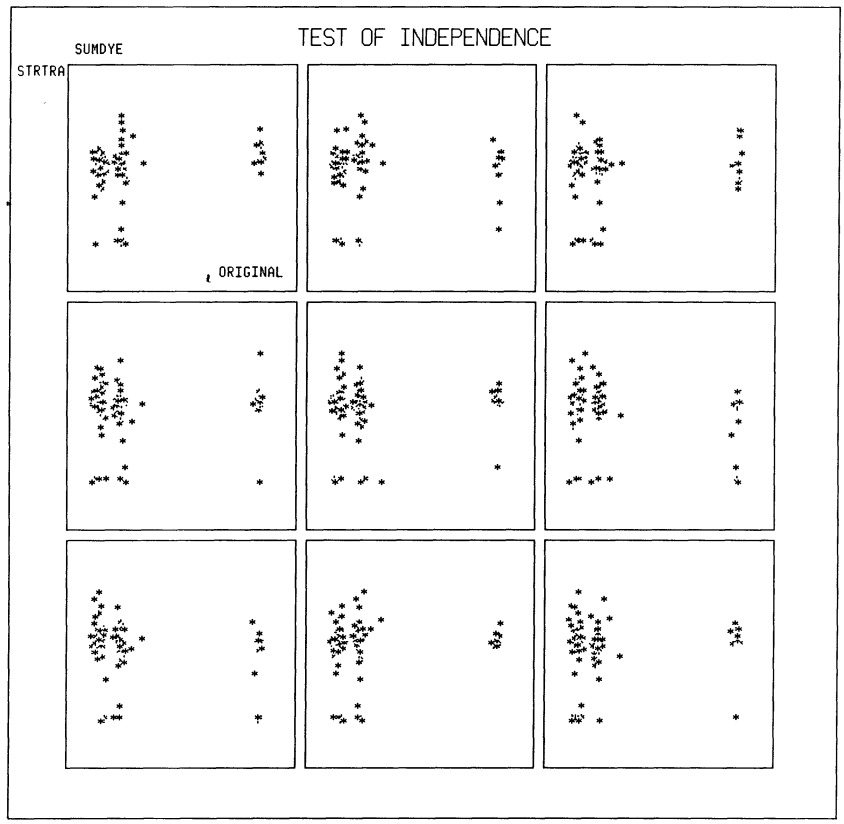


FIG. 10. *Permutation test: parallel views.*

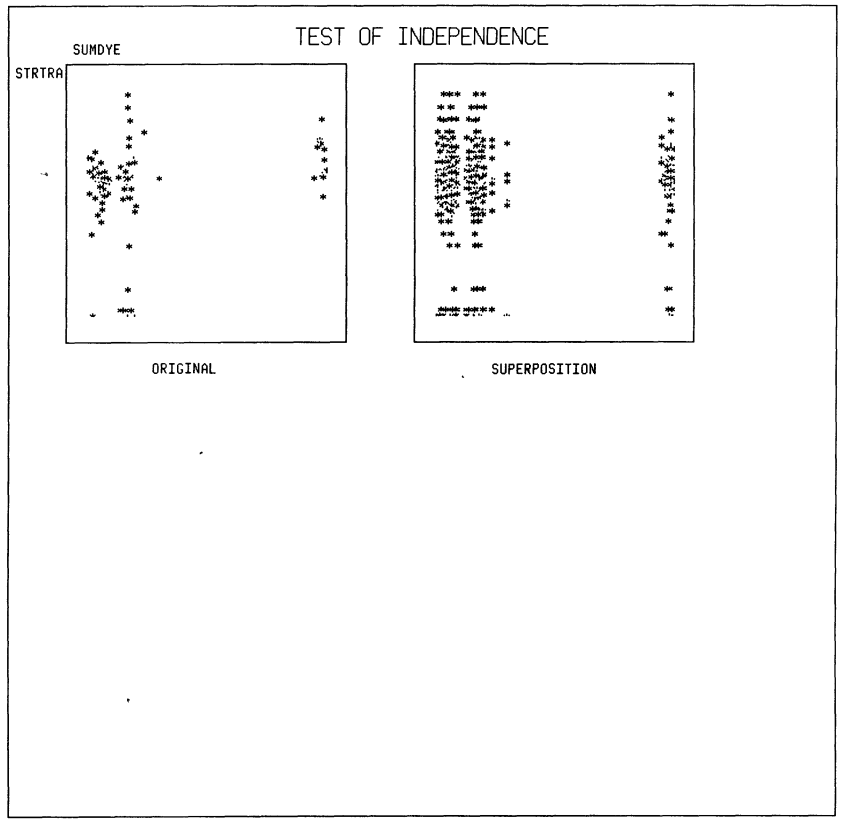


FIG. 11. *Permutation test: superposition.*

Let us conclude our summary so far with a methodological remark. Surely, these insights could equally well have been obtained by inspecting just 2D projections on the original variables by means of scatterplot matrices? We are not convinced. In particular, would we not have overlooked the second finding in the $29 \times 29$ array?

### 5.3 Color Strength: Economics of Measurement

One of the main results of the above analysis was the decision to change the measurement procedure for the color strength. However, we have also seen that the three measured strengths are highly correlated (see Figure 6), and one might have thought that a measurement of one of the three variables would have been sufficient.

Moreover, the three variables measuring color strength can be brought into a natural ordering. Indeed, transmission spectra (STRTRA) are measured directly in slurry, whereas polyester has to be dyed in order to be able to take remission spectra (STRREM). Visual judgments (STRVI) rely on the same dyeings as remission spectra, but additionally suffer from varying environmental and human conditions. We therefore used Successive OrthoGonalization (*SOG*) (see Section 3.5) to investigate whether there is more information in STRVI than in STRREM, and more in STRREM than in STRTRA. Fortunately, the plots of the residuals versus the original variables show no distinct structure, with the exception of the clustering caused by the bimodality of the strength (see Figure 12).

In addition, a graphical test on point symmetry about zero for the two successive residuals was carried out to test for bivariate normality. This hypothesis is supported, since the corresponding strategic random fluctuations appear to be qualitatively the same as in the original plot (see Figure 13). We conclude, therefore, that because the residuals seem to contain no further information, measurement of the strength by transmission (in fact the fastest and most economical method) might be sufficient in the future.

### 5.4 Hue: Successful Prediction

One main goal of the analysis is to find predictors for strength, hue and brightness. Having discussed strength, let us now concentrate upon the hue, as derived from measured remission spectra for daylight (HUEREM) and artificial light (HUEREMAL), and as judged visually by dyers (HUEVI; HUEVIAL).

Since the colorimetric variables HUEREM and HUEREMAL are numerical, and we are interested in

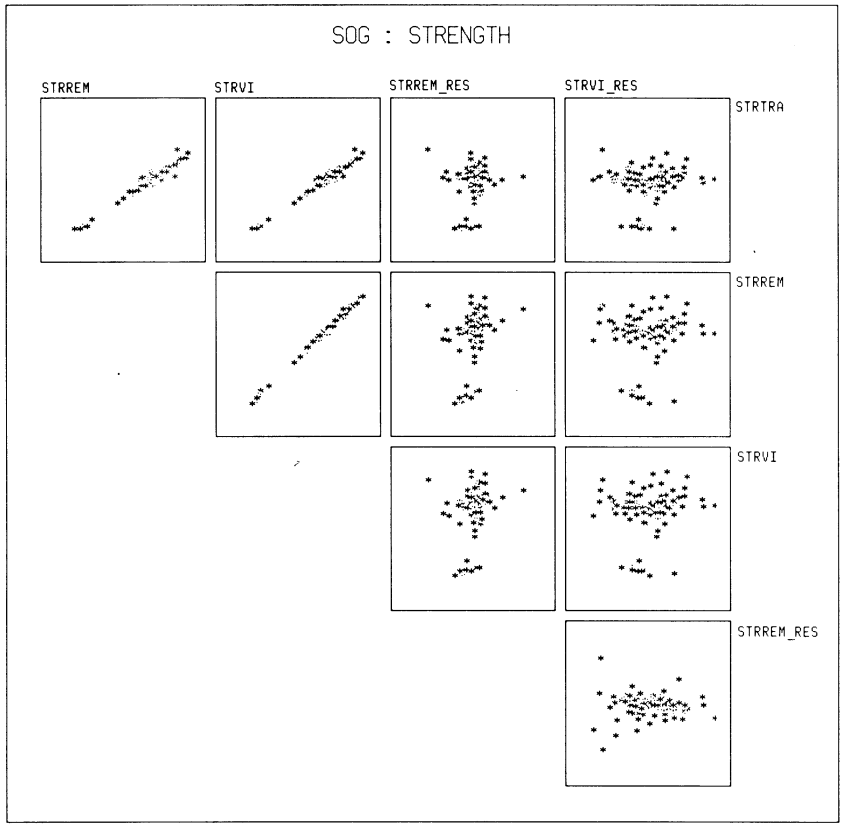


FIG. 12. *Upper half of the scatterplot matrix of the strengths and the residuals of a Successive OrthoGonalization. STRREM_RES: STRREM orthogonalized with respect to STRTRA. STRVI_RES: STRVI orthogonalized with respect to STRTRA, STRREM.*

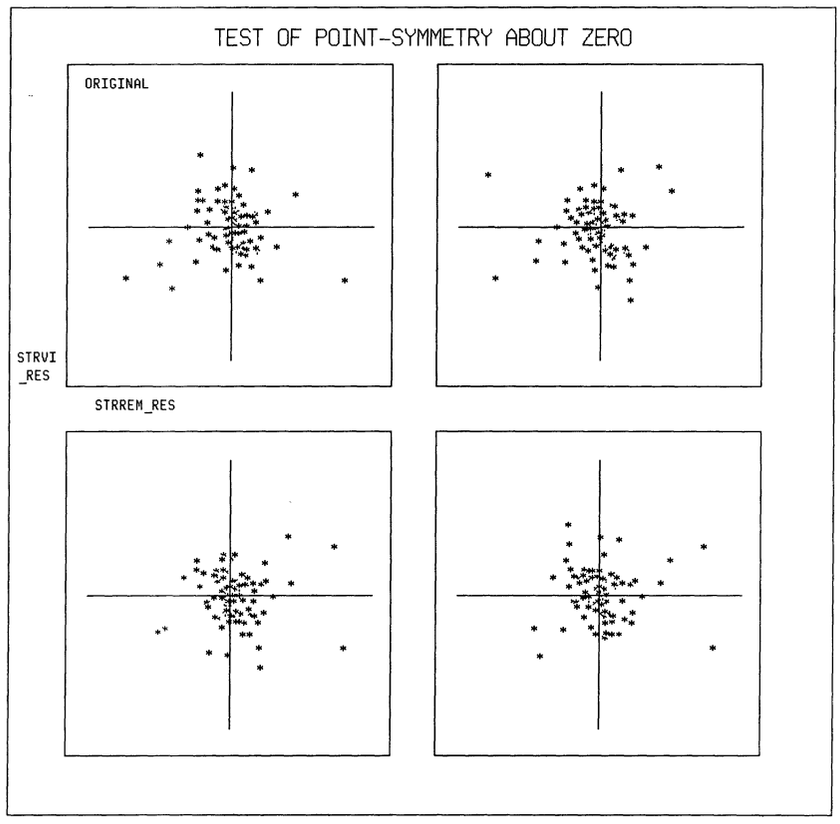


FIG. 13. *Test on point-symmetry: successive residuals of an SOG on strengths.*

prediction of the hues by means of the analytical variables, again CCA is the method of choice. Since the second canonical correlation coefficient is already low (.66), this resulted in a 1D prediction rule: one linear combination of the analytical variables predicting one linear combination of HUEREM and HUEREMAL (see Table 4a).

However, before we attempt any interpretation, let us check whether one could possibly get closer relationships by nonlinear pretransformations (see Section 4.6.1). For this, the two groups of variables were pretransformed individually by each of inversion ($1/x$), log-, and logit-transformation before CCA was applied. In fact, such transformations do not improve linear predictability (see Figure 14).

Thus, interpretation by simplification of the two linear combinations appears to be interesting (see Section 3.6). The corresponding importance criterion leads to the choice of the three variables TERTMC, SUMGRD and SUMRED as predictors for HUEREMAL only (see Tables 4b and 5).

Since we have omitted three analytical properties (SECMC, SEC/TE, LMBDAC) having high correlation with the three chosen ones (TERTMC, SUMGRD, SUMRED) (see Table 6), a new CCA on the selected variables appears to be necessary before rounding to three significant digits (see Section 3.6). The correspondence between projections on the original and simplified canonical components is now seen to be high (see Figure 15).

Moreover, the simplified canonical component of the analytical properties is easily interpretable. Since TERTMC represents the "green" main component, green and red have opposite influence on HUEREMAL, green more so than red. This simply relates to the fact that hue differences are expressed in "greener" and "redder" relative to the standard (see Table 2), and that batches are mainly produced on the greener side (see also Figure 16 illustrating the corresponding visual classification).

In contrast to the colorimetric hue variables, the visual judgments HUEVI and HUEVIAL are ordinal variables. It was therefore decided to choose CDA as the method of analysis. Note that the usefulness of CDA for ordinal variables is sometimes questioned, since the natural class ordering might not be preserved by the corresponding values of the discriminant coordinates. In our applications, however, this would not lead to any interpretational problems. Let us concentrate here on HUEVIAL because of its correspondence to HUEREMAL "modeled" by CCA. In order not to ask too much, somewhat extreme batches are not put into separate classes. This resulted in four classes only (see Figure 16). The first discriminant coordinate

TABLE 4
*Canonical correlation analysis*

| | Canonical linear combinations | | | | |
|---|---|---|---|---|---|
| | (a) CC1 | CC2 | (b) Importance criterion | Initial simplification | CC1S New CCA on selected variables |
| Canonical correlations | | | | | |
| Analytical variables | 0.93 | 0.67 | | | 0.86 |
| 1. TERCUP | −0.026 | | −0.014 | 0.00 | |
| 2. DNANDI | 0.000 | | 0.000 | 0.00 | |
| 3. DNBZDI | −0.007 | | −0.007 | 0.00 | |
| 4. SECMC | −0.148 | | −0.474 | 0.00 | |
| 5. TERTMC | −0.416 | | −1.407 | −0.42 | −0.012 |
| 6. PRIMSC | −0.197 | | −0.109 | 0.00 | |
| 7. SECSC | −0.097 | | −0.036 | 0.00 | |
| 8. DNSEC | −0.179 | | −0.053 | 0.00 | |
| 9. TERTSC | −0.414 | | −0.398 | 0.00 | |
| 10. SUNKUV | 0.015 | | 0.021 | 0.00 | |
| 11. SUMUV | 0.035 | | 0.052 | 0.00 | |
| 12. SUMRED | 0.159 | | 0.605 | 0.16 | 0.012 |
| 13. SUMGRD | 0.460 | | 1.759 | 0.46 | −0.007 |
| 14. SUNKDY | 0.032 | | 0.044 | 0.00 | |
| 15. SUMDYE | −0.001 | | −0.014 | 0.00 | |
| 16. TOTORG | 0.001 | | 0.005 | 0.00 | |
| 17. SEC/TE | −0.018 | | −0.006 | 0.00 | |
| 18. LMBDAC | −0.075 | | −0.030 | 0.00 | |
| Hue measurements | | | | | |
| 24. HUEREM | 0.419 | | 0.053 | 0.00 | 0.525 |
| 25. HUEREMAL | −0.751 | | −0.149 | −0.75 | |

The loadings CC1 of the first canonical components of the CCA on the analytical variables and the measured hues are given together with their importance criterion for each variable and the resulting (initial) simplification. A new CCA was performed on the selected variables giving the loadings CC1S.

then represents nearly 95% of the differences between the within-class means, the second nearly 5% (see Table 7a).

Moreover, the first coordinate successfully separates greener, and "near-standard" batches from redder ones, and keeps the natural class ordering, whereas the second tends to separate "slightly greener" and "near-standard" batches from the rest (see Figures 17 and 18). However, at the level of individual batches it is obvious that only the redder batches are clearly separated from the others (see Figure 17). Moreover, the separability of the class means is supported by resampling, after the elimination of three observations in the original dataset (see Figure 19c). In fact, a first trial using the whole dataset revealed bad separability of classes 1, 2, 3 and a clearly nonnormal distribution of the individual class means (see, e.g., class 4 in Figure 19a) caused by one outlying observation. A second trial after the elimination of this outlier improved both separability and normality (see Figure 19b). After the elimination of two more influential observations, the third and final resampling was carried out.

Using the resulting reduced dataset, a new CDA and the corresponding simplification was carried out. The simplification of the first discriminant coordinate resulted in the selection of the 4 variables SECMC, TERTMC, SUMRED and SEC/TE (see Tables 7b, c and 8). The interpretation of the simplified coordinate is similar to the interpretation of the simplified first canonical coordinate of the analytical variables of the corresponding CCA.

As for CCA, a new CDA on the selected variables was necessary before rounding to three significant digits, because of the high correlations of the selected variables with some of those not selected (see Tables 6 and 7c).

At this stage, the producer's proposals regarding analytical predictor variables for hue, namely SEC/TE and LMBDAC, reappeared in the discussion. How much worse a prediction than the first, say, canonical variable for HUEREMAL, or the first discriminant coordinate for HUEVIAL, would they deliver? For HUEREMAL, the correlation with the simplified canonical variable was negligibly higher than the correlations with SEC/TE and LMBDAC (0.86 versus 0.85 and 0.84, see Tables 4b and 6). For HUEVIAL, a scatterplot matrix representing the two proposals, the first discriminant coordinate and its simplification, is used for checking, where the different HUEVIAL classes are indicated by their class

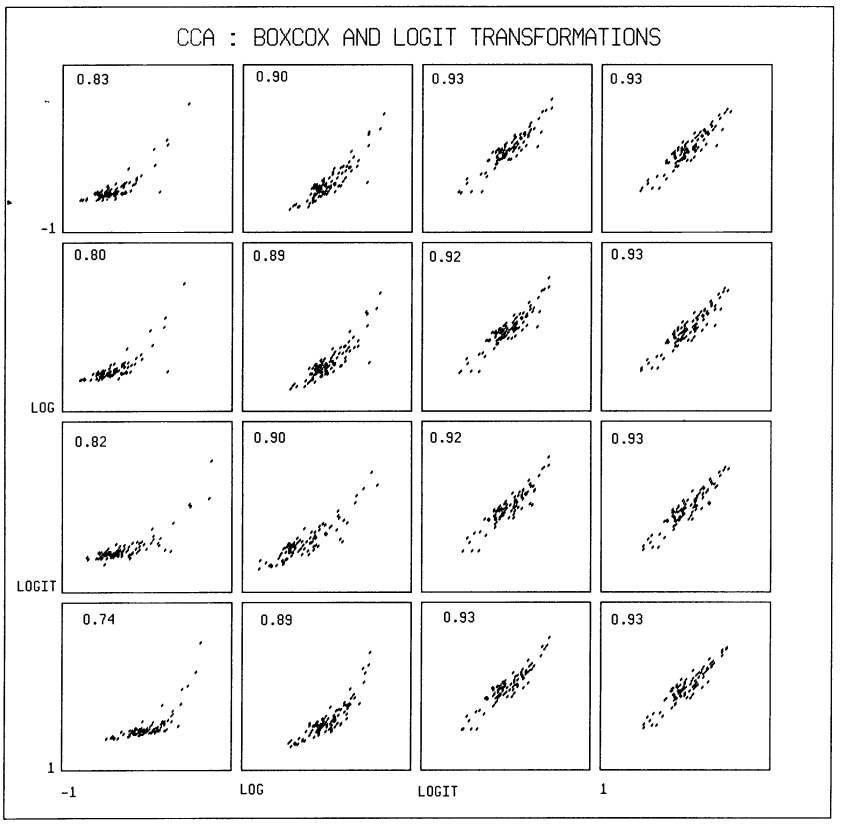


FIG. 14. *CCA between transformed analytical variables and transformed hue measurements: numerical and graphical correlation of first canonical coordinates.*

TABLE 5
*Canonical correlation analysis: Simplification: Selection of variables*

| Selection of analytical variables | | | | Selection of hue variables | | | |
|---|---|---|---|---|---|---|---|
| $p$ | $R$-square | # variables | | $p$ | $R$-square | # variables | |
| 00 | 1.0000 | 18 | | 00 | 1.0000 | 2 | |
| 10 | 0.9216 | 5 | | 10 | 1.0000 | 2 | |
| 20 | 0.9216 | 5 | | 20 | 1.0000 | 2 | |
| 30 | 0.9050 | 3 | ⇐ selected simplification | 30 | 1.0000 | 2 | |
| 40 | 0.8467 | 2 | | 40 | 0.9492 | 1 | ⇐ selected simplification |
| 50 | 0.8467 | 2 | | 50 | 0.9492 | 1 | |
| 60 | 0.8467 | 2 | | 60 | 0.9492 | 1 | |
| 70 | 0.8467 | 2 | | 70 | 0.9492 | 1 | |
| 80 | 0.8452 | 1 | | 80 | 0.9492 | 1 | |
| 90 | 0.8452 | 1 | | 90 | 0.9492 | 1 | |

Simplification leads to the choice of the 3 analytical variables TERTMC, SUMRED and SUMGRD as predictors for the one hue variable HUEREMAL. (See Section 3.6 for a discussion of $p$ and goodness-of-fit. The measure used here, R-square, is the usual regression goodness-of-fit statistic.)

numbers (see Figure 20). From this, it seems that the first discriminant coordinate and its simplification do not separate the classes much better, especially as SEC/TE is included in both coordinates (see Table 7c).

The basic outcome of the hue analysis by OMEGA can therefore be summarized as follows. Essentially, the producer's predictors SEC/TE, LMBDAC could not be improved on very much. On the other hand, the alternatives found using OMEGA appear to be easily interpretable, both for hue measurements and for visual judgments.

### 5.5 Brightness: Unsuccessful Prediction

In contrast to the hue, the producer had no proposal for predictors of the brightness. Thus, for the brightnesses only CCA and CDA were carried out, essentially in the same way as for the hues. Unfortunately,

TABLE 6

*Canonical correlation analysis: Simplification: Correlations (analytical variables and hue measurements; highly correlated variables only)*

| | SECMC | TERTMC | SUMRED | SUMGRD | SEC/TE | LMBDAC | HUEREM | HUEREMAL |
|---|---|---|---|---|---|---|---|---|
| SECMC | 1.00 | −0.69 | 0.99 | −0.72 | 0.88 | −0.83 | 0.63 | 0.76 |
| TERTMC | −0.69 | 1.00 | −0.72 | 0.97 | −0.93 | 0.87 | −0.63 | −0.81 |
| SUMRED | 0.99 | −0.72 | 1.00 | −0.75 | 0.89 | −0.88 | 0.65 | 0.78 |
| SUMGRD | −0.72 | 0.97 | −0.75 | 1.00 | −0.92 | 0.93 | −0.62 | −0.82 |
| SEC/TE | 0.88 | −0.93 | 0.89 | −0.92 | 1.00 | −0.92 | 0.66 | 0.84 |
| LMBDAC | −0.83 | 0.87 | −0.88 | 0.93 | −0.92 | 1.00 | −0.66 | −0.85 |
| HUEREM | 0.63 | −0.63 | 0.65 | −0.62 | 0.66 | −0.66 | 1.00 | 0.90 |
| HUEREMAL | 0.76 | −0.81 | 0.78 | −0.82 | 0.84 | −0.85 | 0.90 | 1.00 |

Since the analytical variables TERTMC, SUMRED and SUMGRD chosen by the simplification procedure (see Table 5) are highly correlated with the analytical variables SECMC, SEC/TE, and LMBDAC a new CCA on the chosen variables appeared to be necessary (see Table 4).

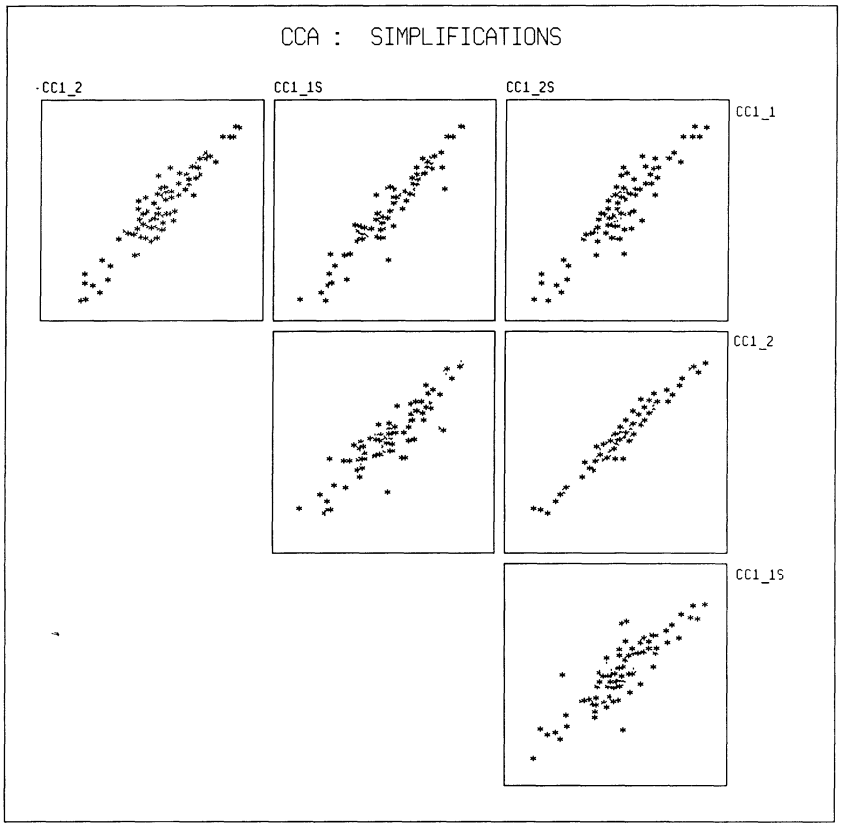


FIG. 15. *Upper half of the scatterplot matrix of the first canonical coordinates of the analytical variables* (CC1_1) *and the hue measurements* (CC1_2) *and their simplifications* CC1_1S, CC1_2S.

CCA for the colorimetric brightness variables BRIREM and BRIREMAL did not deliver useful predictors (maximum canonical correlation = 0.70), nor did CDA for the visual judgments BRIVI and BRIVIAL. We present the outcomes of the CDA for BRIVI. The data included three kinds of visual judgments: near-standard, slightly brighter, and brighter. CDA was not able to distinguish any two (see Figure 21).

This (negative!) finding initiated a closer look by the producer into the definition, measurement and reporting of the analytical variables. In fact, it turns out that not all of the original analytical measurements are being routinely reported, certain components being aggregated before being reported. This loss of information might be one reason for the bad predictability of the brightness and this possibility is now being investigated.

### 5.6 Judgments versus Measurements

The second main question to be answered by the analysis concerned the possibility of replacing visual

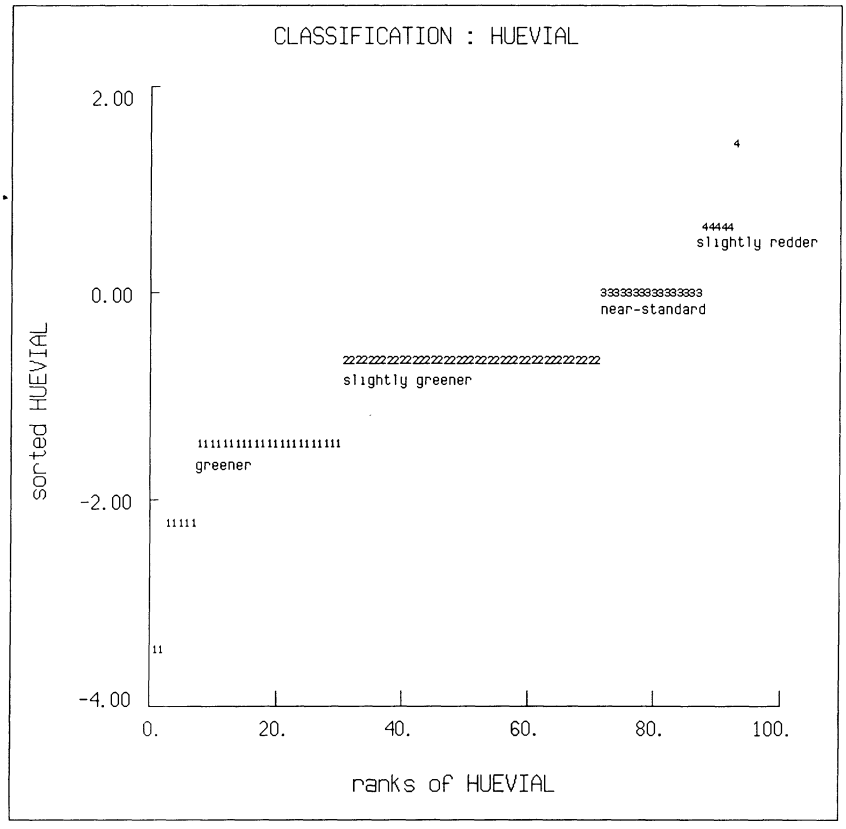


FIG. 16. *Classification of HUEVIAL.*

TABLE 7
*Canonical discriminant analysis: HUEVIAL prediction by analytical variables*

| Variables | Discriminant coordinates | | | | (c) Importance criterion 1.dim | Initial simplification | CD1S new CDA on selected vars |
|---|---|---|---|---|---|---|---|
| | (a) CD1 | CD2 | (b) CD1o | CD2o | | | |
| Explained variation | 94.5% | 99.8% | 92.9% | 99.7% | | | |
| 1. TERCUP | 0.011 | −0.302 | 0.272 | −0.457 | 0.096 | 0.00 | |
| 2. DNANDI | −1.142 | −0.409 | −1.122 | −0.409 | −0.803 | 0.00 | |
| 3. DNBZDI | −0.013 | −0.032 | −0.002 | −0.030 | −0.002 | 0.00 | |
| 4. SECMC | −26.262 | −4.718 | −13.822 | 24.837 | −28.369 | −13.82 | −0.850 |
| 5. TERTMC | 0.305 | −15.035 | 3.308 | −13.283 | 6.981 | 3.31 | 0.802 |
| 6. PRIMSC | −25.052 | −5.346 | −12.738 | 24.034 | −4.815 | 0.00 | |
| 7. SECSC | −25.476 | −4.210 | −12.921 | 25.520 | −4.252 | 0.00 | |
| 8. DNSEC | −26.207 | −5.546 | −14.180 | 23.361 | −3.876 | 0.00 | |
| 9. TERTSC | −0.713 | −15.555 | 2.232 | −13.805 | 1.985 | 0.00 | |
| 10. SUNKUV | −4.663 | −2.899 | 0.470 | 1.712 | 0.552 | 0.00 | |
| 11. SUMUV | −4.836 | −2.678 | 0.214 | 1.952 | 0.270 | 0.00 | |
| 12. SUMRED | 20.295 | 0.896 | 12.817 | −24.129 | 29.743 | 12.82 | −0.640 |
| 13. SUMGRD | −3.411 | 13.433 | −1.087 | 16.575 | −2.575 | 0.00 | |
| 14. SUNKDY | −4.993 | −2.937 | 0.064 | 1.735 | 0.090 | 0.00 | |
| 15. SUMDYE | 0.009 | 0.010 | 0.012 | 0.010 | 0.338 | 0.00 | |
| 16. TOTORG | −0.072 | −0.116 | −0.091 | −0.115 | −0.570 | 0.00 | |
| 17. SEC/TE | 28.298 | 16.126 | 30.630 | 17.756 | 5.399 | 30.63 | 22.124 |
| 18. LMBDAC | −6.685 | −5.313 | −8.165 | −7.319 | −1.967 | 0.00 | |

CD1 are the loadings of the first discriminant coordinate of a CDA with classifier HUEVIAL and all analytical variables as predictors based on the whole dataset; CD1o are the corresponding loadings based on the reduced dataset without 3 outliers. By means of the importance criterion, an initial simplification of CD1o was obtained. A new CDA was performed on the selected variables giving the loadings CD1S.

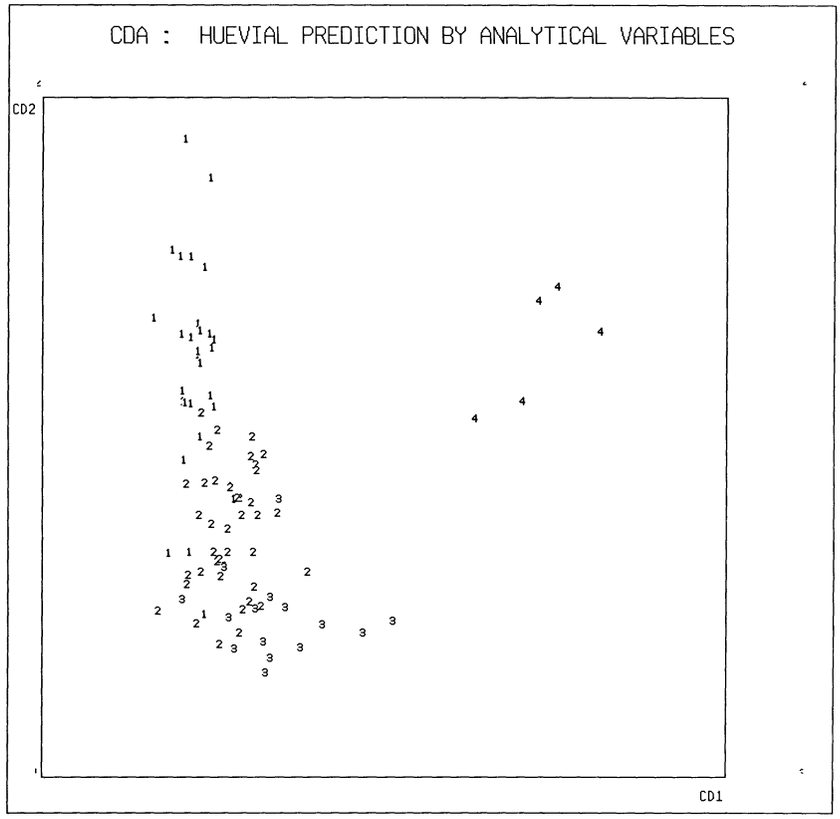


FIG. 17. *First two discriminant coordinates* CD1, CD2 *of a* CDA *with HUEVIAL as classifier and analytical variables as predictors.* 1: *greener*; 2: *slightly greener*; 3: *near-standard*; 4: *slightly redder.*

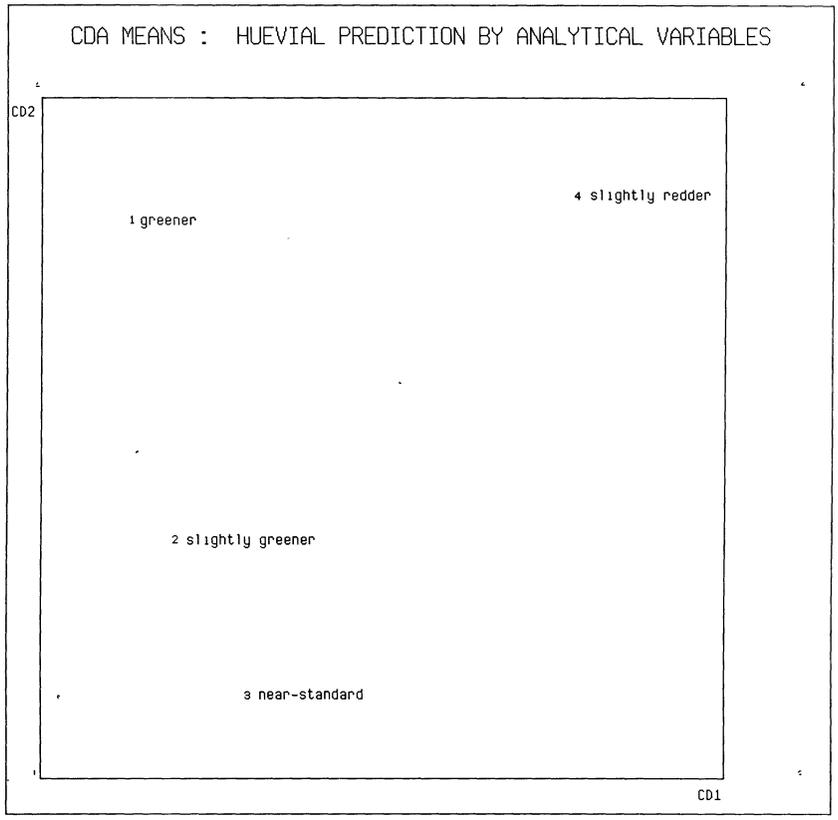


FIG. 18. *Class means of HUEVIAL in the first two discriminant coordinates* CD1, CD2.

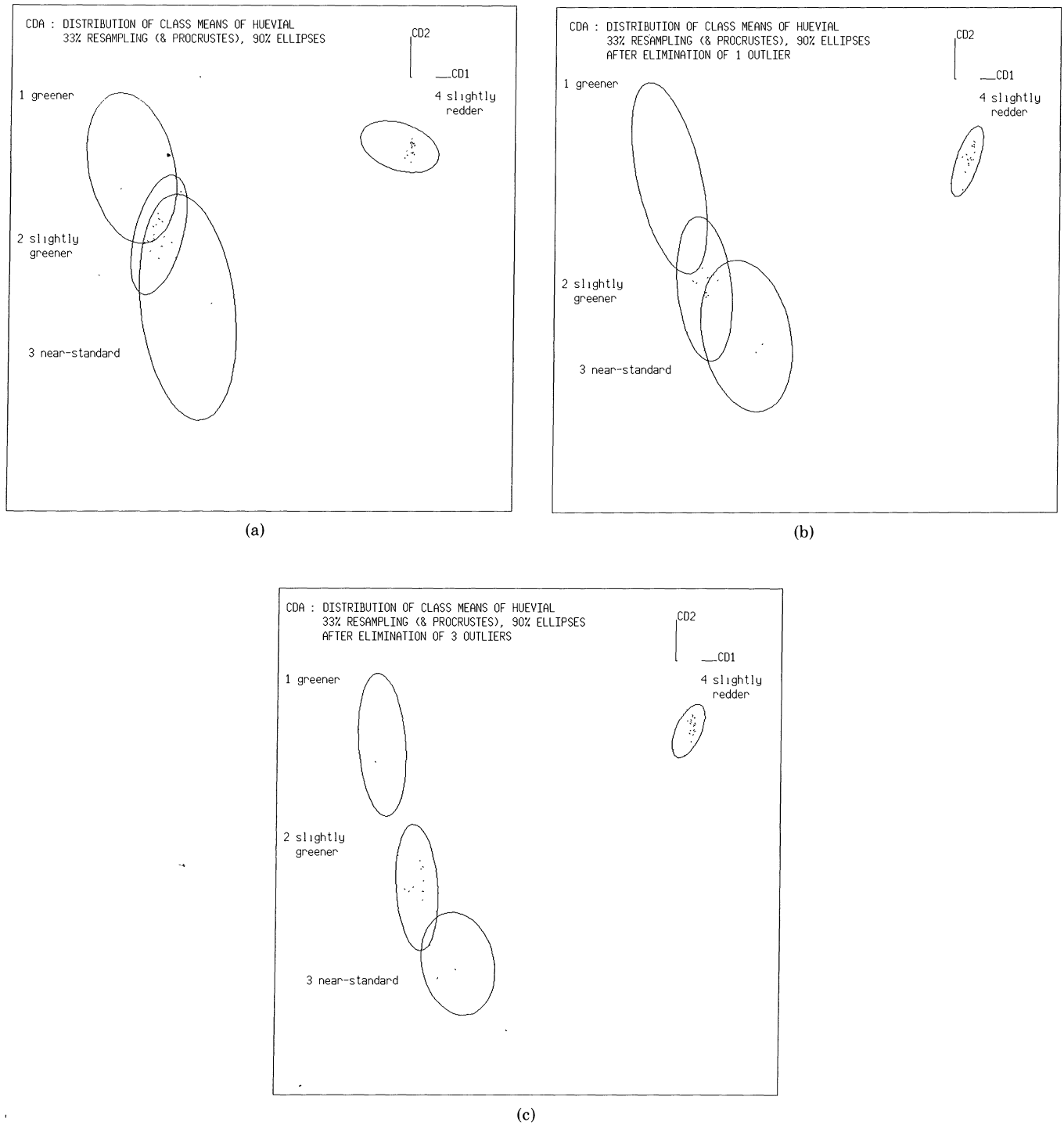


FIG. 19. *Stability test of the first two discriminant coordinates* CD1, CD2 *by resampling and optimal Procrustes transformation.* (a) *based on the whole dataset*; (b) *based on the dataset, reduced by* 1 *observation*; (c) *based on the dataset, reduced by* 3 *observations.*

judgments by technical measurements. Because the human impression is accepted as constituting the final decision, we must try to predict visual judgments by colorimetric measurements. For each of the four visual judgments of hue and brightness, one CDA was carried out using all four colorimetric variables as predictors. The ideal behind this is that on the one hand the two sorts of light, daylight and artificial light, under which the dyestuff is scrutinized, should provide related judgments and that on the other hand hue and brightness are not independent variables as defined by the colorimetric algorithm. With these CDAs the best class

separation is delivered for HUEVIAL, where three or four classes can be distinguished with acceptable misclassification rates (see Figure 22b).

For the other judgments only two classes appear to be separable (see Figures 22a, c, d): greener from near-standard and redder for HUEVI; near-standard and slightly brighter from brighter for BRIVI and BRIVIAL. Indeed, artificial light seems to be preferable to daylight for prediction in that only for artificial light do machine and human environmental conditions appear to be comparable.

Thus, it seems, that the relation between the technically measured and the visually judged hue and brightness is not really close, perhaps with the exception of HUEVIAL. One reason for this poor predictability could be the fact that, due to differing environmental conditions, the colorist's impression of the same dyestuff might differ. Another reason could be the definition of the colorimetric algorithm, calculating colorimetric variables from the remission spectrum. As a result of the OMEGA findings, experiments investigating colorimetrics have been initiated by the producer.

TABLE 8
*Canonical discriminant analysis: Simplification: Selection of variables (1. dimension)*

| $P$ | $R$-square | # variables | |
|---|---|---|---|
| 00 | 1.0000 | 18 | |
| 01 | 0.9960 | 14 | |
| 02 | 0.9713 | 11 | |
| 03 | 0.8813 | 10 | |
| 07 | 0.8805 | 8 | |
| 09 | 0.8749 | 7 | |
| 14 | 0.8740 | 6 | |
| 15 | 0.8719 | 5 | |
| 17 | 0.8331 | 4 | ⇐ selected simplification |
| 19 | 0.5809 | 3 | |

Four variables have to be selected by simplification to be close to the first discriminant coordinate. (See Table 5 for a discussion of $p$ and $R$-square.)

## 5.7 Summary of the Example

The main results of the OMEGA for the producer can be summarized as follows.

The measurement procedure for the coloristic strength has had to be changed to enable prediction by analytical variables. Instead of the 3 strength measurements possibly only the most economic measurement, by transmission spectra, may be used.

Hue prediction cannot be substantially improved beyond the performance of the producer's proposed predictors; brightness prediction seems not to be possible without additional information. Full reporting of

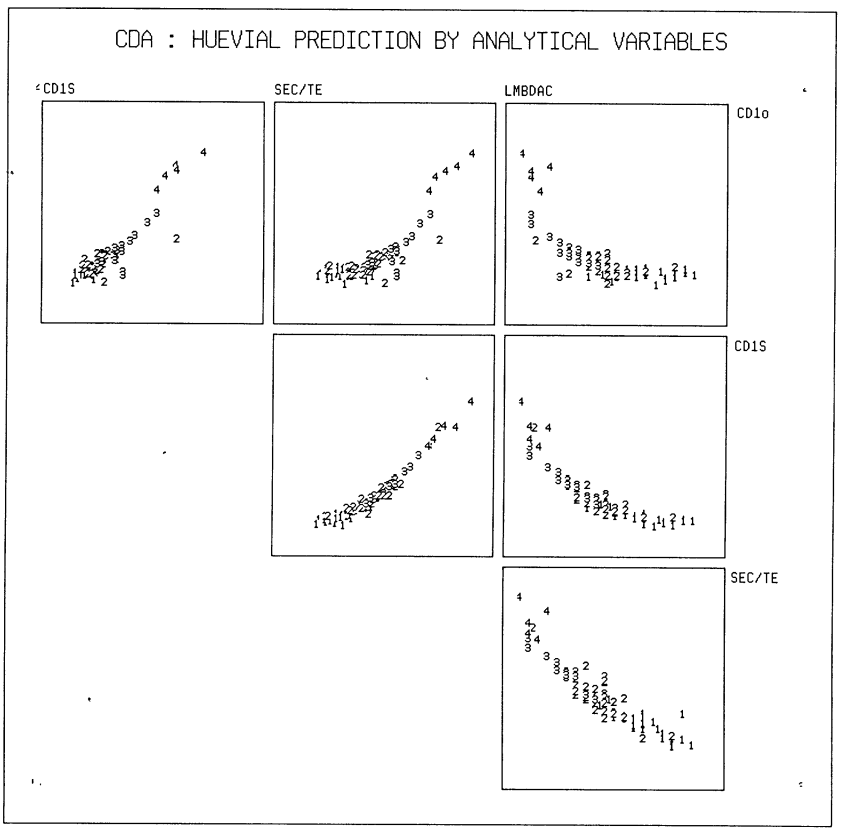


FIG. 20. *Comparison of the first discriminant coordinate* CD1o *and its simplification* CD1S *with two proposed predictors for HUEVIAL.* 1: *greener*; 2: *slightly greener*; 3: *near-standard*; 4: *slightly redder.*

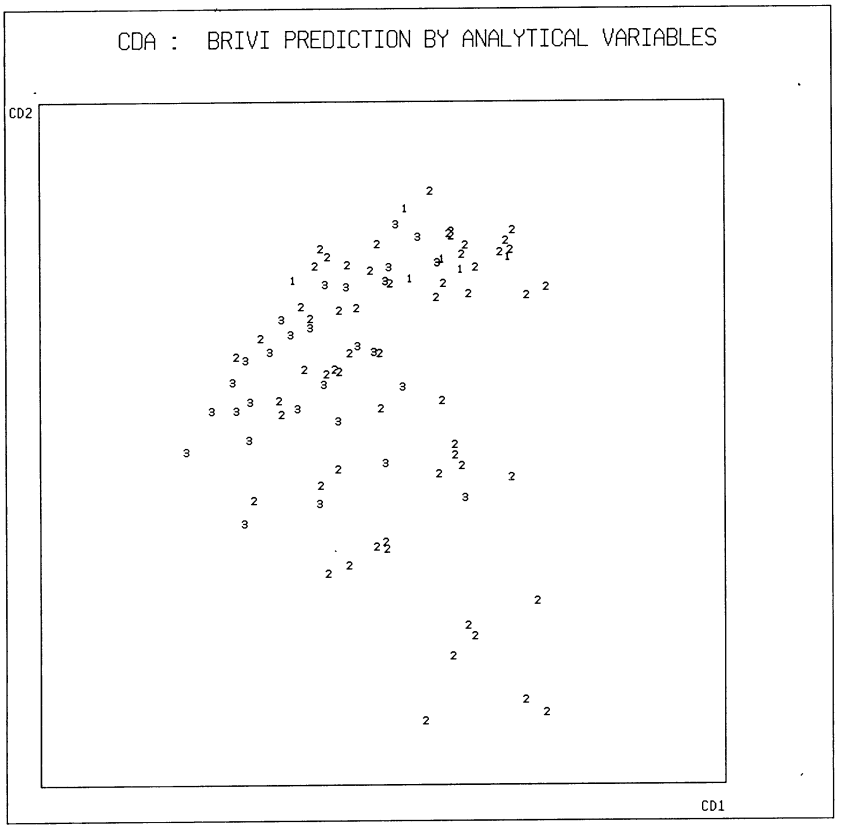


FIG. 21. *First two discriminant coordinates* CD1, CD2 *of a* CDA *with* BRIVI *as classifier and analytical variables as predictors.* 1: *near-standard*; 2: *slightly brighter*; 3: *brighter.*

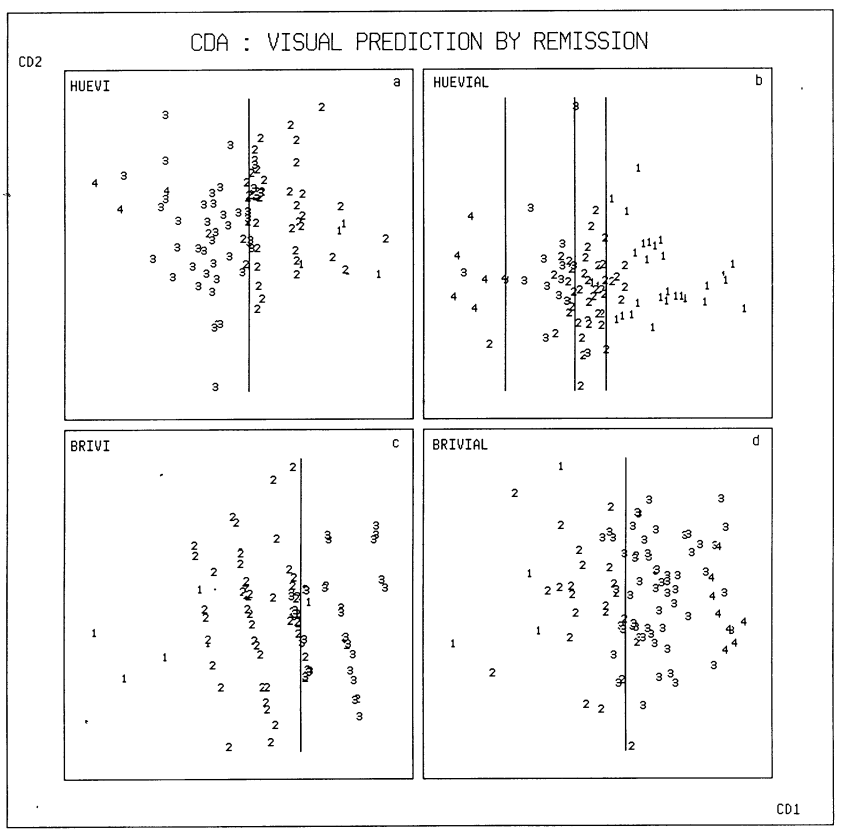


FIG. 22. *Prediction with* CDA *of visual judgments of hue and brightness by measurements of hue and brightness. Lines are optimally separating classes with regard to the first discriminant coordinate* CD1.

all measured analytical variables might help, rather than aggregating a number of them as at present.

A replacement of visual judgments by colorimetric variables is not possible given the current measurements and data quality. A closer look into the definition of the colorimetric variables and experiments with the colorists are suggested.

## 6. CONCLUDING REMARKS

Reconsidering the presented example session in the light of routine application, it should be reemphasized that all actions can be directed by interactive user control (see Section 4.2), i.e., online. In particular, the choice of the dimension reduction method and of the corresponding input variables, the dimension reduction, the resampling, the simplification and the presentation of the results can be directed by means of menus. Indeed, it is this simple control that makes the OMEGA pipeline so attractive.

Of course, the concept is very general and there is scope for introducing further methods. In particular, not only is dynamic graphics developing rapidly, but so also is research on algebraic exploratory methods, e.g., looking for optimal nonlinear pretransformations (see, e.g., the GIFI-methods in De Leeuw, 1984; ACE in Breiman and Friedman, 1985), and for other dimension reduction criteria (e.g., projection pursuit; see Friedman and Tukey, 1974; Huber, 1985). Some of these techniques are being tested at the moment with a view to including them in the OMEGA pipeline.

## APPENDIX 1: ALGEBRAIC TECHNIQUES

### A1.1 Notation

In the following sections, let:

$A :=$ $m \times h$-matrix of observations in original coordinates, sample mean per variable = 0. If there are two sets of variables, we write $A_l$, $l = 1, 2$, for each set of variables.

$C := A'A$ = "covariance" matrix. If there are two sets of variables, we write $C_{ll} := A_l' A_l$ = within-group covariance matrix, $l = 1, 2$, for each set of variables.

$C_S := D_C^{-1} C D_C^{-1}$ = "correlation" matrix,

$D_C :=$ diag(standard deviations).

### A1.2 Principal Components Analysis

*Principal components analysis on covariances* (*PCA-COV*) consists of a sequence of optimization problems

$$v_i' C v_i = \text{maximum over all } v_i' v_i = 1,$$

and

$$v_i' v_j = 0, \quad i \neq j, \quad i, j = 1, \cdots, h.$$

The solution of this problem can be obtained by a singular-value-decomposition (SVD): $C = VD^2V'$. Here the columns of $V$ are the required orthonormal sets of vectors $v_i$ representing the new coordinates, and $D$ is a diagonal matrix. The new coordinates are called *principal components*, and the corresponding representation of the observations can be obtained by $U := AV$. Thus, $D^2 = U'U =: \text{diag}(d_1^2, \cdots, d_h^2)$ is the diagonal matrix of the sample "variances" of the principal components. Since the $d_i$ are in decreasing order of magnitude, a natural *dimension reduction criterion* is given by: Choose the first $k$ such that

$$\sum_{i=1}^{k} d_i^2 \Big/ \sum_{j=1}^{h} d_j^2 \geq 0.95 \quad \text{(say)}.$$

The projections of the observations to the "biggest" $k$-dimensional subspace of $h$D space are then given by the rows of the matrix

$$U_k := \text{first } k \text{ columns of } U.$$

Note that, not only are the new coordinates orthogonal, but also the projections with respect to them are orthogonal, since $U'U$ is diagonal.

*Principal components analysis on correlations* (*PCA-COR*) consists of a sequence of optimization problems

$$v_{Si}' C_S v_{Si} = \text{maximum over all } v_{Si}' v_{Si} = 1,$$

and

$$v_{Si}' v_{Sj} = 0, \quad i \neq j, \quad i, j = 1, \cdots, h.$$

The solution and the dimension reduction is analogous to PCA-COV. The resulting matrices will be denoted $V_S$, $D_S$, $U_S$.

### A1.3 Canonical Discriminant Analysis

Let

$c :=$ $m$-vector of classification of observations,

$B :=$ pooled between-classes covariance matrix:

$b_{ij} := 1/(L-1) \sum_{l=1}^{L} m_l (\bar{a}_i^l - \bar{a}_i)(\bar{a}_j^l - \bar{a}_j)$, where

$L :=$ number of different values in $c$ = number of different classes,

$m_l :=$ number of observations in class $l = 1, \cdots, L$,

$\bar{a}_i^l :=$ mean of elements in column $i$ of $A$ belonging to class $l$,

$\bar{a}_i :=$ mean of all elements in column $i$ of $A$,

$W :=$ pooled within-classes covariance-matrix,

$W := 1/(m-L) \sum_{l=1}^{L} (m_l - 1) S_l$, where

$S_l :=$ sample covariance matrix of observations in class $l$.

*Canonical discriminant analysis* (*CDA*) then consists of a sequence of optimization problems,

$$v_{Bi}' B v_{Bi} = \text{maximum over all } v_{Bi}' W v_{Bi} = 1,$$

and

$$v'_{Bi} W v_{Bj} = 0,$$
$$i \neq j, \quad i, j = 1, \cdots, h_B := \min(L-1, h).$$

The solution of this problem can be obtained by a singular-value decomposition of

$$B_T := U_W^{-1} B (U_W^{-1})',$$

where

$$W =: U_W U'_W$$

is the Choleski decomposition of $W$. Let the SVD of $B_T$ be

$$B_T = V_{BT} D_{BT}^2 V'_{BT}.$$

Then the columns of $V_B := (U_W^{-1})' V_{BT}$ are the required $W$-orthogonal set of vectors $v_{Bi}$ representing the new coordinates. The new coordinates are called discriminant coordinates and the corresponding representation of the observations can be obtained by

$$U_B := A V_B.$$

Thus

$$U'_B U_B = \frac{m-L}{m-1} I_{hB} + \frac{L-1}{m-1} D_{BT}^2$$
$$=: D_B^2 =: \operatorname{diag}(d_{B1}^2, \cdots, d_{Bh_B}^2)$$

is the diagonal matrix of sample covariances of the discriminant coordinates. Since the $d_{Bi}$ are in decreasing order of magnitude, i.e., single coordinate class separation gets worse and worse, a natural dimension reduction criterion is given by: Choose the first $k$ such that

$$\sum_{i=1}^{k} d_{Bi}^2 \Big/ \sum_{j=1}^{h_B} d_{Bj}^2 \geq 0.95 \quad \text{(say)}.$$

The projections of the observations to the "best separating" $k$-dimensional subspace of $h$D space are then given by the rows of the matrix

$$U_{Bk} := \text{first } k \text{ columns of } U_B.$$

Note that the discriminant coordinates are $W$-orthonormal, but that the projections with respect to them are orthogonal in the Euclidean sense, since $U'_B U_B$ is diagonal.

### A1.4 Canonical Correlation Analysis

Let

$C_{12} := A_1' A_2 =$ between-group covariance matrix.

*Canonical correlation analysis* (*CCA*) then consists of a sequence of optimization problems,

$$v'_{1i} C_{12} v_{2i} = \text{maximum over all } v'_{li} C_{ll} v_{li} = 1,$$

and

$$v'_{li} C_{ll} v_{lj} = 0,$$
$$i \neq j, \quad i, j = 1, \cdots, h_{12} := \min(h_1, h_2), \quad l = 1, 2.$$

The solution of this problem can be obtained by a singular value decomposition of

$$C_{12T} := U_{1T}^{-1} C_{12} (U_{2T}^{-1})',$$

where

$$C_{ll} =: U_{lT} U'_{lT}$$

are Choleski decompositions of $C_{ll}$, $l = 1, 2$. Let the SVD of $C_{12T}$ be

$$C_{12T} =: V_{1T} D_C V'_{2T},$$

then the first $h_{12}$ columns of $V_l := (U_{lT}^{-1})' V_{lT}$, $l = 1, 2$, are the required $C_{ll}$-orthonormal sets of vectors $v_{li}$ representing the new coordinates. These new coordinates are called *canonical variables* of group $l$, and the corresponding representation of the observations can be obtained by

$$U_l := A_l V_l, \; l = 1, 2.$$

Thus $U_l' U_l = I$, and $U_1' U_2 = D_C$ is the "diagonal" matrix with the sample correlations between the $v_{1i}$ and $v_{2i}$, $i = 1, \cdots, h_{12}$, the so-called *canonical correlations* $d_{Ci}$, as the first entries. Since the $d_{Ci}$ are in decreasing order of magnitude, i.e., the sample correlation becomes smaller and smaller, interest in high predictability may lead to the following *dimension reduction criterion*: Choose the last $k$ such that $d_{Ck} \geq 0.90$ (say). The projections of the observations to the "best predictable" $k$-dimensional subspaces are then given by the rows of the matrices

$$U_{lk} := \text{first } k \text{ columns of } U_l.$$

Note that the two sets of canonical variables are $C_{ll}$-orthonormal, respectively, but that the projections with respect to them are orthogonal in the Euclidean sense, not only inside the groups but also for the noncorresponding canonical variables in different groups. Indeed, only projections on corresponding pairs of canonical variables are correlated. This may motivate corresponding dimension reductions to 2D.

### A1.5 Successive Orthogonalization

*Successive orthogonalization* (*SOG*) consists of a sequence of orthogonalization steps, such that the projection on the $i$th new coordinate is orthogonal to the projections on the first $i-1$ new coordinates.

The solution to this problem can be obtained by a Choleski decomposition of $C$: $C = T D_L^2 T'$, where $T$ is a lower triangle with 1s on its diagonal, and $D_L$ is a diagonal matrix. Here the columns of $(T')^{-1}$ are representing the new coordinates, and the corresponding

representations of the observations can be obtained by

$$U_T := A(T')^{-1}.$$

Thus, $D_T^2 := U_T'U_T =: \text{diag}(d_{T1}^2, \cdots, d_{Th}^2)$ is the diagonal matrix of the sample "variances" of new coordinates. Since the $d_{Ti}$ are in decreasing order of magnitude, a natural dimension reduction criterion is given by: Choose the first $k$ such that

$$\sum_{i=1}^{k} d_{Ti}^2 \Big/ \sum_{j=1}^{h} d_{Tj}^2 \geq 0.95 \quad \text{(say)}.$$

The projections of the observations on the corresponding $k$-dimensional subspace of $h$D space are then given by the rows of the matrix

$$U_{Tk} := \text{first } k \text{ columns of } U_T.$$

Note that the new coordinates are not orthogonal in the Euclidean sense, but the corresponding projections are, since $U_T'U_T = D_T^2$.

### A1.6 Simplification

Simplification is achieved by the specification of an appropriate importance criterion, i.e., a criterion which helps to decide if a variable is needed for an adequate simplification. The criterion for *PCA-COV* may be motivated as follows. It is well known that the first principal component $v_1$ delivers the best 1D approximation to the data in that

$$\| A - R_1 \|_F^2 \geq \| A - Av_1v_1' \|_F^2$$

for all rank 1-matrices $R_1$, where $\| A \|_F^2 := \text{trace}(A'A)$. Now, the (normalized) simplification $v_{1S}$ of $v_1$ may be chosen to be a "near optimizer" by guaranteeing that

$$\| A - Av_{1S}v_{1S}' \|_F^2 \leq \| A - Av_1v_1' \|_F^2 + \varepsilon,$$

$$\varepsilon > 0 \text{ prefixed.}$$

One can show that this is equivalent to

$$0 \leq 2d_1^2(v_1'(v_1 - v_{1S})) - (v_1 - v_{1S})'A'A(v_1 - v_{1S}) \leq \varepsilon.$$

Since $v_1 - v_{1S}$ is small, the first term on the left-hand side dominates. Since one is interested in simplifying as much as possible, this obviously leads to setting the loadings with the smallest absolute values to zero as long as the restriction holds, leaving the other loadings untouched. Note that this is somewhat related to ideas of Jolliffe (1972) for identifying important original coordinates.

*PCA-COR simplification* works analogously using $v_{Si}$ instead of $v_i$.

*CDA simplification* works analogously to PCA-COV simplification, setting those elements of $v_{Bi}$ equal to zero with

$$|v_{BiSk}| < p\left(\max_j |v_{BiSj}|\right) \Big/ 100,$$

where

$$v_{BiS} := \text{diag}(W)^{0.5}v_{Bi},$$

$$\text{diag}(W) := \text{diag}(w_{11}, \cdots, w_{hh}).$$

Thus, $v_{BiS}$ delivers the importance criterion for $v_{Bi}$. Note, that the full transformation to the PCA-situation would have been

$$v_{BiT} = U_W'v_{Bi}, \quad \text{where } W = U_W U_W'.$$

But since $W$ is positive definite, and thus diagonal dominant, $\text{diag}(W)^{0.5}$ can be used as an approximation to $U_W$, in particular if within-class covariances are much smaller than variances. Moreover, note that multiplication with $\text{diag}(W)^{0.5}$ corresponds to elementwise multiplication by the pooled within-class sample standard variations.

*CCA simplification* works analogously to PCA-COV simplification, applied to each group $l = 1, 2$ of variables individually, setting those elements of $v_{li}$ equal to zero with

$$|v_{liSk}| < p\left(\max_j |v_{liSj}|\right) \Big/ 100,$$

where

$$v_{liS} := \text{diag}(C_{ll})^{0.5}v_{li}, \quad l = 1, 2.$$

Thus, $v_{liS}$ delivers the importance criterion for $v_{li}$. The motivation works analogously as with CDA.

### A1.7 Resampling and Procrustes Transformation

The application of the algebraic techniques and the subsequent dimension reduction always leads to projections that are orthogonal in the Euclidean sense. Hence, it is sufficient to illustrate the optimal Procrustes transformation only for PCA-COV.

*Equivalent projections for PCA-COV.* Let $U_{kI}$, $U_{kIS}$ be $k$D projections of the subsample of observed objects with indices $i \in I$, having based PCA-COV upon the whole sample and upon the objects in the complement of $I$, respectively. Then

$$U_{kIST} := cU_{kIS}R + b$$

is optimally matching with $U_{kI}$ under translation, rotation/reflection and scaling, where

$$c := \text{tr}(S)/\text{tr}((U_{kIS} - MU_{kIS})'(U_{kIS} - MU_{kIS}))$$

: scaling;

$R := ZW'$ : rotation/reflection, i.e., orthogonal transformation;

$b := MU_{ki} - cMU_{kiS}R$ : translation.

Herein, $W$, $S$, $Z$ stem from the SVD:

$$(U_{kiS} - MU_{kiS})'(U_{ki} - MU_{ki}) = ZSW',$$

and

$$M := \frac{1}{m}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix}.$$

Note that in case of total matching $U_{kiST} = U_{ki}$, i.e.,

$$c(U_{kiS} - MU_{kiS})Z = (U_{ki} - MU_{ki})W$$

(compare CCA).

## APPENDIX 2: HARDWARE AND SOFTWARE ENVIRONMENT

This is definitely not the place to discuss which hardware and software allow for what parts of the proposed OMEGA pipeline (see, instead, Cleveland and McGill, 1988; Weihs, 1988). But let us briefly discuss the merits and drawbacks of the computer environment where the pipeline is implemented at the moment in CIBA-GEIGY: the software package ISP with the graphics system DGS/SGS (Dynamic/Static Graphics System) (ISP, 1988) running on Apollo workstations.

Concerning interactive user controls, ISP only offers menus, stop control and animated control. Lasting control is, e.g., available in the system Plot windows (Stuetzle, 1988). Naturally, 3D rotation is offered by ISP. But the effect of interpolation has to be simulated by static techniques, e.g., by parallel viewing at intermediate phases of the movement. As a dynamic method interpolation is offered by the system VISUALS (Young, Kent and Kuhfeld, 1988). The static representation by disconnected arrows is available in S-PLUS (1988). Concerning viewport transformations, the only drawbacks of ISP/DGS are that expanding and shrinking are possible only in both directions simultaneously, and that shifting is possible only in one direction at a time. Here, the Data viewer (Buja, Asimov, Hurley and McDonald, 1988) seems to be most complete. Concerning scatterplot matrices, the only restriction is the nonavailability of linked histograms, which are, for example, offered by S-PLUS (1988). On the other hand, rectangular scatterplot matrices are nicely implemented in ISP/DGS. Concerning brushing in scatterplot matrices, ISP only offers coloring in lasting mode. Here again, S-PLUS is more complete. Parallel views can only be generated using the Static Graphics System ISP/SGS. There, dynamic operations like brushing and rotation are not available. Full implementations of parallel views are available in the Data viewer and in Plot windows.

Despite its restrictions, ISP-SGS/DGS appears to be a flexible tool for the implementation of the OMEGA pipeline on Apollo workstations and even on IBM-PC/AT or -PS2. An alternative software package on workstations may be S-PLUS. The Data viewer and Plot windows are surely the most complete systems, but they run only on the exotic and expensive Symbolics Lisp machines. Other systems, like VISUALS, running on IBM-PC/XT/AT, and the well-known MacSpin (Donoho, Donoho and Gasko, 1988), running on Macintoshs, are not powerful enough for the kind of analyses we have been considering.

## ACKNOWLEDGMENTS

The authors would like to thank Barbara Weeks and Amy Racine-Poon for contributing to the search for a suitable acronym for the strategy, W. Stahel for proposing pointwise distributions of predictions and M. Berres for proposing the static equivalent to plot-interpolation. We thank G. Anders for providing the data, for proposing predictors and for basic analyses of similar data which motivated our data analysis. Last, but not least, we would like to thank W. Baumeister for the fruitful discussions about the relevance of the outcomes of our example, and A. F. M. Smith for reviewing the paper.

## REFERENCES

ARBUCKLE, J. and FRIENDLY, M. L. (1977). On rotating to smooth functions. *Psychometrika* **42** 127–140.

ASIMOV, D. (1985). The grand tour: A tool for viewing multidimensional data. *SIAM J. Sci. Statist. Comput.* **6** 128–143.

BECKER, R. A. and CLEVELAND, W. S. (1988). Brushing scatterplots. In *Dynamic Graphics for Statistics* (W. S. Cleveland and M. E. McGill, eds.) 201–224. Wadsworth, Belmont, Calif.

BECKER, R. A., CLEVELAND, W. S. and WEIL, G. (1988). The use of brushing and rotation for data analysis. In *Dynamic Graphics for Statistics* (W. S. Cleveland and M. E. McGill, eds.) 247–275. Wadsworth, Belmont, Calif.

BECKER, R. A., CLEVELAND, W. S. and WILKS, A. R. (1988). Dynamic graphics for data analysis. In *Dynamic Graphics for Statistics* (W. S. Cleveland and M. E. McGill, eds.) 1–50. Wadsworth, Belmont, Calif.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26** 211–243.

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlations (with discussion). *J. Amer. Statist. Assoc.* **80** 580–619.

BUJA, A., ASIMOV, D., HURLEY, C. and MCDONALD, J. A. (1988). Elements of a viewing pipeline for data analysis. In *Dynamic Graphics for Statistics* (W. S. Cleveland and M. E. McGill, eds.) 277–308. Wadsworth, Belmont, Calif.

CHAMBERS, J. M., CLEVELAND, W. S., KLEINER, B. and TUKEY, P. A. (1983). *Graphical Methods for Data Analysis*. Wadsworth, Belmont, Calif.

CIE (1978). Recommendations on uniform color spaces, color-difference formulas, psychometric color terms. Supplement No. 2 to CIE Publication No. 15 (E-1.3.1) 1971/(TC-1.3) 1978.

CLEVELAND, W. S. and MCGILL, M. E., eds. (1988). *Dynamic Graphics for Statistics*. Wadsworth, Belmont, Calif.

CLEVELAND, W. S., MCGILL, M. E. and MCGILL, R. (1986). The shape parameter of a two-variable graph. *Proc. ASA Sec. Statist. Graphics* 1–10.

DE LEEUW, J. (1984). The GIFI system of nonlinear multivariate analysis. In *Data Analysis and Informatics III* (E. Diday, M. Jambu, L. Lebart, J. Pagès and R. Tomassone, eds.) 415–424. North-Holland, Amsterdam.

DONOHO, A. W., DONOHO, D. L. and GASKO, M. (1988). MACSPIN: Dynamic graphics on a desktop computer. In *Dynamic Graphics for Statistics* (W. S. Cleveland and M. E. McGill, eds.) 331–351. Wadsworth, Belmont, Calif.

EASTMENT, H. T. and KRZANOWSKI, W. J. (1982). Cross-validatory choice of the number of components from a principal component analysis. *Technometrics* **24** 73–77.

FISHERKELLER, M. A., FRIEDMAN, J. H. and TUKEY, J. W. (1988). PRIM-9: An interactive multidimensional data display and analysis system. In *Dynamic Graphics for Statistics* (W. S. Cleveland and M. E. McGill, eds.) 91–109. Wadsworth, Belmont, Calif.

FRIEDMAN, J. H. and TUKEY, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **C-23** 881–890.

GABRIEL, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58** 453–467.

GNANADESIKAN, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York.

GOODALL, C. and THOMA, H. M. (1987). Interpolation of multivariate data. *Proc. ASA Sec. Statist. Graphics* 64–67.

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–475.

ISP (1988). *ISP User's Guide*. Artemis Systems, Carlisle, Pa.

JOLLIFFE, I. T. (1972). Discarding variables in a principal components analysis. I: Artificial data. *Appl. Statist.* **21** 160–173.

KRZANOWSKI, W. J. (1987). Selection of variables to preserve multivariate data structure, using principal components. *Appl. Statist.* **36** 22–33.

SIBSON, R. (1978). Studies in the robustness of multidimensional scaling: Procrustes statistics. *J. Roy. Statist. Soc. Ser. B* **40** 234–238.

SIBSON, R. (1979). Studies in the robustness of multidimensional scaling: Perturbational analysis of classical scaling. *J. Roy. Statist. Soc. Ser. B* **41** 217–229.

S-PLUS (1988). *S-PLUS User's Manual*. Statistical Sciences, Seattle, Wash.

STUETZLE, W. (1988). Plot windows. In *Dynamic Graphics for Statistics* (W. S. Cleveland and M. E. McGill, eds.) 225–245. Wadsworth, Belmont, Calif.

WANG, C. M. and GUGEL, H. W. (1988). A high performance color graphics facility for exploring multivariate data. In *Dynamic Graphics for Statistics* (W. S. Cleveland and M. E. McGill, eds.) 379–389. Wadsworth, Belmont, Calif.

WEIHS, C. (1988). Dynamic graphical methods in multivariate exploratory data-analysis: An overview. Technical Report 8802, Mathematical Applications, CIBA-GEIGY, Basel.

WOLD, S. (1978). Cross-validatory estimation of the number of components in factor and principal component models. *Technometrics* **20** 397–405.

YOUNG, F. W., KENT, D. P. and KUHFELD, W. F. (1988). Dynamic graphics for exploring multivariate data. In *Dynamic Graphics for Statistics* (W. S. Cleveland and M. E. McGill, eds.) 391–424. Wadsworth, Belmont, Calif.

# Comment

**A. Buja and C. Hurley**

Reading the authors' paper was very gratifying for us: as it happens, we have been working on the integration of multivariate analysis and graphical data analysis as well. We are delighted to observe that our separate efforts converged to some degree. While we may differ in details of implementation, human interface and computing philosophy, our independent efforts indicate a certain necessity in the idea of marrying classical multivariate analysis and the more recent high-interaction graphics tools. A paper by us on this subject is in press in the *SIAM Journal on Scientific and Statistical Computing* (Hurley and Buja, 1990). It is based on the Ph.D. thesis of Hurley (1987). The multivariate methods which we considered were the same as those of the authors with the exception of their successive orthogonalization procedure. The authors carried certain ideas of visual inference and assessment considerably further than we did (for now, we have not gone beyond what is documented in Buja, Asimov, Hurley and McDonald, 1988). On the other hand, we may claim a tighter coupling of multivariate analysis and graphics, as we will show below.

*A. Buja is a member of the technical staff, Bellcore, 445 South Street, MRE 2Q-362, Box 1910, Morristown, New Jersey 07960-1910. C. Hurley is Assistant Professor, Department of Statistics, The George Washington University, Washington, DC 20052.*

## MULTIVARIATE ANALYSIS (MA) AND GRAPHICAL METHODS

A basic motivation behind the authors' and our endeavor is the simple insight that MA allows us to generate a wealth of potentially illuminating data projections. Curiously, the first attempts at combining interactive graphics with automatic methods for finding informative projections were based on projection pursuit rather than classical MA. Surely, the latter can be interpreted as a subset of the former, but this view does not do justice to MA. It is more useful to interpret MA as a set of methods for changing coordinate systems in a data-driven way. One reason for the initial lack of interest in the graphical and explor-