The Response Surface as Explorative Tool in Multiple Correspondence Analysis

Gilbert Saporta
Conservatoire National des Arts et Métiers, France
saporta@cnam.fr

Giuseppe Giordano
Dip. di Scienze Economiche e Statistiche
Università di Salerno, Italia
ggiordan@unisa.it

Outline

1. Introduction and background
2. Aim
3. Tools
4. Methods
5. Case study: Insurance Data
6. Concluding remarks

Introduction

- In many application fields the reduction of a multidimensional construct is addressed through the use of multidimensional data analysis methods

- The use of techniques such as Principal Component Analysis or Correspondence Analysis is nowadays very common for the analysis of survey data and it is made even easier by standard procedures in the most known statistical packages: SAS, SPAD, SPSS, etc.

Background

- In *soft*-model analysis, when multivariate data are structured as *response* and *predictor* variables, several multivariate techniques are better suited for both explorative and synthesis aims
- The use of *Reduced Rank Regression* (Anderson 1951), *Partial Least Squares* (Wold, 1982, 1985), *Preference Mapping* (Carroll, 1972) etc. are widespread in the fields of marketing, chemometrics, psychometrics, sensometrics, ecology, etc…

Background

All these different techniques share the powerful capability of geometrical interpretation and graphical representation...

Dynamical graphics represents *"the Visual Revolution in Computer Science"* (Carr, 1998)

...Is there a new dimension in multivariate techniques?

Objective

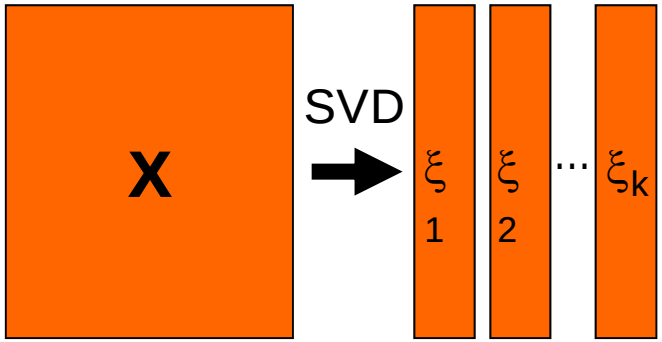
- We propose to use the Response Surface Methodology (RSM, Box and Wilson, 1951) as a further graphical resource to analyse and interpret the results of a multidimensional data analysis

- We aim at showing how this kind of representation may enrich the interpretation of factorial technique results

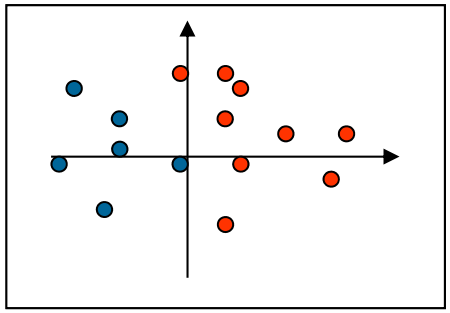
- The basic idea consists in an overlapping representation of traditional factorial plan with a response surface derived from an idiosyncratic "response variable" :

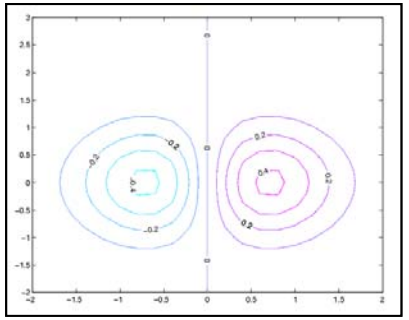
The Response Surface Factorial Map

The Idea



Plot

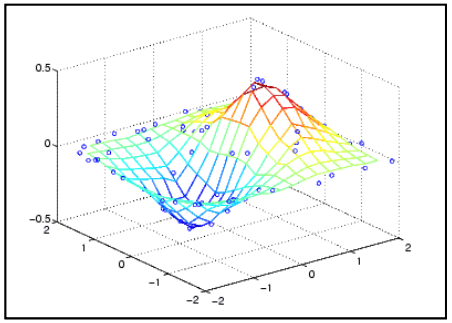




Auxiliary Information
Illustrative Variables
Response Variables
Functional Constraints
……..

Y

$f(\xi_1,\xi_2)$



Response Surface and DOE

In the framework of Design of Experiments the use of the Response Surface Methodology allows to analyse the relationships between the response variable and a set of input factors

The analysis consists of successive steps of experimentation, modelling, data analysis and optimization

The aim is to obtain an accurate approximation of the response surface and to identify an optimum design region

Most designs use a quadratic response surface *(Box-Behnken, Central-Composite Design, etc.)*

Response Surface and DOE

- Several statistical packages allows to carry out the RSM (for example, the *ADX* menu in *SAS/QC* system)
- The typical graphical output is the three-dimensional representation of the surface and the *Contour Plot*. A contour plot is a graphical technique for representing a three-dimensional surface by plotting constant slices, called contours, on a two-dimensional format
- Given a value for the response y, lines are drawn for connecting the (x_1, x_2) coordinates where the y value occurs. These lines are the iso-response values

Response Surface and DOE

- The independent variables are usually restricted to a regular grid. The actual techniques for determining the correct iso-response values are rather complex and almost always computer generated…
- Detailed routines can be found in *Matlab*, *SAS*, *S-plus*, *Statgraphics* and in many other general purpose graphics and mathematics programs for example: the *Q-hulls* in Matlab…

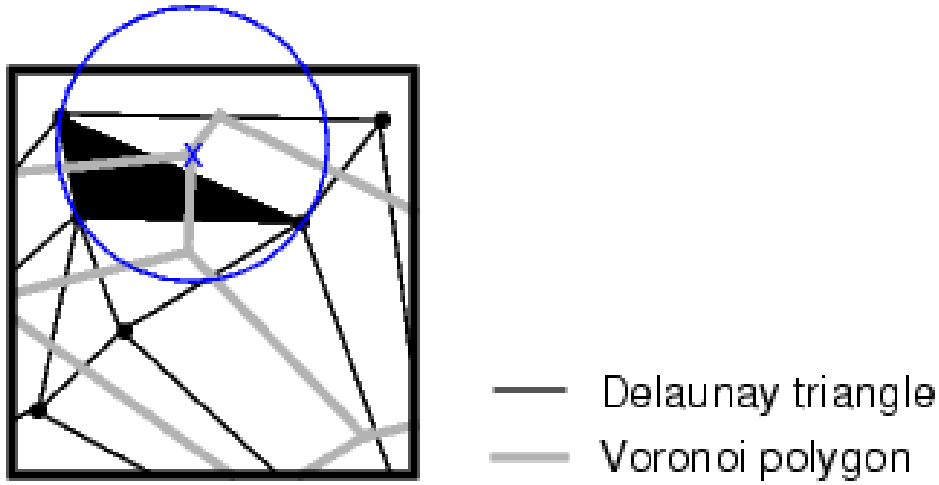
Q-hull routine: http://www.qhull.org

Delaunay Complex

Given a set of points, the Delaunay complex is the set of lines linking each point to its neighbors (useful in finding the Convex Hull)

Voronoi Polygon

The Circle inscribing a Delaunay's triangle has its center at the vertex of a Voronoi polygon



Voronoi Diagram

For each point p in S, consider the hyperplane tangent to the paraboloid in R^{d+1} at p: This hyperplane is represented by :

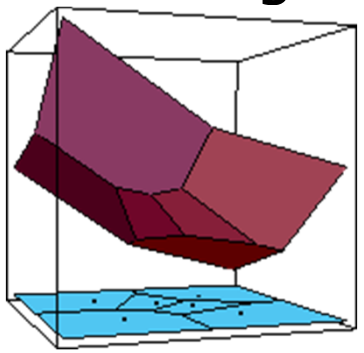
$$\sum_{j=1}^{d} p_j^2 - \sum_{j=1}^{d} 2p_j x_j + x_{d+1} = 0.$$

by replacing the equality with inequality ≥ for each point p, we obtain a system of n inequalities.

The polyhedron P in R^d of all solutions x to the system of inequalities is a lifting of the Voronoi diagram to one higher dimensional space

Komei Fukuda, *Polyhedral computation FAQ*, Swiss Federal Institute of Technology, Lausanne and Zurich, Switzerland

http://www.ifor.math.ethz.ch/staff/fukuda/polyfaq/polyfaq.html



Factorial Maps

- Let's assume as an initial dataset the coordinates obtained by a factorial technique through Singular Value Decomposition…

  Generalised Principal Component Analysis

  (Greenacre, 1984, Appendix A)

Multiple Correspondence Analysis

- We deal with Multivariate Indicator Matrix
- Data are collected so that, for each statistical unit, we may have:
  - Categorical data coding: $Z \in \{0,1\}$

    +
  - Covariate real-valued data: $y \in$?

MCA and Response Surface

- The contour plot is generated for any two factors. Typically, this should be the two most important factors as determined by MCA

- A matrix of all pairwise contour plots can also be generated for a number of important factors (similar to the scatter-plot matrix)

The Response Surface Factorial Map (RSFM)

We call this kind of graphical representation *Response Surface Factorial Map* (*RSFM*)

According to the different kinds of response variables, we distinguish between *Internal Analysis* and *External Analysis*

- In the Internal Analysis, the response variable is used to better understand the results of the factorial technique
- In the External Analysis, an outer information is introduced in the analysis

The Internal Analysis: Quality of Representation

the surface represents information about the quality of the **analysis results**

we can represent different kinds of information, for example, a function of the *relative contribution*

External Analysis: the response variable

The term **External** is referred to the circumstance that the response variable does not contribute to the definition of the principal axes

- … to represent the response of a continuous variable not used in the factorial decomposition
- …to enhance the visualization of illustrative continuous variables

Application fields

As a data analysis interpretative tool
(Quality of representation)

Exploring relationships between latent and measurable variables
(Principal Component Regression)

Showing constrained Solutions on the Factorial Map
(Technological, Economical or Production Frontiers, …)

……

The Procedure (Matlab v.7.0 code)

```
function pcsurf(F1, F2, Z, dens);
tick1 = (max(F1)-min(F1)) / dens;
tick2 = (max(F2)-min(F2)) / dens;
t1 = (min(F1)-tick1) : tick1 : (max(F1)+tick1);
t2 = (min(F2)-tick1) : tick2 : (max(F2)+tick2);
[XI,YI] = meshgrid(t1,t2);
ZI = griddata(F1,F2,Z,XI,YI);
contour(XI,YI,ZI), hold
plot3(F1,F2,Z,'.r'), hold off
figure;
surfc(XI,YI,ZI), hold
plot3(F1,F2,Z,'.y'), hold off
```

Data analysis: Insurance Data (1992)

- Sample: 1106 Belgian policy holder
  (individuals and companies)
  - no claim in the last year, "the good ones": 556
  - one or more claims in the last year, "the bad ones": 550
- Database SPAD: « Assurbin.sba »
- Method: DISQUAL Strategy (Saporta, 1976)
- Goal: Representation of the Scorecard function on the factorial plan of MCA

Variable selection – by P. Périé, CISIA

SELECTION OF CASES AND VARIABLES
ACTIVE CATEGORICAL VARIABLES
9 VARIABLES 20 ASSOCIATED CATEGORIES

Usage type (2): ***Professional; Private***	(2 CATEGORIES)
Insured type (3): ***Male; Female; Company***	(3 CATEGORIES)
Language (2): ***French; Flemish***	(2 CATEGORIES)
Birth coorth (3): ***1890-1949; 1950-1979, Unknown***	(3 CATEGORIES)
Region (2): ***Bruxelles; Other Regions***	(2 CATEGORIES)
Bonus-malus level (2): ***B-M+, other B-M (−1)***	(2 CATEGORIES)
Year of subscription (2): ***<1986; 1986+***	(2 CATEGORIES)
Horse Power (2): ***30-39; 40-349***	(2 CATEGORIES)
Year of vehicle construction (2): ***1933-1989; 1990-1991***	(2 CATEGORIES)

SUPPLEMENTARY CATEGORICAL VARIABLES
1 VARIABLES 2 ASSOCIATED CATEGORIES

---

1 . Accident RC (2): ***0 Claim; 1+Claims*** (2 CATEGORIES)

INDIVIDUALS

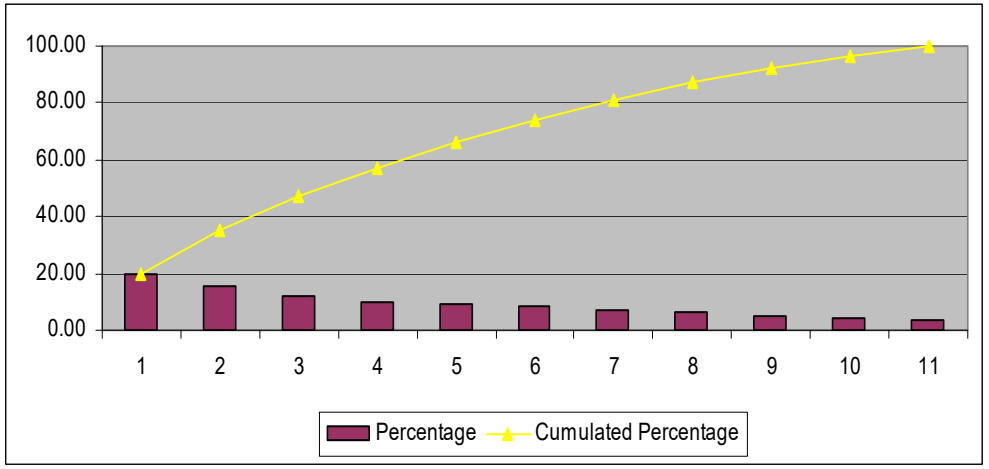
---------- NUMBER ----------WEIGHT ----------

WEIGHT OF INDIVIDUALS: Weight of objects, uniform equal to 1. UNIF
KEPT ..... NITOT = 1106 PITOT = 1106.000
ACTIVE ..... NIACT = 1106 PIACT = 1106.000
SUPPLEMENTARY ..... NISUP = 0 PISUP = 0.000

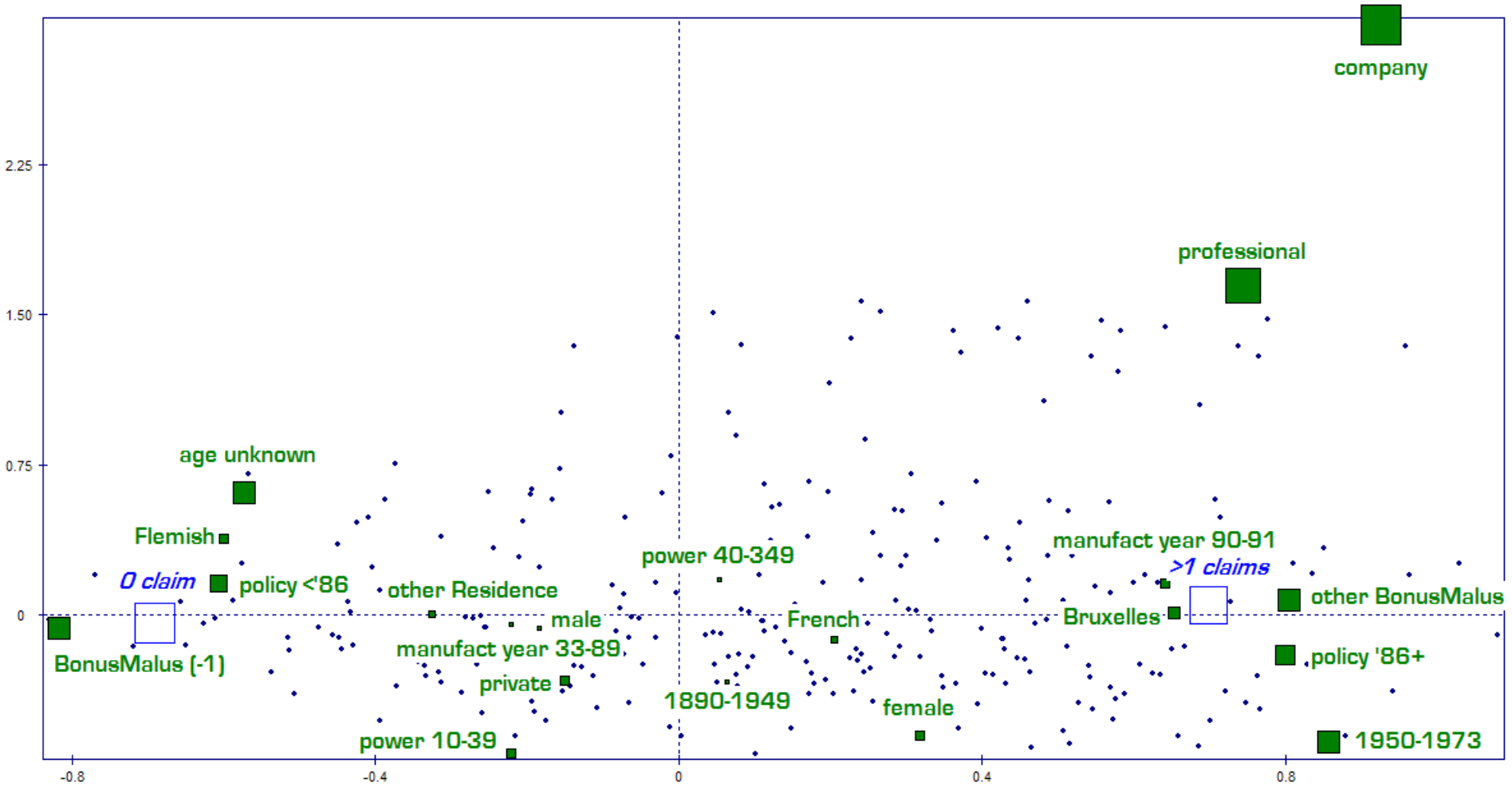
Eigenvalues of MCA

Trace of matrix: **1.22222**

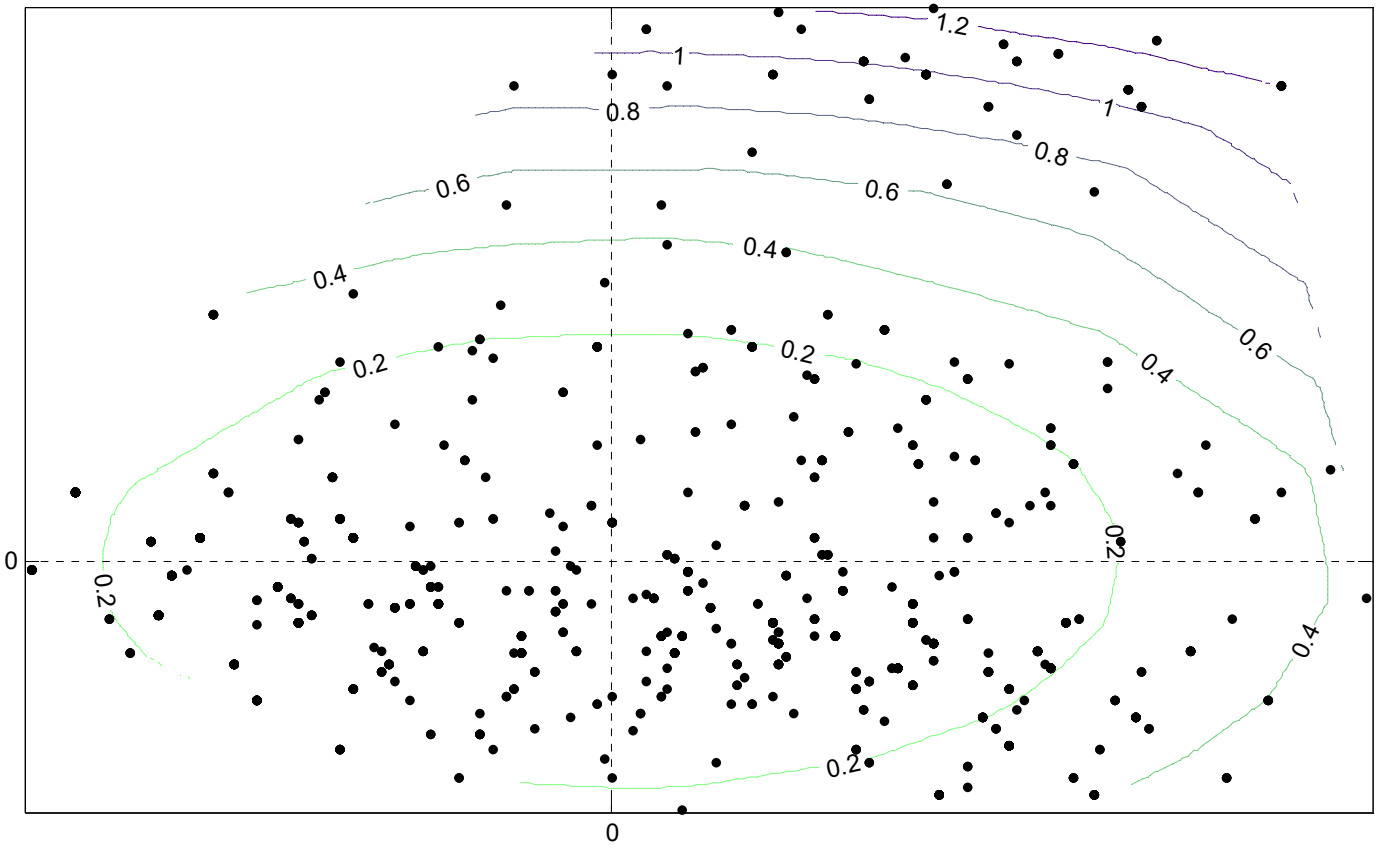
Number	Eigenvalue	Proportion	Cumulative
1	**0.2438**	**19.95**	**19.95**
2	**0.1893**	**15.49**	**35.44**
3	0.1457	11.92	47.36
4	0.1201	9.82	57.18
5	0.1091	8.92	66.11
6	0.0999	8.17	74.28
7	0.0855	7.00	81.28
8	0.0732	5.99	87.26
9	0.0573	4.68	91.95
10	0.0511	4.18	96.13
11	0.0473	3.87	100.00



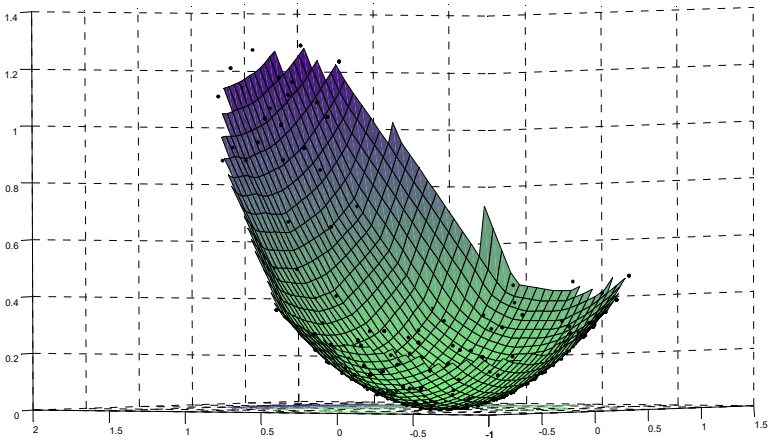
The MCA factorial map

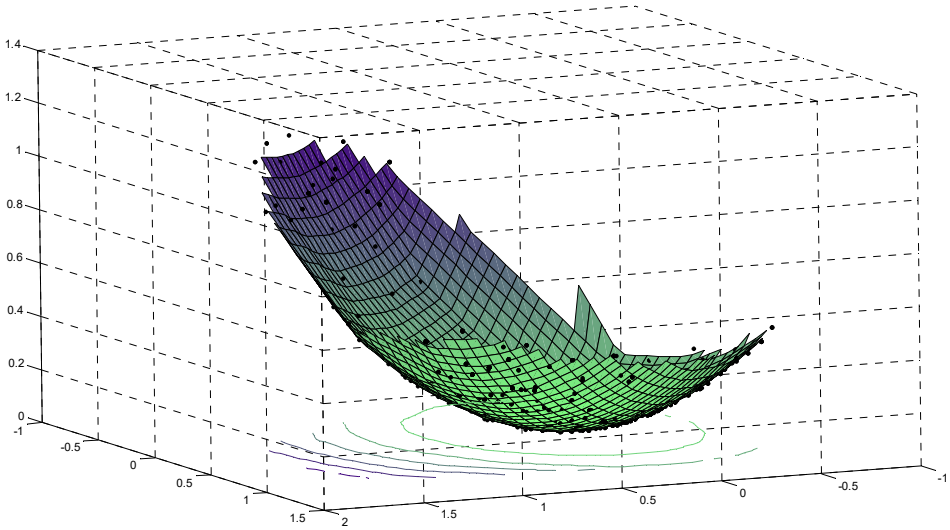


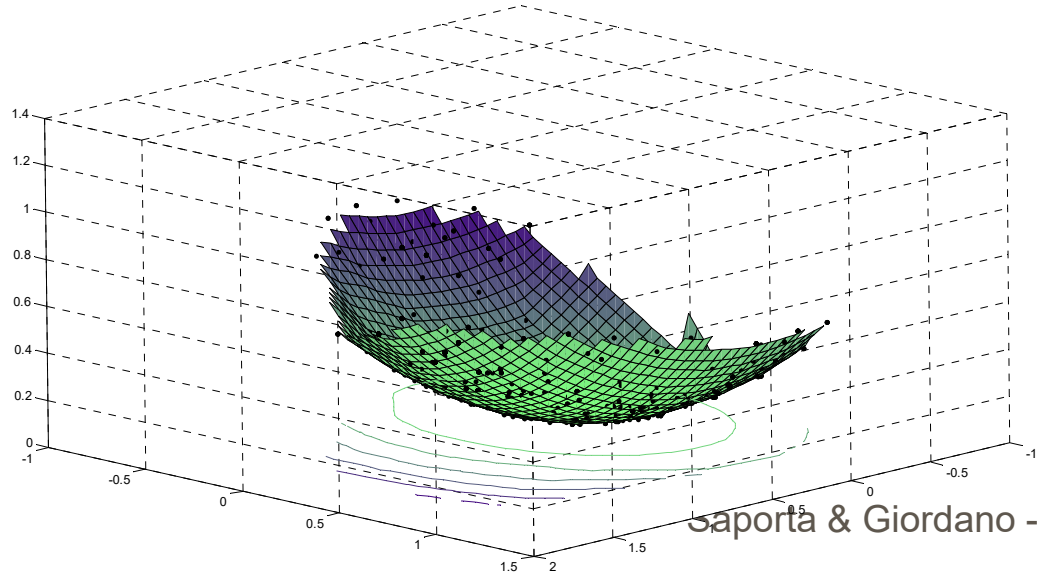
Internal Analysis: Contributions

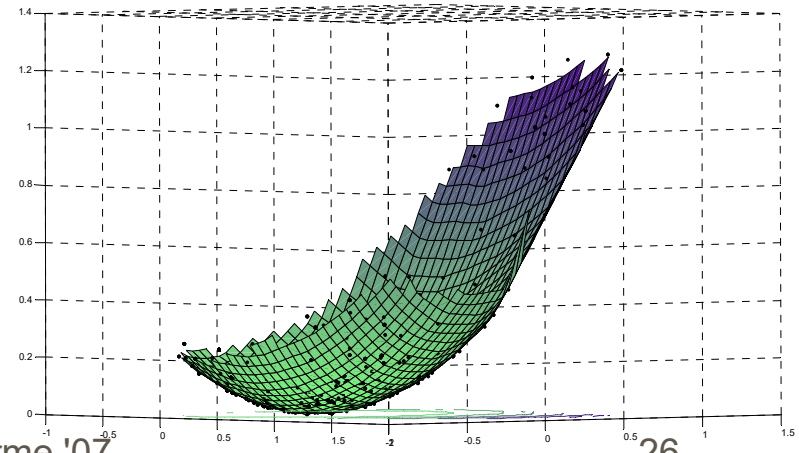


Internal Analysis: Contributions

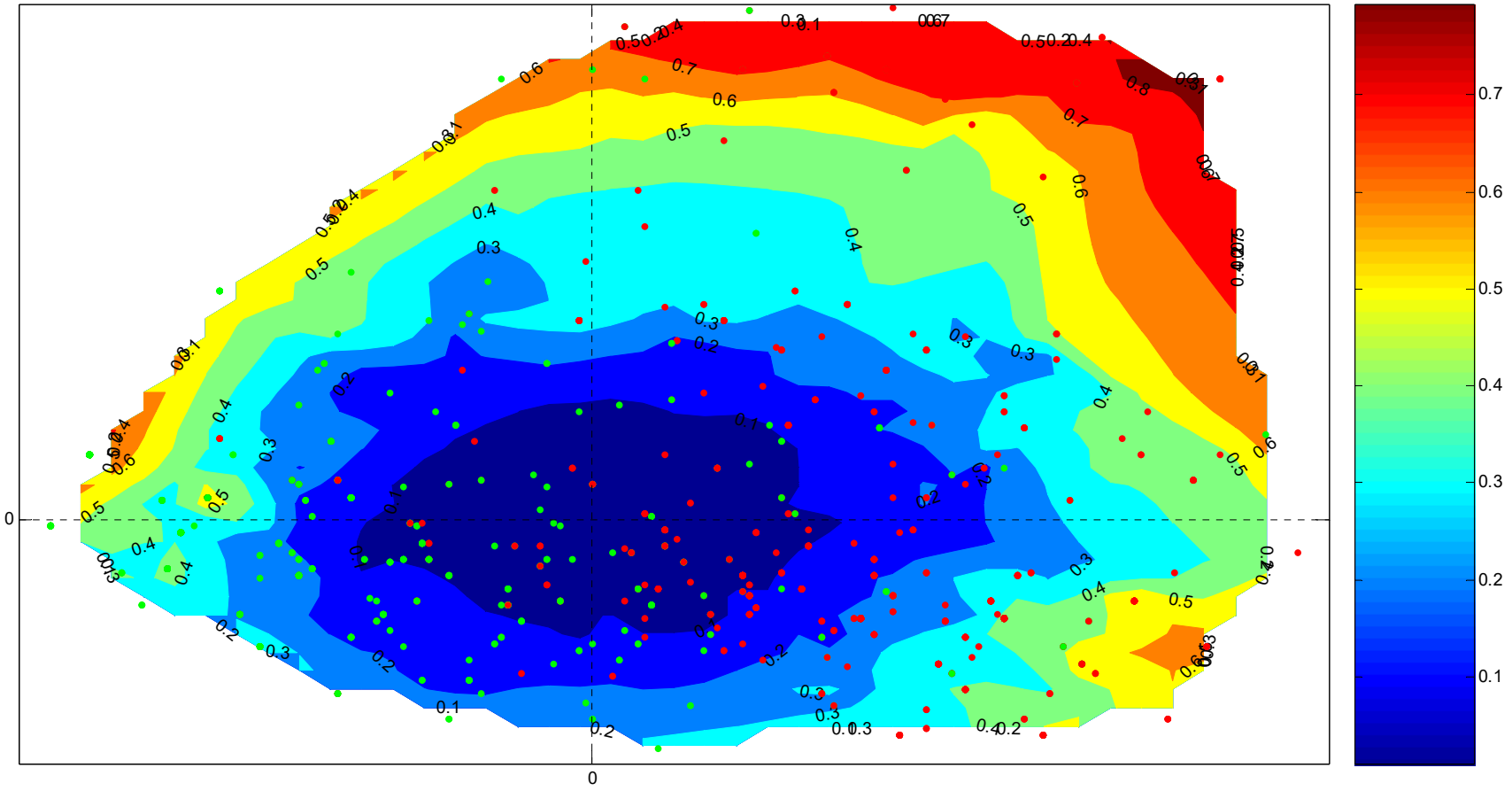




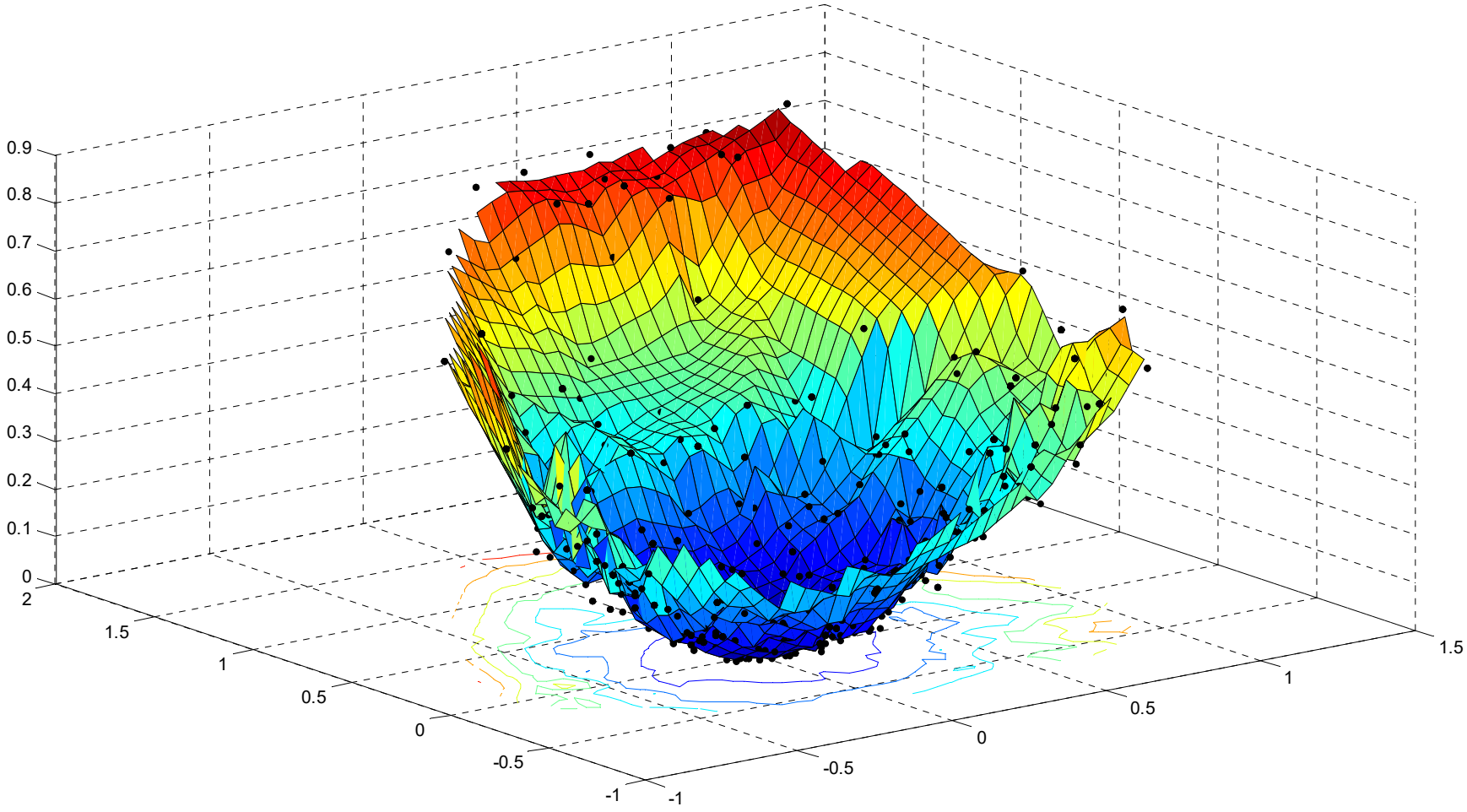




Internal Analysis: Squared Cosines



Internal Analysis: Squared Cosines



External Analysis: Discriminant Analysis on the MCA factors (Disqual)

- **Disqual** was proposed by Saporta (1976) and has been widely used (at least in France) for credit scoring (Bouroche and Saporta 1988)

  It consists of two main steps:

- A multiple correspondence analysis is performed on the array of predictors **Z** with the class variable as a supplementary one.
- A linear discriminant analysis by using factor coordinates as predictors.

DISQUAL: The Scoring

The principle

- The **SCORE** procedure is executed after the discriminant analysis with a two-categories response variable
- It computes the score function, a modification of the discriminant function, to facilitate its interpretation and use.
- Several decision "**zones**" are defined (red, green, middle) according to a given misclassification tolerance rate.

Interpreting aids

According to the "misclassification tolerance rate", different zones can be defined:

- - The "**GREEN**" zone, corresponding to the high scores region. In this region, the misclassified units get an incorrect high score.
- - The "**RED**" zone corresponding to the lower scores, the misclassified units get an incorrect low score.
- - The "**MIDDLE**" zone, where belonging to the two zones is unsettled. This zone can be reduced by increasing the misclassification tolerance rate.

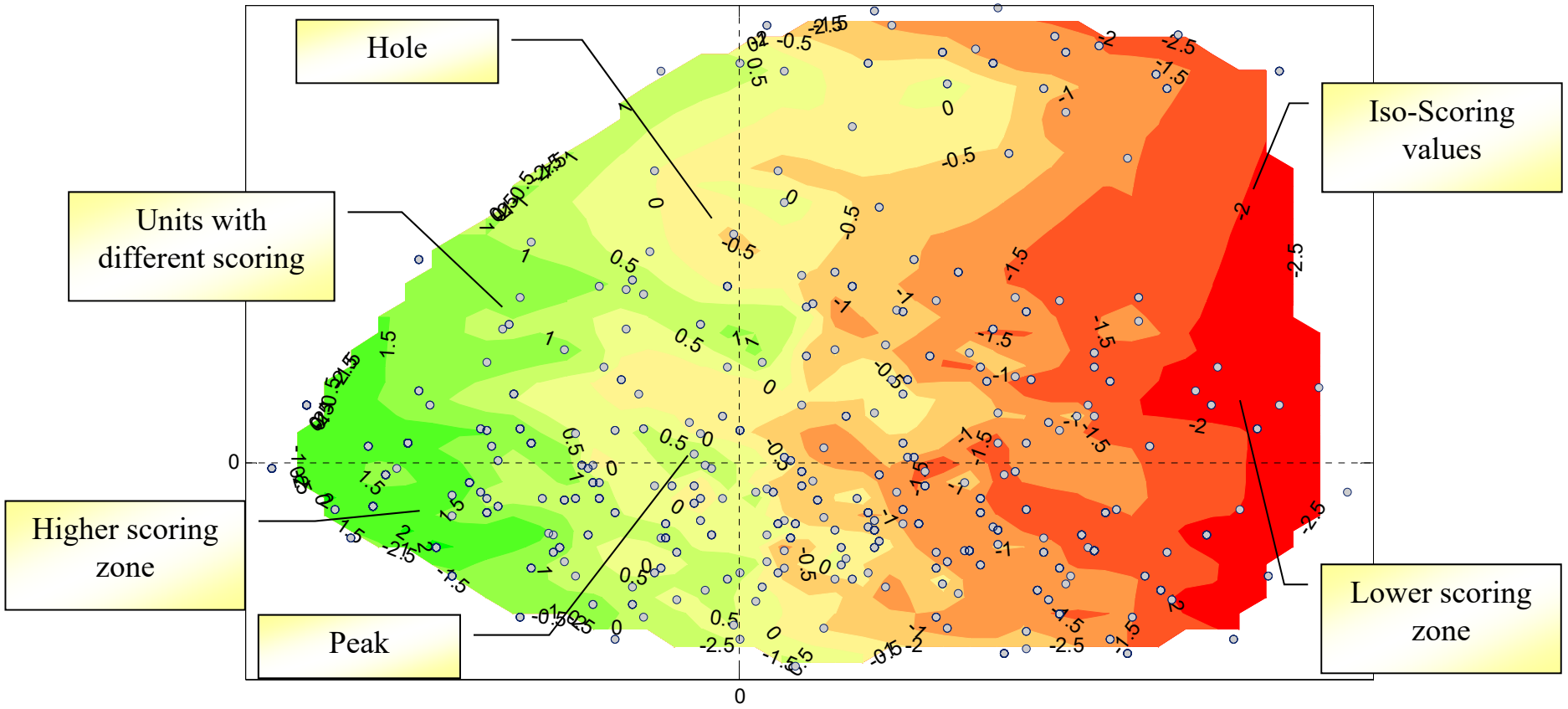
NB:

In this case-study, the high score group denotes the units with low number of accidents

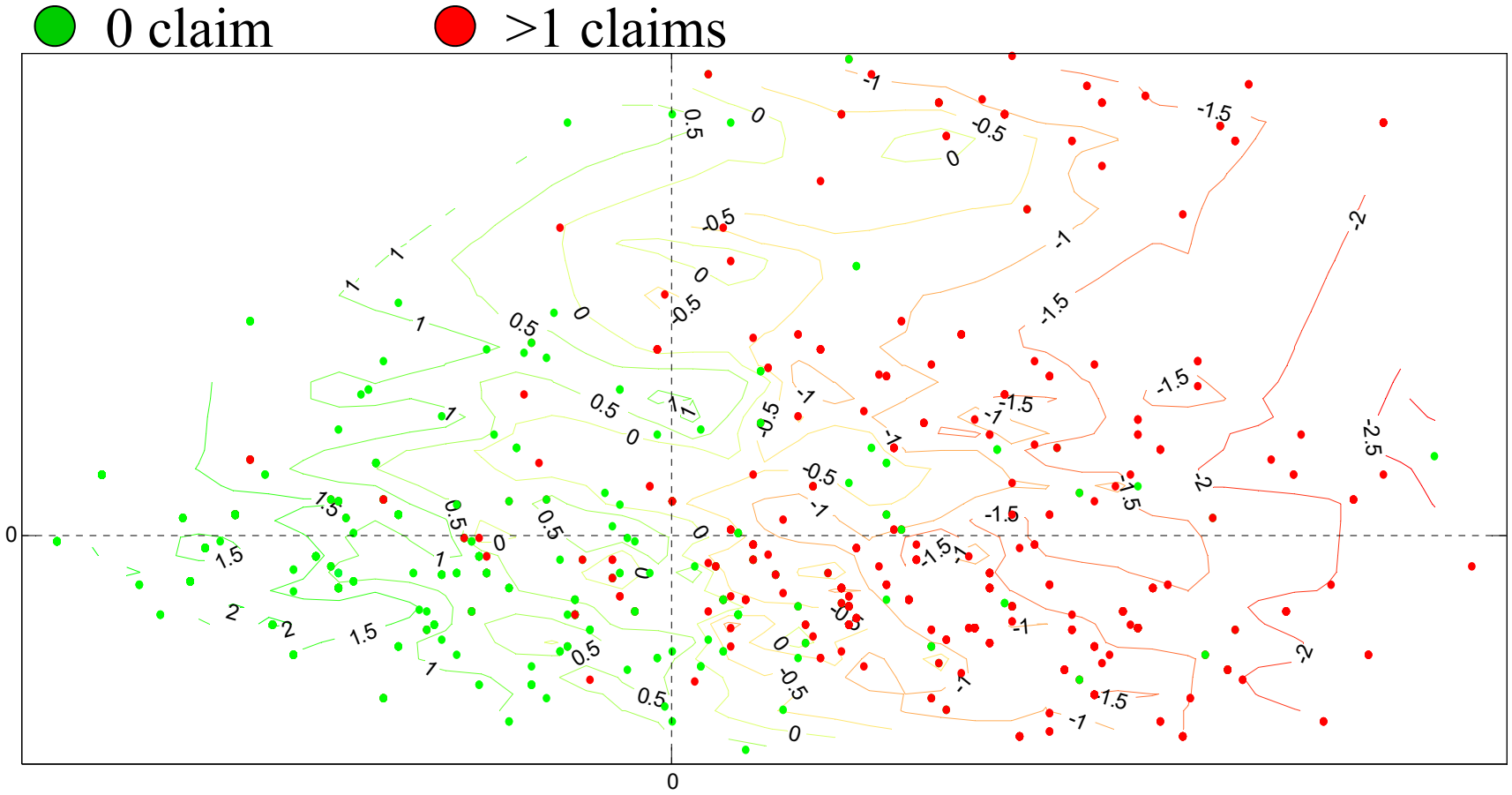
Choosing the parameters

```
function pcsurf(MCA1, MCA2, Score, dens=40);
tick1 = (max(MCA1)-min(MCA1)) / dens;
tick2 = (max(MCA2)-min(MCA2)) / dens;
t1 = (min(MCA1)-tick1) : tick1 : (max(MCA1)+tick1);
t2 = (min(MCA2)-tick1) : tick2 : (max(MCA2)+tick2);
[XI,YI] = meshgrid(t1,t2);
% Building the Grid
ZI = griddata(MCA1, MCA2, Score, XI, YI);
% 2D Contour Plot
contour(XI,YI,ZI), hold
plot(MCA1, MCA2, '.k'), hold off
% 3D Surface Plot
figure;
surfc(XI,YI,ZI), hold
plot3(MCA1, MCA2, Score, '.k'), hold off
```

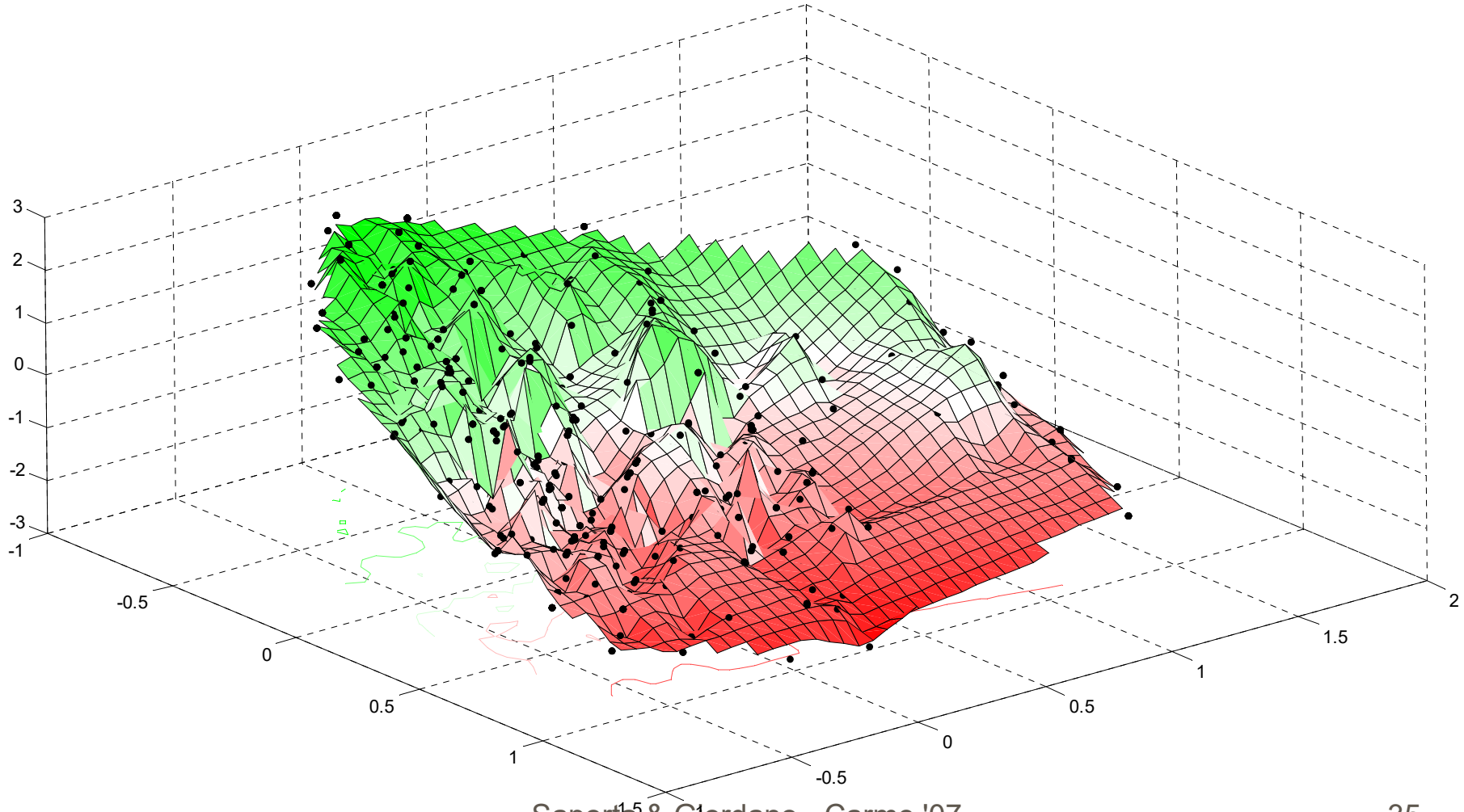
The Contour Plot



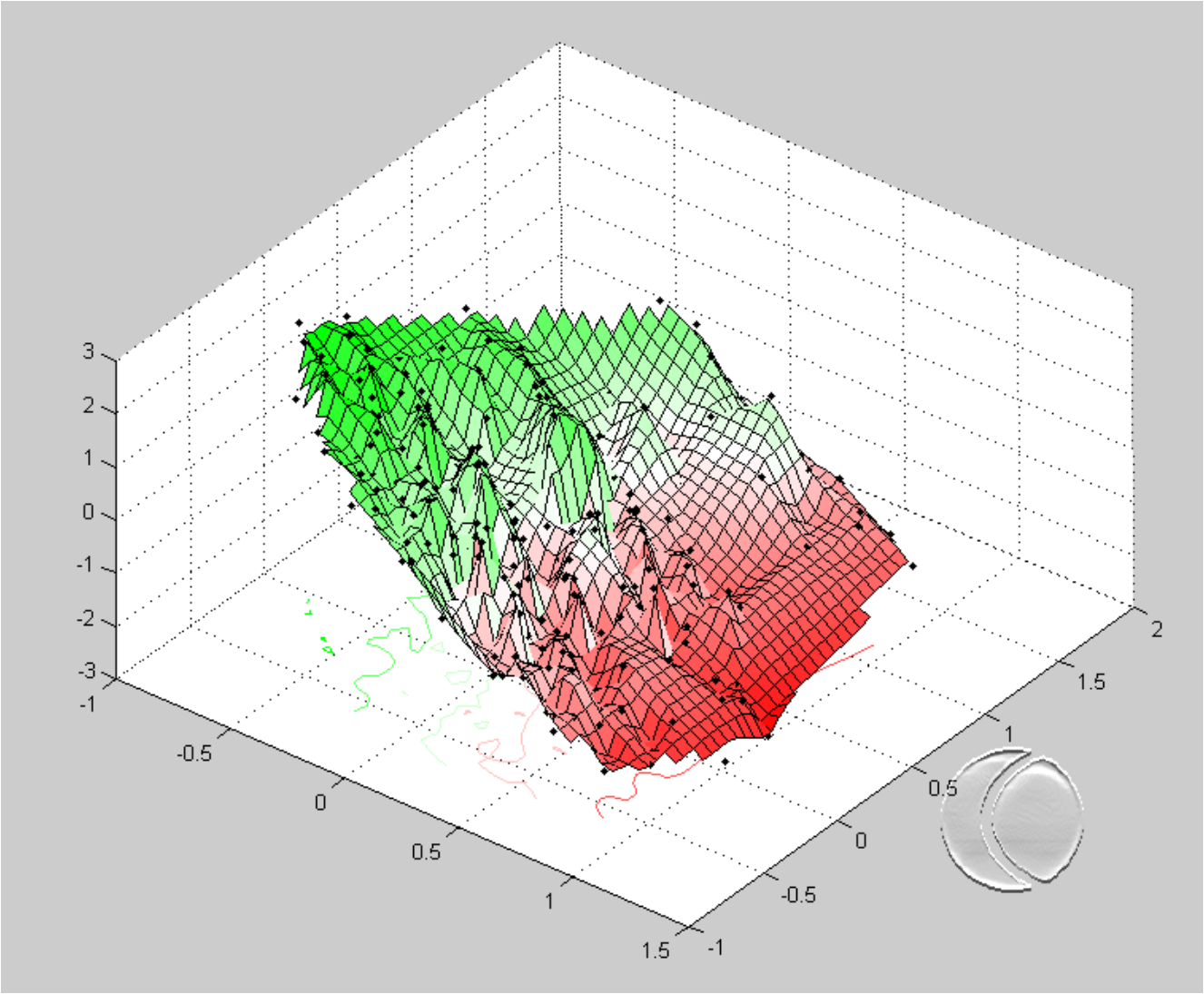
Illustrative variables: Accidents



The Scoring Response Surface

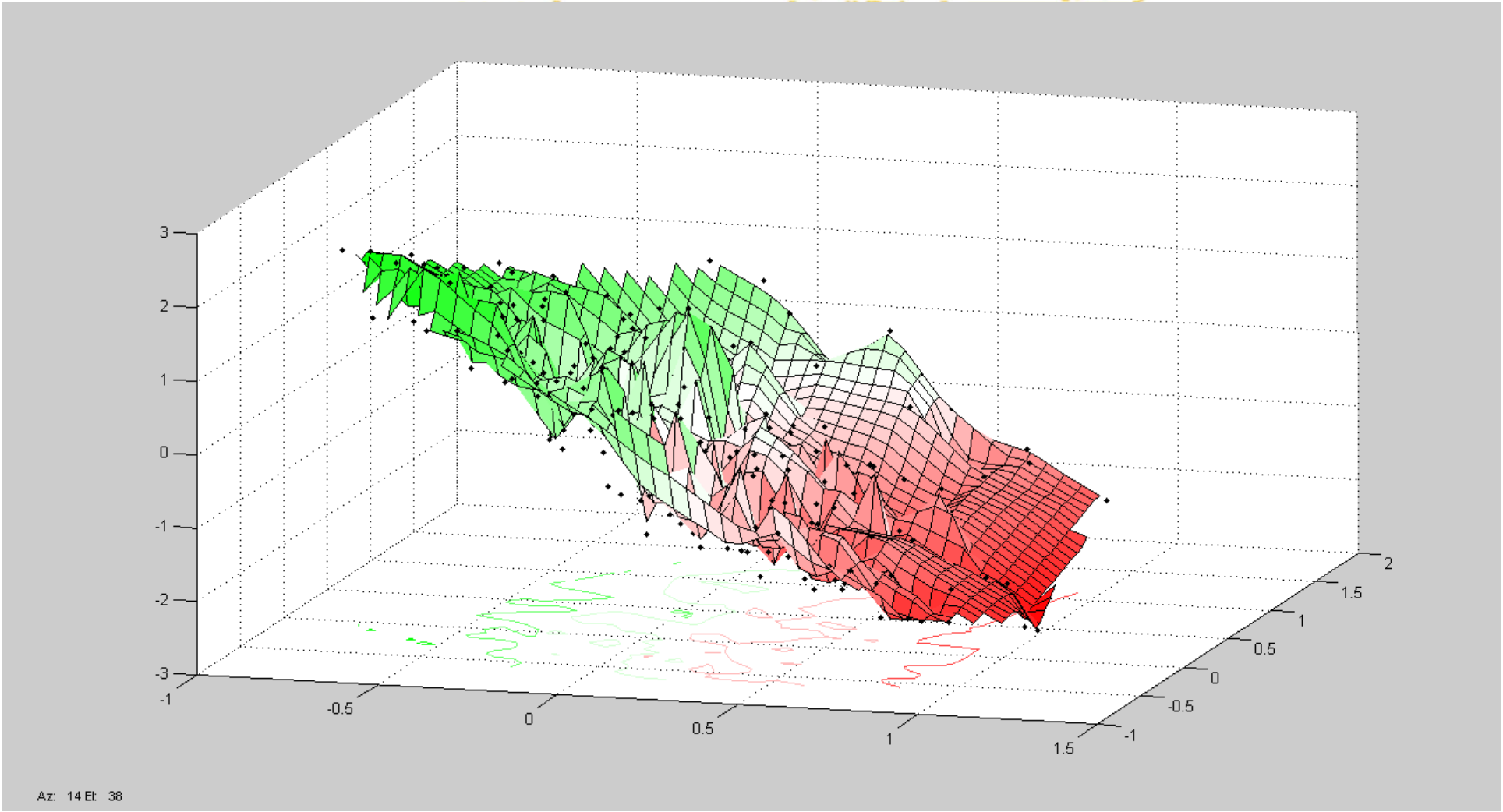


Controlling the Score Levels



The middle zone is affected by the value of the misclassification tolerance rate

Exploring the Surface



Concluding Remarks

- Even
  - Modeling vs/ Exploring
  - Ability to use any metric measurement as response
  - Apply to any kind of factorial techniques
- Odd
  - Exploring vs/ Modeling
  - Overfitting… Why not?
    Smoothing or not smoothing?

- *"…employez des procédures simple, ainsi vous pouvez les critiquer…"*
  Brigitte Escofier-Cordier cited by J. Pagès

Main References

Barber, C. B., Dobkin D.P., Huhdanpaa H.T. (1996), The Quickhull Algorithm for Convex Hulls, *ACM Transactions on Mathematical Software*, Vol. **22**, No. 4, Dec. 1996, 469-483.

Box, G. E. P, Draper, N. R. (1987), *Empirical model building and response surfaces,* New York: John Wiley & Sons.

Carr, D. B. (1998), *Multivariate graphics*, in Encyclopedia of Biostatistics, P. Armitage and T.Colton eds., 2864-2886, Chichester: Wiley & Sons.

Carroll, J.D. (1972), Individual Differences and Multidimensional Scaling, in R.N. Shepard, A.K. Romney, and S.B. Nerlove (eds.), *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences* (Volume 1), New York: Seminar Press.

Danzart M., Sieffermann J.M., Delarue J. (2004). New developments in preference mapping techniques: finding out a consumer optimal product, its sensory profile and the key sensory attributes. *7th Sensometrics Conference*, July 27-30, 2004, Davis, CA.

Giordano, G. (2006), A New Dimension in the Factorial Techniques: the Response Surface, *Statistica Applicata,* Vol. **18**, n.2, 361-375

Greenacre, M. J., Hastie, T. J. (1987), The geometric interpretation of correspondence analysis, *Journal of the American Statistical Association*, **82**, 437-447.

Khuri, A. I., Cornell, J. A. (1996), *Response surfaces: designs and analyses,* New York: Marcel Dekker, Inc.

Saporta, G., Niang, N. (2006), Correspondence analysis and classification, in J. Blasius & M. Greenacre (editors) *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall, 371-392.

Schlich P, McEwan J.A. (1992). Cartographie des Préférences. Un outil statistique pour l'industrie agro-alimentaire. *Sciences des aliments*, **12**, pp 339-355

Wold, H. (1982), Soft modelling: the basic design and some extensions, in: J.-K. Joreskog and H.Wold, eds., *Systems Under Indirect Observation*, **2**, 1–53, North Holland, Amsterdam.