Visual Data Mining : the case of VITAMIN System and other software

Alain MORINEAU – a.morineau@noos.fr

Data mining is an extension of Exploratory Data Analysis in the sense that both approaches have the same goal: the discovery of unknown structure in the data. The chief distinction resides in the size and dimensionality of the data sets, data mining involving much more massive data sets.

Never before data has been generated at such high volumes as it is today. **Visual data mining** is an approach to deal with this growing flood of data. The aim is to combine traditional data mining techniques with information visualization methods to utilize advantages of both approaches. The main advantage of visual data exploration is that the user is directly involved in the data mining process. The utilization of both automatic analysis methods and human perception promises more effective data exploration. Information visualization exploits the phenomenal abilities of human perception to identify structures by presenting data visually, allowing the user to explore the information space, to interact with the data and to draw conclusions.

The **VITAMIN System** (*Visual daTA MINing System*) is designed to help the user in the analysis process of large survey data and time series data. The software has been developed by ATKOSoft SA with the contribution of UNINA-DMS and ONS, in the framework of the Information Societies Technology (IST) granted by the European Commission.

This paper has been written in the context of the validation activities for the VITAMIN Software in the framework of the IST project. The first part is a quick description of the main functionalities in VITAMIN System, trying to stress the originality of this new approach. In the second part, we give a few examples of other software, the contents of which differ because either they address other kind of data, or they implement other kind of visualisation techniques.

PART 1
VitaminS Project

VITAMIN S approach

VITAMIN S is a software for statistical visualisation, combining existing and new graphical methods. Usually when using classical statistical methods based on large computational efforts, it is not easy to extract significant knowledge from the data. Their output often range from very synthetic and sometimes cryptic results, up to thousands of numerical tables.

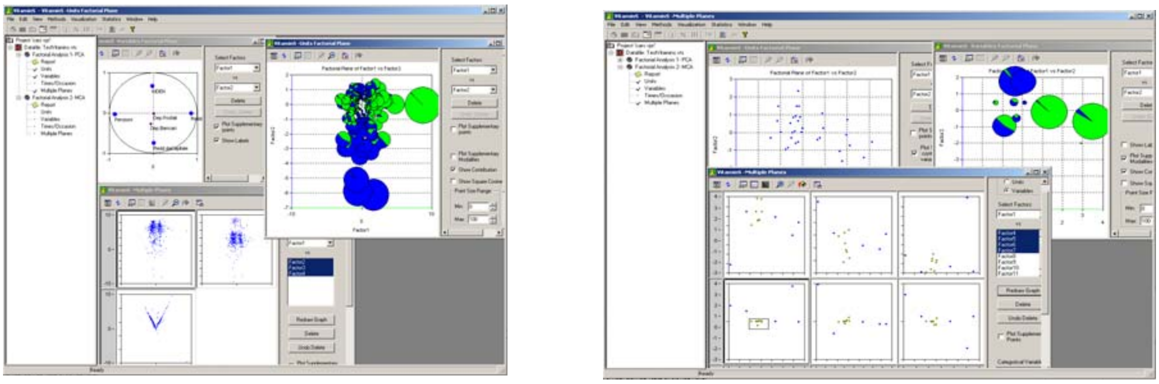
The main goal of the system is to perform visual data mining on large data sets according to the needs of National Statistical Institutes. The system provides the user with:

- classical and innovative visualisation tools;
- an interactive framework to explore multivariate data;
- linked displays, so that several visualisation methods can be used at the same time during the analysis process;
- graphical representation linked with numerical analysis where it is possible;
- combined with interactive result reports.

In data mining, the most useful information is frequently difficult to identify by the mean of classical statistical methods. The innovative idea introduced in VITAMIN S consists in giving the users the capability to define the analysis parameters (variable, individual, options, etc.) by pointing directly on the graphical representation.

Visualisation methods contribute to a more effective data analysis process: looking directly on the visualised data gives a better understanding of the data than looking on several (large) numerical tables.

Starting from an initial graphical representation, users with no special skill in statistics are able to mine their data and look for interesting patterns. The system drives users towards the most appropriate analysis procedures with respect to the data and the software is expected to prevent developing improper analysis.



VITAMIN S visualisation examples

The visualisation methods selected for the VITAMIN S software are mainly related to the following statistical methods:

- Principal Component Analysis (PCA)
- Multiple Correspondence Analysis (MCA)
- Non Hierarchical Cluster Analysis
- Hierarchical Cluster Analysis

The selection of the above methods derived from the needs of the end users and from the potentiality of these methods in analysing and visualizing huge amount of data. Some other methods seem to have been explored for successive developments of the VITAMIN S project. Concerning the time series datasets, the system mainly requires the use of the following statistical methods: ARMA Model and X12-ARIMA combined with Principal Components Analysis.

Based on the above-mentioned methods the following visualisation methods were developed:

- Factorial methods for cases and variable visualisation
- Multiple factorial planes visualisation
- Visualisation of a factorial plane along times/occasions
- Clusters exploration and visualisation on factorial planes
- Interactive dendrogram graphics for clusters exploration
- Visual outliers identification in time series
- Visualisation of the best aggregate time series
- Visual aids for ARMA model identification

Exploring ARMA models with multidimensional data analysis, VITAMIN Software demonstrates innovative ways to use visualisation methods for:

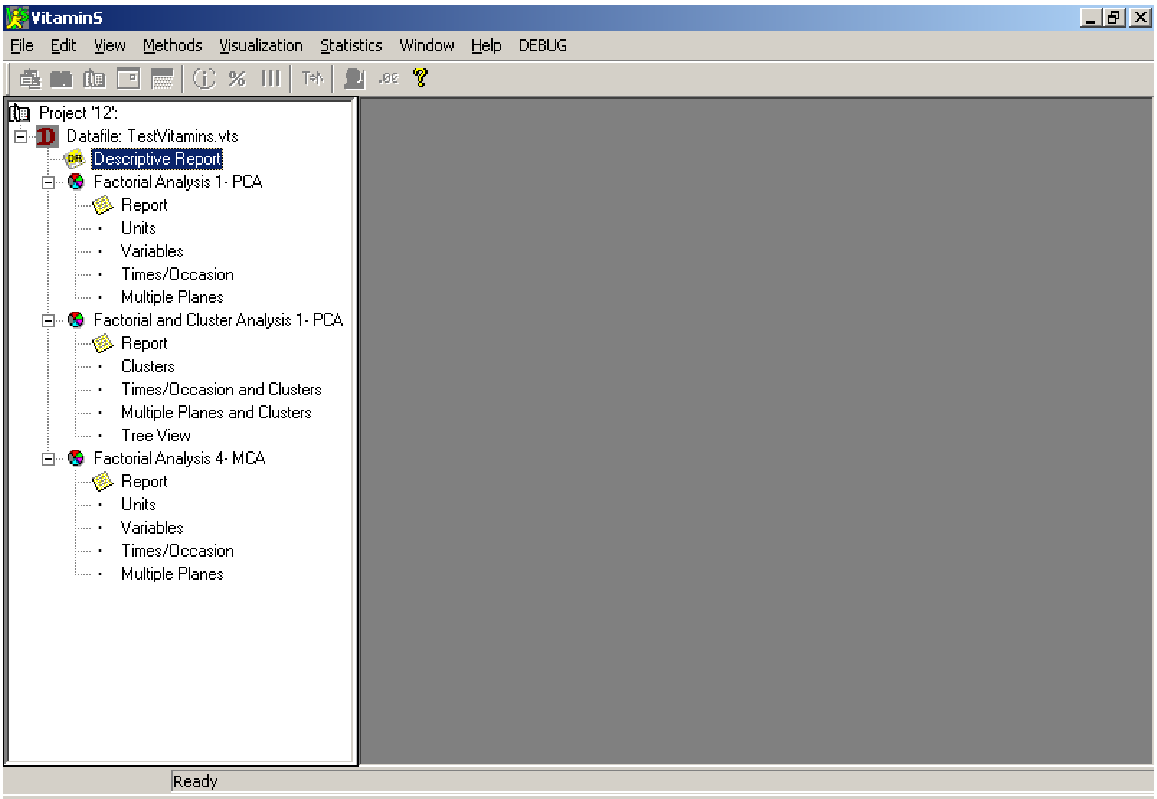
- Adapting statistical concepts to the dynamic global environment
- Processing and analysis of available statistical data using advance and state-of-the-art techniques
- Improved accessibility and user-friendliness in statistics

Main structure of the VITAMIN S software

The main window of the VITAMIN S software is separated into two parts. The left part displays any active project in a tree structure (the “project window”) and the right is the area where result windows appear.

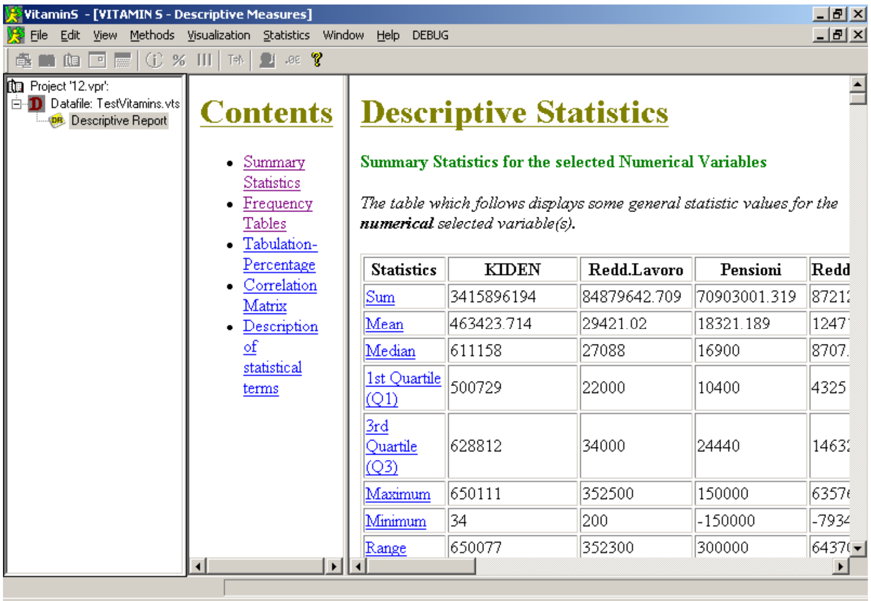
The user selects an analysis method by selecting the **Methods** menu command (or by right clicking the data file name on the **Project tree** and selecting the appropriate method from a pop up menu). The name of the selected method is added to the **Project tree,** below the data file node. Then nodes are added for the **Report** and the several **visualizations** can be associated with the analysis method.

When the user has selected an analysis method, the system prompts the user to define the data subset to be analysed. A wizard is displayed through which the user can define the criteria for extracting a subset from the original dataset and define the parameters specific to the method.


VitaminS
File Edit View Methods Visualization Statistics Window Help DEBUG
Project '12':
Datafile: TestVitamins.vts
Descriptive Report
Factorial Analysis 1 - PCA
Report
Units
Variables
Times/Occasion
Multiple Planes
Factorial and Cluster Analysis 1 - PCA
Report
Clusters
Times/Occasion and Clusters
Multiple Planes and Clusters
Tree View
Factorial Analysis 4 - MCA
Report
Units
Variables
Times/Occasion
Multiple Planes
Ready

VITAMIN S main window

The first report that can be generated using the input data is a **Descriptive Report.**


VitaminS - [VITAMIN S - Descriptive Measures]
File Edit View Methods Visualization Statistics Window Help DEBUG
Project '12.vpr':
Datafile: TestVitamins.vts
Descriptive Report

Contents

- Summary Statistics
- Frequency Tables
- Tabulation-Percentage
- Correlation Matrix
- Description of statistical terms

Descriptive Statistics

Summary Statistics for the selected Numerical Variables

The table which follows displays some general statistic values for the ***numerical*** *selected variable(s).*

Statistics	KIDEN	Redd.Lavoro	Pensioni	Redd
Sum	3415896194	84879642.709	70903001.319	8721
Mean	463423.714	29421.02	18321.189	1247
Median	611158	27088	16900	8707.
1st Quartile (Q1)	500729	22000	10400	4325
3rd Quartile (Q3)	628812	34000	24440	1463
Maximum	650111	352500	150000	6357
Minimum	34	200	-150000	-7934
Range	650077	352300	300000	6437

Descriptive Report

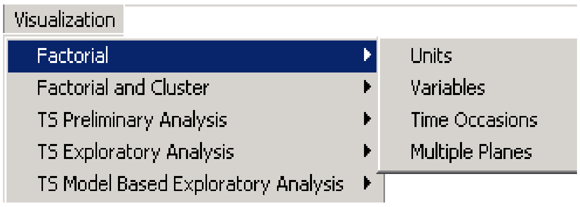
The *Descriptive Report* includes many useful functionalities and calculations. The Summary statistics provides a set of several functions, and a section with the description of the statistical terms is provided at the end of the report:

- Summary Statistics : Sum, Mean, Median, 1st and 3rd Quartile, Maximum, Minimum, Range, Standard Deviation, Variance, Interquartile Range, outliers, Extreme Values
- Frequency Tables
- Cross tables with raw frequencies and percentages
- Correlation Matrix

The user is able to view the HTML report either in his Web Browser or it is loaded in the VITAMIN-S right window for user's convenience.

Many Visualizations of Data Analysis

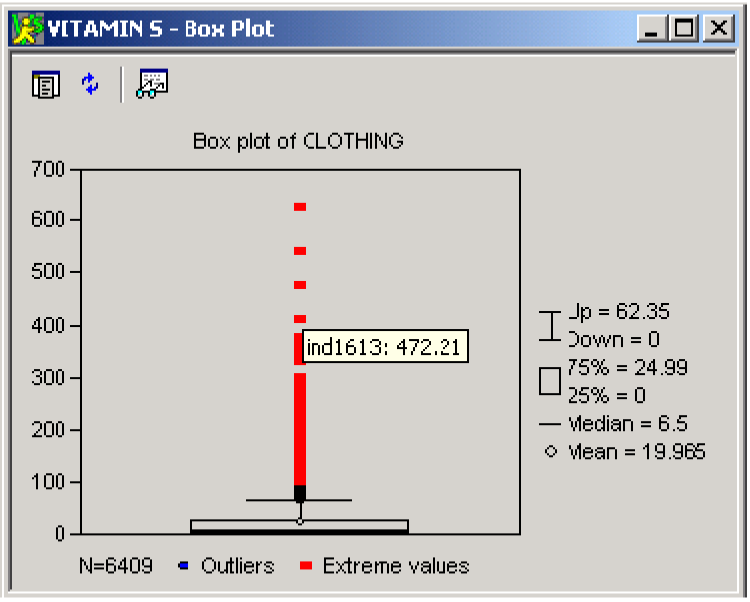
A number of statistical methods are implemented as ActiveX controls. These controls can work with VITAMIN S as well as with any other ActiveX container application such as Excel, Internet Explorer, etc. The user selects one or more graphs that are associated with the current method by selecting the Visualization menu command and then gets the graphs that are associated with the method. This opens a new view containing the visualization graphs and some controls for user interaction.



Visualization menu command (example)

Box Plot

Box plot is a descriptive graph that summarises the information contained in a variable. Some specific measures of location and variability are depicted, helping the user to obtain an image of the variable distribution. The box plot can give information about the distribution of the data, especially with respect to the asymmetry of this distribution. Pointing with the mouse on the outliers or the extreme values, a small window appears showing the respective value of the current point.



Box Plot Outliers

Bar Chart, Pareto Chart, Histogram and Pie Chart

The software provides the user with all usual statistical graphs such as bar chart, Pareto chart, histogram and pie chart, each one with many options and useful ways to customize the display. We can see below the control properties of a Bar Chart.

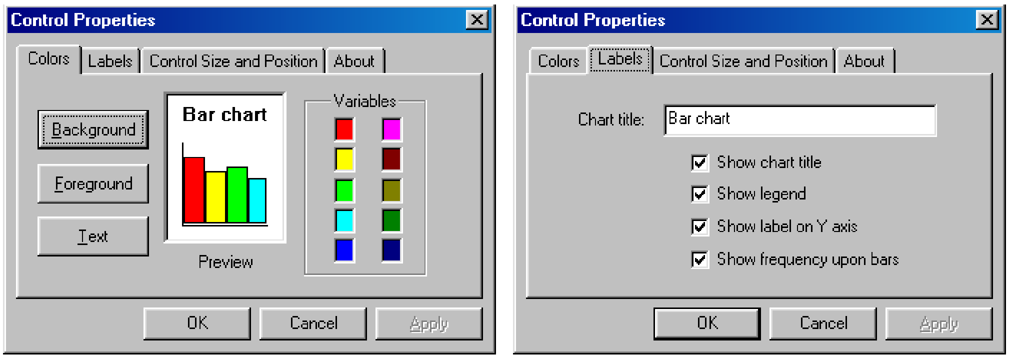


Chart title: Bar chart

☑ Show chart title

☑ Show legend

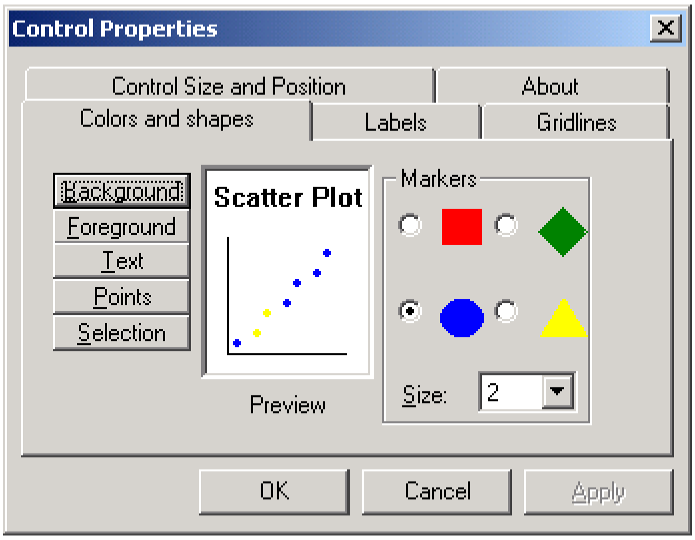
☑ Show label on Y axis

☑ Show frequency upon bars

Control properties of a Bar Chart

Scatter Plot and Scatter Plot Matrix

The determination of the relationship between two variables is achieved through a *scatter plot*. A scatter plot is useful in exploring the relationship between two variables. The user is able to view the graph by selecting **Statistics\Descriptive Graphs\Scatter Plot** menu command. In scatter plot method, the following dialog appears:

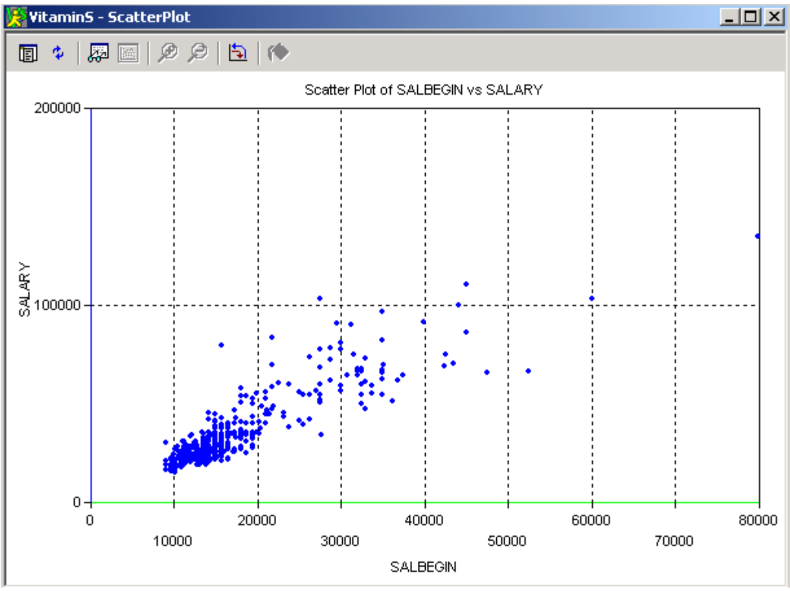


Size: 2

Scatter plot control properties

There are many interactive functionalities:

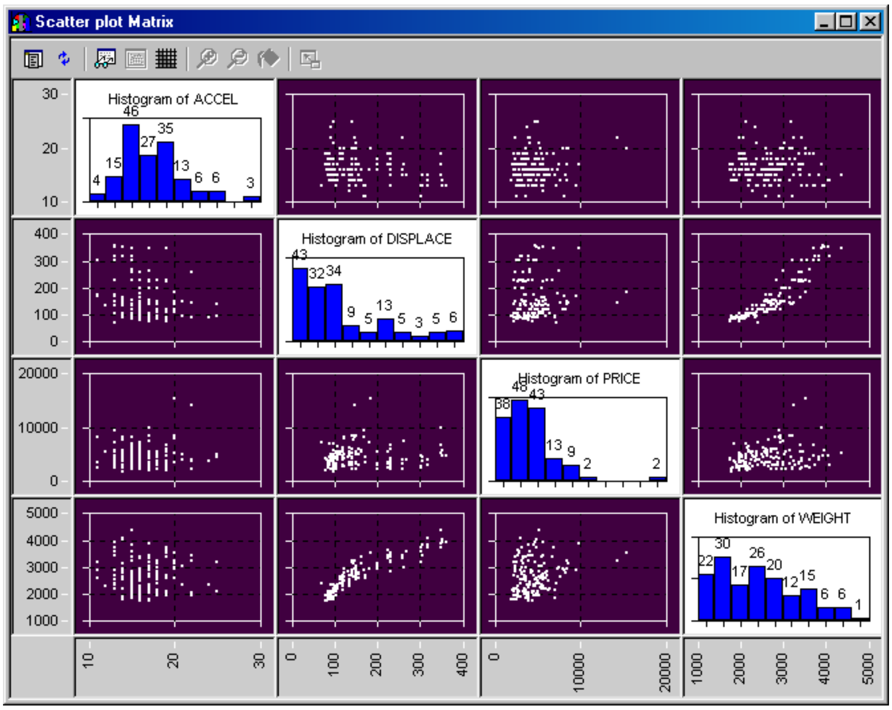
- select a region on the graph containing data points: All data points inside the region are colored with a different colour.
- Paint the selected points in a region by clicking on a toolbar button.
- Select some points and then zoom (unzoom) on them using a toolbar button.
- Change the axis (make the X axis Y and vice versa).
- View the data values that correspond to the selected points.



Scatter Plot

Scatter plot matrix consists of several scatter plots and histograms. The diagonal of the table depicts the histograms of the selected variables. Each scatter plot in the matrix is defined by two variables. Y-axis depicts the variable that appears at the same row of the matrix and X-axis depicts the variable that appears at the same column of the matrix. Scatter plot matrix provides the user with graphs of the relationship between variables.

A pop-up menu opens the **Control Properties** menu command and the user can set several properties associated with the appearance of the scatter plot matrix window.



Scatter Plot Matrix

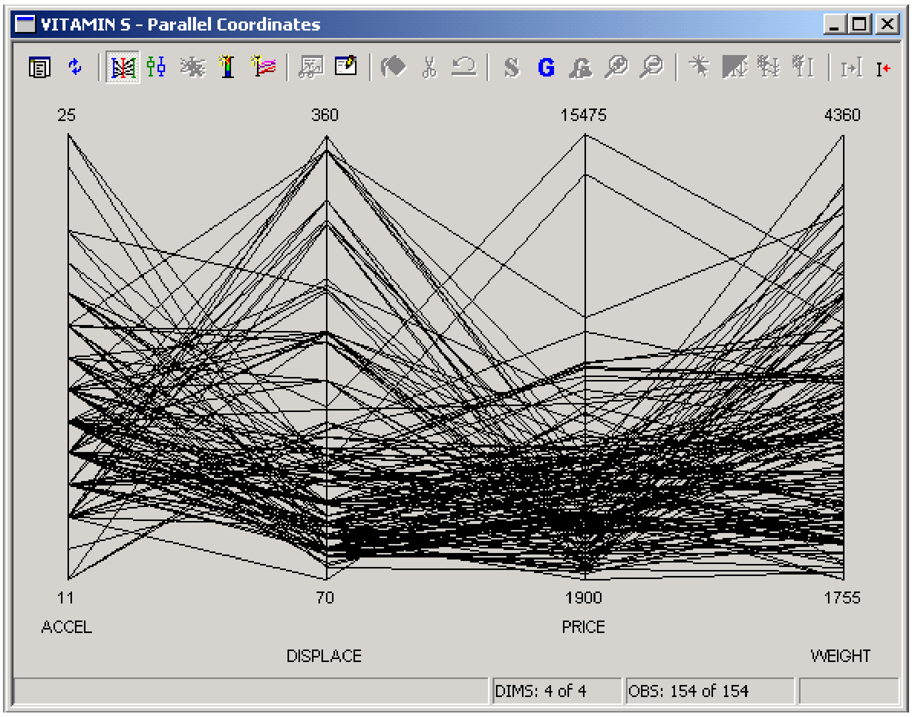
Parallel Coordinates

The user has several ways available for manipulating the data set and its corresponding visualised objects (*dimensions*, *observation*, and *data values* on the graph), as well as a number of operations that can be applied in order to start identifying trends, patterns, and correlations between variables and sets of variables.

An *observation* is a tuple of data values, represented on the parallel coordinates display as a broken line. The user can select it by clicking it with the left mouse button. A *dimension* is represented on the display as a vertical straight line, which spans from the top to the bottom of the visualisation window. The user can select it by clicking with the left mouse button. *Data values* are represented by the vertices of the broken lines of the visualisation. i.e., they are the points where the *dimensions* intersect with the *observations*.

A *region* is an arbitrary user selection on the screen. It may contain any of the above object types. To select a particular region on the screen one has to drag the mouse while the left button is pressed. *Sliders* are a set of observations passing through certain data values range on a dimension. It may be moved across the dimension. It is a visual way to see how data values are changing across the data set variables.

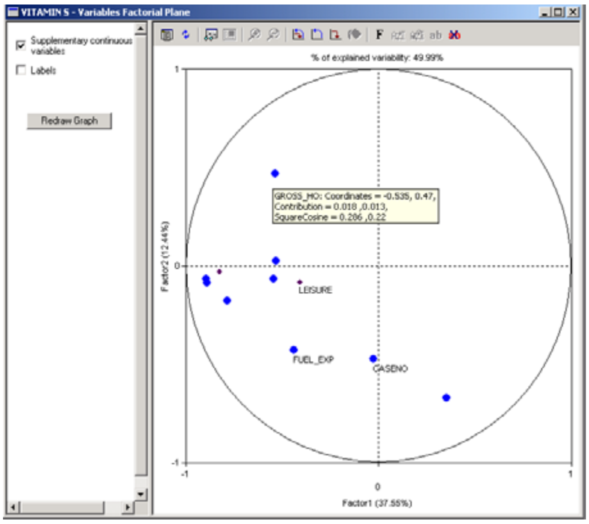
Many operations for the Parallel Coordinates visualization method are available from a tool bar that is available in the parallel coordinates visualisation window. Among them: Selection, Move, Hide/Unhide, Show info, Show only, Rearrange, Zoom, Same scale, Restore scale, Paint, Cut, Summarize, Group, Undo, etc.



Parallel Coordinates

Variables and cases on factorial planes

The factorial plane for any factorial analysis allows the visualization of the variables (columns of the dataset) and the cases (rows) using two-dimensional graphs determined by the factors that best represent the inertia of the cloud of points. This kind of graphical representation can be built using the results of the Principal Correspondence Analysis (PCA) and Multiple Correspondence Analysis (MCA).

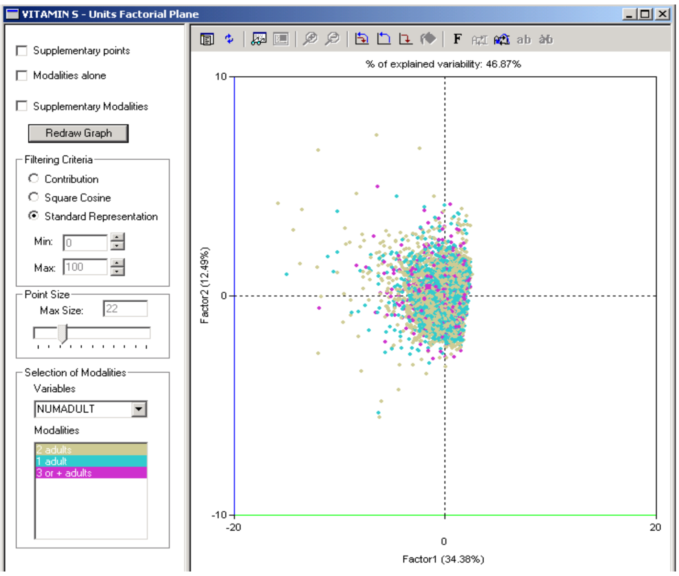

VITAMIN S - Variables Factorial Plane
☑ Supplementary continuous variables
□ Labels
Redraw Graph
% of explained variability: 49.99%
1
0
-1
-1
0
1
Factor2 (12.44%)
Factor1 (37.55%)
LEISURE
FUEL_EXP
CASENO

GROSS_HO: Coordinates = -0.535, 0.47,
Contribution = 0.018 ,0.013,
SquareCosine = 0.286 ,0.22

Information for a specific variable point in PCA

By pointing the mouse cursor on any of the drawn points, an *info note* appears providing information relevant to that point. Generally the same principles regarding the functionalities apply to any graph by using the menu bar located at the top.

The representation of the cases allows the visualization of the observed units on a plane using the factors that best represent the variability of the observations. This representation also allows the visualization of the modalities of categorical variables.



VITAMIN S - Units Factorial Plane

□ Supplementary points

□ Modalities alone

□ Supplementary Modalities

Filtering Criteria

○ Contribution

○ Square Cosine

● Standard Representation

Min: 0

Max: 100

Point Size

Max Size: 22

Selection of Modalities

Variables

NUMADULT

Modalities

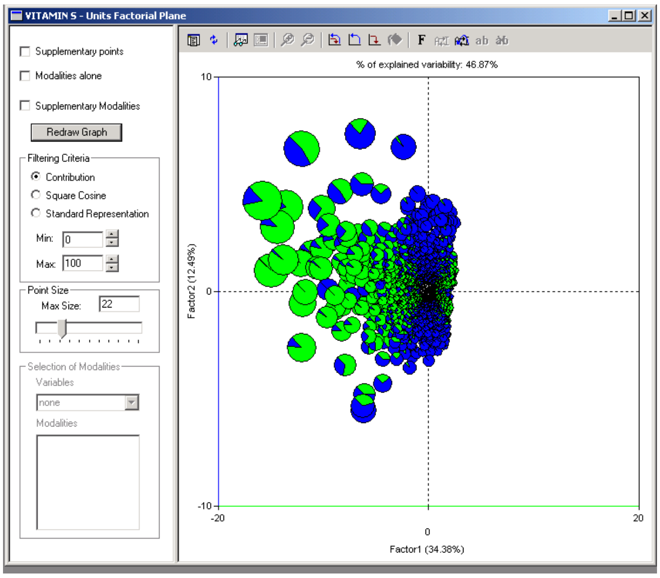
2 adults

1 adult

3 or + adults

Cases with different colors according to modalities of a categorical variable

It is also possible to visualize properties of points on the factorial plans such as contribution or squared cosines as shown on the figure:



VITAMIN S - Units Factorial Plane

□ Supplementary points

□ Modalities alone

□ Supplementary Modalities

Filtering Criteria

● Contribution

○ Square Cosine

○ Standard Representation

Min: 0

Max: 100

Point Size

Max Size: 22

Selection of Modalities

Variables

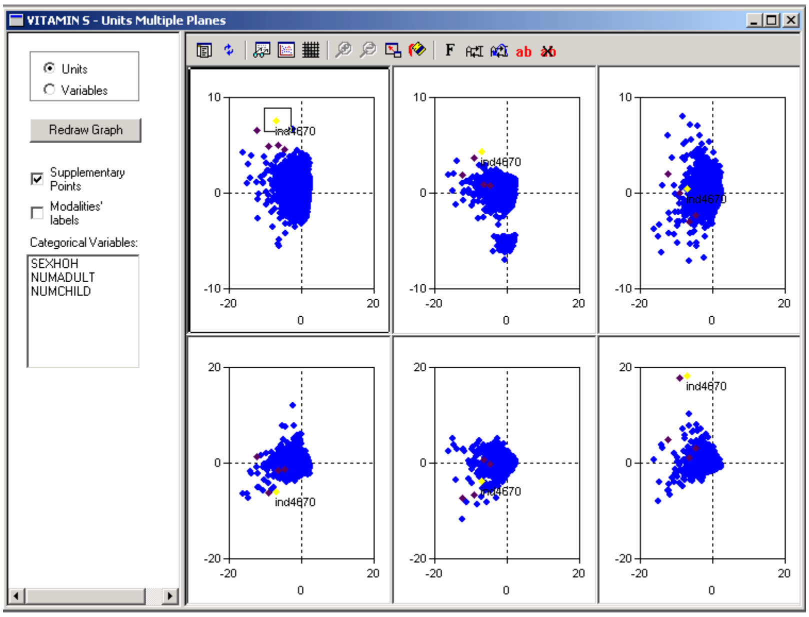
none

Modalities

Units visualization according to Contribution

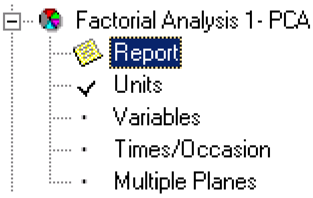
The user can choose a matrix-form representation using a factor on an axis and two or more factors on the other axes to simultaneously visualize the several factorial planes. On the basis of the number of selected factors, the system plots the coordinates of the items into multiple planes view.

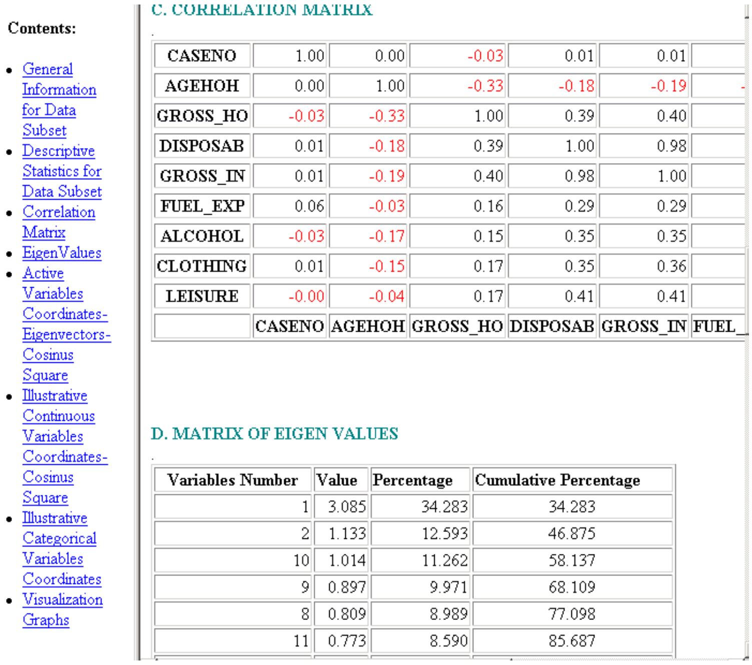
When the user selects points on one factorial plane then the corresponding selected points are simultaneously highlighted on the others factorials planes. Furthermore, when pointing with the mouse on a point of the graph, an info note appears with information related to the selected point.



Multiple Planes with selection

The user is provided with a report on the results of analysis where he can decide the contents of his report:





Contents:

- General Information for Data Subset
- Descriptive Statistics for Data Subset
- Correlation Matrix
- EigenValues
- Active Variables Coordinates-Eigenvectors-Cosinus Square
- Illustrative Continuous Variables Coordinates-Cosinus Square
- Illustrative Categorical Variables Coordinates
- Visualization Graphs

C. CORRELATION MATRIX

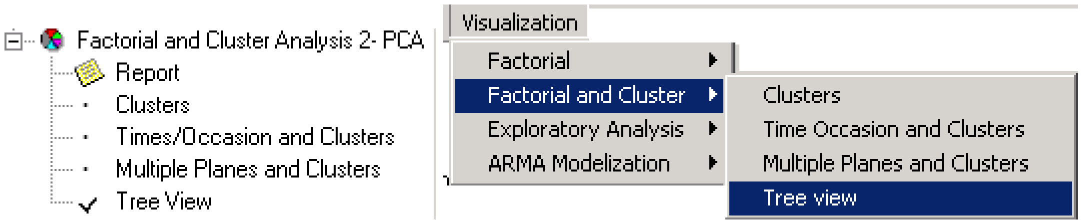
CASENO	1.00	0.00	-0.03	0.01	0.01	
AGEHOH	0.00	1.00	-0.33	-0.18	-0.19	-
GROSS_HO	-0.03	-0.33	1.00	0.39	0.40	
DISPOSAB	0.01	-0.18	0.39	1.00	0.98	
GROSS_IN	0.01	-0.19	0.40	0.98	1.00	
FUEL_EXP	0.06	-0.03	0.16	0.29	0.29	
ALCOHOL	-0.03	-0.17	0.15	0.35	0.35	
CLOTHING	0.01	-0.15	0.17	0.35	0.36	
LEISURE	-0.00	-0.04	0.17	0.41	0.41	
	CASENO	AGEHOH	GROSS_HO	DISPOSAB	GROSS_IN	FUEL_

D. MATRIX OF EIGEN VALUES

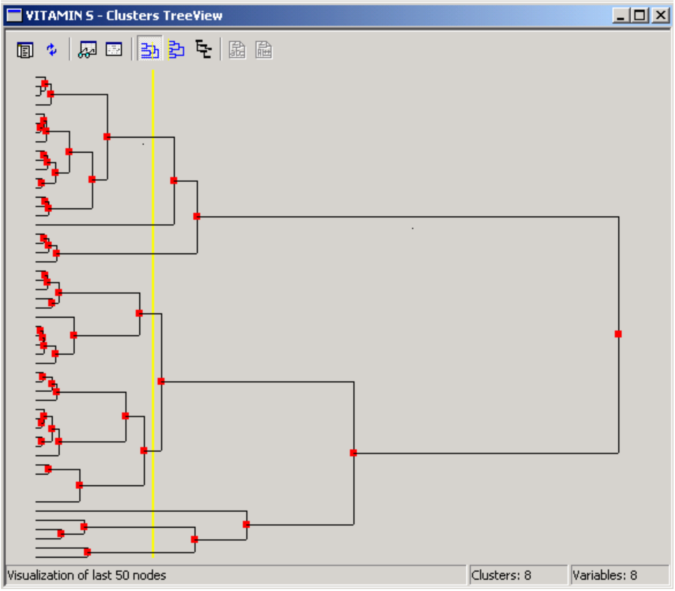
Variables Number	Value	Percentage	Cumulative Percentage
1	3.085	34.283	34.283
2	1.133	12.593	46.875
10	1.014	11.262	58.137
9	0.897	9.971	68.109
8	0.809	8.989	77.098
11	0.773	8.590	85.687

Factorial and Cluster Analysis – Tree View

The results of Cluster Analysis are used for graphical representation. These results are shown in a typical dendrogram representation or in a tree-view representation. The user is able to view the dendrogram representation by double clicking on the node “Tree View” of the project tree.


Factorial and Cluster Analysis 2- PCA
Report
Clusters
Times/Occasion and Clusters
Multiple Planes and Clusters
Tree View
Visualization
Factorial
Factorial and Cluster
Exploratory Analysis
ARMA Modelization
Clusters
Time Occasion and Clusters
Multiple Planes and Clusters
Tree view

Tree View graph selection



Visualization of last 50 nodes
Clusters: 8
Variables: 8

Cutting the dendrogram for a partition

The user can interact with the representations using some tools/controls on the graph window. Such interactions are for example:

- In the tree-view representations, the interactions are directly embedded: the user can collapse (expand) a particular cluster to hide (show) the items included in that cluster.
- In the dendrogram, a menu is displayed by right clicking on the red rectangles. When the user selects the menu command **Parallel Coordinates** he can view a parallel coordinates representation for the units that belong to the selected cluster.

The user is able to get a report on the analysis by clicking on the node “Report” of the project tree, related to the current method. In case of factorial method executed before any cluster analysis, then the report also includes the information about the preliminary analysis. The report contains information about the cluster analysis, e.g. the results of K-means method, the description of the dendrogram (Hierarchical clustering), the test values as well as the graphs of the current method.

Time Series analysis

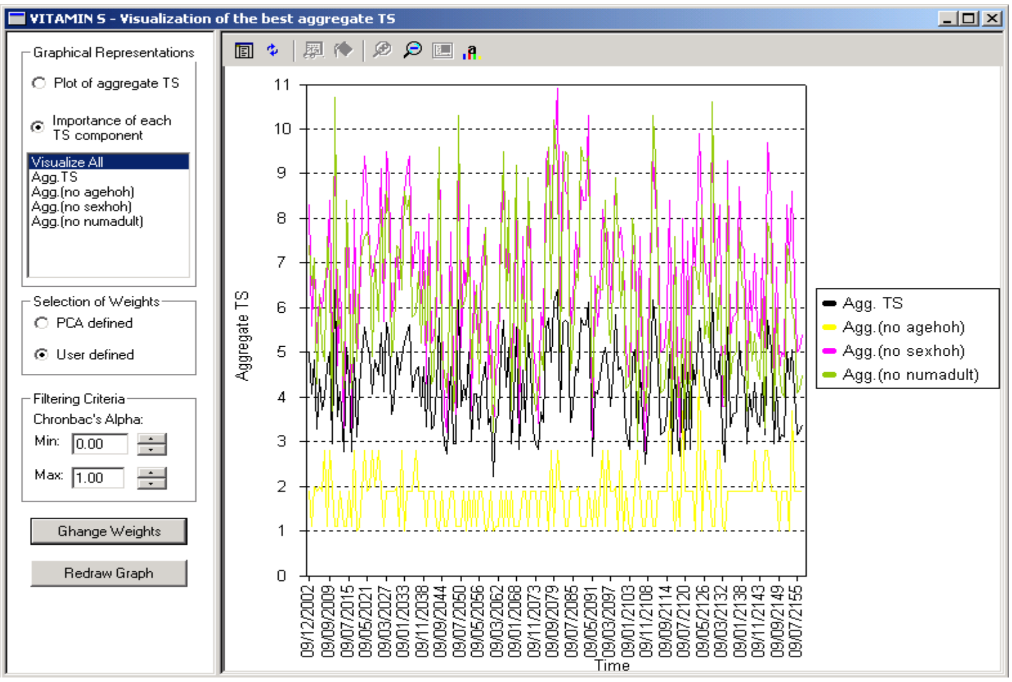
This part of the software is an innovative part compared to traditional Data mining software as we will see in the following. We will give only an quick overview of the methods implemented in VITAMIN System for time series analysis.

Preliminary analysis: Transformed Series

In the case of Time Series dataset, preliminary analysis is applied in order to identify the proper transformations that are to be used on time series in order to identify its components. These transformations are mainly based on differentiation on one hand, and seasonal adjustment on the other hand. In case of seasonal adjustment, we can choose among additive and multiplicative models. Bar charts visualize ratio between the variance of the original time series and of the transformed time series.

Time Series Exploratory Analysis: Best Aggregate and other methods

Result of Principal Component Analysis is used to find a compromise aggregate series based on multiple time series. The coherence and reliability of the aggregate series with the component series is evaluated through Cronbach's α indices. One can plot all the aggregate time series leaving out one component at a time. Visualization is made using the first eigenvector computed in the Principal Components Analysis as weights, or using alternatively weights defined by the user.



VITAMIN S - Visualization of the best aggregate TS

Graphical Representations
○ Plot of aggregate TS
● Importance of each TS component

Visualize All
Agg.TS
Agg.(no agehoh)
Agg.(no sexhoh)
Agg.(no numadult)

Selection of Weights
○ PCA defined
● User defined

Filtering Criteria
Chronbac's Alpha:
Min: 0.00
Max: 1.00

Visualization of the importance of each component time series

Many other operations are possible on time series datasets in VITAMIN System, such that for example **outliers** identification or **clustering** techniques, ARMA models with **multidimensional data** analysis or **dynamic** PCA.

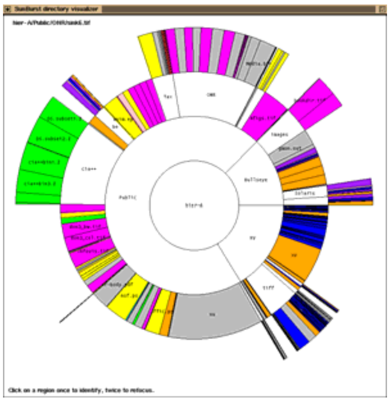
PART 2
Other Visual Data Mining Systems

To explore and analyse large amounts of data in order to extract meaning is the field of *Data Mining* utilizing statistical methods (clustering algorithms, multidimensional techniques...) and Artificial Intelligence methods (Bayesian networks, Kohonen technique, neural networks...). Nevertheless, current data mining tools are far from being simple. In general the complex parameters of these techniques make it difficult to control the mining process. It appears that data mining tools often are difficult to use and the results are difficult to value.

Visualization techniques are continuously improved in order to provide a clearer view on different aspects of the data as well as a clearer view on the results of automated mining algorithms.

Example 1: Visualization of structures

A number of customized methods for visualizing structures have been developed, part of them based on the visualization of hierarchical trees. Some techniques show relationships between elements by special arrangements such as for example *Treemap* (Shneiderman, 1992) or *Sunburst* (Stasko & Zhang, 2000). Other techniques represents relationships by edges. It is the case of the *Hyperbolic Viewer* (Lamping, 1995) which uses the hyperbolic plane for arranging the nodes of the hierarchy by radial layout and projection into the Euclidian space. Another example is the *Magic Eye View* (Kreuzeler, Lopez & Schumann, 2000) where the layout is based on a plane radial layout which is mapped onto a hemisphere, with possibility of focusing to display the context with a smooth transition.



Sunburst directory visualizer

Click on a region once to identify, twice to refocus.

Fig. *Sunburst* image

Example 2 : visualisation of contents

Techniques to visualize the quantitative and qualitative properties of data have to deal with large amounts of multivariate data. We may use *Parallel Coordinates* to map the n-dimensional space onto a two-dimensional plane as proposed in VITAMIN S. *Scatterplot Matrices* arrange bivariate displays in adjacent attributes in matrix form in one image. With icon-based techniques, the attribute values of the elements are mapped to the features of an icon. It is the case of *Chernoff Faces* or *Needle Icons*.

For effective visual data mining, the data analyst must be able to interact and to change the visualizations (ie. the mining parameters) according to his needs. Interaction techniques may be navigation techniques using manual or automated methods to modify the projection of the data on the screen. They may allow the user to adjust the level of detail on parts of the visualization to emphasize some subset of the data. Some techniques provide users with the ability to select and isolate a subset of the displayed data for operations such as highlighting or filtering.

Visual data exploration aims at integrating the human in the data exploration process. Since the user is involved in the exploration process, shifting or adjusting the exploration goals is automatically done if necessary. Visual data exploration should be intuitive and require no understanding of complex statistical algorithms with multiple parameters. The visualization of the data allows the user to come up with new hypotheses and the verification of these hypotheses can be done also via visual data mining or through statistical techniques.

Example 3: Visualization of links: WordMapper

Visual link analysis combines powerful data mining algorithms with visual representation of associations as networks. The results of the analysis are displayed as a graph of linked objects. It allows you to discover hidden structure in large amounts of data. The graphic visualization facilitates the understanding of relationships among items and allows the user to quickly identify interesting patterns for further investigation.

The links between nodes can be analyzed as proximities between cities on a roadmap. This type of analysis is very intuitive and quicker than the search of information in a great number of cross tabulations. This approach does not require particular statistical skills.

WordMapper analysis performed on databases allows the quick identification of the significant connections among items and their visual representation as graphs. The connection nets are easy to interpret and the user's interaction facilitates the navigation within huge amounts of data.

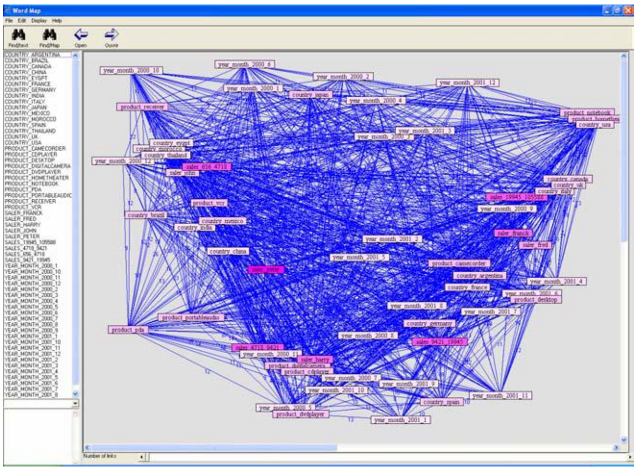
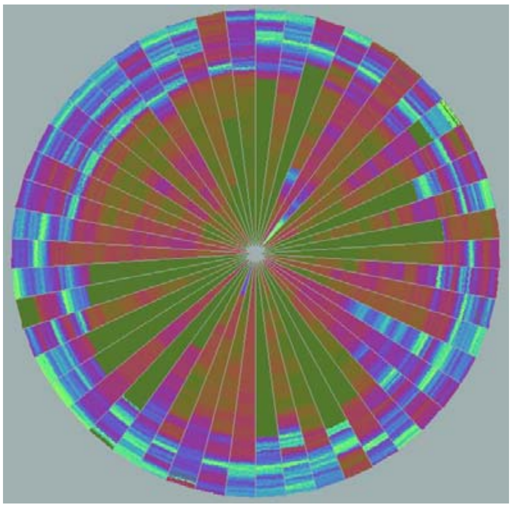


Fig. *WordMapper* image

For example, *WordMapper* creates networks in order to represent the connections among modalities of categorical variables. All the relations are taken into account : an item is associated to all the categories of all variables. The user can navigate in several maps by selecting nodes and opening new relations between items. The software detects homogeneous classes of items, or clusters. These clusters identify general themes. *WordMapper* has four levels of maps : the clusters' map, the sub-clusters' map, the one category-oriented map, the several categories-oriented map. By clicking on the different items, the user goes from one level to the other. He navigates within the network. He can also go back and browse the initial data. A scroll bar can be used to display the most important nodes on the map.

Example 4: Circle Segment

Circle segments is a relatively new technique for visualizing large amounts of high-dimensional data. The technique uses one colored pixel per data value and can therefore be classifed as a pixel-per-value technique (Ankerst *et al.*, 1996). The basic idea of the 'circle segments' visualization technique is to display the data dimensions as segments of a circle. If the data consists of k dimensions, the circle is partitioned into k segments, each representing one data dimension. Inside the segments, the data values belonging to one dimension are arranged from the center of the circle to the outside in a back and forth manner orthogonal to the line that halves the segment. The frst results show that the 'circle segment' technique is very powerful for visualizing large amounts of data, providing more expressive visualizations than other wellknown techniques such as the 'recursive pattern' technique and traditional 'line graphs'. A further advantage of this technique is that it allows the user to control the arrangement of the dimensions, which is important especially for comparing multiple dimensions.



Circle segment image

Example 5: Some comments on Parallel Coordinates

Parallel Coordinates is a multidimensional visualization tool often employed for data visualization. In order to represent a *d*-dimensional point, the basic idea is to draw *d* parallel axes labeling them according to the data variables. A point is then represented by locating the value of each variable along its respective axis and then joining the resulting points by a broken line segment. Many such diagrams are found in the literature. A discussion of the statistical and data analytic interpretations of parallel coordinate displays is given in Wegman (1990).

One advantage of the parallel coordinate display is that it represents *d*-dimensional points in a 2-dimensional planar diagram. In principle, there is no upper limit to the number of dimensions that can be represented. The big idea of the parallel coordinates representation, however, proceeds from its interpretation in terms of projective geometry. Both the Cartesian coordinate display and the parallel coordinate display can be regarded as projective planes. The mappings from the Cartesian coordinate system to the parallel coordinate system can be shown to preserve certain mathematical properties through the projective geometry notion of duality of points and lines.

One extremely useful property is that conic sections in Cartesian coordinates map into conic sections in parallel coordinates. In particular, ellipses in Cartesian coordinates map into hyperbolas in parallel coordinates. Thus, point clouds from high-dimensional ellipsoidal distributions can readily be recognized in parallel coordinates by structures with hyperbolic boundaries. Other dualities of interest include the fact that rotations in Cartesian coordinates get mapped into translations in parallel coordinates and vice versa.

Another useful feature of parallel coordinate displays is the ability to distinguish clusters. Any gap in any slice of the diagram separates two clusters. This ability to separate clusters is an extremely important feature of parallel coordinates.

Other interpretations include ability to detect linear structures and multidimensional modes. One feature also worth mentioning is the ability to compare observations on a common scale. Of all of the multidimensional data representations such as star plots, Chernoff faces, glyphs, and so on, parallel coordinates are unique in their ability to represent common scales on parallel axes. This is the easiest form of measurement comparison for humans to make.

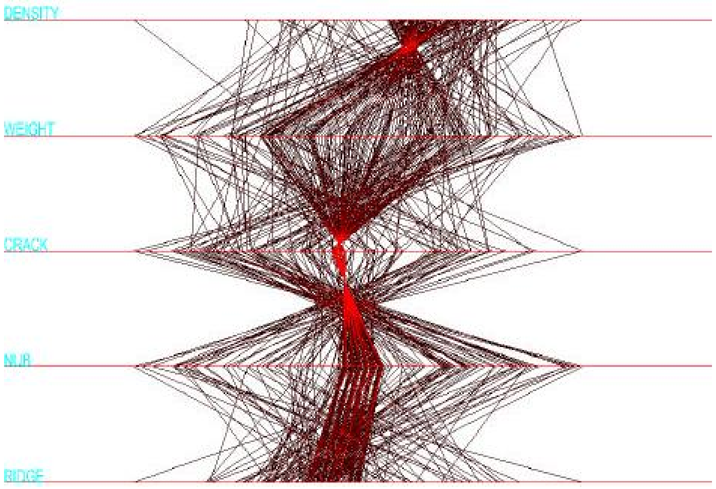


Fig. Parallel Coordinates image (Wegman)

Example 6: The Grand Tour

The Grand tour is an animation of the data. The basic idea of a Grand tour is to look at a data cloud from all possible points of view. It effectively allows for the data analyst to look at the data from all points of view. It allows the human visual system to follow the data cloud in the sense that, as the tour progresses, individual points make only small incremental changes at each step of the tour.

The *d-dimensional* Grand Tour is a generalization of the usual grand tour introduced by Asimov (1985) and Buja and Asimov (1985). It is a continuous geometric transformation of a d-dimensional coordinate system such that all possible orientations of the coordinate axes are eventually achieved. This animation allows for much more structure to be revealed than would be from simply looking at static plots.

The key idea is to find a path through the set of all rotation matrices as a function of a time parameter. Once a rotation matrix is determined, the canonical basis vectors for the coordinate system are rotated by the matrix and then the data cloud is projected into the rotated coordinate system. Coupled with the parallel coordinate display, these two techniques allow for an in-depth study of high dimensional data. A grand tour is in some sense a generalization of a two-dimensional rotation, although it is not a rotation in conventional sense. Partial grand tours can be accomplished by holding one or more variables fixed.

Example 7: Brushing

Ordinary brushing is implemented by drawing a rectangular box and accomplished by brushing a data cloud with a color for the purpose of visually isolating segments of the data. Any data point or line segment, which intersects the rectangular box, is colored with the chosen color.

In data sets where there is considerable overplotting, brushing is potentially misleading, especially where there is an animation such as rotation or Grand tour. *Saturation Brushing* (Wegman *et al*., 1997) is a generalization of ordinary brushing. In saturation brushing, each point is assigned a highly desaturated color (nearly black) and when points are overplotted, their color saturations are added via a computer hardware device for blending two images. Because it is a hardware rather than software implementation, it is usually very fast.

Thus, heavily overplotted pixels have fully saturated colors, whereas pixels with little overplotting remain nearly black. Saturation brushing is an effective method for dealing with large data sets. Coupled with parallel coordinates and the Grand tour, these methods allow for an extremely effective visual approach to large, high-dimensional data.

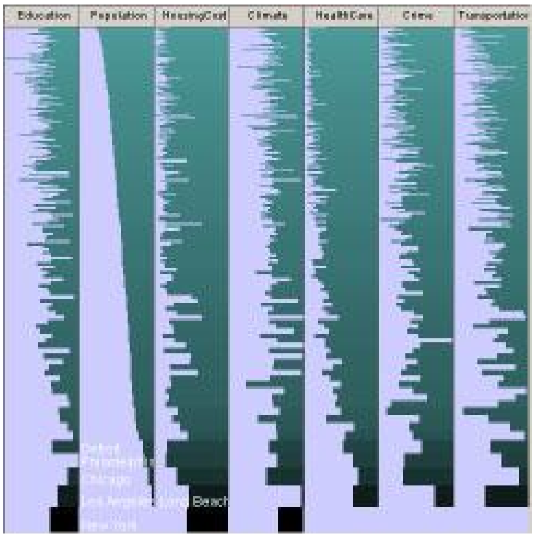
Example 8: ILOG-Discovery System

ILOG Discovery is a visual information exploration tool. It provides simple dialogs to view this data in many different ways, highlight trends, spot general patterns, and pinpoint exceptions to find out what underlying structures govern data. It can handle large amounts of data and produce a wide variety of visualizations with few manipulations, which makes it an interesting exploration tool. One can filter the data sets and then create selections and viewpoints in these visualizations. One can create template visualizations for easy reuse.

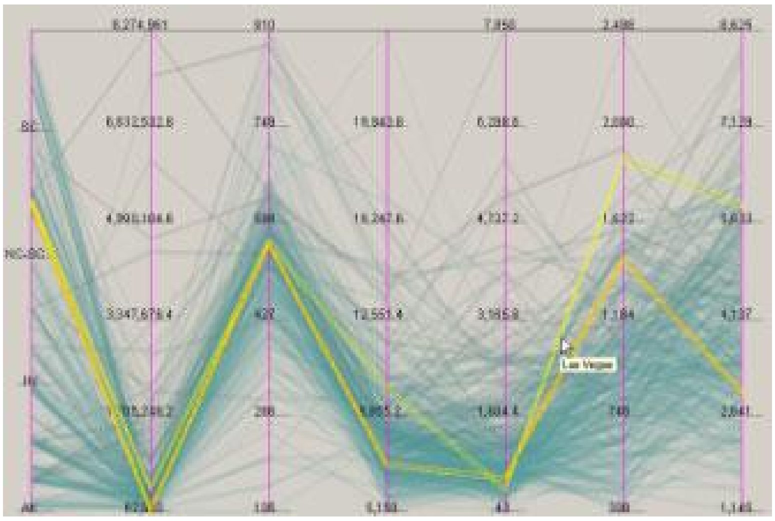
Histograms and Distributions

Parallel histograms and parallel coordinates give a first overall glimpse of the data set when the data consists of numerous numerical or enumerated variables. These visualizations let you spot the general distribution of each variable, find out any outliers or missing data, and detect possible correlations between variables.

- **Parallel histograms** present the data set in a table in which each variable is represented by a small bar whose length is proportional to the variable's value.
- **Parallel histograms with size specification**. If one of the variables can be summed up (for example, if a variable represents a size, a population, or an absolute value referring to each object), then you can make the width proportional to this variable. This emphasizes the objects that have large values for this variable.
- **Parallel coordinates**. Parallel coordinates provide you with information on the various distributions and correlations between columns of the data sets.



Parallel histograms with size specification for huge data sets

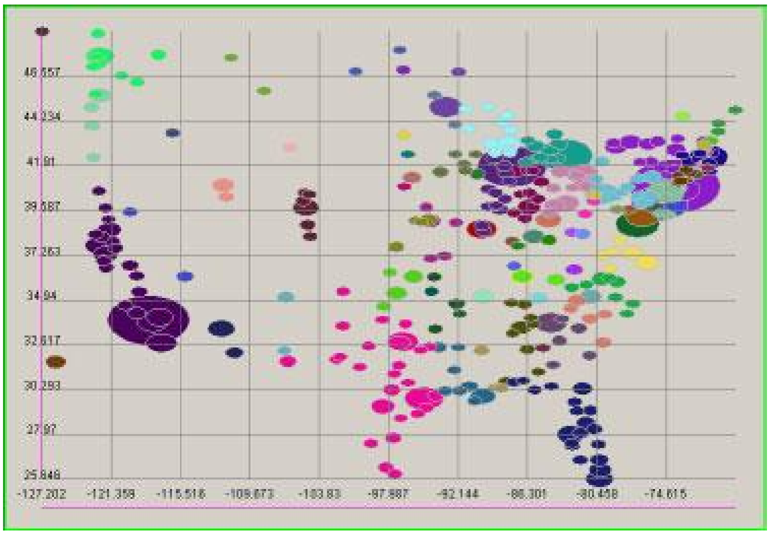


DISCOVERY Parallel coordinates

Graphs

2-D plots and diagrams are useful to compare two major variables, visually assessing their correlation and distribution. Other graphic attributes can be mapped (such as size and colour) to include further variables into the correlation analysis. When you want to focus your attention on a few specific variables, you can use:

- **Simple 2-D graphs**
- **Augmented graphs** with size specification, colours, embedded bar charts, and other features.
- **Interactively filtered graphs**. The software enables dynamic queries, which are particularly useful with 2-D graphs.

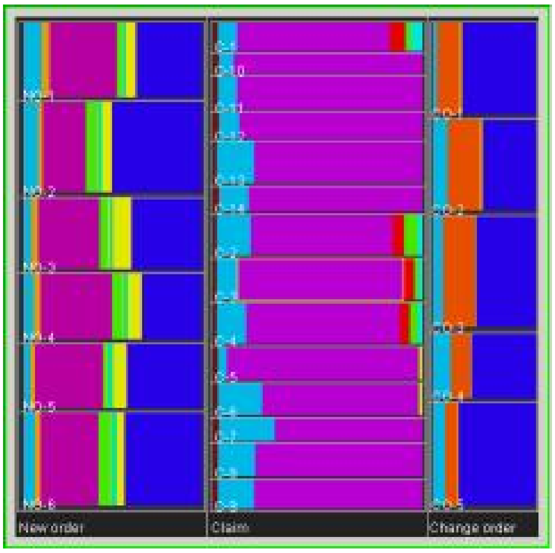


US cities population (by longitude and latitude)

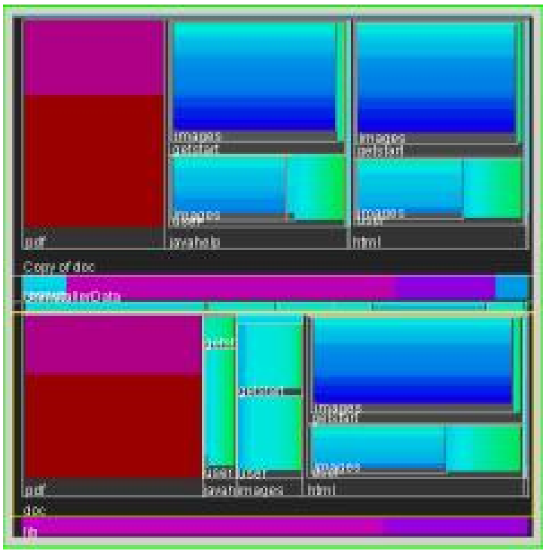
Maps and Hierarchies

Maps of various kinds (grids, treemaps, and others) are useful when the data is hierarchized, or when you want to see how the data is distributed according to one or more dominant criteria:

- **Grids** enable you to view data split into multiple categories, each category being shown as a small square whose representation can be customized to display relevant information on this category.
- **Hierarchized views**. Grids and other maps can be embedded in one another, showing the relative distribution of a subcategory according to a main category.
- **Treemaps** are one kind of hierarchized view that can be produced with the software.



Hierarchized Views



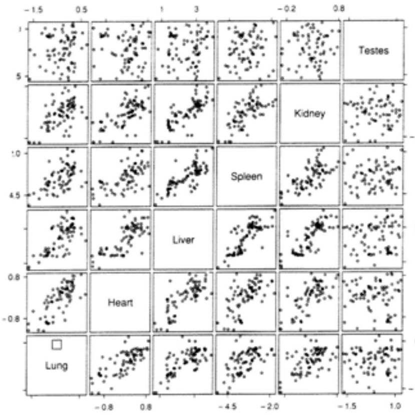
Treemaps

Data Visualization Techniques: Final remarks

Pioneering works of Bertin (1981) and Tufte (1983) focused on visualization with inherent 2D or 3D semantics. Important development of visualization techniques for arbitrary multidimensional data appears in the 1970 in France with Benzécri (1973) using the techniques for dimension reduction: principal components analysis for quantitative data and multiple correspondence analysis for qualitative data. Multidimensional scaling has the same goals but uses the similarity-dissimilarity matrix which prevents the processing of large data sets. Segmentation and aggregation techniques coupled with principal coordinates graphical displays provide a lot of enlightening visualization displays.

Visual Data Mining is the process of finding previously unknown information from a data base, using data base management, data analysis, data mining and data visualization facilities. Since visualization is a critical component, this places greater focus on the human component of the knowledge discovery process. Visual data mining is then inherently interactive: it focuses on refining hypotheses based on results obtained from interacting with the data.

The data analysis part of the process provides discovery functions such as clustering, principal coordinates, regression, decision tree, pattern matching, etc. Automatic discovery algorithms determine significant pattern classifications from the data with minimal user intervention. A more human-centred approach is exemplified by OLAP systems where data is organized into hierarchies of different resolutions. In that context, a drill-down feature is used to access lower-level data from summary results. Other data navigation aids are used such as cross-tabulations, comparing graphs in scatter plot matrices: a popular method is to link data displays together and brush one display to see the effects on the other displays.



Scatterplot Matrices

The visualisation part of the process must provide an adequate user interface to map data, and operations to intuitive forms for the user to interact with data. This visualization field has yielded many techniques for portraying data in 2D and 3D spaces often with animations. They transfer the discovery task upon the user who has to change the display parameters such as viewing orientations, thresholds, opacity or data ranges.

The user part of the process is to perform sequences of linked interactions in order to isolate relevant information that serve to generate and validate hypotheses about the data. The user drives the discovery process by formulating and testing hypotheses and finally drawing conclusions.

The case of huge data

On huge data sets, visualization often works in a batch mode where a cursor interface might not be appropriate for interaction. By nature the visualization process affords greater interactions with data because the graphical display may uncover relationships and information undetected by automatic tools. Then there are visual probes that need to be mapped to the data base for rapid data selection such as dynamic queries for retrieved data (Albher *et al*., 1992).

References

Albher, C., Williamson, C., Schneiderman, B.: *Dynamic Queries for Information Exploration: an Implementation and an Evaluation*. Proceeding of CHI'92, pp. 619-626, 1992.

Ankerst, M., Keim, D., Kriegel, H.: *Circle Segments': A Technique for Visually Exploring Large Multidimensional Data Sets*, Proc. Visualization '96, Hot Topic Session, San Francisco, CA, 1996.

ATKOSoft S.A. 15 Etolias str., 152 31, Athens, Greece. E-mail: info@atkosoft.com, URL: http://www.atkosoft.com

Baudel, T.: *ILOG Discovery Preview*, http://www2.ilog.com

Benzécri, J-P.: *L'Analyse des Données*. Dunod, Paris, 1973.

Bertin, J.: *Graphics and Graphic Information Processing*. Berlin, 1981.

Grimmer J-F.: *WordMapper*, GrimmerSoft, 2003, http://www.grimmersoft.com

Kreuzeler, M.; Lopez, N.; Schumann, H.: *A Scalable Framework for Information Visualization*. Proceedings InfoVis'2000, Salt Lake City, 2000, pp. 27-36.

Lamping, J. et al: *A Focus Context Technique Based on Hyperbolic Geometry for Viewing Large Hierarchies*. ACM Proceedings CHI'95, Denver, 1995, pp. 401-408.

Lebart L., Morineau A., Piron M. (1995). *Statistique Exploratoire Multidimensionnelle*. Dunod, Paris.

Shneiderman, B. *Tree Visualisation with Treemap: a 2D Space Filling Approach*. ACM Transactions on Graphics, Vol.11, No 1, 1992, pp. 92-99.

Stasko, J.; Zhang, E.: *Focus Context Display and Navigation Techniques for Enhancing Radial Space Filling Hierarchy Visualizations*. Proc. IEEE Information Visualization 2000, Salt Lake City, UT, Oct. 2000, pp. 57-65.

Tufte, E. R.: *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983.

Wegman, E.: *Hyperdimentional data analysis using Parallel Coordinates*. JASA, '(1990) vol.85, pp. 664-675.

Wegman, E. and Luo, Q. (1997), *High dimensional clustering using parallel coordinates and the grand tour*. Computing Science and Statistics, vol.28, pp. 352-360.