

Darab János

**BUDAPEST POLYTECHNIC
HUNGARIAN FUZZY ASSOCIATION**

PROCEEDINGS

OF THE

**INTERNATIONAL SYMPOSIUM OF
HUNGARIAN RESEARCHERS**

COMPUTATIONAL
INTELLIGENCE

**MAGYAR KUTATÓK NEMZETKÖZI
SZIMPÓZIUMA**

2000

NOVEMBER 2

BUDAPEST

**BUDAPESTI MŰSZAKI FŐISKOLA
MAGYAR FUZZY TÁRSASÁG**

Visual Query System for Datamining

József Dombi

University of Szeged

Abstract: Visualisation and data mining must connect closely to each other, because the cognitive aspect can be realised only by using visual techniques. Without this, there cannot be an effective communication between people and programs. Currently, manager information systems contain graphical description of reduced data (e.g. in Microsoft Excel), which can only be a part of a real manager information system.

We propose a new generation of visual information system, called DataScope, which has won the Information Technology Prize in 1997 in Brussels and "the best software of the year" prize at Comdex in 1999 in Las Vegas. The program has several properties: digital information is translated into analogue one; queries, an essential function of databases, can be realised; several features can be examined at the same time and only the visualisation capacity of the computers or human perception can limit it; comparison (relation) can be accomplished, i.e. the relation of two or more alternatives can be visualised at the same time.

Keywords: Data Mining, Data Visualisation

1. The Role of Visualization

More and more visual elements appear in communication. First of all, we want to answer why.

1.1. Communication Over the Languages

The process of globalization seems to be evident. The world seeks a new communication method, which is independent from the particular languages. The visual communication seems to be the most effective (for example, more and more symbols appear in the traffic, in hotels, airports, computer programs, etc.).

The whole process can be seen as a return to the ancient culture. Interestingly, writing is still based on visual form in Asia. The speciality of this writing system is the independence of pronounced words and grammar. In other words, the advantage of Chinese letters is that they create a communication which is independent of dialect.

1.2. The Relation of Verbal and Visual Communication

The use of visual approaches to this difficult and complex task is intuitively appealing (we cite the maxim "a picture paints 1000 words" as evidence of this intuition). Few decades ago, mostly the verbal descriptions have been dominant in books and only a few illustrations could have been found as a matter of special interest, while nowadays most books are full of illustrations. There are several examples, when it is difficult to express complicated structures with words, or they cannot be expressed at all, but it will be easy to understand them with the help of a simple picture. During the evolution of science, special visualization methods have been developed. Here we can refer to the mathematical system of notations, the inscriptions of chemical compounds or biological processes. These are visual methods, too.

The other aspect of visualization is that the visual perception is the most effective. It is a massively parallel method. To give only one example, the global optimum can be found immediately by coloring a relief map, while other (mathematical) methods have limited success.

1.3. Human Perception and the Digital World

Nowadays, everybody speaks about digitalization. From technical point of view, it is very effective, since electronic circuits work well on digital information. But the digital information is not for human perception, because usually it does not indicate an evaluation. We do not know what really means if we change a digit in a number from 0 to 1. It is only a small change regarding its visual form, but the change can be enormous depending on the position of the digit.

Recognition of digital information takes more time. The history of digital watches is a good example. They were fashionable only for a short time, because of the human perception method. It is easy to verify that analog signs are recognized better because they indicate an evaluation. Digital communication is not only disappearing for watches, we can also mention the car and airplane cockpits as examples.

1.4. Computers, Visualization and Data Mining

Now, as computers enter into general use and the capacity of the color screen significantly improves, the visualization receives an important role. Currently, geographical information systems, image processing and pattern recognition are the most important visualization trends but the visual database query languages also require more and more significant place. Visual programming is also a quite new trend. [3]

First of all, computers take over the task of previously mechanically accomplished data operation methods. It results in considerable decrease of expenses. The by-product of this process is a huge amount of data, in which typical characteristics and anomalies have been discovered while they were treated by humans. However, as computers take the role of persons, it becomes necessary to get such information from data using advanced computer-based tools and techniques.

There are three levels: *data* from the *information* can be extracted and on the basis of the information, *knowledge bases* can be formed. The purpose of data mining is to obtain information. The main concept is to find relationships in data and give right directions about processes for the executives with the help of the right algorithms.

Visualization and data mining must connect closely to each other, because the cognitive aspect can be realized only by using visual techniques. Without this, there cannot be an effective communication between people and programs. Currently, manager information systems contain graphical description of reduced data (e.g. in Microsoft Excel), which can only be a part of a real manager information system.

1.5. Future of the Cognitive Aspects of Data Mining

Computers can dynamically visualize graphs. Until now, there were only a few applications which utilized the possibility of animation, for example. The reason of this is the adherence to the paper office. By the use of computers, the paper form communication might be succeeded by the electronical data storage, nevertheless it did not happened yet because of the traditions. Probably this process will be accelerated by changing generations.

2. The history of data vision

Transforming figures into charts is a classical toll of data analysis. The descriptive statistics and computer graphics is widely used to illustrate numerical information by producing standard visual representations (bar charts, line graphs, pie charts, etc.) or using some more advanced techniques, for example Andrew's curves [1] Chernoff's faces [5] or Korhonen's harmonic houses [4].

In the early 1970's, two promising techniques were developed [4,5] for visualizing multivariate data by using original variables. Andrew plotted a curve

$$f_i(t) = x_{i1} / \sqrt{2} + x_{i2} \sin t + x_{i3} \cos t + x_{i4} \sin 2t + \dots$$

for each data point $x_i = [x_{i1}, x_{i2}, \dots, x_{im}]$ over the interval $-\pi \leq t \leq \pi$. Thus each observation will be a harmonic curve drawn in two dimensions. In this method, the number of variables are unlimited. The harmonic curves depend on the order in which the variables are used.

Chernoff used a human face to represent each observation graphically. The construction of Chernoff's faces consists of geometrically well defined elements, such as arcs of circles, arcs of ellipses and straight lines. The variables are used as the parameters of these elements. Chernoff original proposal consisted of 18 face parameters. Korhonen's harmonic houses map the variables to the parameters of a house figure. This projection also helps to evaluate the alternatives.

Andrew's approach is too technical, it makes only a few representations of component, backs up only the "holistic" valuation. The main advantage of Chernoff's face that a large number of features can be represented simultaneously. It is difficult to give valuation. To perform the valuation of the alternatives we have to define the "nice face" and good changes should make the face "nicer". This is a very difficult task. There have not been approaches at all in this direction.

Korhonen's harmonic house approach helps us in drawing up valuation. The deformed houses are bad. Nevertheless, here we also have to take care of the right composition. It is difficult to assign criteria so that the nice house is the right alternative. For multicriteria decisions, most features are controlled so that the more (or less) is the value, the better is the alternative. The harmonic house does not comply with these requirements, because sizes are not proportional to beauty. If we build internal relations we can use this process successfully. None of the developed processes are suitable for composing, which is the essential function of

using the databases. Therefore, Korhonen's model makes only possible to compare pairs. We have to mention that the development of classic tools is already present on the area of economical modeling. [3]

To summarize all, we can speak about a successful visualization technique if

1. digital information is translated into analog one;
2. we avoid the pitfall of the aesthetical constitution and return to the abstract model;
3. queries as an essential function of databases can be realized;
4. queries are realizable without typing and learning instructions;
5. we can examine several features at the same time and only the visualization capacity of the computers or the human perception can limit it;
6. the system supports the monotony of features (the goodness of the evaluation is monotone by the value of the features);
7. comparison (relation) can be accomplished, that is, the relation of two or more alternatives can be visualized at the same time;
8. it will be possible to visualize the transformed valuation list of a feature.

3. The DataScope Concept

3.1. Car Example and Types of Fields

We will often refer to an example database containing the following data on 83 cars (the parentheses contain the measurement units): name, price (DEM), performance (HP), cubic capacity (cm^3), consumption (l/100 km) and fuel type used (diesel, normal or super). These fields can be classified to the following categories:

- *Identifier*: The identifier field is used to identify the records. It is important for the identifier field to be unique, or almost unique, to help in identifying the records easily. In the car example, the identifiers are the names of the cars.
- *Numeric field*: This field type can be used to display numeric data. If a database field contains (mostly) numbers, it can be specified as a numeric field. (Examples are price, consumption, etc. in this car example. Numeric fields are represented with a distribution function.
- *Discrete field*: When a field contains mostly non-numeric data (categories), then we refer to it as a discrete field. In the car database, the discrete field is the fuel type (Fuel field). Its possible values are 'D' (Diesel), 'NF' (Normal fuel) and 'SF' (Super fuel).

3.2. Empirical Distribution Function as a Basic Tool for Visualization

If we have a large spreadsheet full with different numbers, it is not easy to understand the semantical meaning of a particular value. To understand it, we have to translate the values to an evaluation (good-bad, high-low, etc.). We can do it only by determining how many values are greater. A good example is that when a child gets a mark, to understand what really does it mean, we usually ask how many children got a better one in the class. Only on this basis we can evaluate the performance of the child, and this is nothing else, but to calculate the value of the empirical distribution function. DataScope uses this empirical distribution function to translate the numbers into evaluation.

A numeric field is represented by its distribution function. The distribution function consists of 'stairs'. The X axis of the coordinate-system goes from the least to the greatest value of the database field, while the Y axis goes from 0 to 100%.

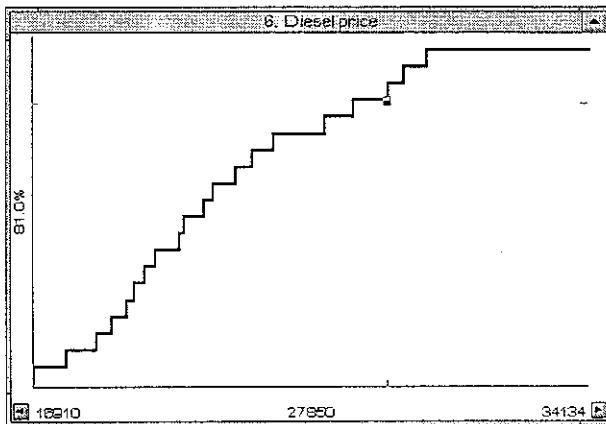


Fig. 1. The empirical distribution function

The empirical distribution function is a good visualization tool because:

- it shows the *sequence of the records*, which is much more informative than the original value;
- it is a transformation from numbers into *evaluation*;
- it transforms the digital information into *analog information*;
- it has the *monotony property*, i.e. a larger value is better (or worse, depending on the sorting order);
- The distribution function is a good tool for modeling *context-dependency*. A specific value has different meaning in different contexts. For example, the 6

liters/100 km fuel consumption can be good for a petrol-powered car but too high for a diesel car.

- It is easy to *compare records* by marking their location on the distribution function curve;
- By displaying several distribution functions simultaneously, we can compare the alternatives by several features easily (e.g. the car is cheap and low-consumption).

Other important properties of the distribution function are:

- By selecting a point on the X axis (a price, in this example), the value of the distribution function shows the percentage of the records that precede this record (e.g. what is percent of the cars are cheaper). In this way, we can see the *relation* of one record to the others.
- If we select a percentage value on the Y axis, we can see the X value that relates to it. In this way, we can easily examine what value is needed to reach a specific place. For example, if we select the percentage value 75% in the Price window, you can see that what price is needed for a car to be placed into the first quarter.
- If the distribution function contains a *long horizontal line*, this means that there are no records in that interval of the X axis. The right side of the above picture shows an example: there are no cars in a large price interval.
- A *long vertical line* means that there are many records that belong to one X value. For example, if we display the cubic capacities of cars, we see that many cars have the value 1599 cm^3 or a little smaller. This is because cars with a cubic capacity of more than 1600 cm^3 are taxed and there is a long horizontal line after this point, which means that there are no cars in a larger interval after 1600 cm^3 .
- An example for modeling the context-dependency on distribution functions is shown in Figure 2. The percentage value of 52.4% relates to 22475 DEM in the Diesel Price window. The same record is automatically selected in the Price window, too. However, the percentage value is higher there (73.5%). This means that 22475 DEM is a middle-level price for a diesel car, but it is a high price if you examine all cars, so diesel cars are generally more expensive.

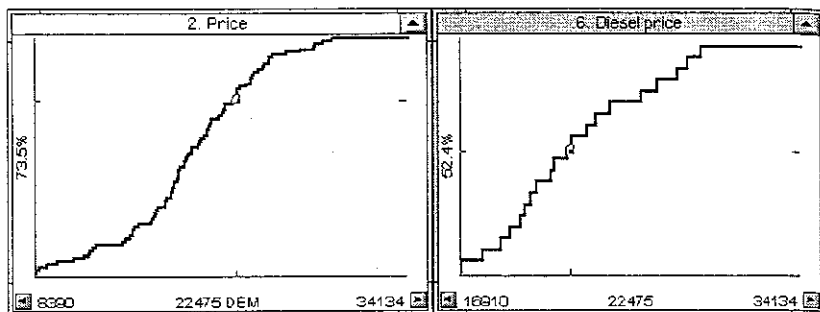


Fig 2. Context dependency with empirical distribution functions

3.3. Relational Diagrams

By using relational diagrams, it is easy to compare two numeric fields. The horizontal/vertical axis of the relational diagram is the vertical axis of the distribution function of the first/second field. Each record is represented by a point, which is placed at the intersection of the values of the distribution functions of the two fields.

The diagonals of the window are important dividing lines. The records that have similar positions by both fields are near this line (for example, cars with an average price/performance ratio). The farther a point from the line, the more exceptional the record is. Whether the record is a good or bad exception can be determined on the basis of the sorting direction of the two fields.

This window type gives you a possibility to notice exceptional records. If you open a relational window, you can see the exceptional points immediately. If you wish to know which record a particular point represents, you can click on the point to select the record and see its identifier in the identifier window. Other windows will show the other features of the record. Naturally, you can open more than one relational window, select an exceptional point in one, and see whether it is exceptional in the other.

The relational diagram can also be used well to examine whether a better choice exists than a specific record. Let us open a relational window of the 'Price' and 'Performance' fields of the car database, and choose an ascending sorting order for both of them (Figure 3). Now, let us examine the car marked 'P' and considering that we want a car with a similar price, but with a higher performance. The price is

on the horizontal axis, so the cars with the same price are on a vertical line. The car marked 'A' costs slightly more, but its performance is much higher. You now need to determine whether this additional performance is worth the higher price, and, of course, you need to examine the other features of car 'A'.

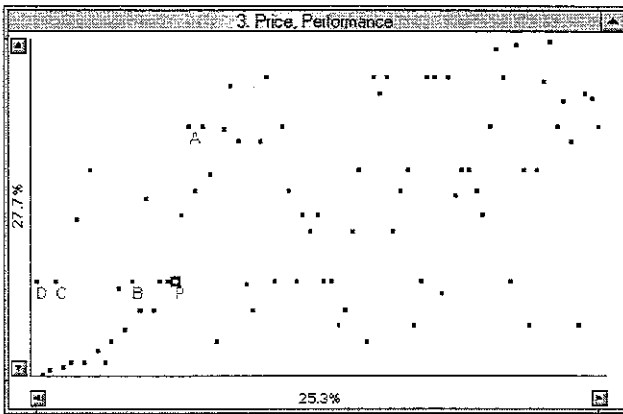


Fig. 3. A relational diagram

3.4. Identifier Window

An identifier field is represented with a list. The sorting order of this list depends on which the active window is in DataScope:

- If the identifier window itself is active, the list elements are in alphabetical order.
- If another field window is active, the list is sorted by the values of the field represented by that window.

3.5. Discrete Field Window

This window type displays the contents of discrete fields, i.e. fields that have only a few different values. Conventional methods are used to display these fields, such as a pie chart or a bar chart.

3.6. Queries

Using the visualization procedures we can also perform queries on the database. Let us see a list of typical questions to be answered:

- Which is the cheapest/most expensive diesel car?
- Do I have an expensive car?

- Which are the cars with consumption lower than 7 l/100 km and price lower than 20000 DEM?
- What is the price of a middle-priced car?
- What is the price of a middle-priced diesel car?
- Are diesel cars generally more expensive than petrol-driven ones?
- What percentage of the cars under 20000 DEM are diesel-driven?
- How many cars satisfy a given condition?
- How many cars have no price data?
- What are the proportions of diesel and non-diesel cars in the database?
- Is there a connection between the consumption and the performance? What are the exceptions?
- I want a car whose price is about 20000 DEM, but I would like it to have a high performance. How can I find these cars?
- I'm definitely not interested in petrol-driven cars. How can I simplify my work?
- How can I locally select all cars that have no price data?
- How can I locally select the top 10 cars in performance?
- I have a discrete field. I know the goodness of the individual categories, so I would like to make a numeric field, where the goodness of each category is represented by a number.

Elementary Query with Local Selections

Users can select records by any field. For example, we can select cars between 15000 and 20000 DEM by the 'Price' field, all Audi's and BMW's by the 'Name' field, or all diesel cars by the 'Fuel' field. This is called *local selection*. You can locally select records by each field. These local selections are independent from each other and can be modified at any time.

All field windows show the local selections made by the field(s) they represent. The identifier windows play a special role, because they always show the local selection of the field that the active field window represents.

The most important thing is that these queries are done with the mouse. There is no need to learn and type any commands or formulas.

Making a Query as Global Selection

A global selection can be made from a logical combination of local selections:

- *Union*: For example, if we select the cars cheaper than 15000 DEM by the Price field and the cars whose consumption is lower than 7 l/100 km by the

Consumption field, we get the cheap *or* low-consumption cars when we create the union.

- *Intersection*: With this feature we can examine 'and' connections. If we have the previously mentioned local selections (cars cheaper than 15000 DEM and with a consumption lower than 7 l/100 km), creating the intersection results in cheap *and* low-consumption cars.

3.7. Summary of Window Elements

The following figures show how the global and local selections appear in the various window types, as well as other window elements:

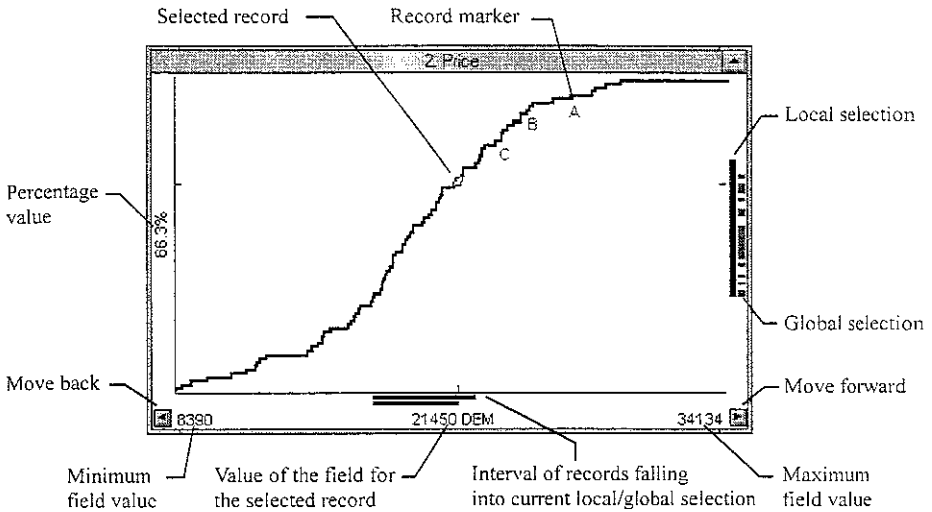


Fig 4. Elements of a numeric window

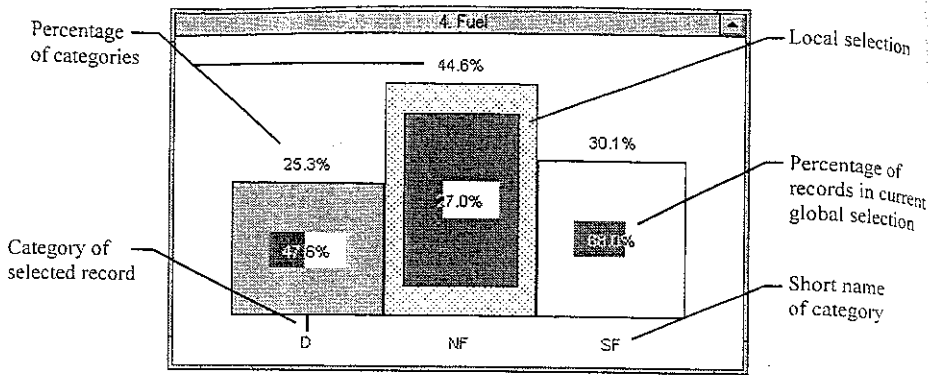


Fig 5. Elements of a discrete window

The selected record is marked with a short line (under area 'D' in the picture). Locally selected categories are marked with a thick border. Inside the bars, you can see what percentages of the individual categories fall into the current global selection.

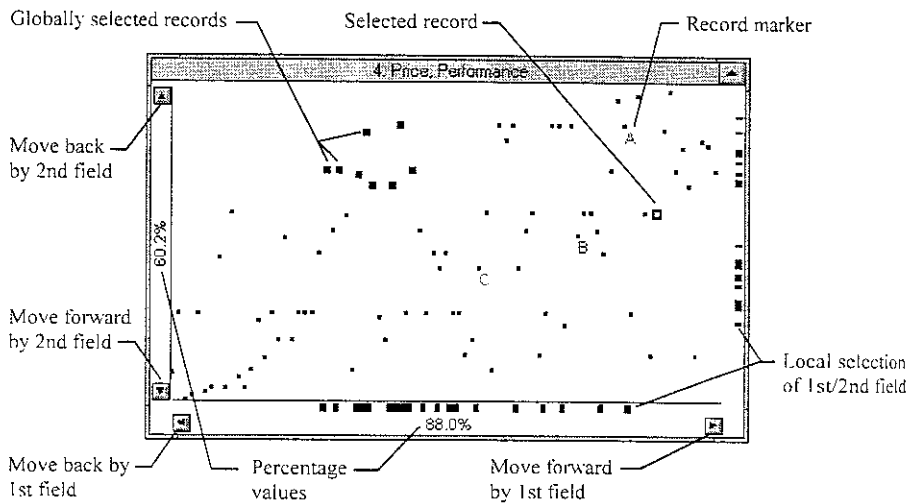


Fig. 6. Elements of a relational window

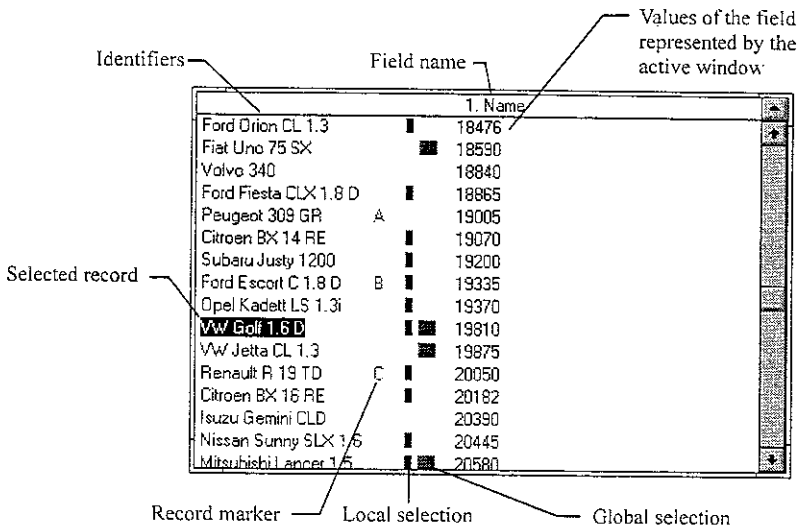


Fig. 7. Elements of an identifier window

4. Summary of DataScope Features

- Up to 16 windows can be opened, so users can analyse the database according to 16 fields (or field pairs) at the same time.
- This software is suitable for analysis of both individual records and record groups.
- Database queries are done with the mouse. There is no need to learn any commands or formulas. The database can be queried interactively, directly from the displayed diagrams.
- Noteworthy records can be assigned a two-letter identifier which appears in all diagrams, so these records can be followed easily.
- One of the most interesting features of DataScope is the full synchronicity. Elements of the database can be examined from many aspects simultaneously. Users can select elements with specific features from one aspect and see their connection from other aspects.
- Numeric data can be analysed in an exciting new way, eliminating the considerable time needed to determine where a record is situated among the others. The value of the distribution function now shows this immediately thus allowing to transfer the numbers to subjective opinions.
- Since numeric information is translated, relational diagrams allow to compare any two numeric fields of the database, thus making searching for relationships easy.

- DataScope can import data by using Microsoft's ODBC standard. This standard allows handling of most standard database types.

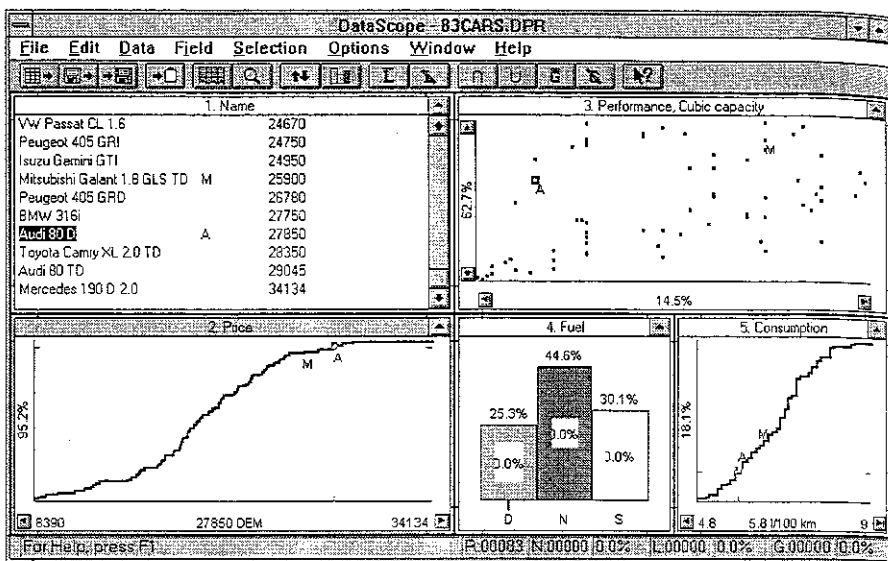


Fig. 8. The DataScope main window

5. References

- [1] Andrews, D.: Plots of high dimensional data, *Biometrics*, 28, (1972) 125-136.
- [2] Kiper, J. D. Howard, E. and Ames C.: Criteria for evaluation of visual languages, *Journal of Visual Languages and Computing* 8, (1997) 175-192.
- [3] Schroeder, W. Martin, K. and Lorenzen B.: *The visualization Toolkit: an object-oriented approach to 3D graphics* (1996) Prentice Hall
- [4] Korhonen, P.: Using harmonious houses for visual pairwise comparison of multiple criteria alternatives, *Decision Support Systems* 7 (1991) 47-54.
- [5] Chernoff, H.: Using faces to represent points in k-dimensional space graphically, *J. Amer. Statist. Assoc.* 68 (1973) 361-368.