FROM DATA COLLECTION TO MATHEMATICAL MODELS: METHODOLOGICAL PATHWAYS

Sergio Camiz[1]

Dipartimento di Matematica Guido Castelnuovo - Sapienza Università di Roma

E-mail: sergio.camiz@uniroma1.it

Abstract: Opposite to physics, the complexity of randomness and factors involved in ecological studies claims for a reflection on the scientific methodology and suggests a methodological pathway, composed by three steps, corresponding to exploratory and confirmatory studies, followed by the mathematical modeling. Following this pathway, the scientist can better control his work, since he becomes aware of the structures involved in the phenomenon and of the reliability of the hypotheses he stated. In particular, in what concerns the data analysis, each step helps in a better planning and organization of the following one. In this paper some aspects of this pathway are introduced and discussed, as well as the interaction between the scientist and the data analyst. Two case-studies are proposed as examples.

Key words: data collection, mathematical models

MSC 2000: 93A30

1 Introduction: physics vs. ecology

In physics the mathematical modeling was used since Galileo Galilei. He first had the idea of the set up of an experimentation. In classical mechanics, the modeling reflected the deterministic cause - effect relation of mechanical objects, and so did the mathematics involved, whose complexity reflected the complexity of the studied phenomena. The success of *deterministic models* was such that other sciences used the same approach, until mayor problems raised both in physics and in other scientific frameworks.

Whereas in physics the problems raised once the dimension of the observed objects were of the same order of the observing tools, from which the Heisenberg indetermination principle derived, in both life sciences and humanities

[1]This paper was written during a visit to I. Vekua Institute of Applied Mathematics of Tbilisi State University, carried out in the frame of the cultural agreement with Rome University La Sapienza. Both institution grants are gratefully acknowledged.

they are due to the uncertainty of the factors that influence the observed effects. In addition, in these frameworks the experimentation is seldom possible and models should express tendencies, due to both the high number of factors involved and the so-called *individual variation*, typical of living beings. As a consequence, the mathematics to be used must take into account these differences, giving raise to either *probabilistic models* or deterministic ones, provided that they limit their explanation to only a part of the total variability observed.

This suggests that an effective modeling should be based on an approach able to give reason of both the aspects taken into account, namely to the explained variation and the remaining unexplained one. So, *data analysis* may be a correct discipline to support this approach. In Camiz (2001) the essentials of a scientific investigation are outlined, together with a methodological pathway that gives reason of the different data analysis tools to be used in each step. In the following both aspects will be reminded, followed by two case studies, useful for the comprehension of the utility of such methodologies.

2 What is a scientific study?

Once a question is asked in a scientific environment, in order to give a correct answer, one starts from the present knowledge, if necessary collecting the literature that concerns the question and studying the way some answers were given to similar questions; then, one checks if the answer may be obtained using the collected knowledge. If it is possible, the answer is *deduced* from the current knowledge.

When this is not the case, one may try to build up an answer. In a scientific environment, this may be done through empirical experience. To this purpose, one must do some observations and collect some data, that will be further analyzed. In order to be found in this way, the searched answer *must* be *contained* within the data, in the sense that the collected data must be able to give an answer to the asked question.

There is a difference between the first and the second case. In fact, in the first case, an answer may be *deduced* from other knowledge in a purely logical way. Several different levels of logic may be considered. It may be the *associative* logic, that is the logic pervading the dreams, whose connections may be own of each single person, although having some universal rules, as psychoanalysis argues. It is in many cases the specific science framework *paradigm*,

that condition the construction of most scientific theories, by grounding them to some very general hypotheses. It may be *mathematical logic*.

Mathematical logic has very strict rules. In fact, both mathematics and mathematical logic may be completely formalized, in order to provide at the same time a language for the description of most natural phenomena, and a tool for *proving* the truth of what is expressed, provided that this description is consistent with the *axioms*. These are properties of the objects of the investigation that are empirically proved to be always true for them. In other sciences, reasoning is more free and the deduction of scientists are less constrained: it is frequent to find some hypothesis considered reliable only by *analogy* with some other, based on the researcher experience.

It must be emphasized that in any case the deduction must be stated on solid bases, otherwise a huge risk of false deductions may arise, leading to false conclusions. It is the case of *subculture*, namely statements apparently reliable but totally denied of truth. This is the case of statements once considered scientifically proved, but later overcome by new ones. In this case the presence of subculture is due to the lag of broadcasting of both information and innovation of scientific statements to the population, as well as their consequent application to the common life. Examples of subcultures are the astrology, some alternative medicine, exotericism, false political issues, etc.

As an alternative, the truth of a hypothesis may be *induced* by experimentation. This may be achieved either by controlling the conditions under which the observations of the phenomenon under study are repeated, so that it is isolated from all possible noise and bias, or observing many times the phenomenon in the reality and connecting it with the conditions present during the observation. In both cases, once the conclusions of the experimentation are drawn, one may proceed by *inference*, that is by generalizing the results to a broader *population*. This is a larger set of objects, individuals, or situations, that share with the observed ones the same characters that condition the results of the experimentation. The consideration of data taken from reality and their critics is actually the *new science*, born in the Renaissance period thanks to Lorenzo Valla, who used a scientific method to deny the power of the Church. At that time the Church power was justified by the *donatio Constantini*, a document that Valla proved to be false, written during medieval period and not in the IV century, as it was pretended to be. Later Galilei used the new science in his grounding of mechanics.

Some particular cautions may be necessary, in order to proceed by gener-

Figure 1: The data analysis position in the knowledge process of a scientist, based on observations.

alization, in particular considering the causal relations between phenomena. In fact, one may distinguish at least among two different causalities. The one usually considered in mathematical terms is *strict causality*, that is any kind of functional relation between cause and effect, typical of most of relations described in classical mechanics. This is usually described through *deterministic models*. For such an identification of the relations, and for their quantification, phenomena may be isolated in the experimentation, in order to avoid the co-occurrence of other causes that may bias both the effects and their magnitude.

There is another kind of causality, namely *weak causality*. In this case, many events may favor the phenomenon under study or be co-occurrent with it. It is the typical case of observations in human sciences, in ethology, and in ecology, since individual behavior, responses, and choices are various and not always dependent by identifiable causes. In addition, an experimentation is not always possible, whereas surveys may be performed. In this context, the searched relations cannot be expressed in functional form, since one may better describe *tendencies*, so that relations assume a *probabilistic* character. Thus, this leads to either *stochastic modelling*, where stochastic variables are involved, or deterministic models, limited to a part of the total data variation, that is singled out and explained by the built model.

It must be emphasized that the observations made by the scientists are

based on their culture, so that the same phenomenon, the same landscape, the same object, may be observed by different scientist in a different way, according to their culture, experience, and the paradigms specific to the science framework in which the investigation takes place. In this way, they *select* the information in which they are interested, that composes the data set collected during their observations. As well, the results derived from the observations are produced and expressed in the frame of their science and contribute to its progress. Nevertheless, there is a phase of elaboration of the collected information that may be considered common to different scientists, because it is based to methods that do not belong to their reference scientific environment. It is the *data analysis* phase and it is situated between the data collection and the interpretation of the obtained results (Figure 1).

3 What is data analysis?

In the context of the experimentation, once a data collection was done, the scientist must study the observed data, to extract from them the contained information, aiming at drawing his conclusion. Data analysis is the set of mathematical and statistical techniques, based on computer science methods via computers use, useful for a scientist to read, synthesize, and understand the data he collected for an investigation.

Data analysis methods are so many, that it became nowadays a discipline in its own, getting an identity different from those that contribute to its methods. Actually, whereas computers are the primary tools for data analysis, computations are performed through mathematical methods, and statistics is taken into account every time one deals with populations, samples, indexes, etc.. Nevertheless, one has to recognize that data analysis, although an intersection of the previous matters, is in some respect different from all of them. In addition, since data analysis must always deal with existing empirical data and their explanation in the frame of an existing scientific discipline, one may not forget that its use cannot be considered outside such frame.

4 Why a methodological pathway?

One may say that, at the very beginning of a study concerning a phenomenon, very little of it, or nothing at all, is known, whereas at the very end, most of

it should be understood, in order to take advantage of the acquired knowledge for any possible use.

Since the acquired knowledge is supposed to become a known truth, it is necessary to take special cautions, in order to proceed from the beginning to the end through a methodological pathway, allowing to state that the results obtained by the investigation may be considered as true.

It is advisable to accept that a scientific study should be previously *designed*. Design a study means first of all check its feasibility and its usefulness, in particular considering the time and the costs involved. Then, the several investigation components, in particular the data collection, the data treatment (manipulation, computation, analysis), and the expected results, must be tied, in order to check for their compatibility. Thus, one must verify the needs of his theoretical updating or improvement, to be able to adequately face the problems he may meet, and consider the opportunity of inter-disciplinary cooperation with some experts. Eventually, timing and action modes may be well organized, in order to exploit at the best the available resources.

In this frame it must be taken into account that both the scientist's approach and the data analysis tools to be used should be different, according to the different possible aims of an investigation. In addition, one may say that, within the same investigation, different phases occur for the formation of the knowledge. To each phase correspond different aims and, as a consequence, a different approach is necessary and different methods and tools are supposed to be used.

5 Three steps for the investigation path

The process of knowledge acquisition going from the very beginning to the very end of an investigation, may be sufficiently well identified as a path with three steps, corresponding to three different methodological phases. It is important to stress their differences to suggest different approaches for each phase, since only their knowledge may indicate to the scientist the caveats to notice, particularly in data analysis, since different tools may be specific to a particular step, and they may not be used in the others without care. The three steps, as proposed by Camiz (1993; 2001) may be defined as follows:

- *Exploratory phase*, in which one defines a frame of reference for his work, defines the aims of the investigation, and begins a data collection. In this

phase the analysis of the data aims at searching structures and relations, in order to have a general idea of the phenomenon that may allow to formulate some hypotheses. Data are thus submitted to *exploratory data analysis*, to recover as much synthesized information as possible, in order to reveal any existing data structure and, in particular, to see whether or not the research aims are reachable on the basis of the collected data.

- *Confirmatory phase*, in which one may want to test whether the stated hypotheses are true or not. The information obtained by the previous analysis is used to setup the experimental design necessary for the organization of a correct sampling suitable for hypothesis testing. This should clarify the relations among characters, among units, and between characters and units. In this phase, *confirmatory analyses* are applied to the data, in order to figure out the assumed relations, through the tests of hypotheses. The sampling requires particular care to allow the statistical inference of the results to the reference population and statistical techniques are used to reveal the significance of the detected relations.

- *Modeling phase*, in which a theoretical and formalized (mathematical) description of the observed phenomenon is given, able of giving evidence on the relations observed. In this phase, data must be used for *calibration*, say the estimation of the correct model's parameters, and to test the effectiveness of the *model*, so that the most different data sets are used to evaluate its application range. *Simulation* is an outcome, either to show the effectiveness of the model to depict the studied phenomenon, or to derive further information from *simulated data*, i.e. data produced by the model and not from direct observation. Another outcome is *forecasting*, i.e. the ability of the model to predict the behavior and the evolution of the phenomenon itself.

The advantage to proceed to a modeling after a better knowledge of the data structure is the acquired ability to tailor the model in order to fit the data in the best way. In this sense, one may roughly consider it a *Marxist* approach to modeling, since the model is derived from the observation, in opposition to the *Hegelian* one, where a model is imposed to the reality. One may wonder if many doubts concerning pollution, glasshouse effects to the climate, and the evolution of particular ecological environment may depend sometimes on a questionable definition of the models used for the interpretation

of the phenomenon.

5.1 Exploratory studies

In the exploratory phase, the scientist defines a framework for the investigation and some aims; consequently, he collects some data, that are then inspected in order to understand their content, in particular aiming at identifying some structures, able to give some general idea of the key-elements of the phenomenon. It is the identification of structures that leads the scientist to the formulation of some hypotheses concerning the relations among the elements composing the phenomenon.

In order to inspect the data, rather than reading them all, a cumbersome work very difficult to carry on successfully, one may apply to some instruments able to reveal the information contained in the data in a synthesized way. Data analysis tools are able to reorganize the data, in order to reveal the structures that may exist. Such structures represent a way to synthesize the information contained in the data, since exploratory data analysis aims at describing the data information content through a *strong synthesis*. In particular, this leads to describe which relations exist among the characters considered during the observation and which resemblance can be detected among the observed units.

For such syntheses the *important* information is produced in graphical form. The hypothesis underlying the exploratory techniques, say their paradigm, is that the important information contained in the data may be extracted via some mathematical techniques, based on *association measures*. Two main graphical approaches are generally used, based on two different mathematical models:

- the *objects* are represented as *vectors* and *clouds of points* on an affine plane, forming scatter diagrams, useful to show the influence of characters on the observations and to find *factors* that best describe these influences;

- the *objects* are represented as *nodes* of a graph, whose *edges* represent *relations*; particular graphs used are *dendrogram trees*, useful to build data *taxonomies*, thus showing structures and suggesting partitions, obtained by cutting the trees.

Both approaches are useful to reveal the structures contained in the data, since they lead to the identification of

- *factors*, the characters that best explain the objects diversity; they may be merely descriptive but they are useful to represent the objects position in their respect; in many cases they are also interpreted as the causes of the diversity.

- *classes*, building homogeneous classes of objects allows to consider the obtained partition as a structure of the considered set.

Summarizing, an exploratory study pathway may be outlined in the following way:

- an argument of investigation is decided;
- references in the literature are searched;
- some data are gathered:
  - by collecting them from the literature;
  - by performing observations:
    - concentrated to the situations of interest,
    - aiming at identifying interesting objects, and
    - their relations with possible causes or other elements.
- data are submitted to exploratory data analysis tools:
  - *descriptive statistics*, to check for consistency of data, existence of outliers, etc.
  - *factor analyses*, to reveal the main sources of information, to be considered as the best references for the representation of data;
  - *hierarchical cluster analysis*, to build a complete set of encapsulated partitions, able to reveal the structure of the objects set; this allows to search later the partitions having optimal qualities;
- strong syntheses are thus searched on
  - *factors* describing objects diversity,
  - *classes* of homogeneous objects.
- the results are translated into the reference framework.

Figure 2: Localization of a plant species in Sardinia, Italy.

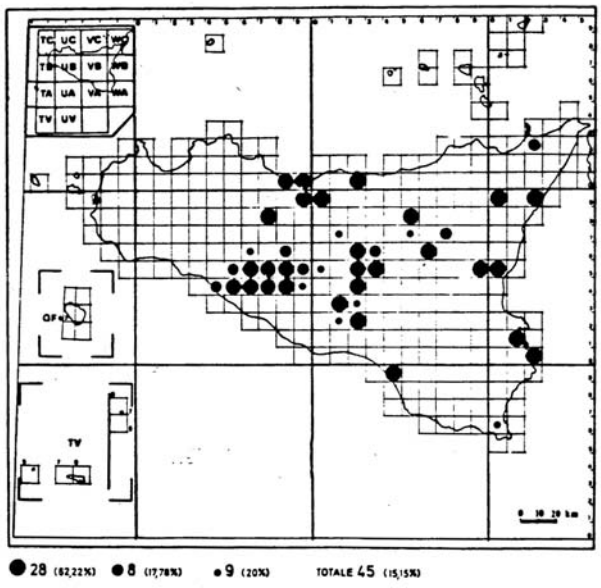

Figure 3: Distribution of a bird species in Sicily, Italy.

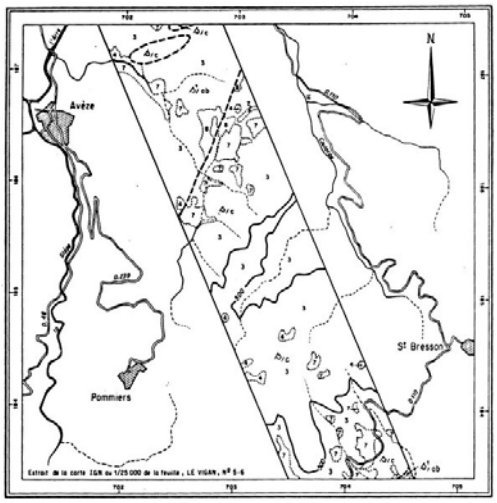

Figure 4: A transect for the study of vegetation in South France.

In the frame of the exploratory studies, several ecological studies may be considered: in particular all those concerning the description of the flora, fauna, and vegetation, as well as their distribution; the description of some environmental ecological characters, such as morphology, pedology, climate, pollution; the studies aiming at identifying types of plant and / or animal communities. For this aim, the sampling has no particular interest, since it is limited to the studied area, so that a map where dots indicate the position of the observation, if necessary included in a grid, may be sufficient: this is the case of the localization of particular plants in Sardinia (Figure 2) or the distribution of birds in Sicily (Figure 3). For studies concerning environmental factors, transects are often taken into account, oriented in the direction of the factor, such as the one considered for the study of vegetation in South France (Figure 4).

The strong synthesis of the phenomenon obtained by exploratory factor analysis and classification is usually a good starting point for further, deeper investigations. Nevertheless, the interpretation of the results of these analyses is sometimes sufficient for a first knowledge of the phenomenon itself.

Some limits of the exploratory analyses should be put in evidence, partly depending on the used tools that do not consider any statistical paradigm, but are limited to the results of mathematical procedures leading to optimal data representations. In fact, since no particular attention is usually paid to data collection in this phase, the obtained synthesis is only a synthetical description of the data at hand. Thus, neither it may be inferred to reference populations nor it may concern other (similar) data.

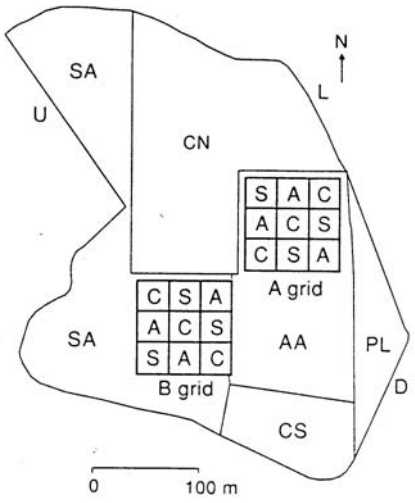

Figure 5: A Latin square sampling in a study area.

It is possible that the proved capacity to help the scientist in formulating hypotheses depends on the *natural* way used for performing the synthesis. In this sense one may understand the Benzécri (1982) statement, that data are studied without attributing to them any interpretative *a priori* model. Anyway, it is a warranty of the interest of the information that results, with the caution that the hypotheses formulated based on this synthesis ought to be validated through other suitable methods.

5.2 Confirmatory studies

In this phase, the scientist is concerned with the need of selecting, among the hypotheses he formulated in the previous phase, which are acceptable and which are not. It is then a phase where *knowledge* is acquired and *decisions* are taken. For this reason, in order to achieve these tasks, it is necessary to ground most of the investigation on statistical methods, the only able to test the formulated hypotheses and decide whether they are acceptable or should be rejected. For this decision a precise statistical procedure should be used (see, e.g., Mood *et al.*, 1974, Droesbeke, 1997).

For this task, the scientist must consider the experimental design for his sampling, in order to get data suitable of statistical treatment. In particular, he may ensure that the hypothesis may be testetd, the important issues concerning the phenomenon are not ignored, all variations of factors are taken into accountin the sampling , and compatibility of aims, data, and analysis

methods is assured. The sampling performed on this base may be either systematic, taken along either transects crossing or grids covering the study area, or random, choosing at random the sites as points in the space, cells of both transects and grids, or random walks. Special samplings, like *latin squares* are useful, when minimizing the number of observations is important (Figure 5).

With these data, the scientist may build some relations and test the hypotheses concerning these relations; as a consequence, the results of the investigation may be inferred to a reference population, with controlled risks of being in error. For this aim, the data are treated through classical and inferential statistics methods, such as parameters estimation, relations and influences between factors and characters, through regressions, identification of homogeneous classes and relations with factors, through analysis of variance, discriminant analysis, etc.. Both point estimation and confidence intervals are among the possible issues of this phase.

In order to achieve this phase, one may test if the stated hypotheses are acceptable, if it is possible the statistical inference of the results, and, eventually, if it is possible to gather several hypotheses in a theoretical structure, thus, something that may lead to the modeling phase. In the confirmatory studies frame one may consider the estimation of a population density, the amount of a pollutant in a region, the response to an ecological factor, the identification of the ecological niche of a species, the relation between some factors and the amount of a pollutant, etc.

It is interesting to consider that a synthetic knowledge acquired through a previous exploratory analysis gets more effective the organization of this phase, since in particular it is possible to identify and fix measure or input errors, identify possible outliers and decide how to handle then, withdraw objects without relation or influence with the study targets, select the character of higher interest, etc.. In addition, some knowledge of the structure of the phenomenon, and in particular of the referenced population, can lead the scientist during the experimental design. It is not to be ignored that exploratory analysis may provide as well a frame of reference for the report of the confirmatory results: something not strictly necessary, but helpful for a nicer and easier description of the study.

5.3 What is modeling?

Once that several hypotheses have been confirmed, the scientist may try to tie them in a theoretical structure. The structure may be formulated in a *mathematical model*, usually a set of equations that describe the relations among the most typical characters that appear in the phenomenon. It is clear that during the formulation of the model, attention must be paid to fit all kind of data collected, no matter how. For this reason, the previous steps are helpful, since enough knowledge of elementary relations has been already built. In any case, the model is usually continuously adjusted with a feedback process, based on tests performed on available data.

A model is a point of connection between empirical data and the theoretical paradigms of the reference discipline, since it justifies the empirical data on the basis of the paradigms but it validates as well the paradigms through empirical data. The use of mathematics for its formulation is due to the fact that «*nature is a book written in the language of mathematics*» (Galileo Galilei). With a consistent mathematical model the phenomenon may be simulated, at least concerning its main characters. In this way one can simulate the data as they were resulting from the observation of the phenomenon in reality. It is clear that simulated data should agree with both exploratory and confirmatory analysis results. This may be a good way for both *calibration* and *validation* of the model.

The construction of a model must observe some constraints: in fact it must be compatible with the data, in the sense that is must fit them at the best. In this sense, the issues raised both in the exploratory and in the confirmatory phases should be taken into account. Albeit constraining, these previous steps are particularly useful, since they may drive towards a *natural* modeling, based on the revealed structure on one side and on the estimates and decisions made possible on the other. In this way, the model is a synthesis of various components, already independently studied in detail, and the paradigmatic meanings are added to a structure already validated.

6 Interaction between scientist and data analyst

As it was said, data analysis domain is now very broad and a specialist is required to effectively exploit it at its best. In addition, new data analysis methods may be required in order to better suit the problems at hand during

an investigation. Unfortunately, in the practical activity, it is frequent that a data analyst is applied only during data computations, when problems arise. So, its role remains very limited, since the investigation aims were already set, the characters were chosen, the sampling was already decided, and the data were already collected. In this case, his role is limited to the best way of computing analyses, something that may be even misleading, with a loss of time and energy, if the previous investigation did not observe the rules necessary to make an effective use of the statistical tools. Instead, the ideal role of data analyst is to be close to the scientist along all his investigation, that may deal in this way:

- the scientist considers a problem and submits it to the data analyst, with his ideas on a possible investigation procedure;
- the analyst translates the problem in his own language, shows his ideas, concerning characters and sampling;
- the scientist performs his experimentation or his survey;
- the analyst treats the data, gets results and shows them to the scientist: together, they translate the results in the scientist's own language;
- the scientist draws his conclusions.

The complexity of the environmental phenomena and the sampling difficulties suggest a strong interaction, in order to achieve an effective scientific study.

7 Case-study 1: pollution bio-indicators

Plants do not move: for this reason their state is highly dependent on the pollution condition of the site where they grow. For the same reason, their state may be an indirect indicator of the pollution state of the site. In addition, they may be subject to cumulate effects, that may not be detected by instruments. The analysis of damage measures of *Pinus pinea* L. leaves (Camiz et al., 2008) was performed, in order to identify the impact of pollution on leaves in a natural environment and to decide which micro-morphological and bio-chemical-physiological measurable parameters may be used as indirect indicators of pollution. In the following, it will be outlined how the limits of the experimental design could compromise the results of the whole study.

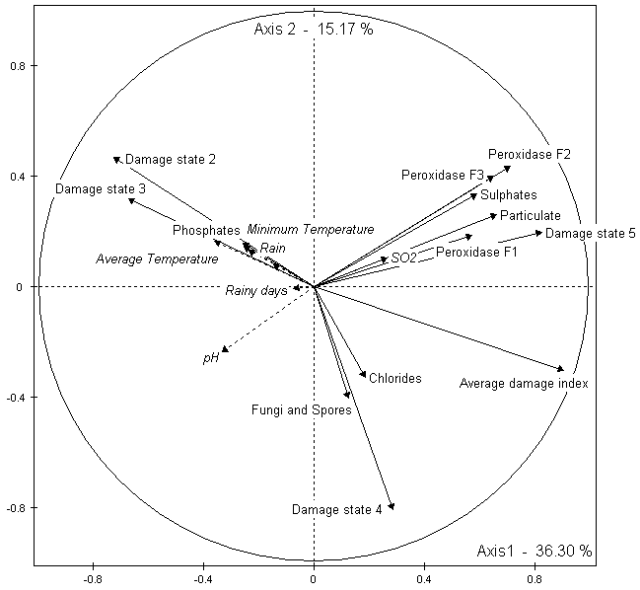

Figure 6: PCA of Pinus pinea damage data: representation of continuous characters on the first factor plane

7.1 Materials and methods

Pinus pinea L. leaves samples where collected from groups of three trees in three sites: in the center of Roma, a highly polluted site close to a road with important cars traffic; near Civitavecchia, in a potentially polluted area, close to the sea but 500 off the wind of a power station; in the Republic President's Castelporziano residence, a protected area close to the sea, considered a natural environment, without pollution. Samples of leaves of three different ages where taken in nine occasions, from February 1988 to February 1989, thus supposing to have an age series of leaves from 1 to 26 months belonging to three generations, namely from 1986, 1987, and 1988.

From the samples the following measures were taken (in parentheses are reported the abbreviations used in the tables):

- damage of *epicuticular waxes*, attributed to five classes (*Sta1 ... Sta5*) of increasing damage;

- relative frequencies of *Fungi and spores* (*Fusp*) and *particulate* (*Part*) in the epistomatic chambers;

- *peroxidase* (*POD*), according to three fractions: *solvable* (*POD1*), *ionic* (*POD2*) and *covalently tied to cellules* (*POD3*);

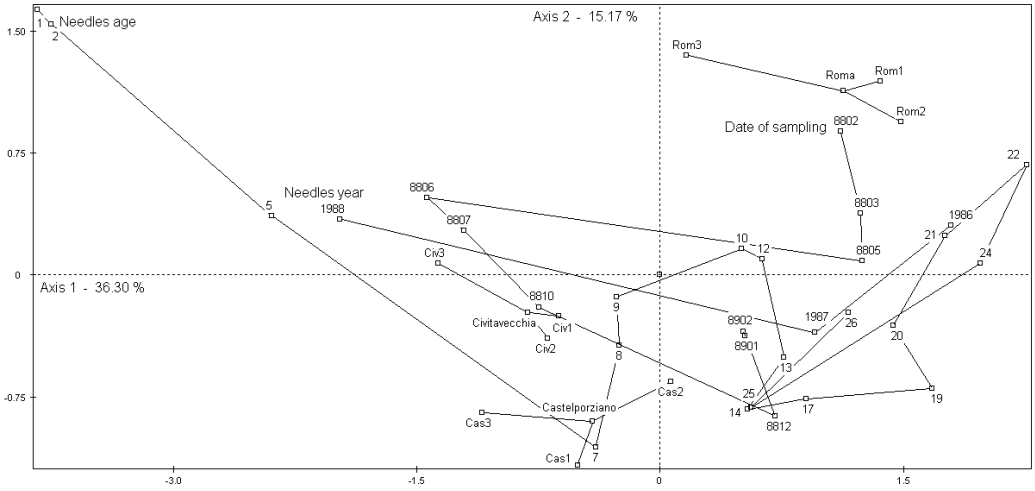

Figure 7: PCA of Pinus pinea damage data: representation of nominal characters on the first factor plane

- *ions* content (*Solf* = sulphate, *Fosf* = phosphate, *Clor* = Chlorides).

The data were submitted first to exploratory analyses, namely Principal Components Analysis of measures (*PCA*, Lebart *et al.*, 1984; Legendre and Legendre, 1983), considering the site characters as supplemental, then to Ascending Hierarchical Classification (*AHC*, Anderberg, 1973, Legendre and Legendre, 1983) based on samples factor scores: groups where then characterized through *typical* characters and measures *sensu* Lebart *et al.* (1995, χ^2 and Fisher's F tests), in order to identify their most important features. Then, for the test of hypotheses, the Analysis of Variance (*ANOVA*, Mood *et al.*, 1974; Legendre and Legendre, 1983) was used, to identify the differences among sites, dates of sampling, and leaves ages: for this aim, the Student-Newman-Keuls *multistage* tests (Miller, 1981) were used. Eventually, simple linear and non-linear regressions (Gallant, 1975; Kennedy and Gentle, 1980) where used, in order to model the variation of the measures in relation with the leaves age.

7.2 The results

The examination of the plane spanned by the first two principal axes of *PCA* (Figure 6) shows that to the first axis mainly contribute (set on the positive side) the *damage stage 5*, with *particulate*, *sulphate* and the three *peroxidase*, that indicate all a potential damage due to human activity, opposite to the *damage stage 2* and *3*, that have higher relative frequency in poorly damaged

samples, settled on the negative side: then, this axis corresponds to a *pollution* factor. To the second axis contribute mainly the damage state 4 with *fungi*, *spores*, and *chlorides*, set on the negative side, opposite to *damage stages 2* and *3*, showing then a factor closer to *natural ageing*. The third axis (not shown here) seems to correspond to an *environmental conditions* factor. The position of the characters on the same factor space (Figure 7) shows that the three sites are well separated: Rome is opposite to Civitavecchia and Castelporziano on the first axis, on the second axis they are set along in an increasing sequence, from Rome to Castelporziano, in accordance with the presence of *chlorides*, *fungi and spores*, whereas Castelporziano is opposite to the other two sites along the third axis, due to more natural conditions as described here by the presence of *phosphates, fungi, spores* and *chlorides*. Every site is surrounded by its three sampled trees, that suggests a homogeneous state of the site.

The *age classes* are in a sequence along the first axis, whereas the position of *sampling dates* is set in a coherent continuity that involves time, contrasting the different periods which are identified by biological cycles: samples picked in summertime are in fact in a lower damage space zone. The distribution of *needles age* along the first axis clarifies the close relationship that exists between the increasing damage (in particular the one due to human activities) and the *needles ages*, even if strongly conditioned by the *site*: samples which are 25 and 26 months old, not present in Rome, are placed on the second axis in the side of natural damage; in addition it must be noted that the pattern of *ages* in the first and second year is very similar, even if the second one is shifted to a higher damage zone, which suggests that the seasonal fluctuation of observed values must be investigated further, together with sampling data pattern.

The classification in seven groups of the leaves samples derived from *AHC* reflects this pattern: on the plane's upper right two groups of young leaves are found, very poorly damaged, then two groups of leaves from Castelporziano, with medium damage, a group from Civitavecchia, with medium-high damage, and finally two groups of Roma leaves, the most seriously damaged. These results were generally confirmed by the *ANOVA* results: site, age class, sampling date, and needles age, showed all significant differences for most indicators.

7.3 The models

According to the results found in the previous analyses and the observation of the scatter diagrams, different models were considered for the different damage indicators, as functions of the leaves age. The different conditions of the three studied sites suggested to consider them separately in the models building. In the following, only three such models will be shown, as an example.

7.3.1 Sulphate

For the sulphate, the evidence of a seasonal variation suggested to add to the linear effects a seasonal one, imposing a period of 12 months. It may thus be expressed as follows:

$$\text{Sulphate} = \text{Intercept} + \text{Slope} * \text{Age} + \text{Width} * \left(\cos \frac{2\pi}{12} * \text{age} + \text{phase} \right)$$

The most interesting parameter is by no means the slope, being around 85, 37, and 0 for Roma, Civitavecchia, and Castelporziano, respectively. This is an important indicator of the cumulate effect of the pollutants, as well as the recovering of the plant during the good season. The amount of variation by the models, as expressed by the R^2 coefficient, is in the three sites around 67%, 33%, and 11%, respectively (Figure 8).

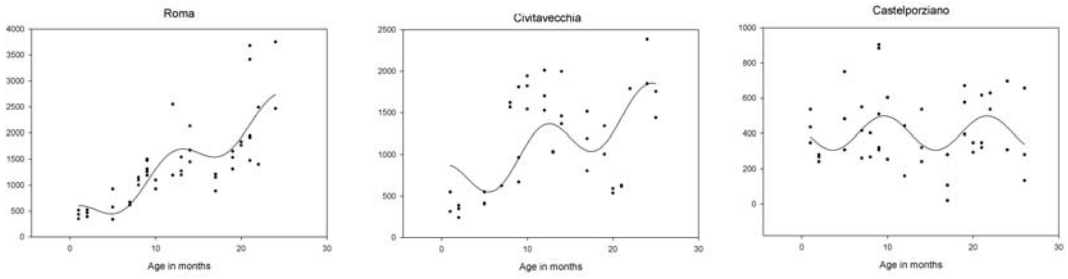

Figure 8: Sulphate values observed (dots) and predicted (line) by the regression model in the three study sites

7.3.2 Peroxidase POD1

For this Peroxidase no seasonal variation was noticeable, but the increase was clearly more intense than linear, so that a quadratic model was chosen, namely

$$\text{Peroxidase} = \text{Intercept} + \text{Slope} * \text{Age} + \text{Quad} * \text{Age}^2$$

In fact, only the quadratic term had a significant contribution, giving R^2 values of around 40%, 29%, and 70%, respectively for Roma, Civitavecchia, and Castelporziano (Figure 9). Actually, the behaviour of Peroxidase in Civitavecchia was not clearly understood.

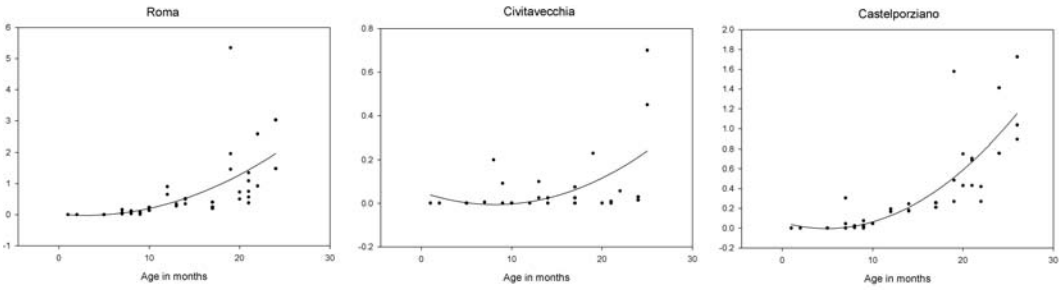

Figure 9: Peroxidase values observed (dots) and predicted (line) by the regression model in the three study sites

7.3.3 Waxes

For each waxes damage state, a gaussian model was fitted, namely:

$$\text{f}(\,\text{age}\,) = \text{Width}\ \text{e}^{\frac{1}{2}\left(\frac{\text{age}-\text{mode}}{\text{stdev}}\right)^2}$$

In Figure 10 the results are reported. From the observation of the graphs is easy to see that the distributions are quite different. The damage state 3 lasts for longer time in Castelporziano than in Civitavecchia and, even worse, in Roma. The damage state 4 has is maximum in Roma at 12 months, whereas in Castelporziano is at 14 and Civitavecchia at 17 months. The damage state 5 both in Roma and Civitavecchia reaches its maximum at 20 months, whereas in Castelporziano it is reached at the end of the period. The highest maxima of state 4 in Civitavecchia and Castelporziano, in comparison with state 5 may reflects the more natural character of these sites.

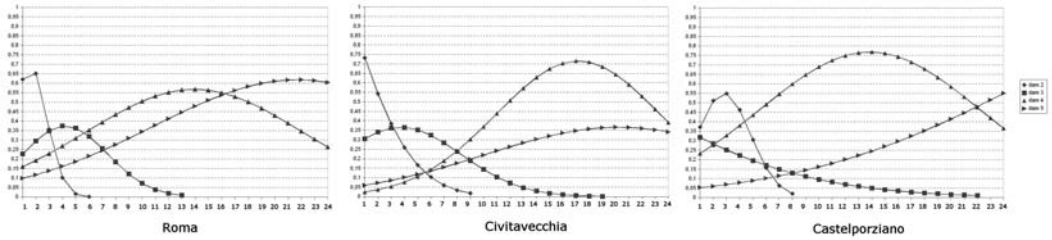

Figure 10: Wax damage stages predicted by the gaussian models in the three study sites

8 Case-study 2: sustainable development and estimation of herbivores diet

The estimation of plant composition of the diets of the herbivores, allows to evaluate the sustainable development of an animal species from the vegetal environment where it settles. The diet composition, in particular of wild herbivores, may not be checked directly, so that it must be indirectly estimated through the examination of non-digested parts of the plants that are found in the faeces. In particular, n-alkanes are saturated hydro-carbides that may be used as diets markers, as Dove and Mayes (1991) underlined, able at estimating the amount of food eaten by an animal. The different distribution of the different n-alkanes in different plant species suggested Dove and Moore (1996) to propose a method (implemented in their program *EATWHAT*) for the quali-quantitative estimation of the diet, i.e. the amount of the different plants eaten. They claimed that its application on their *Australian sheep* data (one diet of three species detected from the distribution of five alkanes) worked very fine, but this was not the case of Castelporziano *fallow-deers*, taken by Focardi and Piasentier (see Camiz *et al.*, in press), whose 13 diets were estimated to be composed only by *deschampsia*, although based on a 12 $\times$ 12 alkane $\times$ species matrix: a diet biologically impossible. Focardi, that worked in Castelporziano, wondered why the method did not work, arguing that it could maybe depend on the quality of the approximation of the alkanes distributions estimates.

8.1 The n-alkanes method

The problem of the diets estimation through the alkanes method may be formalized as follows:

- a herbivore species animal eats some amount of plants belonging to different species;

- every species is characterized by a distribution profile of some alkanes, empirically known, so that one may build a matrix that contains the alkanes profiles present in each species;

- in animals faeces an alkanes distribution profile is found, with higher concentration, due to the digestion of the rest and of a small part of the alkanes themselves;

- the profile of the alkanes in the animals faeces may be detected, whose relation with the diets profile is supposed linear, so that one should estimate the diet as a linear combination of species profiles;

- the diet is composed by non-negative amounts of the considered plants.

For the solution of the problem, the Dove and Moore (1996) model was defined as a *non-negative least squares problem*. Given

- H - the $n \times p$ matrix containing each alkane concentration in each plant species;

- f - the n-dimensional vector containing the alkanes concentration in an animal faeces;

- **r** - the n-dimensional vector containing faecal restoring of every alkane;

- b - the n-dimensional vector containing the faecal concentrations corrected by the restorings, say $b_i = f_i / r_i$ for each i;

- **x** - *unknown* p-dimensional vector, containing the amount of every species eaten;

the solution for **x** is given by the non-negative least squares problem

$$\| H\,\mathbf{x} - \mathbf{b} \|^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{p} h_{i\,j}\, x_j\, b_i \right)^2 = \text{minimum}_{x_j \geq 0,\ 1 \leq j \leq p}$$

The solution through a minimization procedure is due to the fact that the matrix H may be non-square. In addition, according to the different problem conditions, a solution may be always estimated, but it is questionable why a biologically impossible result was found in the case of the *fallow-deers*.

number	eigenvalue	percentage of variance	cumulate percentage	trace = 12.000000000
1	7.58442177	63.20	63.20	**********
2	1.72645357	14.39	77.59	********
3	.99981671	8.33	85.92	*****
4	.79965998	6.66	92.59	****
5	.61960218	5.16	97.75	***
6	.18495612	1.54	99.29	*
7	.04516704	.38	99.67	*
8	.02698427	.22	99.89	*
9	.01112096	.09	99.98	*
10	.00140554	.01	100.00	*
11	.00041187	.00	100.00	*
12	.00000000	.00	100.00	*

Figure 11: The eigenvalues of PCA of the alkanes *times* species matrix.

8.2 Some mathematics

Given the system $H\,\mathbf{x} = \mathbf{b}$, the least squares solution is an approximation of the exact one that, when it exists, may be roughly expressed as $\mathbf{x} = H^{-1}\,\mathbf{b}$. For this the matrix H must have maximum rank. Now, computing *H'H* eigenvalues, where H is the alkanes $\times$ species matrix of Castelporziano, it results that H as rank about 5 (Figure 11), that is the 12 species all belong to a 5-dimensional vector space spanned by the alkanes and that the solution is not unique and may be arbitrarily chosen among ∞^7 possible ones. The solution given by Dove and Moore (1996) is the point with non-negative coordinates on the space spanned by species closest to the orthogonal projection of the diet on this space.

Now, if we look at alkanes, species, and diets on the plane spanned by the first two eigenvectors of *H'H*, that is the first two factors of *PCA* of H (Figure 12), we find that all diets are set very close to the direction of *deschampsia*, probably due to its strong content of alkanes. This plane describes about 77% of total variation, so that it is sufficiently clear that in these conditions the solution given by the program depends on the strong collinearity in the matrix H, that is the strong linear dependence among both its rows and columns, and the strong content of alkanes in the *deschampsia*.

9 Conclusions

The analyses of the pine-tree leaves outlined clear differences due to the different pollution state of the three observed sites. It was confirmed that the pollution may rather be detected through the different indicators *trend*, than

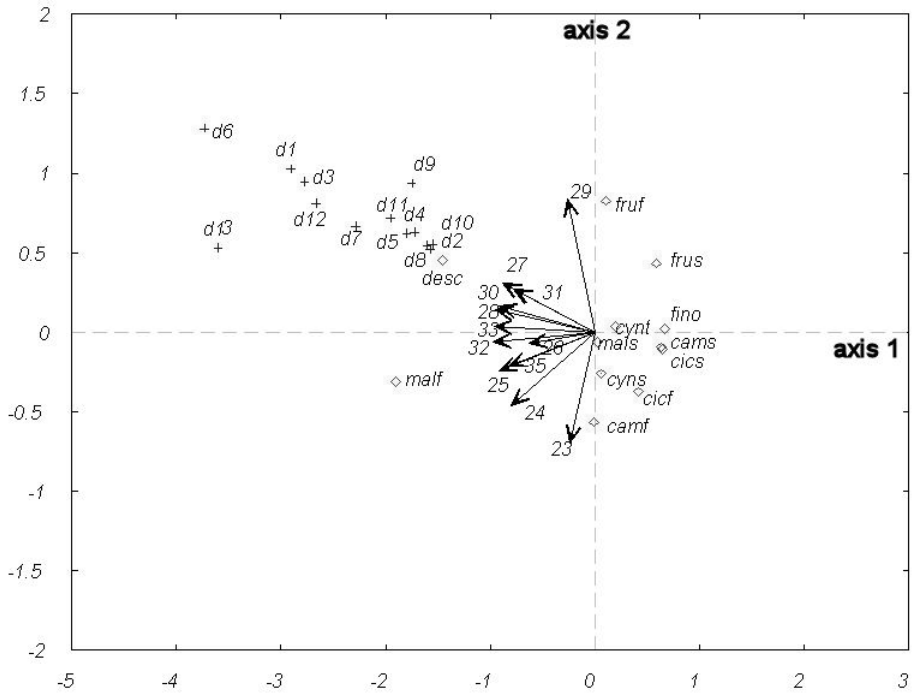

Figure 12: Plant species, alkanes, and diets on the first factor plane of PCA

from point measures, due to accumulation phenomena. Indeed, a true dependence model could not be developed, since the pollutants were not measured. Some indicators showed the plant's restoring capacity, represented by the seasonal fluctuations. So, plants effectiveness as pollution bio-indicators could be proved. The previous use of exploratory methods led toward better solutions, in particular concerning the choice of non-linear models, but mainly to easier interpretation of the results.

Some problems nevertheless seriously damaged the investigation, due to a bad sampling procedure. In fact, the seasonal fluctuation, limited as it was to only one season (observed on three generations of leaves) could not be sufficiently proved.

The problem of the fallow-deers diets, as it was set, cannot have a solution, since the *PCA* of the matrix H showed that the considered alkanes are not able to discriminate sufficiently well the considered species and some other markers may be searched. The problem must be re-stated in a more treatable way. Maybe a better way could be to estimate an *average diet* rather than each one.

The proposed case-studies show that exploratory techniques can be very useful in the investigation of the data structure, prior to the other steps. In fact

they were able both to drive the scientist in the following phases and to help in the understanding of the problems raised in the application of a previously built model. Their use, in particular as the first step in any scientific investigation involving data analysis, helps in preventing problems in the organization of the following steps. It is the author's opinion that the exploratory methods, presently considered as very specialized, should become more familiar to the scientist than both the tests of hypotheses and the modeling, since their use is less subject to the limitations of the other methods, that require a special attention to prevent serious errors.

References

[1] Anderberg, M.R. (1973). *Cluster Analysis for Applications*. New York, Academic Press.

[2] Benzécri, J.P. *et al.* (1982). *L'Analyse des Données.* 2 voll.. Paris, Dunod.

[3] Camiz, S. (1993). «STATIS Ordinations vs. the Juhász-Nagy Models: the Predictability of an Exploratory Tool». *Abstracta Botanica,* 17(1-2): pp. 29-36.

[4] Camiz, S. (2001). «Exploratory 2− and 3−way Data Analysis and Applications». *Lecture Notes of TICMI*, Tbilisi International Centre of Mathematics and Informatics, 2. See also: http://www.emis.de/journals/TICMI/lnt/vol2/lecture.htm.

[5] Camiz, S., A. Altieri, and F. Manes (2008). «Pollution Bioindicators: Statistical Analysis Of A Case Study». *Water Air and Soil Pollution*, n. 194(1-4): pp. 111-139.

[6] Camiz, S., S. Focardi, E. Piasentier, and A. Purpura (in press). «On the Methods of Estimation of the Composition of the Herbivores Diet through the n-Alkanes». Work in progress.

[7] Dove, H. and R.W. Mayes (1991). «The Use of Plant Wax Alkanes as Marker Substances in Studies of the Nutrition of Herbivores: A Review». *Australian Journal of Agricultural Research*, n. 42: pp. 913−52.

[8] Dove, H. and A.D. Moore (1995). «Using a least-square optimisation procedure to estimate botanical composition based on the alkanes of plant cuticolar wax». *Australian Journal of Agricultural Research*, n. 46.

[9] Droesbeke, J.J. (1997). *Éléments de statistique*. Paris, Ellipses.

[10] Gallant, A.R. (1975). «Nonlinear Regression». *The American Statistician*, 29: 73−81.

[11] Kennedy, W.J. and J.E. Gentle (1980). *Statistical Computing*. New York, Marcel Dekker.

[12] Lebart, L., A. Morineau, and M. Piron (1995). *Statistique exploratoire multidimensionnelle*. Paris, Dunod.

[13] Lebart, L., A. Morineau, and A. Warwick (1984). *Multivariate Descriptive Statistical Analysis*. New York, J. Wiley and Sons.

[14] Legendre, L. and P. Legendre (1983). *Numerical ecology*. Amsterdam, Elsevier.

[15] Mood, A.M, F.A. Graybill, and D.C. Boes (1974). *Introduction to the Theory of Statistics*. New York, McGraw-Hill.

Received September, 8, 2008; revised December, 20, 2008 accepted February, 2, 2009.