UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Laurea Magistrale in Scienze Chimiche (LM-54)

Anno Accademico 2024-2025

Corso di Chemiometria

4 crediti di lezione frontale (32 ore)

2 crediti di esercitazioni numeriche in aula incluse all'interno delle lezioni frontali (30 ore)

Prof. Ilario Losito

e-mail: illosdid@hotmail.com

Sito Internet: http://puccini.chimica.uniba.it/~losito/indexChemiometria.htm

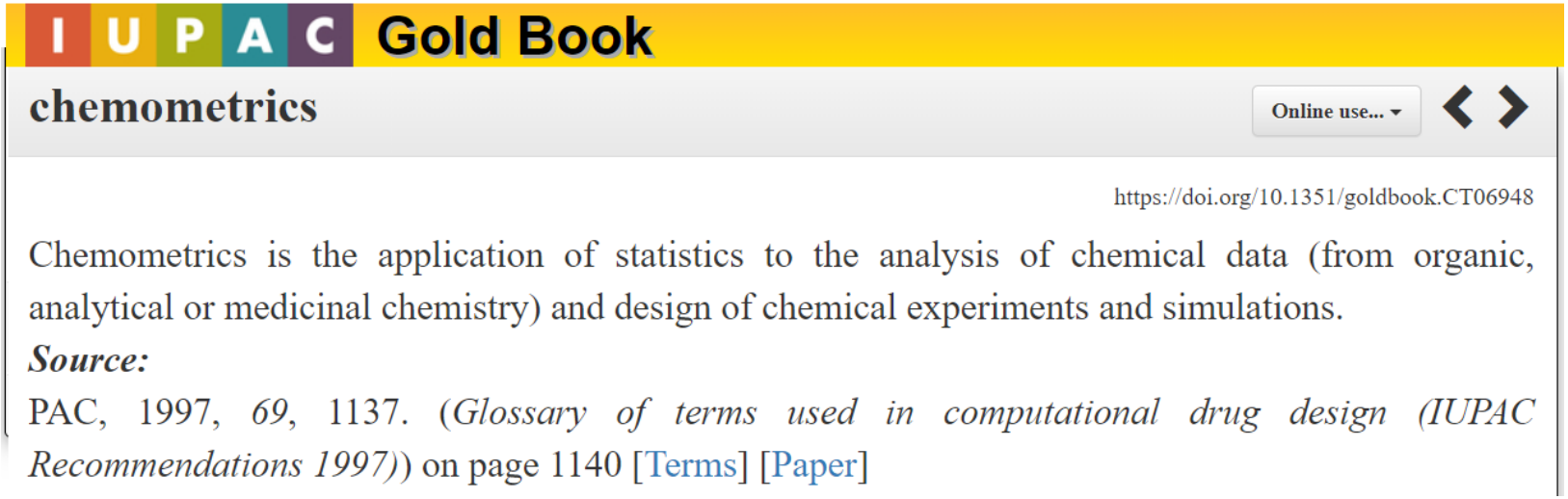
A general definition for Chemometrics

International Chemometrics Society (ICS) has given the following definition for Chemometrics:

A chemical discipline that uses mathematical and statistical methods to:

1) design/select optimal procedures and experiments
2) provide maximum chemical information by analyzing data
3) give a graphical representation of this information

This is the definition of Chemometrics given in the **IUPAC Gold Book**:



The birth of Chemometrics was promoted by Svante Wold, from Umea University in Sweden and Bruce Kowalski, from University of Washington in Seattle, USA.

In June 1974 they sent a letter to the *Analytical Chemistry* journal, proposing to call Chemometrics a scientific sector including all mathematic techniques aimed at elaborating and modelling ensembles of chemical data.

They were also the founders of the **International Chemometrics Society**, sending a letter to all potentially interested colleagues and inviting them to participate to a newsletter.

The letter sent by Wold and Kowalski

Dear Prospective Chemometrician:

Although statistics has been a tool of the chemist for many years, the recent literature shows a substantial increase in the number of novel applications of statistics and non-statistical mathematics to problems in the field of chemistry. The reasons for this increase are

.....

Our first idea is to publish a newsletter containing the membership list and a summary of what we have learned from responses to this letter. Therefore, we ask for your comments, suggestions, and most importantly, your indication of interest in the society. Please send your full name, address, telephone number. It would be very helpful if you could include a bibliography of past and current papers and manuscripts that are in for publication. In addition, we would appreciate your notifying us about prospective members among your colleagues. If all the members cooperate in this endeavor, the first newsletter will be a valuable aid for anyone involved in the application of statistical and mathematical methods in chemistry.

Hoping to hear from you soon

For the Chemometric Society, yours faithfully

Bruce R. Kowalski
Laboratory for Chemometrics
Department of Chemistry
University of Washington
Seattle, Washington 98195

Svante Wold
Research Group for Chemometrics
Institute of Chemistry
Umea University
Sweden

In December 1983 Prof. Forina from University of Genova became the President of the International Chemometrics Society:

CHEMOMETRIC

No 10 (December 1983)

NEWSLETTER

Election of new officers

During the NATO ASI on Chemometrics, the congregation met to select new officers for the Chemometric Society. A nominating committee proposed Professor Michele Forina, Universita Degli Studi, Istituto de Scienze Farmaceutiche, Viale Benedetto XV, 3, 16132 Genova, Italy for president and Dr. Robert Meglen, Director, Analytical Laboratory, University of Colorado-Denver, 1100 14th Street, Denver, CO 802027 USA for secretary. These selections were unanimously approved by 100 chemometricians. I hereby request that the membership approve the selection and that the two officers take office in 1984.

Bruce R. Kowalski,
President

Prof. Forina was a pioneer in the field of Chemometrics and the author of computer programs for some of the earliest applications reported.

PARVUS was the first computer module for Chemometrics application (pattern recognition) developed by Forina, recognized as a milestone in the history of this discipline:

Computer Corner

PARVUS

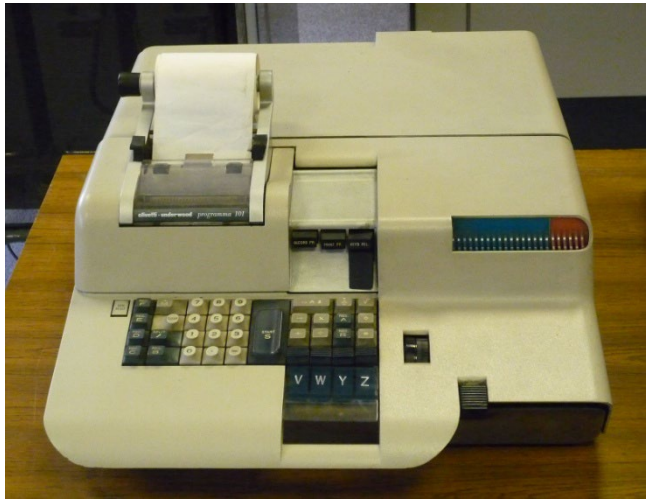
In this contribution, Prof. Forina describes a package for pattern recognition, which to my knowledge must be the most complete such package written directly for microcomputers. It contains almost all display methods and supervised learning methods currently used by analytical chemometrists. Professor Forina was recently appointed President of the Chemometrics Society. He succeeds the President-Founder of the Society, Prof. Kowalski from Seattle.

D. L. MASSART

Trends in analytical chemistry vol 3, n.2, 1984

At the start of his work on Chemometrics Prof. Forina used two among the first personal computers available in the world, Olivetti Programma 101 (P101) and P6060.

Olivetti Programma 101



Model exposed at the National Museum of Computing – Milton Keynes, UK

Model exposed at the Museo Nazionale della Scienze e della Tecnologia «Leonardo da Vinci» - Milan

Olivetti P101 was developed in Italy between 1962 and 1964 and then presented as prototype at the BEMA exposition in New York, in 1965. It was produced between 1965 and 1971 and sold at a launch price of 3200 dollars. About 44000 units were sold worldwide but 90% of them were purchased in North America. 10 P101 were also purchased by NASA and used for calculations in the APOLLO missions.

This "computer" had a maximum RAM of 1920 bits (about 10 million times lower than RAM memories currently used in personal computers).

https://it.wikipedia.org/wiki/Olivetti_Programma_101

Olivetti P6060

Olivetti P6060 was developed in Italy in 1970s and presented at the Hannover Fair in 1975 (the same year in which Apple I was released).

It is considered the first world personal computer that included also a floppy disk (8 inches) unit.

It also included a printer and a plasma display with 32 orange alphanumeric characters.

The RAM memory was 32 or 64 Kbytes.

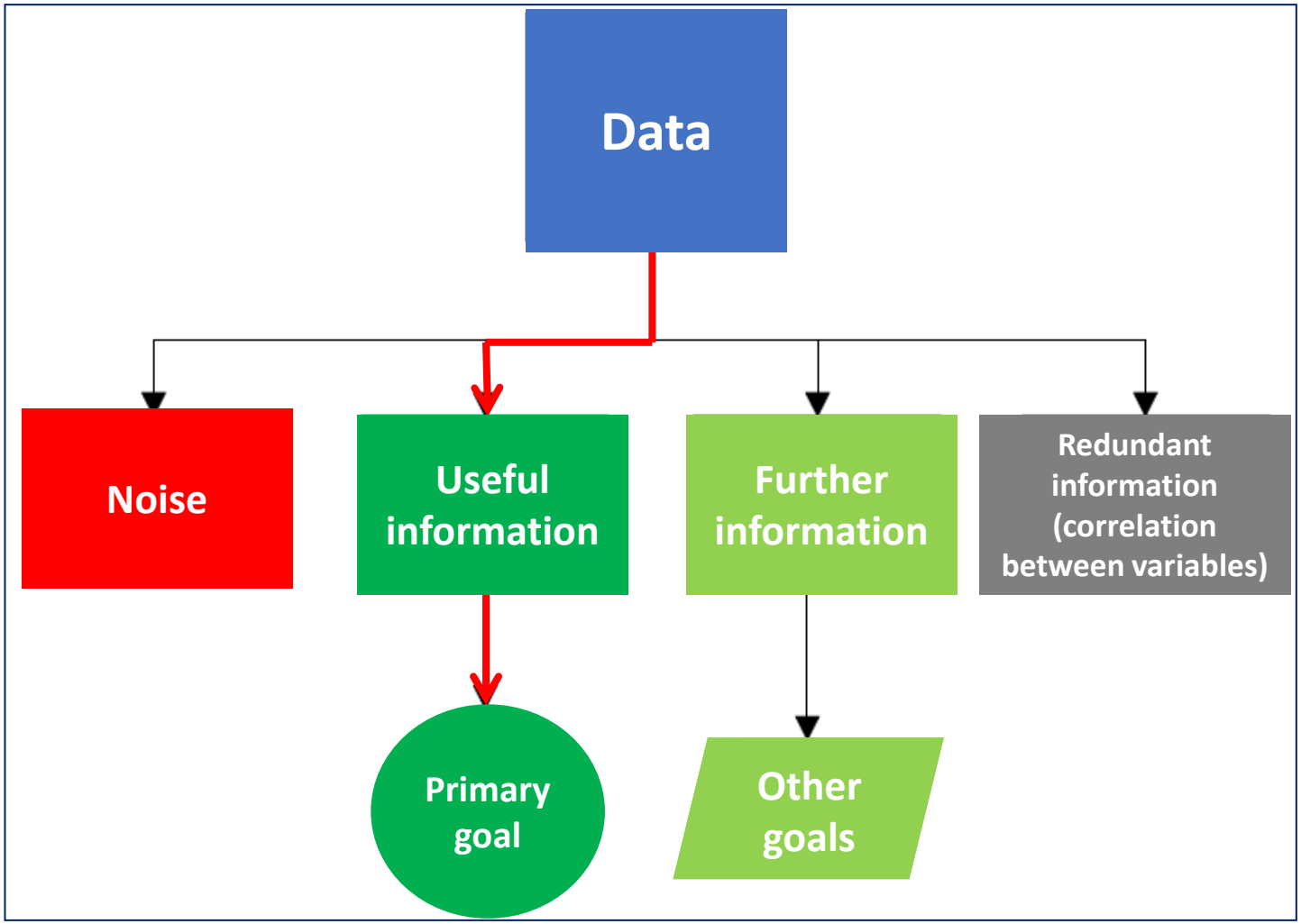
Despite the relatively short dimensions (60 x 60 x 25 cm), the P6060 was very heavy (50 kg!). A lighter model was released afterwards.

https://it.wikipedia.org/wiki/Olivetti_P6060

High-complexity problems often faced by Chemometrics

- ✓ In many real cases a process/system cannot be described by a well-defined theory
- ✓ Usually, many variables are involved but not all of them can be controlled at the desired level of precision and their relevance could not be known
- ✓ Experimental noise might eventually mask the real effects of variables
- ✓ Correlations between two or more variables might be present, thus the effects of variables cannot be studied using a one-at-a-time approach
- ✓ Synergic or antagonist effects may be present
- ✓ Non-linearity between descriptors and response may be present
- ✓ Adequate sampling and/or experimentation are sometimes impracticable

Methods based on Chemometrics try to separate useful information from other types of information contained in data:

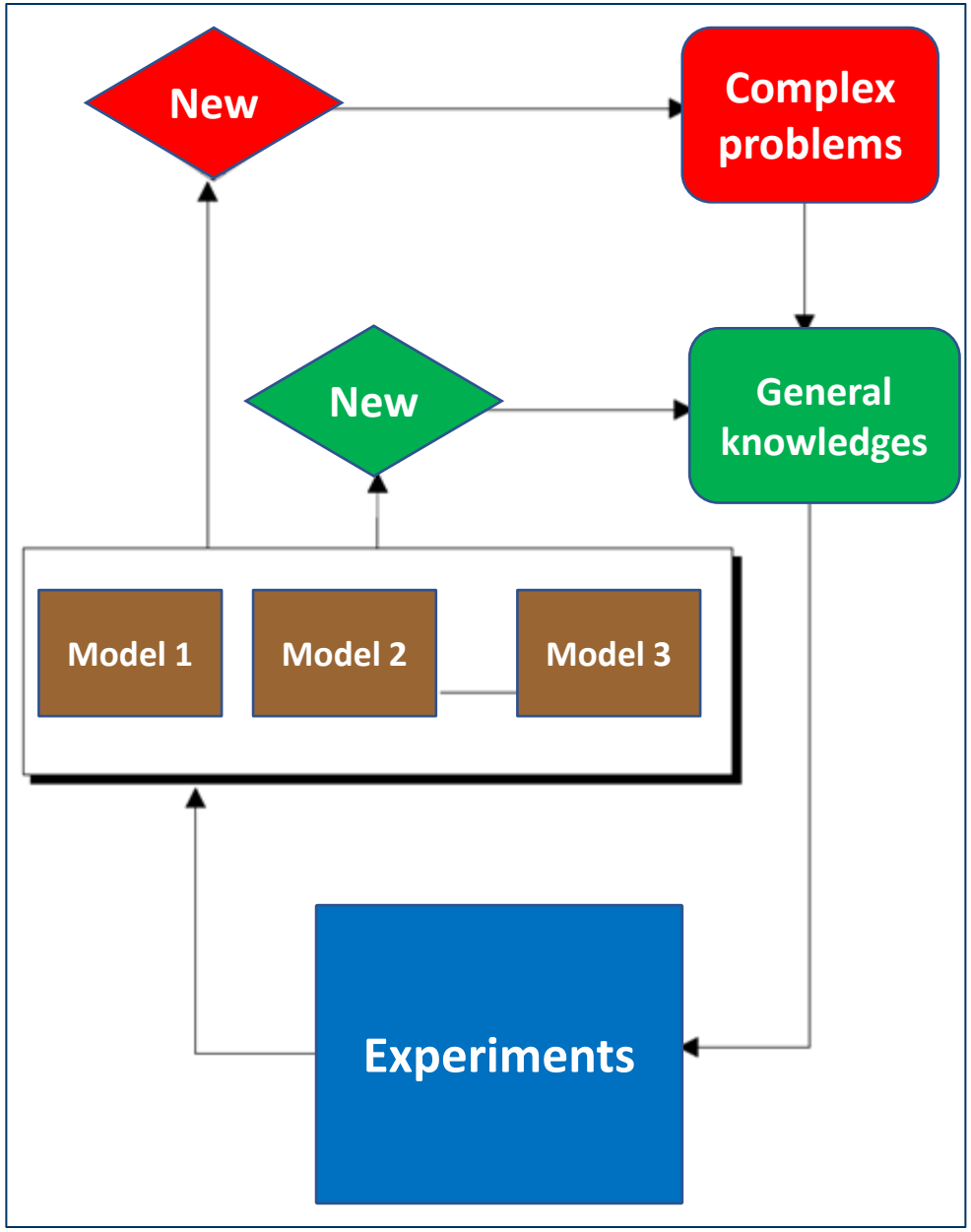


Adapted from R. Todeschini, Introduzione alla chemiometria, EdiSES

In the absence of explicit theories, the approach to high-complexity problems must be reversed, i.e., useful information is extracted from experimental data, including those obtained from purposely designed experiments.

One or more local models are thus built to face the needs posed by a specific problem, yet their acceptance depends on their coherence with basic knowledges and on the adoption of validation criteria to evaluate their reliability.

A qualifying feature of Chemometrics is the systematic use of validation techniques, able to estimate the predictive power of models.



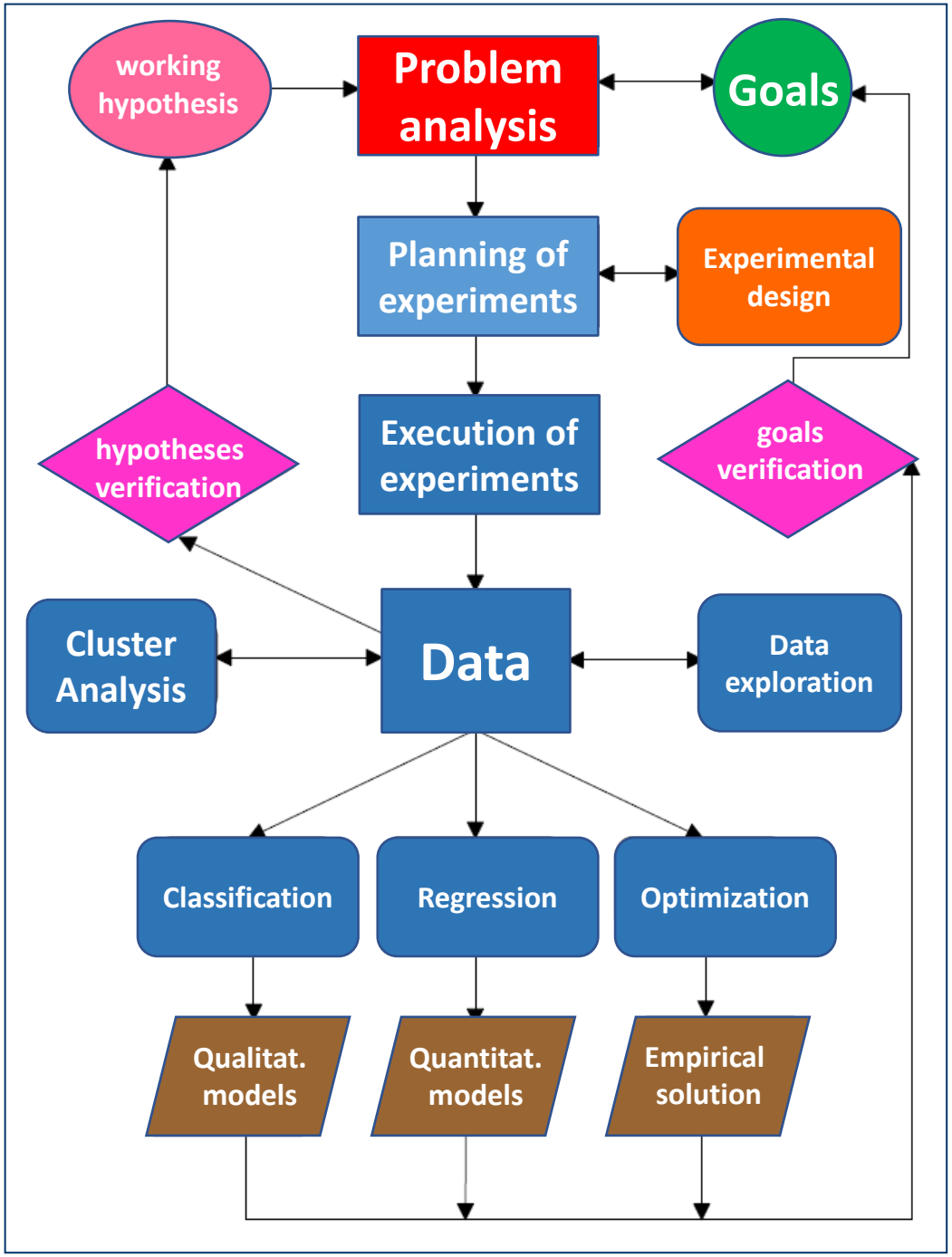
Adapted from R. Todeschini, Introduzione alla chemiometria, EdiSES

Flow chart for the analysis of a complex problem

The ensemble of chemometric strategies and of relationships between different procedures can be visualized in the flow chart on the right, starting from the problem analysis.

Goals are individuated and working hypotheses are formulated, considering also the availability of previous experimental data.

If these are not available, experiments need to be planned.



Adapted from R. Todeschini, Introduzione alla chemiometria, EdiSES

Experimental design

This is a crucial step: a good experimental design enables the attainment of experimental data with a high content of information.
Moreover, experimental design helps in performing experiments under the best possible conditions: minimum number of experiments, minimum costs, etc., thus leading to the best **quality/effort ratio**.

Data exploration

Supposing data are represented by ***n* objects** (samples), each described by ***p* variables** (for example, the concentrations of different compounds), an explorative analysis is performed first: statistical parameters (mean, variance, etc.) are obtained for each variable; correlation between variables is evaluated; the eventual presence of outliers is assessed, etc.

Principal Component Analysis is currently considered as one of the best approaches for the initial exploration of data related to multivariate systems.

Similarity analysis

Cluster Analysis enables the evaluation of similarities between analyzed samples, thus the attainment of additional information on the presence of groups (***clusters***).

Cluster Analysis is inevitably affected by a certain degree of subjectivity, depending on the number and nature of homogeneous groups of samples included in the dataset.
Anyway, if data elaboration confirms the recognition of those groups of samples, they can be raised to the rank of **classes**.

Classification methods

Once objects (samples) are assigned to different classes, classification methods enable the construction of classification models, i.e., mathematical models, depending on variables and classes, that can predict the association of a new sample to one of the recognized classes. If its validity is considered general, the model becomes independent on the specific set of samples adopted for its elaboration (***training set***).

Regression methods

Regression methods are used when another complex problem needs to be solved, i.e., when a mathematical model needs to be developed to predict quantitative values of a variable (for example, the analytical response) from those, known for ***n samples***, assumed by an independent set of variables (the **predictors**).

A typical example of regression is represented by a calibration line in analytical chemistry.

As for classification methods, if a regression method is considered generally valid, it becomes independent on the specific data set employed for its construction.

Optimization

In some cases a more or less consolidated solution exists, thus the goal is simply to improve that solution. Optimization can be the approach adopted for such improvement and usually leads to an empirical solution.

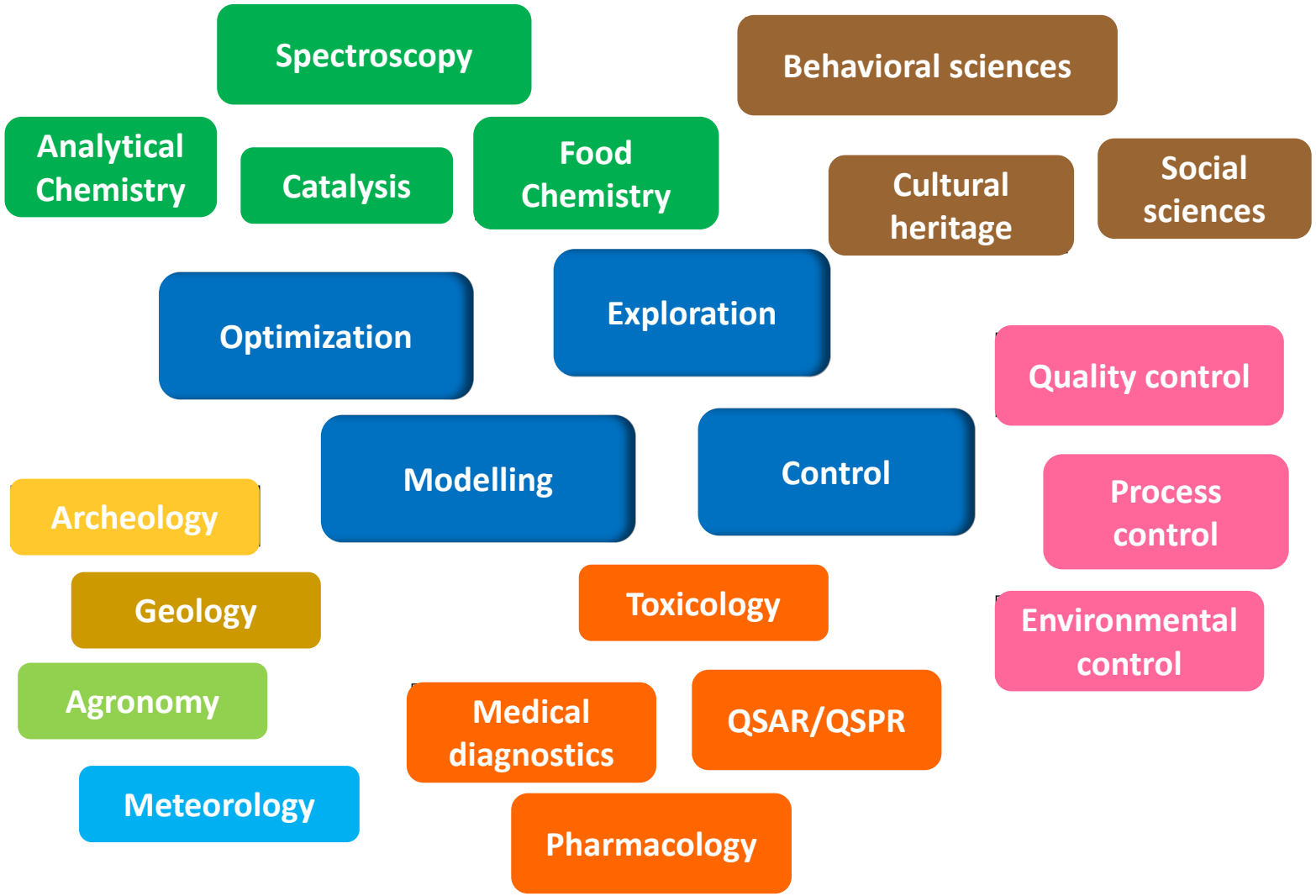
Current Chemometrics include the following main branches:

1) Modelling, classification and regression methods
2) Similarity analysis
3) Principal Component Analysis and related methods
4) Methods for experimental design and optimization
5) Expert systems and artificial intelligence methods
6) Strategies based on neural networks

In any case, several chemometrics methodologies are related to applied statistics and provide the opportunity to test many techniques proposed by statisticians.

Methods designed by mathematicians and statisticians to solve multivariate problems have progressively become useful not only to face problems related to scientific disciplines, like chemistry, pharmacology, environmental sciences, but also for humanistic ones, like social sciences and behavioral sciences, especially when complex data are available.

Main fields of application for Chemometrics



Adapted from R. Todeschini, Introduzione alla chemiometria, EdiSES

Consolidated specific applications

- ✓ Product study and optimization

  Food products, drugs, detergents, emulsions, alloys, fuels, additives, paints, cosmetics, textiles, composite materials, lubricants, reaction mixtures

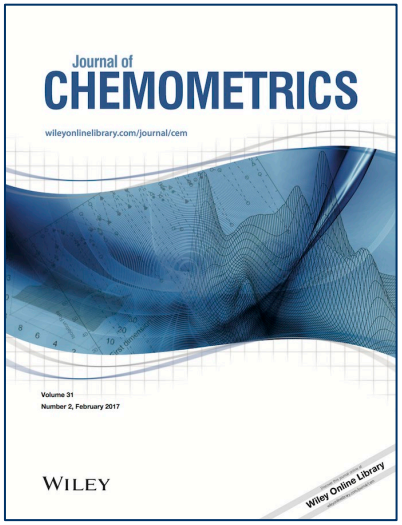
- ✓ Plant and processes optimization

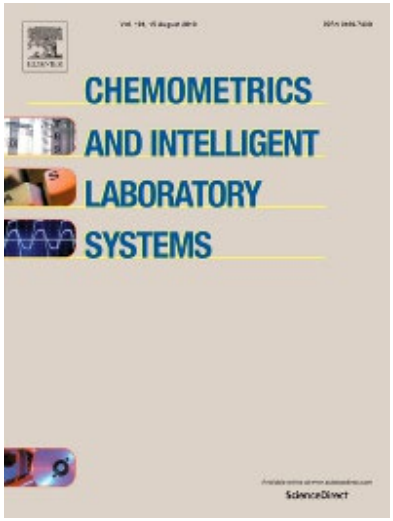
  Chemical reactors, production plants, fermentators, purification plants

- ✓ Control and modelling systems

  Quality control, pollution sources, clinical diagnostics, calibration models, biosensors, analytical procedures, organic syntheses, classification models

Three international scientific journals are currently dedicated expressly to Chemometrics:



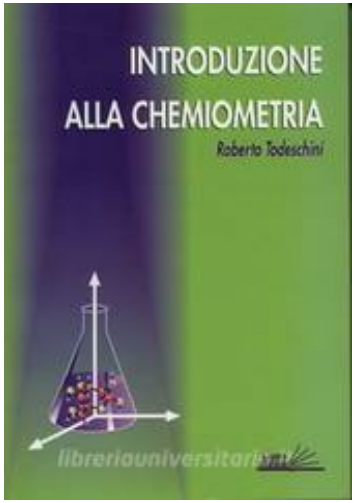


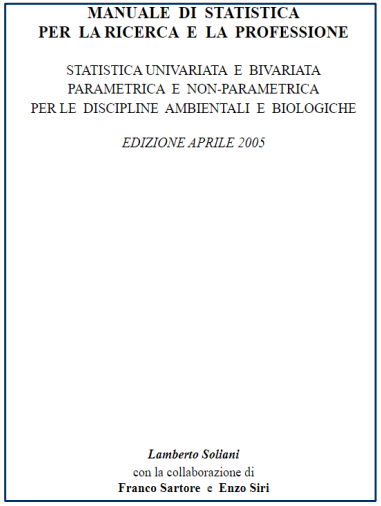
Specific courses are dedicated to Chemometrics in Degree Courses in Chemistry/Chemical Sciences of at least 10 Italian Universities:

Bari, Bologna, Cagliari, Genova, Milano Bicocca, Palermo, Parma, Pavia, Perugia, Piemonte Orientale, Roma La Sapienza

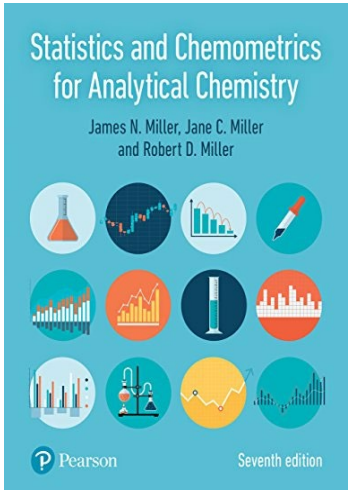
General books available on Chemometrics:

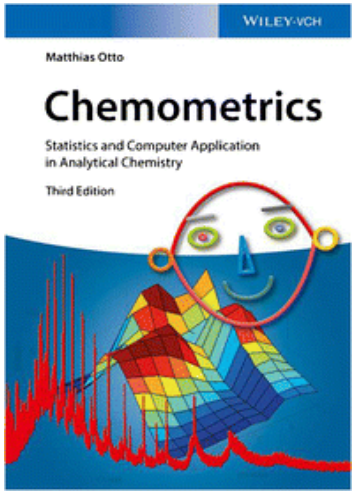
In Italian

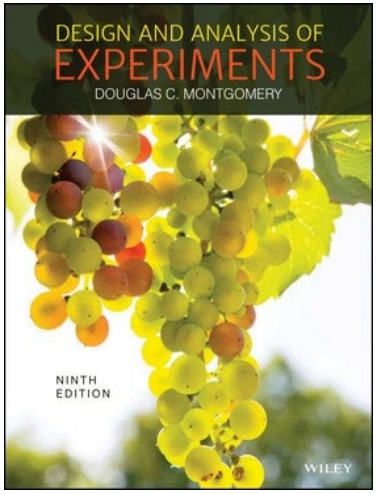




In English







Main topics of the Chemometrics course

1) Introductory section

Continuous random variables, probability density and partition functions. Graphical representation of univariate distributions; measures of central trend and dispersion; shape indexes: symmetry and kurtosis; moments generating function; common types of distribution.
Statistical inference: frequentist and Bayesian confidence intervals.
Hypothesis testing; two one-sided t-test (TOST) for equivalence testing; tests for outliers, tests for comparison between several variances (Hartley's F-max, Bartlett's test, Levene's S test).

2) Analysis of Variance (ANOVA)

One-way ANOVA (fixed- and random-effects models); multiple comparison methods for means: Fisher Least significant difference, Tukey, Tukey-Kramer, Bonferroni, Duncan's multiple range test, Dunnett, Kruskal-Wallis.

3) Experimental design and optimization

Screening designs; complete and fractional factorial design; Plackett-Burman design; response surface designs: composite central design, Box-Behnken design.
Steepest ascent (descent) approaches; Box-type Evolutionary Operation (EVOP); fixed- and variable-step Simplex.
Mixture design. Use of experimental design in separation science.

4) Regression

Ordinary linear regression; model parameters; model evaluation parameters; diagnostic methods for regression.
Non-linear regression; multiple linear regression; multicomponent analysis.

5) Multivariate methods

Multivariate structure of data; pre-treatment of data; variables transformation; data scaling.
Principal Component Analysis; Cluster Analysis: distance measures, hierarchic and non hierarchic methods; Supervised methods: Linear Discriminant Analysis, k-nearest neighbors method.

6) Basic concepts on quality control and sampling

Shewhart and cumulative sum control charts; proficiency tests and collaborative trials; uncertainty; basic concepts of sampling for quantitative analysis; acceptance sampling.