IEC TR 62829-1

Edition 1.0 2019-11

TECHNICAL REPORT

colour inside

Chemometrics for process analytical technologies – Part 1: General provisions, and methods for univariate statistics and chemometric processing of data

IEC TR 62829-1:2019-11(en)

IEC Central Office
3, rue de Varembé
CH-1211 Geneva 20
Switzerland

Tel.: +41 22 919 02 11
info@iec.ch
www.iec.ch

About the IEC
The International Electrotechnical Commission (IEC) is the leading global organization that prepares and publishes International Standards for all electrical, electronic and related technologies.

About IEC publications
The technical content of IEC publications is kept under constant review by the IEC. Please make sure that you have the latest edition, a corrigendum or an amendment might have been published.

IEC publications search - webstore.iec.ch/advsearchform
The advanced search enables to find IEC publications by a variety of criteria (reference number, text, technical committee,…). It also gives information on projects, replaced and withdrawn publications.

IEC Just Published - webstore.iec.ch/justpublished
Stay up to date on all new IEC publications. Just Published details all new publications released. Available online and once a month by email.

IEC Customer Service Centre - webstore.iec.ch/csc
If you wish to give us your feedback on this publication or need further assistance, please contact the Customer Service Centre: sales@iec.ch.

Electropedia - www.electropedia.org
The world's leading online dictionary on electrotechnology, containing more than 22 000 terminological entries in English and French, with equivalent terms in 16 additional languages. Also known as the International Electrotechnical Vocabulary (IEV) online.

IEC Glossary - std.iec.ch/glossary
67 000 electrotechnical terminology entries in English and French extracted from the Terms and Definitions clause of IEC publications issued since 2002. Some entries have been collected from earlier publications of IEC TC 37, 77, 86 and CISPR.

IEC TR 62829-1

Edition 1.0 2019-11

TECHNICAL REPORT

Chemometrics for process analytical technologies – Part 1: General provisions, and methods for univariate statistics and chemometric processing of data

INTERNATIONAL
ELECTROTECHNICAL
COMMISSION

ICS 25.040.40

ISBN 978-2-8322-7584-9

CONTENTS

INTERNATIONAL ELECTROTECHNICAL COMMISSION

_____

CHEMOMETRICS FOR PROCESS ANALYTICAL TECHNOLOGIES –

Part 1: General provisions, and methods for univariate statistics and chemometric processing of data

FOREWORD

1) The International Electrotechnical Commission (IEC) is a worldwide organization for standardization comprising all national electrotechnical committees (IEC National Committees). The object of IEC is to promote international co-operation on all questions concerning standardization in the electrical and electronic fields. To this end and in addition to other activities, IEC publishes International Standards, Technical Specifications, Technical Reports, Publicly Available Specifications (PAS) and Guides (hereafter referred to as "IEC Publication(s)"). Their preparation is entrusted to technical committees; any IEC National Committee interested in the subject dealt with may participate in this preparatory work. International, governmental and non-governmental organizations liaising with the IEC also participate in this preparation. IEC collaborates closely with the International Organization for Standardization (ISO) in accordance with conditions determined by agreement between the two organizations.
2) The formal decisions or agreements of IEC on technical matters express, as nearly as possible, an international consensus of opinion on the relevant subjects since each technical committee has representation from all interested IEC National Committees.
3) IEC Publications have the form of recommendations for international use and are accepted by IEC National Committees in that sense. While all reasonable efforts are made to ensure that the technical content of IEC Publications is accurate, IEC cannot be held responsible for the way in which they are used or for any misinterpretation by any end user.
4) In order to promote international uniformity, IEC National Committees undertake to apply IEC Publications transparently to the maximum extent possible in their national and regional publications. Any divergence between any IEC Publication and the corresponding national or regional publication shall be clearly indicated in the latter.
5) IEC itself does not provide any attestation of conformity. Independent certification bodies provide conformity assessment services and, in some areas, access to IEC marks of conformity. IEC is not responsible for any services carried out by independent certification bodies.
6) All users should ensure that they have the latest edition of this publication.
7) No liability shall attach to IEC or its directors, employees, servants or agents including individual experts and members of its technical committees and IEC National Committees for any personal injury, property damage or other damage of any nature whatsoever, whether direct or indirect, or for costs (including legal fees) and expenses arising out of the publication, use of, or reliance upon, this IEC Publication or any other IEC Publications.
8) Attention is drawn to the Normative references cited in this publication. Use of the referenced publications is indispensable for the correct application of this publication.
9) Attention is drawn to the possibility that some of the elements of this IEC Publication may be the subject of patent rights. IEC shall not be held responsible for identifying any or all such patent rights.

The main task of IEC technical committees is to prepare International Standards. However, a technical committee may propose the publication of a Technical Report when it has collected data of a different kind from that which is normally published as an International Standard, for example "state of the art".

IEC TR 62869-1, which is a Technical Report, has been prepared by subcommittee 65B: Measurement and control devices, of IEC technical committee 65: Industrial-process measurement, control and automation.

The text of this Technical Report is based on the following documents:

Enquiry draft	Report on voting
65B/1062/DTR	65B/1095B/RVDTR

Full information on the voting for the approval of this Technical Report can be found in the report on voting indicated in the above table.

This document has been drafted in accordance with the ISO/IEC Directives, Part 2.

A list of all parts in the IEC 62829 series, published under the general title *Chemometrics for process analytical technologies*, can be found on the IEC website.

The committee has decided that the contents of this document will remain unchanged until the stability date indicated on the IEC website under "http://webstore.iec.ch" in the data related to the specific document. At this date, the document will be

- reconfirmed,
- withdrawn,
- replaced by a revised edition, or
- amended.

A bilingual version of this publication may be issued at a later date.

IMPORTANT – The 'colour inside' logo on the cover page of this publication indicates that it contains colours which are considered to be useful for the correct understanding of its contents. Users should therefore print this document using a colour printer.

INTRODUCTION

Chemometrics is a rapidly developing subject. It was thus felt that a report offering guidance on its application to process analytical applications would both be helpful to all users of such technology and would stimulate specialists in chemometrics to work with users and developers of this technology.

This document does not seek to do other than provide a useful overview and a brief bibliography that enables interested parties to learn about and, hopefully, apply chemometrics in the most useful and appropriate ways for their circumstances. In that sense, it is definitely not prescriptive but constructively critical and seeks to encourage good practice and a wider appreciation.

It also aims at encouraging new research and development, as well as innovation, in applications of chemometrics for process analytical applications by highlighting areas to which such activities might usefully be directed.

Nowadays, the use of chemometric data analysis methods is widespread. Applications are in fields like

- design of statistical/chemometric sampling strategies, design of experiments, design of observational studies,
- design of data collection (including signal processing) protocols, data validation methods and database management (including metadata management),
- quality management, including quality assurance and quality control,
- data analysis and interpretation, not only in the use of multivariate (many variable) methods but also univariate (one variable) and bivariate (two variable) methods,
- process monitoring, optimization and control,
- chemical process and property modelling,
- guiding decision analysis and designing decision analysis methods/protocols in process control and optimization,
- method and instrumentation performance validation (Annex A) and calibration.

Because of the interdisciplinary and multidisciplinary nature of the discipline of chemometrics, it is often possible to be able to make unusual links and thereby solve problems taking cues from disciplines that are as diverse as medical diagnostics, decision sciences and quality assurance.

For example, in diagnosing the likely environmental impact of discharges of waste water from an industrial process, we might want to link toxicity assessment to chemical composition, the route and extent of discharge and the organisms likely to be affected. This might involve establishing a chemometric (mathematical) model of the impact of the discharge, bio-sensing the toxicity of the discharge on-line and relating both to the time, volume and concentration variations in chemical composition and physicochemical properties. This could then be used to assess the predictive reliability of the model and how this might be linked to process control and optimization of the discharge treatment and any associated risk assessment of the discharge process.

Conventionally, process control has involved using control charts for individual variables and this sometimes leads us to false impressions of process behaviour. Since 2010, techniques including both commercial and other software have become available to construct a wide variety of useful multivariate control charts that sometimes reveal "out-of-control" situations not apparent using conventional univariate control charts.

Due to the applicability of chemometric methods to a nearly unlimited number of cases in all fields of measurement and testing, but particularly due to need of using chemometric techniques in process analytical applications, it was felt a necessity to have guidance on the available methods and their appropriate choice.

CHEMOMETRICS FOR PROCESS ANALYTICAL TECHNOLOGIES –

Part 1: General provisions, and methods for univariate statistics and chemometric processing of data

1 Scope

This part of IEC 62829, which is a Technical Report, covers

- a study into the pre-requisites of chemometric (exploratory) data analysis,
- an overview of common data analysis procedures for univariate, bivariate and multivariate data analysis,
- explanations of the basic principles and major application areas of the different methods),
- some recommendations on the selection of an appropriate data analysis strategy.

These recommendations not covered earlier by other guidance documents on the topic are complemented by some advice on the validation of commercial (at the site of installation) and tailored software for process analytical purposes. Recommendations are given on available reference data sets (Annex B) for benchmarking of software implementing the data analysis methods covered (if available). An application example is given.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

No terms and definitions are listed in this document.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

- IEC Electropedia: available at http://www.electropedia.org/
- ISO Online browsing platform: available at http://www.iso.org/obp

4 Fields of application

4.1 Process control and process analytical technologies (PAT)

There is currently a considerable trend to use process analytical technologies (PAT) for reaction monitoring and (direct loop) process control. Current developments in the field of process engineering are not imaginable without PAT, such as modern process design, integrated processes (e.g., reactive separation processes), and intensified processes along with requirements to process control, model-based control, and soft sensing – all involving chemometrics.

The process industry relies on the design, operation, control, and optimization of chemical, physical, or biological processes. This involves creating production facilities that translate raw materials into value-added products along the supply chain. Such conversions typically take place in repeated reaction and separation steps – either in batch or continuous processes. The end products of a chemical production facility are the result of several production steps that are connected not only in a sequential fashion, but also involve recycling of unused raw

materials and by-products, as well as waste treatment stages. Production processes in the process industry are particularly disturbed by variations in feed-stocks and other influences that impact the product quality. An integrated process control approach enables constant product quality and prevents out-of-spec production by effectively compensating for such process variations. In a conventional approach, quality is determined by withdrawing samples from material streams and conducting offline analytics, which is called in-process control or at-line or off-line control. By applying quality by design (QbD, see ICH definition in 4.2) approaches, quality can significantly improve to generate less waste, reduce reprocessing of substandard material, and create products of superior quality.

Today's optimized process design relies heavily on computer aided tools, which account for, for example, mass transfer, thermodynamic, kinetic, and other physical properties of the treated materials. Typically, a sufficient understanding of such properties is available and implemented in dynamic numeric models. Dynamic models are in turn the essential basis for optimized process and plant design. Unfortunately, they are only sparsely used for process control. A definition from Lee (2008) brings this to a contemporary level:

"Cyber-physical systems (CPS) are integrations of computation with physical processes. Embedded computers and networks monitor and control the physical processes, usually with feedback loops where physical processes affect computations and vice versa."

4.2 Physical and chemical properties

Process analytical techniques are extremely useful tools for chemical production and manufacture and are of particular interest to the pharmaceutical, food and (petro-)chemical industries. It can be easily transferred to manufacturing for process control and for quality assurance of final products to meet required product specifications, since it provides dynamic information about product properties, material stream characteristics, and process conditions.

Normally, the quality of the final product is assessed after processing by adequate testing procedures. The rationale behind process analytical technologies (PAT) is to measure, and assess, physical and chemical properties over, and throughout, the production process in order to assure a product which is within the tolerance limits or regulatory restrictions.

The quality of any product or the properties of a material can be described by a complex functional relationship to physical and/or chemical properties of the constituents of the product/material and its temporal changes during processing.

According to ICH Q8(R2), quality is the suitability of either a drug substance or drug product for its intended use (ICH: International Conference on Harmonisation). This term includes attributes such as identity, strength, and purity. Before the launch of the PAT initiative, pharmaceutical production was confronted with challenges like drug shortages due to manufacturing difficulties, process deviations coupled with frequent inconclusive investigations, batch failures and rejections, in-process test debates (e.g., blend uniformity), slow and protracted cGMP (current good manufacturing practice) remediation, warning letters, and others.

The science-based regulatory guidances such as the FDA (US Federal Drug Administration) and ICH PAT guidance have recognized spectroscopic techniques as potentially useful tools on building quality into the product and manufacturing processes, as well as continuous process improvements. The goal of PAT is to enhance understanding and thereby control the manufacturing process.

The common future vision in pharmaceutical production is continuous manufacturing (CM), based on real-time release (RTR), i.e. a risk-based and integrated quality control in each process unit. This will allow for flexible hook-up of smaller production facilities, production transfer towards fully automated facilities (featuring less operator intervention and less down-time), and end-to-end process understanding over product life cycle, future knowledge, and faster product to market.

Both methods, i.e. in-process and final-product testing, intensively use statistical methods as described in this document.

4.3 PAT fields of application

4.3.1 Definition of chemometrics

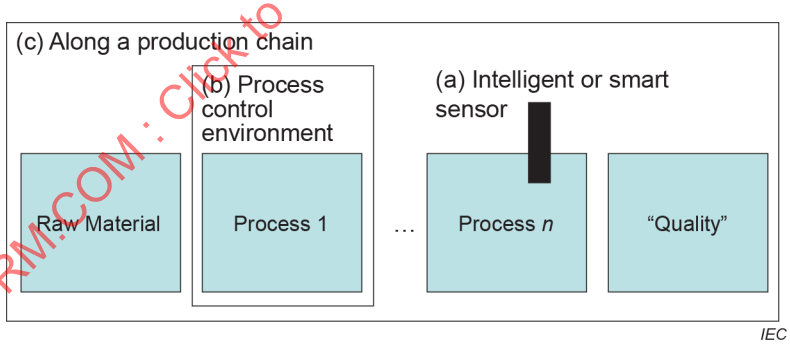
At the dawn of chemometrics, an appropriate definition of the term was that chemometrics is "… the use of statistical, mathematical and other logic-based techniques, together with chemical knowledge to solve chemical problems" (D1) with a clear emphasis on chemical problems, thus explaining the word "chemo" in the term. A definition taken from the Journal of Chemical Information and Computer Sciences (1975), Vol. 15, page 201, defines chemometrics as the "development/application of mathematical/statistical methods to extract useful chemical information from chemical measurements" (D2) thus detailing the aim, namely the extraction of (useful) information from measurement data.

"Chemometrics is a sub-discipline of metrology dealing with the application of mathematical, statistical and other methods employing formal logic to evaluate and interpret (chemical, analytical) data, optimize and model (chemical, analytical) processes and instrumentation, extract maximum (chemical, analytical) information from experimental and observational data" (D3).

To date, no internationally agreed and standardized definition of chemometrics exists. Although all three definitions reflect, in principle, both the intention of, and the instruments used for, chemometrics, definition D3 is the most general. The application of the principles and tools of chemometrics is explicitly not limited to the field of chemical/analytical measurement, so definition D3 may by read and used without the specification "chemical/analytical", deliberately put in parentheses here.

4.3.2 Overview on PAT fields of applications

Fields of application of PAT are described below and visualized in Figure 1.



(a) within an (intelligent or smart) sensor

(b) within a production unit, e.g. process control system, process control environment, or laboratory information management system (LIMS)

(c) along a production chain, e.g. ERP system, data mining

Figure 1 – Different levels of chemometric applications

4.3.3 Chemometrics for sensors

Sensors are the sense organs of process automation. At present there are serious changes in the areas of information and communication technology, which offer a great opportunity for optimized process control and value-added production with dedicated network communicating sensors. These kinds of smart or intelligent sensors (see Figure 1) provide services within a network and use information from the process information systems.

Intelligent field devices, digital field networks, Internet protocol (IP)-enabled connectivity and web services, historians, and advanced data analysis software are providing the basis for the future project “Industrie 4.0”, and industrial Internet of Things (IIoT). This is a prerequisite for the realization of cyber physical systems (CPS) within these future automation concepts for the process industry.

As a consequence, smart process sensors enable new business models for users, device manufacturers, and service providers.

4.3.4 Chemometrics for production units

With the introduction of advanced process analytical technology, the closeness of key process variables to their limits can be directly monitored and controlled and the processes can automatically be driven much closer to the optimal operating limit. Classical, non-model-based solutions reach their limits when sensor information from several sources has to be merged. In addition, their adaptation causes a high effort during the life cycle of the process. This calls for adaptive control strategies, which are based on dynamic process models as mentioned above. Model-based control concepts have also the potential to automatically cope with changes of the raw-materials as well as process conditions.

Chemical process control technology has advanced significantly during the last decades. For world-scale high-throughput continuous processing units such as crackers and separation trains, in most cases classically engineered control solutions (proportional–integral–derivative (PID) controllers, cascade and override structures) have been replaced by model-based techniques, most prominently model-predictive control (MPC) based on linear plant models. However, the engineering and implementation costs of such advanced controllers are still high. For smaller, flexible processes in which varying products or intermediates are manufactured, it is not economic to re-engineer the control concept or to re-model the process for all intended processes.

Advanced control strategies have to be built upon empirical – often data-based – models which describe cause-effect mappings between the degrees of freedom of the process and the product properties in a black-box fashion. Chemometric techniques for the derivation of empirical models – e.g. partial least squares (PLS), principal component analysis (PCA) – are available but currently mostly used for off-line data analysis to detect the causes of variations in the product quality. An automated application along the life cycle is still very limited. The development of such models requires significant experimental work, and the reduction of the effort needed for these experiments is the focus of ongoing research. When such stationary models are available and are combined with dynamic models that describe the times needed for the transition from one steady state to the other, feedback control and iterative optimization schemes can be built that make use of the novel sensors.

The departure from current automation measurement to smart sensor systems has already begun. Further development is based on the actual situation over several steps. Possible perspectives will be via additional communication channels to mobile devices, bidirectional communication, integration of the cloud and virtualization. The cost of connectivity is dropping dramatically, providing powerful potential to connect people, assets, and information across the industrial enterprise. While only providing add-on information, the first cloud services may not require a high availability or real-time capabilities. But when available in the future, even process control tasks will be possible using cloud services, e.g. when complex computing algorithms are needed, which require computing performance and availability.

4.3.5 Chemometrics along a production chain

Current focuses of research are closed-loop adaptive control concepts for plant-wide process control, which make use of specific or non-specific sensors along with conventional plant instruments. Such advanced control solutions could give more information than only control information, such as sensor failure detection, control performance monitoring, and improve simulation-based engineering. At present, such data is typically collected and analysed in an enterprise resource management system (ERP).

Closed-loop adaptive control concepts can be used to optimize global mass and energy balances, local response surfaces that relate specifications of outgoing and ingoing streams to the consumption of energy and the cost of production, or simple dynamic models of the behaviour of the process stream (i.e. delays, settling times, etc.). In such a manner, the plant-wide control of the entire process can be performed by setting targets and constraints on the flow rates and on the properties of material streams.

The plant-wide control scheme is implemented using iterative set-point optimization on the basis of the local models taking into account the dynamic behaviour. When the local controllers are model-based, the response surfaces can be computed from these models. This is not the case if classical control schemes are used, where they must be derived from empirical data.

Such powerful analytics will help optimize both assets and systems. Predictive analytics will be installed to reduce unplanned down-time. Newly available information generated by these tools will lead to new, transformative business models supported by new applications. Instead of offering physical products for sale, companies will increasingly offer products as a service.

5 Pre-requisites of chemometric data analysis

5.1 Data has to be adequate and reliable

Extraction of useful information from measurement results implicitly supposes, and even requires, the data to be adequate and reliable. This imposes certain restrictions on data acquisition. Issues to be tackled before the start of measurement campaigns for the acquisition of large amounts of data are briefly described in 5.2 to 5.10.

5.2 Data representativeness

Data representativeness is understood as validity of the data obtained on a certain selection of sub-units from a manifold to characterize the latter. Sub-units may be single products, samples taken from a reactor, sub-samples taken from a plot of land at different horizons, prospective measurements taken in an oil, gas or gold field, the atmospheric environs of a huge city, or anything related. The manifold is the entity which the samples are taken from, whichever applies.

Representativeness refers to the fitness-of-purpose of decisions (not measurements) taken on a manifold. A single product or a single sample can easily be qualified using adequate methods of analysis, but propagating this result and the corresponding decision on the fitness-for-purpose of the whole manifold remains arbitrary.

Good sampling procedures providing reliable estimates for a manifold consisting of many sampling units (defined, for example, by the final packaging unit size for the consumer) are well described leading to factor analysis procedures as described in the ANOVA (see 7.1.3).

Visman (1969) and Ingamells (1973) developed specific sampling constants for defining the amount of sub-sampling from a manifold of matter. These can be used to characterize the manifold by

- a sum parameter, in the simplest form the average over a certain number of sum-samples,
- a value distribution which might be spatial or time-dependent requiring a (much) larger number of observations.

Although it requires more data than the above methods, the Krige (1951) approach may be helpful. For prospecting of value-carrying metals (like gold or platinum), this might be acceptable; for all other manifold-investigation tasks, prior information should be included in the final assessment. However, none of the statistical procedures can guarantee to find the legendary needle in a haystack.

5.3 Data acquisition

In gathering data from process instrumentation, one should be seeking to obtain data of adequate quality and quantity for further processing, display and finally reliable interpretation within the context of process monitoring, control and perhaps feedback for process adjustment. In order for data to be appropriate and "fit for purpose", the data users need to think very carefully about the specification of the data at each stage in data processing from initial capture to final application of the interpretation. Issues relating to the various biases and uncertainties that may accumulate through these various stages need addressing.

Process instrumentation may have sensors and transducers for direct measurement or may be more complex involving spectrometric or chromatographic systems for online or off-line measurement of either batch or continuous processes with either discrete or continuous physical sampling, either taken from a single location along a chain or gradient or from a two-dimensional surface or out of a three-dimensional space. The immediate output from sensors, transducers or more complex instrumentation is commonly analogue and may be conditioned by filtering and/or amplifying and subsequently digitized for further processing. Alternatively, the immediate output may be digital, as with photon or ion counting systems.

An important part of the design of the data gathering process involves the careful consideration of (a) the nature of the signal (analogue or digital, range, frequency, noise extent and types, etc.), (b) how it is gathered, (c) how it is conditioned and (d) how it is amplified. These various processes may seriously affect the usability of the data gathered for process monitoring, control or adjustment.

5.4 Data management

Metadata is a class of data describing data, such as process measurement data acquired from process instrumentation. For example, the time and date when a measurement was made, what variable was measured, in what units the variable was measured, from where and from what the specimen was sampled, under what conditions the specimen was taken, time delays and time lags in the sampling system, etc. Reliable recording of metadata associated with measurements made using process instrumentation, whether for calibration or process monitoring, is essential. It is also crucially important that the links between measurement data and associated metadata are preserved intact through all data collection, data management and data analysis processes. Otherwise, interpretation of such measurement has little or no value in process monitoring, control or feedback adjustments.

5.5 Databases versus spreadsheets

In the electronic recording of data and associated metadata from instrumentation, various options may be available from which to choose.

1) Data may be recorded on a data collection device and subsequently transferred manually by a data clerk onto an electronic record.
2) Data may be transferred automatically into a computer-software-based spreadsheet.
3) Data may be transferred automatically into a computer-software-based relational database management system in which data from an observation (involving one or more variables) made using process instrumentation and the associated metadata linked to that observation are considered as a single entity. This enables analysis to be performed with queries designed in a structured query language (SQL) or via well designed on-line analytical processing (OLAP) systems as is now being implemented in various types of data mining.

Option 1, involving manual entry to electronic records, requires data quality and validation checks to ensure adequate correction of transcription errors, whose incidence may be as high as 5 % of data entries.

Option 2, involving automatic data transfer to spreadsheets, is less satisfactory than Option 3 for the following reasons: Although some spreadsheet software packages have facilities for

security and auditing, many inexperienced users do not appreciate the value of these and they tend not to be used or are implemented in a less than satisfactory manner. Spreadsheets are easy to corrupt, whether accidentally or deliberately. Accidental corruption may occur, for example, as a result of instrumentation faults or signal transmission errors or corruption may be deliberate and fraudulent.

Because it is not straightforward to protect spreadsheets from accidental or deliberate corruption and because the auditing of such errors is not straightforward or reliable, it is difficult to ensure data validity and quality.

Metadata needs to be properly linked to the data with which it is associated and this link is much less secure, reliable and more difficult to protect in spreadsheets.

In well-designed database software, facilities are available to set many levels of data protection to limit corruption and to have various forms of data validation (see below), although the latter types of data screening require care in their design so as to create data flagging rather than data rejection. This then enables rigorous forms of data audit to be implemented.

5.6 Data quality

The quality of data may be expressed in terms of an uncertainty budget. Included in that budget will be contributions from sources of random variation in measurement and sampling (components of the precision contributions) and sources of bias in measurement and sampling (accuracy).

These should be checked periodically, preferably in a planned and strategic way, with additional opportunities for unplanned checking as well, for example using various kinds of calibration protocols. The data management system protocols should include ways of checking data quality and that quality management protocols are being actively and reliably used.

5.7 Data validation

Data validation as data is entered into a database may include checks on whether data values are in the expected range. Data should never be deleted or rejected merely because they are outside the expected range. Rather, it should be flagged as outside the set range to enable or initiate validity checks. The data may actually be valid and outside the expected range because of changes in process behaviour and thus provide a useful warning of such changes.

5.8 Data corruption

Data may be prone to accidental corruption as a result of failure or faulty behaviour of sensors and transducers, electrical interference, faulty transmission, incorrect setting of signal conditioning, amplification, integration, operator error, etc. The integrity of sensor and transducer behaviour, signal acquisition, etc., should be checked on a planned basis with opportunities for unplanned checking for accidental corruption. Quality management of operator compliance with standard operating procedures should be included in the quality management system to allow for data corruption checks to be readily made.

5.9 Data security and fraudulent data detection

Means of protection of process measurement/monitoring systems and database management systems from accidental or deliberate corruption should be available and implemented, along with ways of establishing the level of security and its continued integrity. There are many useful ways of evaluating data quality and validity. These include exploratory and chemometric data analysis techniques (both univariate and multivariate), as well as data mining techniques. These can be used to highlight outlier data, unusual data patterns, non-randomness and other forms of data behaviour that may suggest possible fraud. Having highlighted the unusual aspects of a data set, those may then be investigated to determine

whether these are genuinely unusual or arise as a result of accidental or deliberate data corruption. Non-randomness may arise from faulty equipment or connections, poor operating practices, shifts in process behaviour or deliberately faulty data recording, for example. Statistical analysis of the patterns of such non-randomness may be useful in suggesting causes (examples of fraudulent data recording have been found in the past in hazardous waste treatment and waste disposal process control).

Sometimes, statistical process monitoring data may be replicated from past measurement data sets to give an appearance of good quality process control when no such activity was actually being undertaken. There are notorious recent examples of this in nuclear waste reprocessing industry and others have been found in hazardous waste characterization. This kind of fraudulent activity may be difficult to detect because, superficially, the data presented appear to be valid and reliable if the fraud was carefully constructed. To avoid such fraud, regular comparison checks of current data versus historical data should be undertaken. Statistical graphics tools are especially valuable here.

5.10 Data management for data mining

Data mining or knowledge discovery in databases (KDD) as a subject has become very important in connection with the need to extract useful information from extremely large collections of data and associated metadata. In many areas of process measurement, there exists the possibility of acquiring very large quantities of measurement data simultaneously associated with many variables as a function of both of space and time. The data mining approach requires attention to the choice of data management and storage systems that may be used, including databases and data warehouses.

For process measurement data, consideration needs to be given to the requirement for the data to be analysed using chemometric and/or statistical software as well as data mining software. It is very important to ensure that such data is accessible via ODBC or OLE DB for ODBC connections, enabling the accessibility to access any relational data. Likewise, data mining with a data warehouse should follow the OLE DB for OLAP standard to enable data access via a variety of software systems. Currently, the most frequently discussed standard is OPC unified architecture (OPC-UA).

6 Pre-requisites of chemometric data analysis

6.1 Technical requirements of chemometric data analysis

Data analysis methods are grouped into classes according to the principles laid down in 5.2 and 5.3. Within each group, the most commonly used techniques are described, which does not exclude the application of techniques specifically tailored for a particular purpose. Description of the methods/techniques covered by Clause 5 follows the layout developed in ISO/TR 10017, i.e. each method is described in four subclauses tackling what it is, what it is used for, the benefits of application, and limitations and cautions (where applicable).

6.2 Data dimensionality

Data subject to any of the data analysis tools described in Clause 7 may be of quite different dimensionality.

As far as measurement data are concerned, they are taken on objects of interest. These objects normally display a (large) number of (quite) different properties. Depending on the task, one or more of these properties have to be defined, thus specifying one or more measurands.

The most trivial case is a single measurement of a single measurand (property) for a single object. The generated data does not allow any further statistical processing.

Any measurement is subject to measurement uncertainty. Therefore, one would rather take a certain number of replicate measurements of a single measurand of a single object. The resulting data format is a set of values all referring to a single measurand, commonly also called a univariate data set.

Repeating the measurement of a single measurand for more than one object results in a two-dimensional data matrix with columns representing the replicates, and rows referring to the objects under investigation. The data format already allows for a comparison between the objects, revealing or not revealing significant differences.

Since the data still refer to a single measurand, they are univariate.

The same data format results from a measurement of a single measurand of a single object under different specified (e.g. environmental) conditions. The data format allows a search for significant influences of the specified conditions on the measurand.

As only a second measurand of the same object is taken into consideration, the data format becomes multivariate. In the general case, multivariate data represent a number of replicate measurements of a number of measurands of a number of objects under a number of specified conditions (four-dimensional data matrix). Any sub-group (e.g. one object but different conditions, or same conditions but a number of objects) provides three-dimensional data matrices.

Although formally belonging to the multivariate case, bivariate data sets are of major importance for regression analysis, i.e. the search for a known or assumed functional relationship between two measurands. Therefore, this data format is handled and described separately.

6.3 Method classification

Most often, data analysis techniques are classified according to the dimensionality of the data they are applied to. This document formally also follows this classification. However, one might wish to look at data analysis techniques from the perspective of their aim, or, in other words, the general principle they implement for the given situation and data dimensionality.

The aims of at least the majority of available data analysis techniques are the following.

- Description of a data set concerning probability distribution, dispersion, and an appropriate central value.
- Search for major impacts on a single or a number of measurands. The general principle is factorization, both with unknown functional dependence (ANOVA, MANOVA, factor analysis) and with an assumed functional relationship (GLM, multivariate curve resolution, regression). Mixed forms are possible.
- Search for similarities between objects described by one or a number of measurands. The general principle is distance calculation and grouping with a specified measure. The latter may be a pre-defined tolerance (ranking, figures-of-merit) or a tolerance defined by the data set (hierarchical and k-means clustering, KNN). Mixed forms are possible.
- Search for constructs with a minimum number of explanatory variables, describing the major features of the multiple-measurand data set. The general principle is data reduction using both linear (PCA, discriminant analysis (DA), PLS) and non-linear constructs involving cross-products (e.g. numerical self-learning algorithms like artificial neural networks (ANN)).

Description of data distribution, dispersion and central value normally refers to univariate data; in multivariate cases it is applied per variable or per measurand. Hypothesis testing is not a data analysis technique in the sense of this document, however it may be part of any of the methods, at least as a demonstration of significance of the effect under investigation.

The interrelation between the classification perspectives is shown in Table 1, which displays the aims of the data analysis techniques and the data formats these techniques are most often applied to, or which they do require.

Table 1 – Data analysis techniques and data formats

Data analysis technique	Univariate data	Bivariate data	Multivariate data
Descriptive statistics	X		
Hypothesis testing	X	X	X
Factorization	X		X
Regression		X	
Grouping			X
Data reduction			X

6.4 Data pre-processing

6.4.1 Filtering

Filtering mainly aims at noise removal before a univariate or multivariate treatment of data. Kalman filters are widely used for any (electromagnetic) signal pre-processing; they may be both hardware- and software-based implementations. The latter may be applied also to data streams originating from measurements other than current, voltage or frequency.

Most often, smoothing procedures combine getting rid of measurement noise (in this sense they are also filters) with making the original data stream (partially) describable in a mathematical sense. See 6.4.2.

6.4.2 Smoothing

Smoothing is a type of fitting a (set of) sufficiently flexible mathematical basis functions (e.g. polynomials or basic signal forms with a series of overtones) to the full or (most often) segments of the data stream.

It can be done using several routines (e.g. Savitzky-Golay algorithm or using wavelets). Methods based on local alignment (such as correlation-optimized warping) are relevant, for example, for NMR and chromatographic applications.

Alternatively, bucketing can be used to split the full data set into segments (buckets) and the integral of each segment is used as a replacement for the original intensities. The buckets' width is a very important parameter for subsequent multivariate analysis. Some variations of the method are available, including rectangular bucketing, point-wise bucketing, variable size bucketing and advanced bucketing.

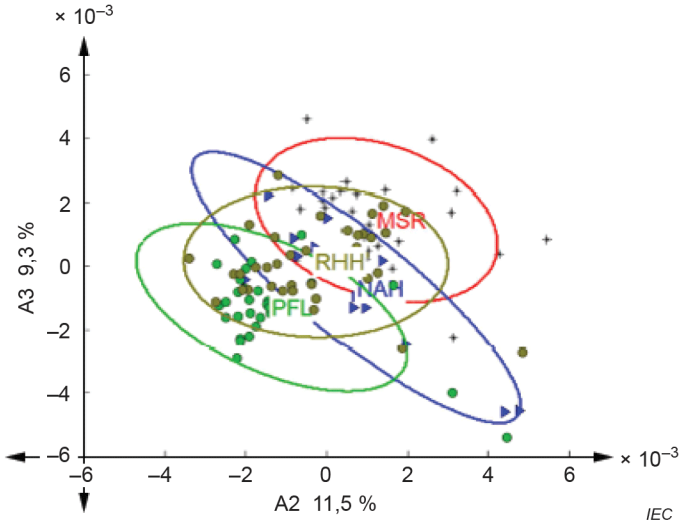
6.4.3 Data reduction

Data reduction facilitates and accelerates chemometric analysis. Elimination of regions with zero intensities as well as regions of solvent and internal reference signals is recommended. Bucketing or taking the average of several data points can be further used for this purpose. In either case, all spectra used for multivariate modelling and validation are processed with the same procedure.

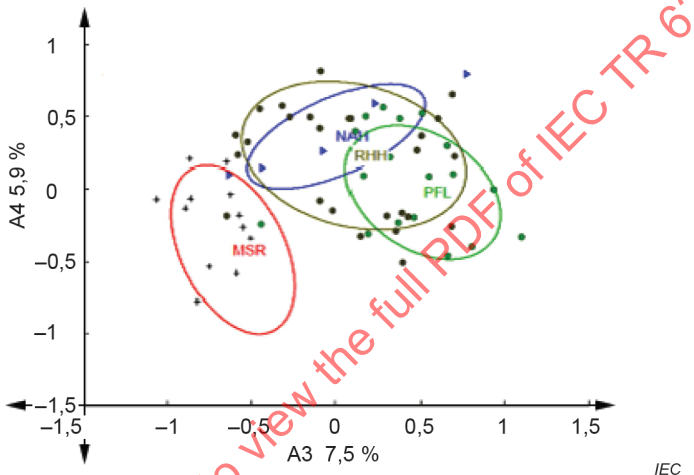
For selecting the most significant regions of the data available, variable selection methods such as clustering of latent variables or evolving window zone selection can be used. For multivariate calibration applications, one should only consider regions informative for the envisaged measurand.

Pre-processing can also involve mean-centring and scaling the variables. The mean-centred matrix is obtained by subtracting the mean intensity for each of the variables from single sets of data. Second, different types of scaling (scaling to unit variance, Pareto scaling) or, alternatively, element-wise transformations (e.g. log transformations) can be used. Mean-centring is recommended for PCA applications.

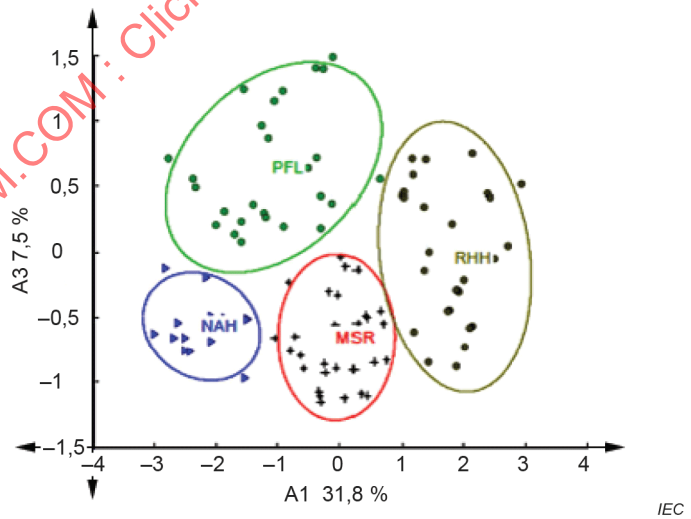
EXAMPLE Figure 2 shows the influence of pre-processing techniques for classification of the geographical origin of wine.



a) mean-centring



b) auto-scaling



c) scaling to unit variance

MSR, NAH, PFL and RHH are abbreviations for different wine-growing regions in Germany.

SOURCE: EUROLAB Technical Report No. 1/2015, Guide to NMR Method Development and Validation – Part II: Multivariate data analysis. Reproduced with permission.

Figure 2 – Influence of pre-processing techniques for classification of the geographical origin of wine

7 Methods of chemometric data analysis

7.1 Univariate analysis

7.1.1 Descriptive statistics

7.1.1.1 What it is

The term descriptive statistics refers to procedures for summarizing and presenting univariate quantitative data in a manner that reveals the characteristics of the distribution of data. Multiple measurands (multivariate data) are treated variate-wise unless it is known that the measurands are not independent but reveal (strong) correlation.

The characteristics of data that are typically of interest are their central value and spread or dispersion. Another characteristic of interest is the distribution of data. Quantitative measures are skewness (describing symmetry) and kurtosis (describing flatness of the distribution maximum).

The information provided by descriptive statistics can often be conveyed readily by a variety of graphical methods, including

- bar- and pie-charts illustrating fractions of data within specified limits,
- box-and-whisker plots indicating both asymmetric (skewed) data distributions and extreme-value situations,
- histograms depicting the distribution of data,
- probability plots demonstrating the compliance with, or the deviation from, an assumed distribution.

Descriptive statistics (including graphical methods) are implicitly invoked in many of the techniques cited in this document and should be regarded as a fundamental component of data analysis.

7.1.1.2 What it is used for

Descriptive statistics is used for summarizing and characterizing data. It is usually the initial step in data analysis. The characteristics of sample data may serve as a basis for making inferences regarding the characteristics of populations from which the samples were drawn, with a prescribed margin of uncertainty and level of confidence.

7.1.1.3 Benefits

Descriptive statistics offers an efficient way of summarizing and characterizing data, and a convenient way of presenting such information. The technique is potentially applicable to all situations that involve the use of data. It can aid the interpretation of data, and is valuable in decision-making.

7.1.1.4 Limitations and cautions

The quantitative measures of the characteristics of sample data provided by descriptive statistics are subject to limitations of the sample size and sampling method employed. These measures cannot be assumed to be valid estimates of characteristics of the population from which the sample was drawn, unless the underlying statistical assumptions are satisfied.

Exploring a data set involves examining its shape and looking for unusual data without any assumptions about any underlying distribution. In a general view, this requirement is making the problem underdefined: Judging about "unsual" data without any assumption about the distribution is not sensible, while judging about a distribution without exactly knowing which data are, or are not, members of this distribution turns out to be equally difficult or impossible.

Assumptions are needed, and plenty of approaches exist, mainly focused on excluding or down-weighting most distant data, or re-grouping the data and handling the groups as a manifold.

EXAMPLE Latter value analysis (introduced by John W. Tukey) is powerful but simple. Data exploration using tools from non-parametric or distribution-free statistics, based on the sign test and on Wilcoxon's signed rank test are also valuable here.

7.1.1.5 Application examples

Descriptive statistics has useful applications in almost all areas where quantitative data are collected. It provides information about the product, the process or some other aspect of quality management measures in force.

Examples are

- summarizing key features of product characteristics,
- describing the performance of some process parameter, and displaying the distribution of process parameters,
- illustrating measurement data, such as equipment calibration data,
- providing general or key parameters for process control.

7.1.2 Hypothesis testing

7.1.2.1 What it is

Hypothesis testing is a decision, at a given level of confidence, about a population based on observations from a sample of the population. Tests to be executed depend on an assumption of the (underlying) population. The classical, conventional statistical approach to data relies upon the central limit theorem to justify the overwhelming dominance of the normal distribution (also called Gaussian) in methods promulgated by such conventions. The probability distribution of the population is described by

$$p = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

with μ being the mean and σ the standard deviation of the population. Any limited set of data can be characterized by an estimate x of the mean and an estimate s of the standard deviation, where a statistic follows the so-called Student distribution

$$f(t) = \frac{\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{2}$$

Under these assumptions, a broad spectrum of hypotheses may be tested. Most obviously, the comparison of data sets with pre-defined values, of different sets of data between each other, and of sets with regard to their inherent spread are basic. Thus, the basic (and major) hypothesis tests are the following.

t test: Testing for significant difference between i) the means and a reference value ii) two data sets (difference of means) or iii) difference between pairs of measurements.

F test: Testing for significant difference between the spreads of two data sets (difference of s).

Basic steps for questioning a hypothesis test are:

- formulate the question,
- select the test,
- decide on one- or two-sided test,
- choose the level of significance,
- define null and alternative hypothesis,
- determine the critical value,
- evaluation of the test statistic using the appropriate equations,
- decisions and conclusions.

The statistical decision-making decides between two hypotheses, namely the following.

– Null hypothesis H_o: It implies that there is no difference between the observed and known value, other than that which can be attributed to random variation ($\mu = x$).

– Alternative hypothesis H_1: It is the opposite of the null hypothesis, i.e. there is a difference between values ($\mu \neq x$), where x is a sample mean, and μ is a true value.

Certainly the null hypothesis may refer to more than one total (expected) mean: this particularly occurs in some of the multivariate statistical applications.

Decision rules follow (with exceptions, see, for example, the multiple Grubbs test) the principle that if a critical value Q_{crit} is exceeded by the statistic Q calculated (t, F, z, other...), $Q_{crit} > Q$, the considered difference between the objects compared is significant. The critical value is calculated at the level of significance desired (most commonly 5 %, in rare cases 1 %).

Most commercially available software implements (direct or with additional installation of macros or add-ons) tools for carrying out the above tests. This also refers to ANOVA.

7.1.2.2 What it is used for

Hypothesis testing (both t and F) is used for

- the comparison of data sets with pre-defined values,
- the comparison of different sets of data between each other,
- the comparison of sets of data with regard to their inherent spread,
- testing assumptions on the normality of a data set (three-fold: normality, skewness, and kurtosis),
- testing the likelihood of a functional relationship in modelling a particular coincidence of pairs of sets of measured data,
- testing the performance of large-dimensional data array reduction algorithms like PCA, PLS, and MCR,
- (statistical) testing of any tailor-made assumptions on behaviours of measured feature variables, and/or dependencies between those features.

EXAMPLES Quality of different series of products; different-operator performance on the same instrument; differences between measurements taken on the same object at different days/under different conditions.

Applications of the (simple) F test are in the analysis of variance (ANOVA, see 7.1.3) which extends the differentiation between groups of data originating from different objects.

7.1.2.3 Benefits

Given the non-ideal character of any measurement device, the hypothesis test as described above provides statistical evidence on the significance or non-significance of any difference

between values. Depending on the kind of values, different layouts apply, and different techniques are recommended.

EXAMPLE A limiting or governing factor in the production of a good may be identified by paired t-tests; however, and depending on the number of groups in the analysis, an ANOVA may be appropriate.

7.1.2.4 Limitations and cautions

Despite the above, statistical evidence is not always recognized. Whether or not statistical evidence may help to provide juridical evidence heavily depends on the legislation, and former sentences of higher courts in the particular country concerned. Even the question of whether a certain legal limit imposed on a substance/compound in a certain matrix is exceeded or not may be treated differently by a statistician and a court. Obeying the rules of statistical decision-making does not guarantee prevalence in lawsuits, while not obeying the same rules will most probably lead to disaster (with, in certain cases, devastating commercial consequences).

Furthermore, the approach may show weaknesses since the statistical measures of location and variability such as the arithmetic mean and the standard deviation are considered as useful when adequately justified by careful analysis of a data set, i.e. if it can be proven that it actually follows a normal distribution (see above).

EXAMPLE The above problems are common in trace analysis, especially if the limit of detection (LOD) and/or quantitation (LOQ) result in serious truncation of the assumed data distribution. Monitoring processes using trace component measurements should similarly take account of such LOD/LOQ problems. These may vary if the size varies for the sample (number of measurements per batch, for example) or specimen (e.g. the weight or volume taken for testing or the optical path-length or slit-width used in spectrometry).

7.1.2.5 Application examples

Groups of application examples are given in 7.1.3.2.

7.1.3 Analysis of variance (ANOVA)

7.1.3.1 What it is

ANOVA tries to develop a total variation of a set of data, expressed by the sum of squared deviations SS, into variations (sums of squared deviations) which may be attributed to a particular influencing factor. Depending on the number of factors recognized to be significant, this development may be elaborate. Formula (3) shows the development in a more symbolic way, more factors may be considered, and the correlation terms are not expressed explicitly:

$$SS_{\text{total}} = \sum_{i,j,k\ldots}^{I,J,K\ldots} (x_{i,j,k,\ldots} - \bar{\bar{x}})^2 = \sum_{i}^{I} (x_{i,j,k,\ldots} - \bar{x}^I)^2 + \sum_{k}^{K} (x_{i,j,k,\ldots} - \bar{x}^K)^2 + \ldots + \text{corr} \qquad (3)$$

where

$\bar{\bar{x}}$ is the total mean of the data set;

the $\bar{x}^I$ (including other superscripts) are the means of the data attributed to a certain factor;

corr are the remaining terms of the development describing the cross influences of the factors.

Most often in use is a one-factorial ANOVA which decides upon differences between objects, given a certain repeatability of the measurements. The variance contributions from the objects (most often called between-unit variance) and the repeatability (within-unit variance) are as follows.

Total sum of squares

$$SS_{\text{total}} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_i \right)^2 \tag{4a}$$

can be partitioned into

$$SS_{\text{within}} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(\overline{Y}_{ij} - \overline{\overline{Y}} \right)^2$$

$$SS_{\text{between}} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{\overline{Y}} \right)^2 \tag{4b}$$

7.1.3.2 What it is used for

ANOVA is used to compare sets of data attributed to different objects distinguished by one or a number of features. These features are normally called factors and may be of actually any character or origin. Thus, ANOVA can be applied to any data that can be grouped by a (or a number of) particular factor(s).

EXAMPLES Quality of crops grown on different fields, in different parts of a region or country; quality of goods produced on (parallel) conveyor lines; (categorized) taste or smell of, for example, food products; the impact of different operators, different instruments used, days or time-of-the-year on the results of a specified measurement.

An ANOVA separates different sources of variation, i.e. the variation within a group from the variations between groups (one-factorial design). It compares the estimates of variance for the factors with those of the residuals and decides, on the basis of F tests (see 7.1.3), about significant differences between groups or, in other words, the significant influence of the distinguishing factor(s).

7.1.3.3 Benefits

ANOVA is one of the mightiest tools in classical univariate data analysis. It provides clear indications on the significance or insignificance of influencing factors on the results obtained for different objects under different conditions, where the list of those different conditions may include nearly everything feasible in a measurement process.

As there is only a suspicion that groups of measurement results have been obtained on probably different objects or under probably different conditions, it is recommended to carry out, before any further assessment, an ANOVA with an appropriate number of factors. Results help distinguishing between significantly different objects and/or sets of influential conditions. This paves the way for the consolidation of data sets into a final result (plus a corresponding statement of uncertainty) which forms a reliable basis for making decisions, e.g. a pass/fail assessment versus a legal, commercial or otherwise agreed limit.

Note that there is a multivariate analogue considering several property characteristics at a time.

7.1.3.4 Limitations and cautions

Analysis of variance remains in the domain of classical statistics, meaning that ANOVA works only properly if

- the data sets, at least those belonging to a specific factor, follow a Gauss distribution,
- the data sets do not contain (very) significant outliers.

Furthermore, an insufficient repeatability may invalidate the decision power of the method. More in detail, a repeatability estimate may (also) depend on the object and/or the measurement conditions. Thus, an average repeatability may not always be a good basis for assessing the rest of the total variance.

Robust (i.e. less distribution-dependent) versions exist but are less understood, in particular concerning the implicit rejection or exclusion of data points classically identified as outliers.

7.1.4 General linear models

7.1.4.1 What it is

General linear models (GLM) is used for modelling the dependence of a single output variable from several input variables, with no restrictions to the number of the latter.

It is a mixture of an ANOVA considering any of the input variables as a factor, and a regression (see 7.2.1) using the simplest feasible model:

$$f(x) = a_0 + \sum_{p}^{P} a_p \cdot x_p) \tag{5}$$

From this simple model, statistical characteristics are determined allowing decisions on the significance of certain influential factors.

EXAMPLE A microplate array may have a significant sensitivity decrease from one upper/lower edge towards the other one due to the production process. GLM may reveal the effect, and classify it according to the ANOVA criteria.

7.1.4.2 What it is used for

GLM may help in significantly improving the quality of calibration and prediction. The dependencies on the different variables are partially random and sometimes go into different or even opposite directions and thus may cancel out. The resulting GLM model may therefore be much more reliable, and reveal less variation.

EXAMPLE In a laser-induced plasma spectroscopy (LIBS) analysis of cast iron, the nickel signal may be recorded at different wavelengths. These data can be conveniently combined via a GLM model.

7.1.4.3 Benefits

GLM helps to understand the influences of different (physical) sources of variation and deviation. It makes predictions more reliable. GLM also combines results obtained on a single measurand under different measuring conditions (wavelength, temperature, gas chromatography (GC) temperature programme, column selection, etc.) into a single result.

7.1.4.4 Limitations and cautions

All limitations as mentioned under 7.1.3 and 7.2.1 apply. In particular, one should be aware of the fact that not all measurands depend linearly on the influential factors.

7.2 Bivariate analysis

7.2.1 Regression analysis

7.2.1.1 What it is

7.2.1.1.1 General

Regression analysis defines parameters of a pre-selected model intended to fit experimentally obtained data under the general assumption that the experimentally obtained data depend, in

a certain describable manner, on one or more pre-selected independent influential factors (variables). Regression thus tries to find a set of parameters p_p (p = 1, ..., P) projecting a model G describing the dependence of a vector of y_k dependent variables (k = 1, ..., K) on a set of x_m independent variables (m = 1, ..., M), given a set of experimentally obtained data sets $\{\vec{Y}_n, \vec{X}_n\}$ with n = 1, ..., N.

$$\vec{y} = \vec{G}_{\vec{p}}(\vec{x}) \rightarrow \{\vec{Y}_n, \vec{X}_n\} \tag{6}$$

A basic pre-requisite of regression analysis is the number of available experimentally obtained data being (much) larger than the number of parameters of the pre-selected model (i.e. $N >> P$). This takes advantage of the (at least partially) random variability of any measurement which cancels out given a sufficient number of those data.

Multivariate approaches with mostly linear model functions will be handled in a subsequent part of IEC 62829. However, 7.2.1 considers bivariate cases for all kinds of model functions.

The most commonly used regression techniques refer to bivariate analysis, i.e. a dependent variable in functional relationship with respect to an independent one, use different minimization criteria, and are briefly described under 7.2.1.1.2 to 7.2.1.1.4. In these cases, Equation (6) becomes

$$y = g(x) \rightarrow \{Y_n, X_n\} \tag{7}$$

The parameters of the model function are determined by minimization of the sum of local metrics defining, in a certain way, the coincidence of a point exactly obeying the model function with the closest experimentally obtained data point. Most commonly used metrics are described below, however robust or correlation metrics are feasible.

NOTE In the presence of outliers or other unusual data in the experimentally obtained set, robust and resistant regression methods allow us to reliably test assumptions on functional relationships between variables. Some useful methods include Tukey's Three-Group Resistant Line Regression, Tukey's Biweight and Stepweight Regressions, the Theil-Kendall Median of Pairwise Slopes Regression and Hettmansperger's Rank Regression methods. Detailed explorations of the residuals from these regressions help in assessing whether relationships exist and whether they are linear or nonlinear.

Regression analysis also calculates statistics for assessing the overall reliability and the goodness-of-fit of the model selected with respect to the experimentally obtained data, and referring to the corresponding metric defined. Most often, the variances and covariances of the model function parameters are expressed in, and condensed to a variance-covariance matrix. Any (measurement) uncertainty calculation should refer to the full variance-covariance matrix.

7.2.1.1.2 Ordinary least-squares regression

This type of regression considers the local distances of the points $f(X_n)$ from the actually measured Y_n. The minimization criterion is

$$\vec{p} = \arg\min_{\vec{p}} \sum_{n}^{N} \left(Y_n - f(X_n)\right)^2 \tag{8}$$

For all model functions linear in parameters (such as all polynomials), Equation (8) has an analytical solution that can be calculated directly from an inversion of the normal equations. For non-linear model functions, the system of normal equations derived from Equation (8) has to be solved iteratively. Numerous efficient approaches such as simplex, steepest descent, Markward-Levenberg and others exist, and are comprehensively described in textbooks.

NOTE This type of regression analysis is most often used.

7.2.1.1.3 Weighted least-squares regression

This type of regression considers the local distances of the points $f(X_n)$ from the actually measured Y_n, weighted by the individual or somehow estimated uncertainties of these differences. Uncertainties of the independent variables are not taken into account. The minimization criterion is

$$\vec{p} = \arg\min_{\vec{p}} \sum_{n}^{N} \frac{\left(Y_n - f\left(X_n\right)\right)^2}{u\left(Y_n\right)^2} \tag{9}$$

where the $u(Y_n)$ are the uncertainty estimates (based on whatever is ruled out) of the Y_n. For the solution of Equation (9), basically the same as mentioned in 7.2.1.1.1 applies. The particular case of signal-proportional uncertainties $u(X_n) = \alpha \times X_n$ is handled in literature and leads to a case of 7.2.1.1.1. Note that most of the analysers used (not only in chemical analysis) obey the signal proportionality.

7.2.1.1.4 Generalized least-squares regression

This type of regression considers the local distances of the points $f(x_n)$ from the actually measured Y_n, weighted by the individual or somehow estimated uncertainties of these differences. Uncertainties of the X_n are taken into account by defining x_n pertaining to the function, having a difference from X_n, and their own $u(x_n)$. The minimization criterion is

$$\vec{p} = \arg\min_{\vec{p}} \sum_{n}^{N} \frac{\left(X_n - x_n\right)^2}{u\left(X_n\right)^2} + \frac{\left(Y_n - f\left(x_n\right)\right)^2}{u\left(Y_n\right)^2} \tag{10}$$

The solution to Equation (10) is always iterative, even for functions linear in parameters. Several International Standards make reference to this kind of regression (ISO 6143:2001, ISO 6974-1:2012, ISO 6974-2:2012, ISO 10723:2012, etc.). Particular solutions for power and exponential functions exist (B_LEAST; BAM 2001). The most obvious limitation to application is an underestimation of the accompanying uncertainty.

7.2.1.2 What it is used for

Major applications of regression according to 7.2.1.1.2 to 7.2.1.1.4 are

- calibration of instruments,
- determination of unknowns using the earlier defined calibration functions,
- description of processes, such as chemical reactions, dissolution, or evaporation from surfaces,
- modelling of objects, such as spectra, pattern, and shapes.

All four applications are widely in use, in the more complicated situations most often on the basis of ordinary least-squares regression.

For the description of processes and the modelling of objects, quite often sets of basic functions, i.e. Gaussian or Lorentz profiles, harmonics (overtones and undertones of sinusoids), or defined exponentials are used. Thus, the function $f(X)$ to be regressed might be split into a linear combination of basic functions $g\left(x, p(x)\right)$ according to Equation (11), with the a being weighting coefficients:

$$f(x) = \sum_{p}^{P} a_p \cdot g(x, \vec{p}) . \quad (11)$$

7.2.1.3 Benefits

Regression according to 7.2.1.1.2 to 7.2.1.1.4 is a reliable tool for attaining the objectives listed before, given the limitations are duly taken into account. Predictions (determinations) within the distances between sampling points $\{Y_n, X_n\}$ are reliable up to an uncertainty which may and should be calculated from the covariance matrix of the model function parameters (plus any additional influences not related to or caused by statistics).

7.2.1.4 Limitations and cautions

Independently of whether regression according to 7.2.1.1.2 to 7.2.1.1.4 is carried out on averages of replicate measurements or the single data obtained by experiment, regression assumes that measurements taken at a single X_n follow a normal distribution. For deviations and, in particular, outliers, see the note in 7.2.1.1.1.

When using approaches according to 7.2.1.1.2 to 7.2.1.1.4, consider that the reliability of the selected model function strongly depends on its selection. Given the same experimentally obtained data set $\{Y_n, X_n\}$, several model functions or model function types may be fit-for-purpose. Deciding which of those is most fit for the purpose is not a statistical problem but a matter of experience and prior knowledge.

If scientifically proven models exist but are not prevailing in an "everything is possible" regression analysis, the proven model should be given preference although some additional uncertainty contributions might be needed to be taken into account. These, most probably, come from method insufficiency, method incomparability, or insufficient implementation.

When all the above may be excluded on a reliable basis, statistics may help to assess the reliability of the selected model function. First of all, an ANOVA (7.1.3) gives indications about the percentage of total variation which can be attributed to modelling, and is applicable to any kind of model functions. It is case-dependent to decide whether a certain percentage of variation explained by the model might be sufficient to justify the model; however, a minimum of 70 % (around the 1-sigma of a Gauss distribution) might seem sensible.

Practitioners often use correlation analysis to justify the claim that the quality of their data analyses by conventional methods is good. This is often further justified by the quoting of the coefficient of determination R^2 for the regression. This can be dangerously misleading and totally unjustified. Users should

- inspect graphs/plots (both overall and residual) of the model function against the experimentally obtained data,
- check, and compare with critical values, the overall χ^2 value for the fit,
- use statistical tests, e.g. the Mandel test, to define whether deviations from the selected model function are significant.

If these tests are positive in the sense that pre-defined values are exceeded, select another model function.

7.2.2 Time series analysis

7.2.2.1 What it is

A time series is a sequence of data points made over a continuous time interval, out of successive measurements across that interval, using equal spacing between every two consecutive measurements, or with each time unit within the time interval having at most one

data point. Time series are very frequently plotted via line charts. Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data.

Techniques include

- autocorrelation analysis to examine serial dependence,
- spectral analysis to examine cyclic behaviour which need not be related to seasonality,
- separation into components representing trend, seasonality, slow and fast variation, and cyclical irregularity,
- curve fitting.

To some extent the different problems (regression, classification, fitness approximation) have received a unified treatment as supervised learning problems.

The general representation of an autoregressive model is

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t \tag{12}$$

where the term ε_t is the source of randomness and is called white noise. It is assumed to have the following characteristics:

$$\begin{aligned} E\left[\varepsilon_t\right] &= 0 \\ E\left[\varepsilon_t^2\right] &= \sigma^2 \\ E\left[\varepsilon_t \varepsilon_s\right] &= 0 \,\forall t \neq s \end{aligned} \tag{13}$$

Parameters are determined using regression techniques as described under 7.2.1.

7.2.2.2 What it is used for

Time series forecasting is the use of a model to predict future values based on previously observed values, and is the extension of regression analysis (7.2.1) to future expectations. Time series data have a natural temporal ordering, and models will often make use of the natural one-way ordering of time so that values for a given period will be expressed as deriving direction from past values. Time series analysis can be applied to real-valued, continuous data, discrete numeric data, or discrete symbolic data (i.e. sequences of characters).

In the context of statistics, econometrics, quantitative finance, seismology, meteorology, and geophysics, the primary goal of time series analysis is forecasting. In the context of signal processing, control engineering and communication engineering it is used for signal detection and estimation, while in the context of data mining, pattern recognition and machine learning time series analysis can be used for clustering, classification, query by content, anomaly detection as well as forecasting.

7.2.2.3 Benefits

Statistics about a sample of a population might be used to characterize the whole population, and other related populations, but this is not necessarily the same as prediction over time. When information is transferred across time, often to specific points in time, the process delivers forecast information by classification, regression analysis (method of prediction), signal estimation, and segmentation.

7.2.2.4 Limitations and cautions

Basically the same limitations as to regression analysis apply. Models for time series data can have many forms and represent different stochastic processes. When modelling variations in the level of a process, three broad classes of practical importance are the autoregressive (see above) models, the integrated models, and the moving average models. These three classes depend linearly on previous data points. Combinations of these ideas produce autoregressive moving average and autoregressive integrated moving average models.

In recent work on model-free analyses, wavelet transform based methods (for example locally stationary wavelets and wavelet decomposed neural networks) have gained favour. Multiscale (often referred to as multi-resolution) techniques decompose a given time series, attempting to illustrate time dependence at multiple scales.

However, any forecast on the basis of previous data is unpredictable, best seen from the stock exchange crisis in 2008 where indications were pointing up, but real rates were plunging to extents where even “serious” finance organizations went bankrupt.

Annex A
(informative)

Advice on software validation for process analytical applications[1]

A.1 General

Activities and issues that need IT control within a laboratory's management system with respect to ISO/IEC 17025 can be grouped into two fields:

- general activities of a laboratory that need to be controlled according to the standard in general, whether they are conducted with or without a computer system; for these activities, a laboratory needs to define policies and procedures. In the case of using a computer system, this has to comply with the system requirements;
- special requirements that have to be met for the software and systems used. Software and computer system are validated and/or verified. It is assumed that a laboratory will have measures to comply with the general requirements of ISO/IEC 17025. Therefore, this guidance focuses on the special requirements concerning software and computer system validation, including:
  - identification and interpretation of computer and software clauses in ISO/IEC 17025,
  - implementing computing systems in the laboratory,
  - different categories of software,
  - risk assessment including security,
  - verification and validation of software,
  - electronic documents handling, transmission and archiving (Clause 7),
  - usage of computer networks in connection with the measurement process,
  - security.

A.2 Basic recommendations

Implementation and validation of hardware and software depends on the kind of hardware or software installed. Table A.1 categorizes different types of software in five different categories.
Table A.1 contains examples of groups of programmes as well as examples of specific programmes. There are other ways of categorizing software, e.g. COTS (commercial off-the-shelf), MOTS (modified off-the-self) and CUSTOM; these categories are also indicated in Table A.1.

Before new software, computers, equipment containing computers, etc. are introduced in a laboratory, the risk connected with such an introduction should be assessed. A risk assessment should be performed to evaluate the extent and content of validation and/or verification required. Such an assessment may include, but not be restricted to

1) identification of possible events which may result in a non-compliance with respect to ISO/IEC 17025 (e.g. non-conforming results),
2) estimation of the likelihood of such events,
3) identification of the consequences of such events,

[1] Main issues taken from the EUROLAB Technical Report No. 2/2006.

4) ways of avoiding the events (e.g. by use of check standards or reference materials in calibration/testing),
5) costs, drawbacks, benefits, etc. occurring when the ways in 4) are chosen,
6) decision on activities,
7) office or testing software.

The outcome of the risk assessment is used to determine the extent of the validation of both software and test and calibration methods.

Table A.1 – Categories of software

Category	Types	Groups of programmes, examples	Examples of programmes[a]
1 (COTS)	Operating systems	Operating systems	Windows, LINUX
2 (COTS)	Firmware, (COTS)	Embedded software, built-in software	Instruments, voltmeters, tensile testing machines
3 (COTS)	Standard software packages, Commercial off the shelf	E-mail programmes, word processors	Word, Excel (as a table only), etc., Outlook, Internet explorer, Acrobat, Stock instrument controlling software used “out of the box”.
4 (MOTS)	Configured software packages (MOTS, modified of the shelf)	Programmes as a programming and configuration environment. Tailoring & Customization is needed prior to use.	Excel formulae, Labview, Lab-windows, Labtech Notebook, Mathcad
5 (CUSTOM)	Custom or bespoke software	Custom written software using software programming tools. Includes Word/Excel documents with macrocode (VBA code).	Applications written in C++, SQL+, Java Visual Basic, XML, LabVIEW, LabWindows and other languages. Application may also in some cases be considered as Custom.

[a] Trade names given in this table are examples of suitable products available commercially. This information is given for the convenience of users of this document and does not constitute an endorsement by IEC of these products.

The validation/verification of software is dependent on whether it is bought or custom built. The extent of effort required should be based on risk assessment. Purchased software should be checked (verified) to confirm it usability in the user environment. This is typically evident by acceptance testing against manufacturer specifications and/or user requirements. Custom built or modified software is validated to the extent necessary. Table A.2 provides an overview.

Table A.2 – Software validation levels

V0	Manufacturer’s documentation
V1	Requirement specification
V2	Design and implementation (coding)
V3	Inspection and structural testing (“white-box” testing)
V4	Installation
V5	Acceptance test (“black-box” testing)
V6	Operation and maintenance

A.3 Software validation

According to ISO/IEC 17025:2017, 7.11, all software used for handling calibration or test data shall be validated, except for software in categories 1 to 3 in Table A.1, where the "validation" is limited to an acceptance test. However, all software used for testing or calibration shall be capable of achieving the accuracy required and comply with relevant specifications. Thus, validation phases V1 and V4 shall be completed even for software in category 3. The level of validation depends on the software type and its application. Other validation reports as well as any history of error-free operation may be included as part of the laboratory validation of a software product.

Figure A.1 shows the different paths, depending on the software category, for introduction of new or revised software in a laboratory. It is clear that new versions of software need to be checked/validated before being introduced in the laboratory. The extent of the validation depends on the software and its use as well as on the risk connected with its use.

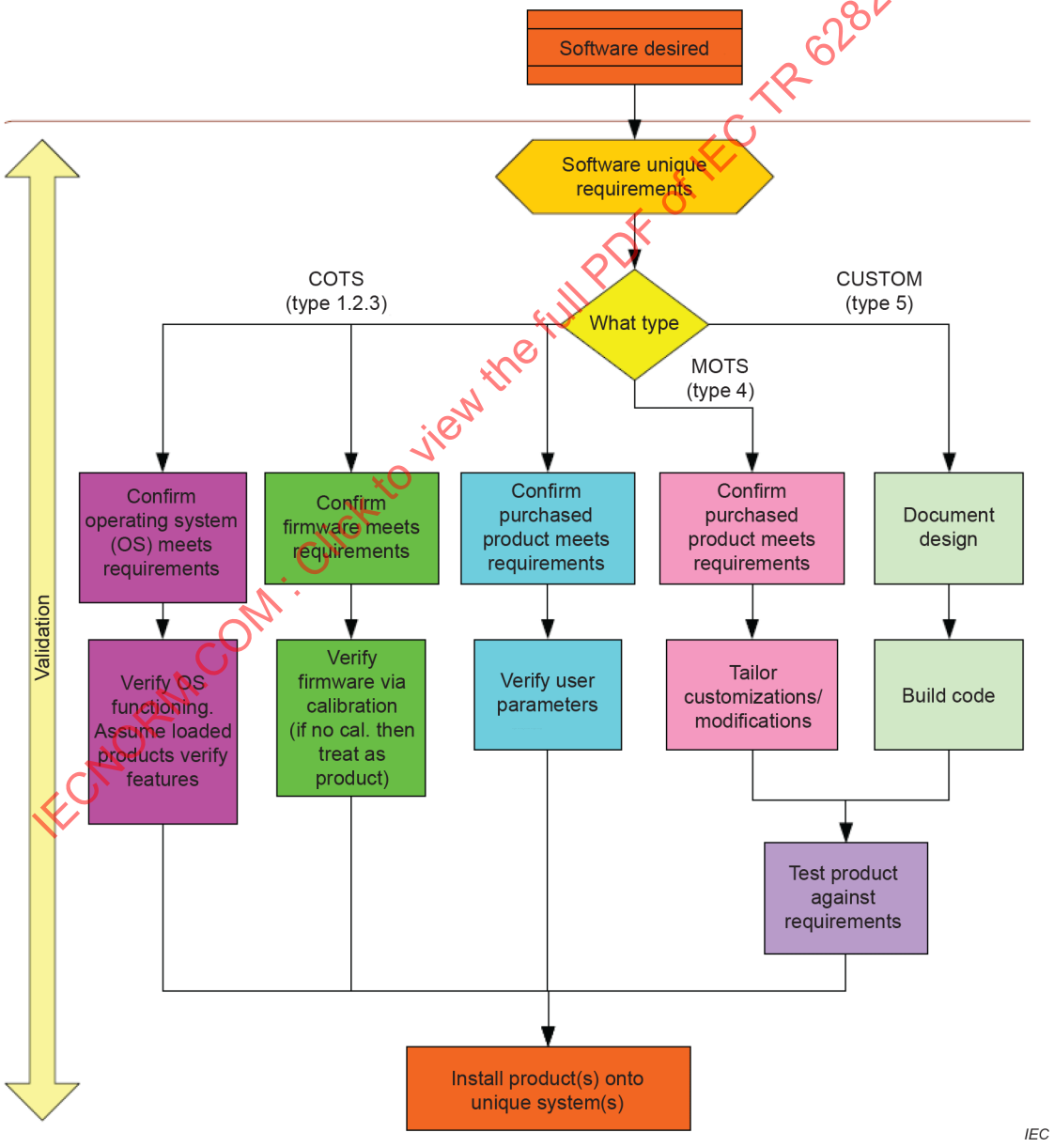


Figure A.1 – Different paths for the introduction of new software in a laboratory