Société de Calcul Mathématique SA
(Mathematical Modelling Company, Corp.)
Tools for decision help
since 1995

Quality of Information:

detecting aberrant data, reconstructing missing data

I. General description of the topic

Any company or institution has data, related to its activity. Usually, they are hard to collect, but they are precious:

- They allow previsions, such as an anticipation of the sales, by product or by sector, they allow the definition of the resources needed for production, logistics, sales, and so on;
- They allow the identification of various difficulties, that is the situations where a risk may appear: in some cases, things did not work correctly;
- They allow proper information of the shareholders, of the customers, of the public, on a factual basis.

However, despite their strategical importance, databases are quite often of poor quality: many aberrant data, many missing data. At worst, the whole set may lose any credibility.

For many years, our Company has developed probabilistic techniques, which allow the improvement of the whole Information System:

- detecting aberrant data;
- reconstructing missing data.

A. Detecting aberrant data

The percentage of aberrant data in databases is often high (more than 10%) and sources of mistakes are numerous:

- Human origin: mistakes in the transcription of the data (mistakes in the units, in the dates, in the various fields); they are hard to control;

- Mistakes due to the measurement instruments (poor calibration, insufficient precision).

During the years 2010-2016, we performed several tasks for the Nuclear Energy Agency (specialized Agency of the OECD, in Paris), concerning databases in the nuclear sector. We identified anomalies which may concern a single situation (isolated singularity) or a whole set of data, which are coherent between themselves, but with a tendency which differs clearly from other points.

We conceived methods for automatic detection, with a number of false alarms which was quite small. This work was published.

When the database has been verified using such methods, the information may be put at the disposition of the users. They will see "reliability indicators" which give an information about the quality of the data. The users may choose to work on all data, or to restrict themselves to the data which have been treated.

B. Reconstructing missing data

Almost all databases have "holes", that is missing data; the reasons may be:

- The person in charge was not present;
- The sensor did not work;
- The budget had disappeared;
- The data were collected but then destroyed or lost.

In 2007, in the framework of a contract with "Veolia Environnement", West Region of France, we had to reconstruct the debits of 19 rivers in Vendée, over 37 years, with 50% of missing data! The probabilistic techniques we used for that are detailed in our book "Probabilistic Methods for the Reconstruction of Missing Data" (in French, reference below).

But there is a positive side to missing data: one spares money! Appropriate methods, deciding in advance not to collect some data and explaining how to reconstruct them, allow to reduce the number of sensors and of measurements.

II. Our recent realizations

1. Books

Bernard Beauzamy and Olga Zeydina: "Probabilistic Methods for the Reconstruction of Missing Data" (in French: Méthodes probabilistes pour la reconstruction de données manquantes). SCM SA, ISBN: 2-9521458-2-2, ISSN: 1767 – 1175, April 2007.

Olga Zeydina and Bernard Beauzamy: Probabilistic Information Transfer (in English). SCM SA. ISBN: 978-2-9521458-6-2, ISSN: 1767-1175. May 2013.

2. Publications

[1] Bernard Beauzamy, Hélène Bickert, Olga Zeydina (SCM), Giovanni Bruna (IRSN): Probabilistic Safety Assessment and Reliability Engineering: Reactor Safety and Incomplete Information. Proceedings of ICAPP 2011 Nice, France, May 2-5, 2011 Paper 11399
http://scmsa.eu/RMM/ART_2011_ICAPP_11399.pdf

[2] Emmeric Dupont (NEA), Bernard Beauzamy (SCM), Hélène Bickert (SCM), M. Bossant (NEA), Carmen Rodriguez (SCM), N. Soppera (NEA): Statistical Methods for the verification of databases. Publication de la Nuclear Energy Agency de l'OCDE, 2011.
http://www.oecd-nea.org/nea-news/2011/29-1/29-1-int-e.pdf#page=31

[3] O. Zeydina (SCM), A.J. Koning (NEA), N. Soppera (NEA), D. Raffanel (SCM), M. Bossant (NEA), E. Dupont (NEA), and B. Beauzamy (SCM): Cross-checking of large evaluated and experimental databases, Science Direct, Nuclear Data Sheets 120 (2014) 277–280.
http://www.scmsa.eu/archives/NEA_SCM_2014.pdf

[4] F. Godan (SCM), O. Zeydina (SCM), Y. Richet (IRSN), B. Beauzamy (SCM): Reactor Safety and Incomplete Information: Comparison of Extrapolation Methods for the Extension of Computational Codes. Proceedings of ICAPP 2015 Nice, France, May 3-6, 2015, Paper 15377.
http://scmsa.eu/archives/ART_IRSN_SCM_15377.pdf

[5] Emmeric Dupont (CEA): Exfor: Improving the quality of International Databases. NEA News, 2014, 32.1, page 28.
http://www.scmsa.eu/archives/EXFOR_NEA_News_2014_32.pdf

[6] Achim Albrecht (ANDRA) and Stephan Miquel (SCM): Modelling soil and soil to plant transfer processes of radionuclides and toxic chemicals at long time scales for performance assessment of Radwaste disposal. Geophysical Research Abstracts, Vol. 17, EGU2015-10476-1, 2015
http://www.scmsa.eu/archives/ART_Albrecht_Miquel_Modelling_Soil_2015.pdf

[7] Gottfried Berton (SCM) : Verification of the databases EXFOR and ENDF. Nuclear Energy Agency, JEFF Meetings - Session JEFF Experiments, November 28 - December 1, 2016.
http://www.scmsa.eu/archives/SCM_NEA_JEFF_Meeting_2016_11.pdf

[8] Gottfried Berton, SCM SA, and Oscar Cabellos, NEA : Checking the resolved resonance re-gion in EXFOR database. JEFF Meetings - Session JEFF Experiments, November 20 - 24, 2017. http://www.scmsa.eu/archives/SCM_NEA_JEFF_Meeting_november_2017.pdf

3. Contracts

In the following contracts, the detection of erroneous data and the reconstruction missing data played an essential role.

- Veolia Environnement, 2005: Analysis of the lack of water in Vendée.
- European Environment Agency, 2006-2015: Probabilistic Methods for Water Quality (SCM won an international framework contract on this question in 2006 ; it was valid for four years and was renewed in 2012 for four more years).
- Veolia Environnement, Région Ouest, 2007: Detecting malfunctions in sensors networks (quality of water).
- Veolia Environnement, Région Ouest, 2007-2009: Building a panel of consumers, in order to anticipate the needs for water consumption.
- Institut de Radioprotection et de Sûreté Nucléaire, 2007-2011: Applications of the Experimental Probabilistic Hypersurface (a method invented by SCM) to safety problems for nuclear reactors.
- International Stainless Steel Forum, 2008: General analysis of the information system and recommendations with respect to the statistical analysis of the data.
- Réseau Ferré de France (French Railways), 2008-2013: Statistical analysis of the reasons why the trains are late (region of Paris) and recommendations for corrections.
- Agence de l'Eau Artois-Picardie, 2008: Probabilistic study of the situations where water in the rivers is of good quality.
- Groupe Novalis-Taitbout, 2008: Critical analysis of some dispositions related to employment.
- Snecma Propulsion Solide, 2009: Probabilistic methods for reliability.
- Caisse Centrale de Réassurance, 2009: Probabilistic study related to rivers flows.
- Fédération des Établissements Hospitaliers et d'Aide à la Personne, 2009: Developing an information system.
- Areva, 2010: Probabilistic methods for the analysis of a site for nuclear waste.
- Brigade des Sapeurs Pompiers de Paris (Paris Firemen), 2010: Probabilistic study related to the intervention of Firemen (their database was found to be of extremely good quality).
- Agence Nationale de l'Habitat, 2010: Probability laws related to payment delays.
- Nuclear Energy Agency (OCDE), 2010-2012: Detecting aberrant data in nuclear databases.
- ArcelorMittal, 2011-2012: Probabilistic methods for the hierarchy of parameters in an industrial process.
- GDF-SUEZ, 2012-2013: General analysis of the quality of data, gas ditribution.
- Areva, 2012-2013: Analysis of the uncertainties in an industrial process.
- Air Liquide, 2012: General reliability analysis.
- Institut de Radioprotection et de Sûreté Nucléaire (Nuclear Safety), 2012: Statistical analysis of environmental data.
- DCNS (Nuclear submarines), 2013: Probabilistic methods in order to improve a welding equipment.

- RFF (French railways), 2013: Tool for decision help: where to spend the money in order to improve the regularity of trains.
- Caisse Centrale de Réassurance, 2013-14: repartition of the damages linked with natural accidents.
- COSEA (High Speed Train Sud Europe Atlantic), 2013: Probabilistic estimate for extreme floodings.
- Coop de France déshydratation, 2013: Statistical analysis for organic components
- Monceau Assurances, 2013-2014: Improving the tarification.
- Nuclear Energy Agency (OCDE), 2014, 2015, 2016: Looking for erroneous data in large nuclear databases.
- Poste Immo, 2014: Tool for decision help: where to spend the money in order to improve the general energy consumption of the buildings belonging to the French Post Office ?
- Secrétariat Général pour l'Administration, Ministère de l'Intérieur, Région Est, 2016 and 2018 : Analysis of the quality of data for crisis management.
- RATP, 2016-2018 : Analysis of the quality of information in the situation of emergency braking
- Taxis G7, 2016-2017 : Correcting the data collected by drivers
- SEDIF, 2017 : Analysis of the data associated with leaks
- Monceau Assurances, 2016-2018 : Improving the commercial strategy
- Bureau de Recherches Géologiques et Minières, 2018 : Finding thresholds for pollution in soils