Société de Calcul Mathématique SA
Outils d'aide à la décision
depuis 1995

La qualité et le traitement des données recueillies :

Un enjeu de survie pour les entreprises

I. Présentation du besoin

Toutes les entreprises recueillent des données relatives à leur activité, ne serait-ce que pour établir le bilan annuel. On y ajoute souvent des données provenant de l'enregistrement des process industriels, de la logistique, de l'état des stocks, du mouvement du personnel, etc. Elles sont faciles à recueillir : on met des capteurs partout, qui délivrent des Go de données chaque jour ; on stocke tout cela dans des disques durs organisés pour la circonstance. Après quoi, que fait-on des données ainsi recueillies ? Dans l'immense majorité des cas, la réponse est très simple : RIEN !

Dans l'immense majorité des cas, les entreprises n'ont même pas vérifié que les données ainsi recueillies étaient exploitables. C'est l'analogue de bouts de papier que vous entassez au fond d'une armoire, au fur et à mesure de leur arrivée : le jour où l'on en a besoin, allez donc retrouver le document pertinent !

Pourtant, ces données sont essentielles pour la survie des entreprises, car elles recensent toutes les difficultés auxquelles l'entreprise a été exposée dans le passé et auxquelles elle risque d'être exposée dans l'avenir.

Un exemple concret pour faire comprendre ceci : pour PSA, en 2020, nous avons montré que le risque de défaillance d'un fournisseur de rang 1 était mal pris en compte. Le retard (15 j) dans la fourniture d'une pièce de valeur < 1 € a compromis pendant plusieurs semaines la fabrication d'un modèle de véhicule.

Notre recommandation de principe à toutes les entreprises, à toutes les institutions :

Une fois par an, analysez les données que vous avez recueillies, en vous demandant :

- *si elles sont de bonne qualité ;*
- *si elles répondent à un besoin.*

Entrons maintenant dans l'analyse technique du sujet :

II. Traitement de données anciennes

Les données collectées doivent permettre :

- L'identification des difficultés, des situations à risque : on voit que dans certains cas, les choses ne se passent pas correctement ;
- La prévision : anticiper les ventes par produit ou par secteur, dimensionner la production, la logistique ou les effectifs ;
- L'information des actionnaires, des clients, du public, sur une base factuelle et incontestable.

Malgré leur intérêt stratégique, les bases de données sont très souvent imparfaites : nombreuses données aberrantes, nombreuses données manquantes. Au pire, l'ensemble tout entier peut perdre toute crédibilité. Il y a souvent des changements de logiciels, de méthodes de mesure, et les données précédentes n'ont pas été converties.

Depuis 1995, la SCM SA développe des méthodes probabilistes, sans hypothèses factices, pour traiter les données anciennes. Nos analyses et conclusions peuvent être soumises aux Autorités dans le cadre d'une démonstration de sûreté. Elles servent également à l'évaluation des risques pour un projet nouveau (durée de retour de phénomènes climatiques, par exemple).

Nous distinguons trois axes de travail, qui sont complémentaires :

- La détection de données aberrantes ;
- La reconstruction de données manquantes ;
- La régularisation de données "bruitées".

Traitons-les dans l'ordre.

A. Détection des données aberrantes

Le pourcentage de données aberrantes est souvent élevé (plus de 10 %) et les sources d'erreurs sont multiples :

- Origine humaine accidentelle : erreurs de report lors de la transcription des données (erreur d'unité, de date, de champ), très difficiles à limiter ;
- Origine humaine volontaire : la fraude. Voir à ce sujet notre fiche de compétences "lutte contre la fraude" : https://www.scmsa.eu/fiches/SCM_Lutte_contre_la_fraude.pdf
- Instruments de mesure (mauvais calibrage, précision insuffisante).

En 2010-2012, 2014, 2015, 2016, 2017, pour l'Agence pour l'Énergie Nucléaire (Agence spécialisée de l'OCDE), nous avons identifié des anomalies concernant une seule donnée aberrante (singularité isolée), ou un ensemble de données cohérentes entre elles, mais avec une tendance différente de celle de l'ensemble des points. Nous avons mis en place des méthodes de détec-

tion automatique, présentant un taux de fausses alertes aussi faible que possible. Ces travaux ont donné lieu à publications (voir ci-dessous).

Une fois la base de données vérifiée par ces méthodes, l'information peut être mise à la disposition des utilisateurs : ils verront des "indicateurs de fiabilité", qui renseignent sur la qualité des données. Ils pourront choisir de travailler sur l'ensemble, ou bien d'utiliser uniquement les données les plus fiables.

B. Reconstruction des données manquantes

La quasi-totalité des bases de données présente des "trous" :

- Absence de la personne en charge de la mesure ;
- Panne de l'instrument de mesure ;
- Panne de budget ;
- Effacement, destruction, usure, etc.

En 2007, la SCM SA, dans le cadre d'un contrat avec Veolia Environnement, Région Ouest, a reconstitué les débits de 19 fleuves en Vendée, sur 37 années, avec 50 % de données manquantes. Les méthodes probabilistes que nous avons mises au point à cet effet sont détaillées dans notre livre "Méthodes probabilistes pour la reconstruction de données manquantes".

Mais les données manquantes ont aussi un aspect positif : elles permettent de faire des économies ! Des méthodes appropriées, destinées à décider d'avance quelles données ne seront pas collectées et comment l'ensemble sera reconstitué, permettent donc d'économiser des capteurs, des mesures et des ressources humaines.

C. Régularisation de données bruitées

Prenons l'exemple d'une structure qui se déforme très lentement (un pont, un bâtiment), ou bien d'un mobile dont on cherche à mesurer la position. Si les données sont rares et peu précises, la série temporelle ainsi obtenue est "bruitée" ; elle ne montre pas clairement la déformation ou le mouvement attendus. Il est nécessaire de "régulariser" les données, mais sans faire d'hypothèse de modèle (sans supposer par exemple que le mouvement est uniforme dans le temps, ni que les erreurs consécutives sont indépendantes ou gaussiennes), parce que de telles hypothèses sont inacceptables pour les Autorités de Sûreté. Les méthodes purement probabilistes que nous mettons en œuvre permettent de régulariser les données et d'obtenir une présentation qui pourra être soumise aux Autorités de Sûreté.

III. Nos réalisations récentes

1. Livres

[IEPE] Bernard Beauzamy : Introduction à l'étude des Probabilités Expérimentales. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA, ISBN 979-10-95773-02-3, ISSN 1767-1175. Relié, 192 pages. Janvier 2023.

[RDM] Bernard Beauzamy et Olga Zeydina : Méthodes probabilistes pour la reconstruction de données manquantes. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA, ISBN : 2-9521458-2-2, ISSN : 1767 – 1175, avril 2007.

[PIT] Olga Zeydina et Bernard Beauzamy : Probabilistic Information Transfer. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA. ISBN : 978-2-9521458-6-2, ISSN : 1767-1175. Relié, 208 pages, mai 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 region 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. Contrats

Avant toute analyse, par principe, nous commençons par vérifier la qualité des données, selon les principes décrits ci-dessus : la validité des conclusions en dépend évidemment. En particulier, dans tous les contrats ci-dessous, la détection de données aberrantes et la reconstruction de données manquantes ont joué un rôle essentiel :

- Agence Européenne de l'Environnement, 2006-2013 : Méthodes probabilistes pour la qualité de l'eau
- Veolia Environnement, Région Ouest, 2007 : Détection de dysfonctionnements dans les réseaux de capteurs
- Veolia Environnement, Région Ouest, 2007-2009 : Constitution d'un panel de consommateurs et prévision des consommations d'eau potable
- Institut de Radioprotection et de Sûreté Nucléaire, 2007-2011 : Applications de l'Hypersurface Probabiliste aux problèmes de sûreté des réacteurs nucléaires
- Réseau Ferré de France, 2008-2013 : Etude statistique concernant les causes des retards des trains en Ile de France
- Agence de l'Eau Artois-Picardie, 2008 : Etude probabiliste concernant la qualité des eaux de rivière et caractérisation des situations de bonne qualité
- Groupe Novalis, 2008 : Analyse critique de l'efficacité de certains dispositifs d'aide
- Snecma Propulsion Solide, 2009 : Méthodes probabilistes pour la fiabilité
- Caisse Centrale de Réassurance, 2009 : Etudes probabilistes relatives aux débits des rivières
- Fédération des Établissements Hospitaliers et d'Aide à la Personne, 2009 : Développement d'un système d'information
- Areva, 2010 : Méthodes probabilistes pour l'étude d'un stockage de déchets radioactifs
- Brigade des Sapeurs-Pompiers de Paris, 2010 : Etude statistique relative aux interventions
- Agence Nationale de l'Habitat, 2010 : Lois de probabilité relatives aux délais de paiement ;
- Nuclear Energy Agency (OCDE), 2010-2012, 2014, 2015, 2016, 2017 : Détection de données aberrantes dans les bases de données
- Air Liquide, 2011 : Construction d'un "indice de proximité" entre pipe-lines
- ArcelorMittal, 2011-2012: Méthodes probabilistes pour la hiérarchisation des paramètres dans un process industriel
- GDF-SUEZ, 2012-2013 : Analyse générale de la qualité des données, distribution du gaz
- Areva, 2012-2013 : Analyse des incertitudes dans un process industriel
- Air Liquide, 2012 : Analyse générale de fiabilité ; interopérabilité de plusieurs bases de données
- IRSN, 2012 : Analyse statistique préliminaire de données de radioactivité dans l'environnement
- DCNS, 2013 : Méthodes probabilistes pour l'amélioration d'un procédé de soudage

- Caisse Centrale de Réassurance, 2013-14 : Ventilation des sinistres "catastrophes naturelles"
- COSEA (Ligne à Grande Vitesse Sud Europe Atlantique), 2013 : Estimation de la durée de retour de crues extrêmes
- Coop de France déshydratation, 2013 : Réalisation d'un outil d'analyse des COVNM
- Monceau Assurances, 2013-2014 : Amélioration de la politique tarifaire
- Poste Immo, 2014 : Amélioration de la politique énergétique des bâtiments de La Poste
- Secrétariat Général pour l'Administration, Ministère de l'Intérieur, Région Est, 2016 et 2018 : Analyse de la qualité des données pour la gestion des crises
- RATP, 2016-2018 : Modélisation du comportement des trains en situation de freinage d'urgence
- Taxis G7, 2016-2017 : Correction des données utilisées par les chauffeurs
- SEDIF, 2017 : Etudes d'algorithmes pour le réseau
- Monceau Assurances, 2016-2018 : Amélioration de la politique commerciale
- Bureau de Recherches Géologiques et Minières, 2018 et 2019 : Méthodes mathématiques robustes pour la détermination de seuils de pollution dans le sous-sol
- Atlandes, 2018 : Comptage des véhicules sur les bretelles de sortie d'une autoroute
- Coop de France Déshydratation, 2019 : Analyses statistiques
- Transporteur, 2019 : Analyses statistiques des données de position émises par des containers
- Orano Mining, 2019 : Hiérarchisation de paramètres intervenant dans un process industriel
- Groupe Atlantic, 2019 : Analyse probabiliste des appels au Service Après-Vente
- PSA, 2020 : Analyse critique des seuils de réassurance
- Coop de France Luzerne, 2019 : Analyses statistiques et comparaisons entre usines
- Coldway Technologies, 2020 : Réalisation d'une démonstration de sûreté
- Ministère de l'Intérieur, SGAMI, 2020 : Appui méthodologique relatif au Télétravail
- Atlandes, Autoroute A63, 2020 : Statistiques relatives aux Poids Lourds
- SARP Industries, 2021 : réglage d'un four
- Monceau Assurances, 2020-2021 : Politique de réassurance "catnat" ; analyse critique des données relatives aux tempêtes
- Bouygues Energies & Services, 2022 : Appui méthodologique à la conception d'un système d'information "Dysfonctionnements et Maintenances"
- Befesa Valéra, 2022 : Hiérarchisation des paramètres intervenant dans le réglage d'un four
- Léon Grosse, 2022 : Analyse du risque "grêle"
- SNCF, 2023 : Appui méthodologique aux plans d'inspection des rails
- Coop de France Luzerne, 2023 : Analyses statistiques
- Agence Nationale des Titres Sécurisés, 2023 : Anticipation des demandes en Titres Sécurisés
- Peptinov, 2023 : Traitement probabiliste de données épidémiologiques
- Cristal Union, 2023 : Méthodes probabilistes pour la comparaison d'essais de biocides
- Airbus Beluga Transport, 2024 : Mise en place d'un Système d'Information "Missions"
- Coopération Agricole "Luzerne de France", 2024 : Homogénéisation de bases de données
- SNCF, 2024 : Analyse d'une approche probabiliste de valorisation des risques associés aux coûts des projets

- Bureau de Recherches Géologiques et Minières, 2024 : Outils mathématiques pour la cartographie des pollutions
- Société SNF, 2024 : Evolution des températures, précipitations, phénomènes extrêmes, sur 7 sites dans le monde
- Ville de Villiers le Bâcle, Essonne, 2024 : Calcul de la probabilité de retour de pluies extrêmes