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

La conception des bases de données :

Précautions, Principes et Méthodes

Depuis la création de la SCM en 1995, nous avons rencontré très peu de bases de données qui soient correctement réalisées et dont l'exploitation ne nécessite pas un important traitement préalable. Voici des exemples :

- Base de données "incidents trains" de la SNCF, un million de lignes par an : forte proportion de données inexploitables ou incohérentes ;

- Bases de données "post Tchernobyl", expertisées par la SCM pour le compte de l'IRSN : 20 années d'observations (taux de radioactivité, taux de cancers, etc.), mais données mal enregistrées, formats incohérents. Exploitation finale presque impossible ;

- Base de données clients, pour une institution de prévoyance : très nombreuses erreurs de saisie, informations incomplètes (on ne connaît pas les dates de début ou de fin des contrats). Exploitation finale presque impossible ;

- Base de données de positions GPS de containers, pour un transporteur aérien, 2019 : extractions très difficiles, nombreuses données incohérentes, pas de finalité clairement définie.

Bien que ces BDD aient coûté fort cher à constituer, leur exploitation finale est décevante.

Seule exception notable : la base de données d'interventions de la Brigade de Sapeurs-Pompiers de Paris, sur laquelle nous avons travaillé en 2010, qui était d'excellente qualité et contenait moins de 1% de données aberrantes.

Pour la constitution d'une base de données, les règles suivantes devraient être observées :

1. Réfléchir d'abord à ce que l'on souhaite enregistrer

Beaucoup d'organismes enregistrent n'importe quoi, avec n'importe quel pas de temps (toutes les minutes, tous les jours, etc.). Ceci n'a pas de sens : la nature de l'information collectée dépend de ce que l'on veut en faire. Par exemple, pour Veolia Transport, pour la conception d'un réseau de bus dans une ville donnée, nous avions fait la suggestion de deux BDD, une grossière et une précise :

- La BDD grossière contient les densités de population et les densités d'emplois sur chaque carré de 400 m de côté ;
- La BDD précise contient les "points of interest" (lycées, hôpitaux, etc.) avec leurs coordonnées et leur fréquentation.

De manière générale, il faut commencer par réfléchir à l'exploitation que l'on veut faire de la base de données. Par exemple :

- Exploitation annuelle, dans un but de tableau de bord d'entreprise (très grossier) ;
- Exploitation instantanée, pour commander des pièces détachées (très fin).

Entre les deux, de nombreuses variantes sont possibles. Par exemple, pour Veolia Environnement Région Ouest, nous avons constitué un "panel" de consommateurs, pour anticiper la consommation d'eau chaque trimestre : il suffit de connaître les relevés tous les trois mois.

Les bases de données "environnement" (taux de pollution, etc.) se contentent en général d'une moyenne annuelle, si l'exploitation est politique.

Il ne faudrait pas croire que la BDD fine soit toujours préférable à la BDD grossière ; c'est tout l'inverse. Si la BDD est trop fine, elle comporte une énorme quantité de références diverses, dont on ne sait pas quoi faire. Nous avons eu un contrat avec une fédération d'hôpitaux : il s'agissait d'étudier l'impact de changements tarifaires. Mais si on descend au niveau de la nomenclature des cotons-tige, on ne peut savoir quelle information est pertinente.

2. Mettre en place des systèmes d'aides et d'alerte lors de la constitution de la BDD

Notre expérience est catégorique : une fois qu'une fiche est mal remplie, c'est irrattrapable, car elle est noyée au milieu de toutes les autres. Il faut donc aider la personne qui réalise les fiches :

- **Par des guides**

    1. Le format de date doit être imposé ;
    2. La date de fin d'une opération doit être postérieure à celle de début ;
    3. Les grandeurs sont positives, etc.

Tout ceci peut être vérifié de manière informatique au cours de la saisie, et un menu d'aide apparaît : "fiche mal remplie, vérifier xxx".

- **Par des vérifications**

Des vérifications simples peuvent permettre de détecter les anomalies : cohérence entre l'amont et l'aval, entre deux capteurs voisins, entre une journée et la veille, cohérence d'un remboursement d'assurance avec le coût à neuf du véhicule, etc. Elles peuvent parfaitement être automatisées. Notre expérience est ici que ces mesures simples permettraient d'éviter 90% des erreurs commises.

Nous avons développé des méthodes probabilistes robustes pour la détection de données aberrantes et la reconstruction des données manquantes, qui sont développées dans deux livres [RDM] et [PIT] ; voir ci-dessous.

Nous recommandons à quiconque développe une base de données de la vérifier au moins une fois par an. A la fin de chaque année, on devrait passer au crible les données qui viennent d'être enregistrées, et se poser les questions vues plus haut : sont-elles de bonne qualité ? sont-elles pertinentes ? etc.

Nous préciserons cette recommandation de la manière suivante : il sera bon de se doter d'une base de données "tampon" (c'est-à-dire intermédiaire) qui accueillera les données fraîchement recueillies ; elles ne seront versées dans la base principale qu'après vérification.

Il est bon de confier la réalisation de la base de données finale à quelqu'un spécialement désigné, et dont c'est le métier. Il ne faut pas confier cette tâche aux ingénieurs "terrain" : ils ont autre chose à faire, en particulier corriger les dysfonctionnements des équipements et la constitution d'une base de données, pour eux, est d'intérêt secondaire.

3. Livres édités par la SCM

[RDM] Bernard Beauzamy et Olga Zeydina : Méthodes probabilistes pour la reconstruction de données manquantes. ISBN 2-9521458-2-2, ISSN 1767-1175. SCM SA, avril 2007.

[PIT] Olga Zeydina et Bernard Beauzamy : Probabilistic Information Transfer (en anglais), ISBN 978-2-9521458-6-2, ISSN 1767-1175, SCM SA, avril 2013.

4. Fiches de compétences associées

Qualité de l'Information
https://scmsa.eu/fiches/SCM_Qualite_Information.pdf

La définition d'un système d'information
https://scmsa.eu/fiches/SCM_Systeme_Information.pdf

5. Réalisations récentes

- Brigade des Sapeurs-Pompiers de Paris, 2010 : Etude statistique relative aux interventions
- Fédération des Etablissements Privés et d'Aide à la Personne (FEHAP), 2010 : Outil de simulation et d'investigation des modifications tarifaires

- Société Grande Paroisse, 2010 : Constitution et analyse de bases de données
- Novalis-Taitbout, 2010 : Analyse du système d'information
- Agence Nationale de l'Habitat, 2010 : Lois de probabilité relatives aux délais de paiement
- Nuclear Energy Agency (OCDE), 2010 : Détection de données aberrantes dans les bases de données
- PSA Peugeot Citroën, 2011 : Etudes statistiques
- Réseau Ferré de France, 2011 : Analyse des causes des retards des trains et optimisation des décisions d'investissement
- FEHAP, 2011 : Statistiques sur les Etablissements
- Nuclear Energy Agency (OCDE), 2011-2012 : Détection de données aberrantes dans les bases de données
- Espaces Ferroviaires, 2012 : Constitution d'une base de mots-clés à propos des opérations immobilières
- CITEPA, 2012 : Détection de données singulières dans un ensemble de données environnementales
- Air Liquide, 2012 : Bases de données de fiabilité
- GDF SUEZ, 2012 : Evaluation des incertitudes dans la comptabilité du gaz
- IRSN, 2012 : Analyse statistique de données de radioactivité dans l'environnement (tritium dans l'eau de pluie)
- Agence Nationale des Titres Sécurisés, 2013 : Retour d'expérience sur le passeport biométrique et analyse des fraudes
- IRSN, 2013 : Appui Méthodologique à l'Evaluation des Ecarts de Bilan de Matières Nucléaires
- DCNS, 2013 : Analyse préliminaire de "non-qualités" sur un site de production
- Espaces Ferroviaires, 2013 : Analyse des risques liés aux opérations immobilières
- 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
- Poste Immo, 2014 : Outils d'aide à la décision pour les économies d'énergie
- Nuclear Energy Agency, 2014 : Détection de donnés aberrantes dans les bases de données
- IRSN, 2014-2015 : Création d'un outil logiciel pour l'aide à la comptabilité de matières nucléaires
- Direction Générale Energie Climat (MEDD), 2014-2015 : Lien probabiliste entre trafic et émission de polluants
- Centre Technique des Institutions de Prévoyance, 2014-2015 : Appui technique et analyse critique de dossiers
- FEHAP, 2015 : Participation à la création d'un tableau de bord pour les dirigeants de la Fédération
- Nuclear Energy Agency, 2015 : Verification of the databases EXFOR and ENDF
- Carrefour, 2016 : Etudes statistiques
- Nuclear Energy Agency, 2016 : Méthodes mathématiques pour la vérification des bases de données
- L'Oréal, 2016 : Etude des données disponibles pour les accidents de la route entre le domicile et le lieu de travail
- COSEA, 2016 : Etudes statistiques relatives à la turbidité de l'eau

- SGAMI/Est, 2016 : Documentation relative à la gestion des situations de crise
- SNCF/Transilien, 2017 : Analyse critique de modèles de représentation des déplacements ; réalisation d'un outil de simulation
- Monceau Assurances, 2017-2018 : Amélioration de la politique commerciale
- Syndicat des Eaux d'Ile de France, 2017 : appui méthodologique
- Nuclear Energy Agency, 2017 : Méthodes mathématiques pour la vérification des bases de données
- COSEA, 2017 : Etude statistique relative à la turbidité de l'eau
- Réseau de Transport d'Electricité, 2017-2018 : Analyse de maintenances préventives
- SNCF Mobilités, 2018 : Estimation de flux de voyageurs au voisinage du bipôle Nanterre-La Défense
- Atlandes, 2018 : Comptage des véhicules sur les bretelles de sortie d'une autoroute
- Eramet, 2018-2019 : Amélioration d'un process industriel
- SARP Industries, 2019 : Hiérarchisation des paramètres intervenant dans un process industriel
- Coop de France Déshydratation, 2019 : Analyses statistiques liées à l'environnement
- 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
- Coop de France Déshydratation, 2020 : Constitution de bases de données pour les producteurs de luzerne
- Eiffage Rail, 2020-21 : Aide à la constitution d'un système d'information "reporting d'incidents"
- 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 ; analyse de la qualité des données recueillies
- Atlandes SA, 2022 : Analyse statistique relative aux trajets des poids lourds ; analyse de la qualité des données recueillies
- RATP, 2022-2023 : Analyse du coût des programmes
- SNCF, 2023 : Appui méthodologique aux plans d'inspection des rails
- Coop de France Luzerne, 2023 : Analyses statistiques
- 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