

**Projet FIRST Hautes-Ecoles
PHOEBUS**

Développement d'une application innovante de techniques avancées de data mining pour l'optimisation d'une installation de production d'énergie.

A ce jour, le stockage croissant des données industrielles couplé au développement de nouvelles techniques algorithmiques (réseaux de neurones, forêts aléatoires, réseaux bayésiens,...) ont permis l'essor d'une nouvelle branche des statistiques : le data mining. Parmi ses applications industrielles, l'analyse des données historiques apparaît primordiale en vue d'optimiser une installation de production d'énergie.

D'une durée de deux ans, ce projet a pour objectif de fournir aux industriels confrontés à une gestion énergétique complexe des outils performants d'aide à la décision. Le marché du papier étant particulièrement préoccupé par le coût de l'énergie et par les contraintes environnementales, le développement d'un prototype de ces solutions se fera en collaboration avec une papeterie installée en région wallonne.

Promoteur du projet

La Haute Ecole Robert Schuman via son Centre de Recherche (CRISIA)

Partenaires de la recherche

- **Partenaires industriels :** PEPITe S.A. (Spatiopôle - Liège)
Burgo-Ardenne S.A. (Virton)
- **Partenaire scientifique :** L.A.S.S.C. (Institut de Chimie – ULg)
Laboratoire d'Analyse et Synthèse des Systèmes Chimiques



Chercheur

Matthieu SAINLEZ

Ingénieur Civil en Mathématiques Appliquées
Ingénieur Industriel, orientation Mécanique

UCL (2008)
HERS (2005)

Objectifs du projet

En partant des données historiques collectées sur les sites de production, le but est de mettre en œuvre des outils innovants de modélisation et d'analyse prédictive.

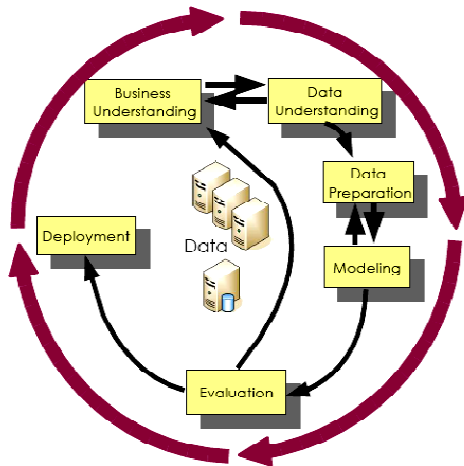
La volonté est de se limiter dans un premier temps à l'analyse détaillée de la chaudière de régénération des liqueurs dans le procédé Kraft. De cette manière, l'apport du data mining pourra être illustré via l'analyse des émissions atmosphériques.

Ensuite, il y aura lieu d'inclure les données périphériques à la chaudière de manière à permettre une modélisation du procédé global de régénération. Cette modélisation sera orientée vers une meilleure compréhension énergétique du procédé.

Les papeteries génèrent des coproduits qui sont déjà valorisés pour produire la vapeur nécessaire à la fabrication du papier. Cependant, il existe des marges potentielles inexploitées pour améliorer le volume de vapeur généré et ainsi produire davantage d'énergie électrique.

Le Data Mining

Le « data mining » (traduit littéralement par fouille de données) est l'application des techniques de statistiques, d'analyse des données et d'intelligence artificielle à l'exploration et l'analyse sans a priori de grandes bases de données informatiques, en vue d'en extraire des informations nouvelles et utiles pour le détenteur de ces données.



Une terminologie ciblée s'est développée autour de ce concept, on parle de « Knowledge Discovery in Databases » (KDD) qui caractérise le processus clé du data mining.

Le KDD se décompose en différentes étapes allant de la préparation et la transformation des données de départ à la mise en œuvre des outils informatiques appropriés.

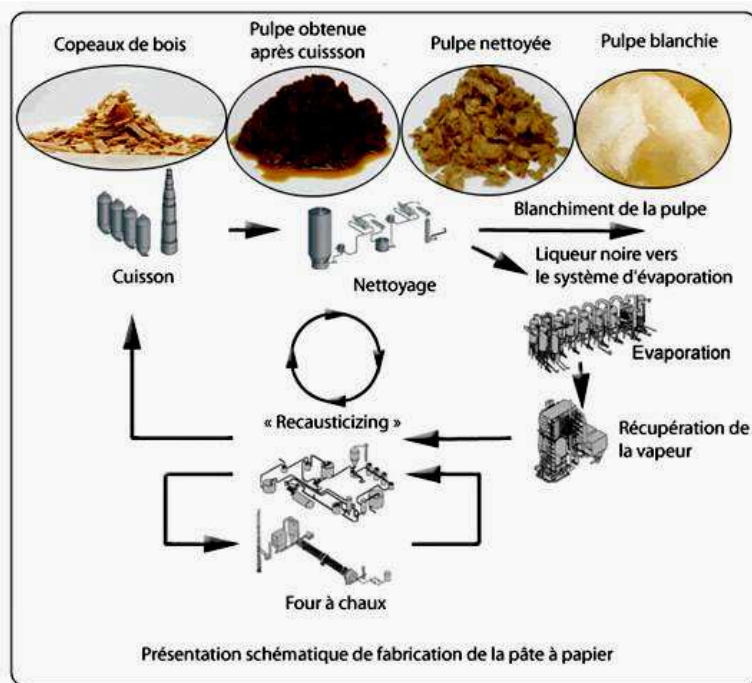
On distingue les techniques descriptives (ou exploratoires) qui visent à mettre en évidence des informations présentes mais cachées par le volume des données et les techniques prédictives (explicatives) qui extrapolent de nouvelles informations à partir des informations présentes.

Citons, entre autres méthodes, les arbres de décision et de régression, les réseaux de neurones, la régression multilinéaire, les règles d'associations, les réseaux bayésiens, les méthodes de clustering...

Le procédé papetier

Parmi les procédés de fabrication de pâtes à papier, les procédés chimiques sont basés sur la dissolution de la lignine et son extraction des parois des fibres du bois pour obtenir des fibres peu dégradées, longues et souples.

Le procédé au sulfate de sodium (ou procédé Kraft) est le plus utilisé dans le monde pour de nombreuses matières lignocellulosiques (bois, plantes annuelles). Dans ce procédé, les copeaux de bois sont imprégnés d'une solution aqueuse de NaOH et de Na₂S: la liqueur blanche. La liqueur extraite du lessiveur contenant les composés éliminés de la paroi est appelée liqueur noire. La cuisson est réalisée dans un réacteur vertical où les copeaux descendent par gravité et rencontrent les diverses liqueurs de cuisson.



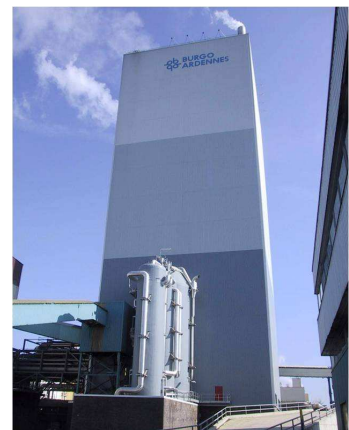
La « régénération » consiste à retransformer la liqueur noire issue de la cuisson des copeaux en liqueur blanche.

Les liqueurs noires sont concentrées par évaporation et calcinées sous atmosphère réductrice pour transformer le sulfate en sulfure. Le salin récupéré est dissous pour produire la liqueur verte.

La soude est régénérée par ajout de chaux dans cette liqueur et le CaCO₃ est calciné dans un four à chaux pour régénérer cette dernière.

La chaudière de régénération constitue généralement la source la plus importante d'émissions atmosphériques: les oxydes d'azote (NO_x), le dioxyde de soufre (SO₂), les composés de « soufre réduit total » (TRS), le monoxyde de carbone (CO) et les poussières.

Le corps de la chaudière est composé de deux parties principales: la partie basse, en atmosphère réductrice, régénère les minéraux de la liqueur noire alors que l'oxydation complète des différents composés organiques s'effectue dans la partie supérieure.



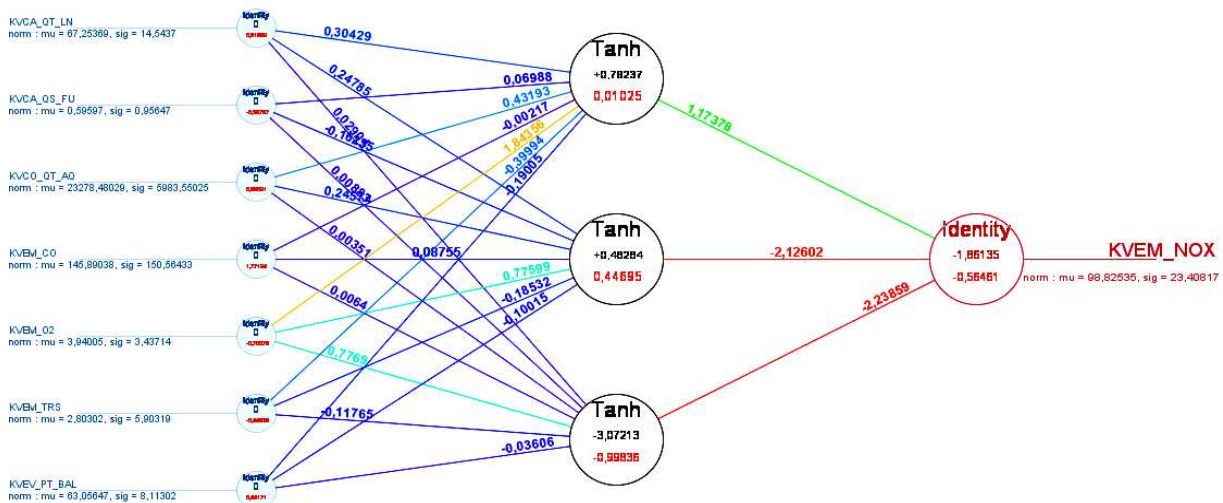
Analyse des NO_x

Une phase du projet consiste à analyser les émissions polluantes de la chaudière de régénération et plus particulièrement les oxydes d'azote (NO_x).

Il s'agit d'établir un modèle statistique permettant de prédire le niveau d'émission de NO_x en fonction de variables « explicatives », c'est-à-dire celles qui expliquent au mieux les variations de NO_x parmi les nombreuses variables récoltées (p.ex. les températures, pressions, les débits de fuel, de liqueur noire, d'air injecté,...). Une étape importante consistera à sélectionner un nombre adéquat de variables explicatives qui constitueront les entrées du modèle, tout en évitant les dangers liés au sur-apprentissage (ou sous-apprentissage) et à la colinéarité.

Parmi les techniques d'apprentissage supervisé, les réseaux de neurones sont principalement utilisés en régression non-linéaire : le modèle fournit des prédictions en utilisant un ensemble de variables d'entrées et une fonction non-linéaire de transition.

Voici un réseau de neurones calculé via le logiciel PEPITO[®] sur base de sept variables explicatives qui constituent les entrées du réseau auquel on demande de prédire les émissions d'oxydes d'azote (mesurées en mg/Nm^3).

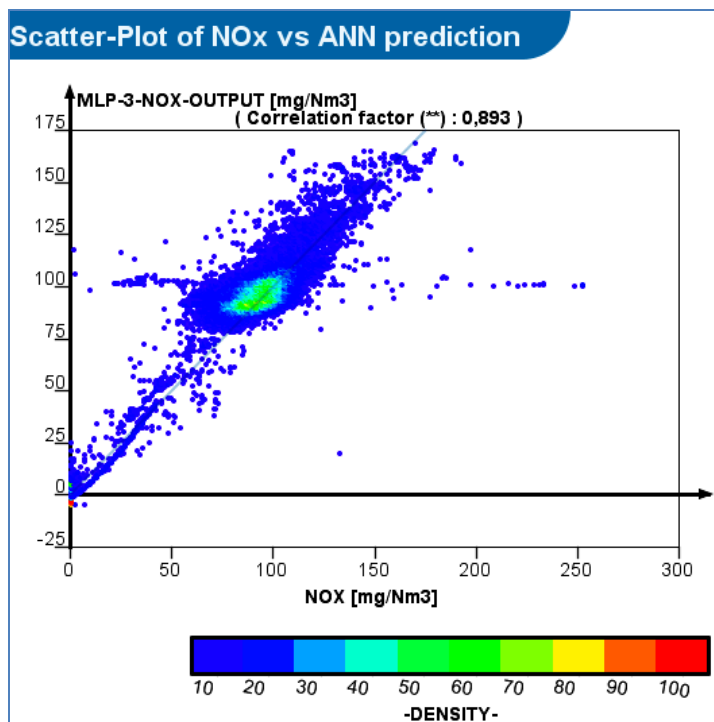


Succinctement, un réseau de neurones peut être vu comme un système complexe consistant en plusieurs unités organisées sous différents couches : la couche des entrées, la couche cachée et la sortie à prédire.

Une couche est un ensemble de neurones n'ayant pas de connexion entre eux. La couche d'entrée lit les signaux entrant (un neurone par variable explicative) et la sortie fournit la réponse du système. Entre les deux, une ou plusieurs couches cachées participent au transfert (la dimension de la couche cachée est un paramètre de la méthode). Un neurone d'une couche cachée est connecté en entrée à chacun des neurones de la couche précédente et en sortie à chaque neurone de la couche suivante.

Les principaux réseaux se distinguent par l'architecture de leur graphe, son niveau de complexité et par la fonction de transition utilisée. Dans cet exemple, on parle de perceptrons multicouches à une couche cachée de trois neurones.

Pour sa part, la fonction de transition opère une transformation d'une combinaison affine des signaux d'entrée, cette combinaison étant déterminée par un vecteur de poids associé à chaque neurone et dont les valeurs sont estimées dans la phase d'apprentissage. Les fonctions d'activation utilisées sont souvent non linéaires pour la couche cachée (la tangente hyperbolique dans cet exemple).



L'apprentissage consiste à calculer les pondérations optimales des différentes liaisons par rétropropagation : on corrige les poids des liaisons en fonction de l'erreur de prédiction obtenue en sortie du réseau. L'algorithme d'optimisation utilisé est celui de la descente de gradient.

Une fois l'apprentissage effectué, il convient d'injecter dans le réseau l'échantillon conservé pour les tests et d'en analyser la réponse prédite.

Dans le cas présent, il y a une corrélation de 0.89 entre les valeurs observées et les valeurs prédites. Un histogramme des résidus (super-) gaussien ainsi qu'une convergence de résultats entre les deux échantillons nous conforte sur l'adéquation du modèle.

La modélisation par réseaux de neurones est très performante mais son interprétation mais n'est pas toujours évidente. Néanmoins, la méthode apporte un point de vue complémentaire et une analyse plus approfondie peut être réalisée via des arbres de décision, des réseaux bayésiens, un partitionnement de données (clustering),...Ce sont là toutes des méthodes qui vont permettre de valoriser les bases de données industrielles par la recherche d'informations pertinentes.

Références

- *Data Mining et statistique décisionnelle* par Stéphane Tufféry – Editions Technip 2007
- *Fabrication des pâtes (J6900)* par Michel PETIT-CONIL – Techniques de l'Ingénieur (2008)
- *Schéma de la fabrication de la pâte à papier* – url : <http://essperans.fr/blog/?p=9>
- *Data Mining et Statistique* par Philippe Besse et al – Journal de la Société Française de Statistique, 142, 5-35 (2001).
- *Apprentissage Statistique & Data Mining* par Philippe Besse – Publications du laboratoire de statistique et probabilités, Université Paul Sabatier (2008).