Université de Poitiers – U.F.R. Sciences
Master Sciences pour l'Ingénieur
Spécialité : Génie des Systèmes Industriels

Année Universitaire 2011 - 2012

RAPPORT de STAGE

En vue de l'obtention du grade de Master

Fonderie du Poitou Fonte

Analyse statistique des paramètres procédé pour l'étude du défaut de soufflure

Stage effectué entre le 05/03/2012 et le 30/09/2012

Réalisé par
PERABÒ Stefano

Responsable de Stage
KOSTER Jean-Yves
Fonderie du Poitou Fonte
Z.I. de Saint Ustre – B.P. 042
86220 Ingrandes-sur-Vienne

Table de matière

1 Présentation de l'entreprise

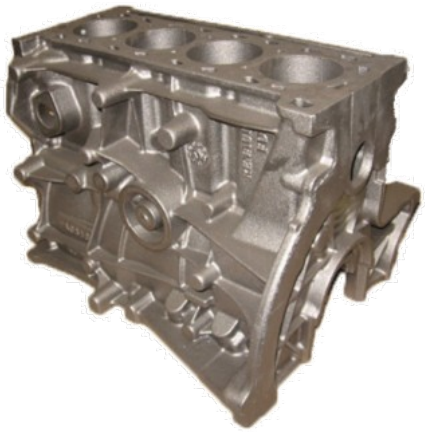
La Fonderie de Poitou Fonte (FPF) est spécialisée dans la production de blocs cylindres en fonte grise pour l'industrie automobile depuis 1981. L'usine est située à Ingrandes sur Vienne, dans le département de la Vienne, à 300 Km au sud de Paris. En 1999, FPF est devenue une filiale du groupe italien Teksid lequel regroupe 8 fonderies à travers le monde (France, Italie, Pologne, Portugal, Mexique, Brésil, Chine).

Ses produits, destinés aux moteurs essence et diesel, couvrent une gamme de cylindrées allant de 1200 cm3 à 2200 cm3, qui correspondent à des poids de carter compris respectivement entre 25 Kg et 65 Kg. Ils sont vendus sur le marché français mais aussi exportés en Espagne, Pologne, Italie et Inde, et équipent les véhicules Renault, Fiat, Lancia, Tata, Suzuki, Dacia et Nissan. Les principaux concurrents sur ce segment de marché sont les fondeurs allemands Brhül et Fritz Winter, et l'espagnol Fagor.

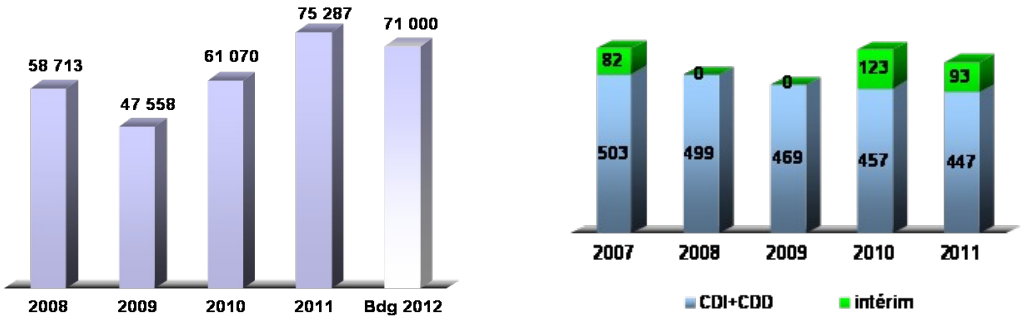
FPF a entrepris une démarche qualité totale, validée par la certification ISO/TS 16949 version 2002. Elle développe une politique environnementale qui lui a permis l'obtention de la certification ISO 14001, et est actuellement engagée dans la mise en place d'un système de management de la santé et de la sécurité au travail qui devrait aboutir à la certification OHSAS 18001 en 2013. En outre, un programme d'innovation à été introduit, développé par Fiat Auto Production System et nommé WCM (World Class Manufacturing), visant la réalisation de changements durables et systématiques de la manière de produire. Ce programme consiste en un ensemble de principes et méthodes visant l'amélioration de la productivité, de la qualité, de l'efficience et du niveau de service.

L'entreprise utilise un procédé de fusion électrique (3 fours à induction d'une capacité de 42 t chacun et d'une puissance totale de 12 MW) et sa capacité de fusion maximale s'élève à 130 Kt par an. Elle est pionnière dans l'utilisation du procédé Disamatic pour la coulée des carters cylindres en position verticale (une seule autre usine au monde, au Japon, utilise ce même procédé) et qui permet de réaliser une production en grande série (80 Kt tonnes par an). Elle est en outre la seule en France à être équipée d'un système de recyclage du di-methyl-ethyl-amine, un gaz catalyseur largement utilisé en production, très volatile et dangereux pour la santé.

En 2011 FPF employait 540 personnes, dont 93 intérimaires (81% d'ouvriers, 16% d'ATAM et 3% de cadres). 75000 tonnes de blocs cylindres ont été livrés, correspondant à 1.8 millions de pièces, pour une chiffre d'affaire de 105 M€.



− Vue aérienne de la Fonderie de Poitou Fonte (gauche) et exemple de bloc cylindre (droite) −



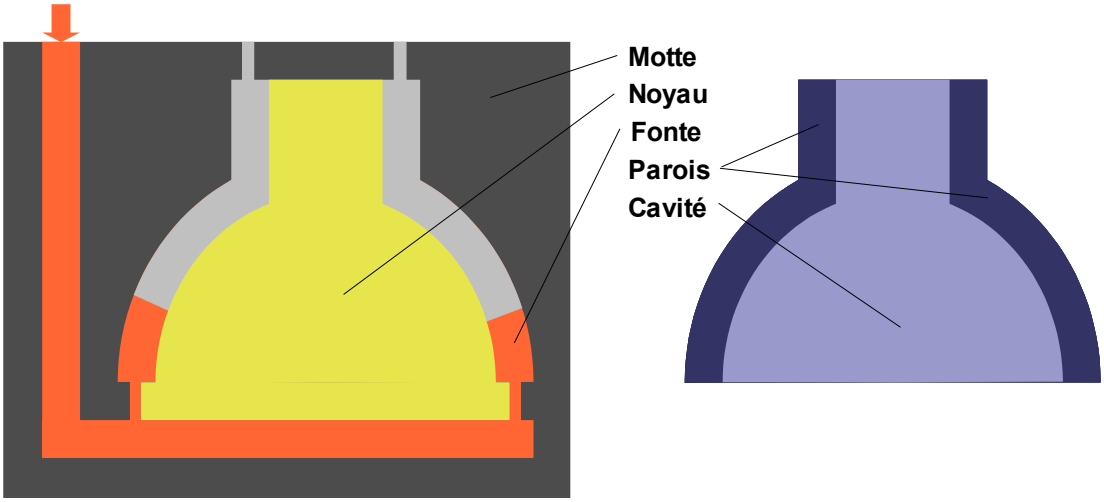
− Évolution de la production annuelle de FPF en tonnes (gauche) et du nombre d'effectifs (droite) −

2 Description du sujet de stage

2.1 Le contexte

Ce paragraphe décrit rapidement le processus de production des carters cylindres à FPF. Il sert d'introduction et vise à clarifier au lecteur les raisons qui ont amené la Direction de FPF à proposer ce sujet de stage.

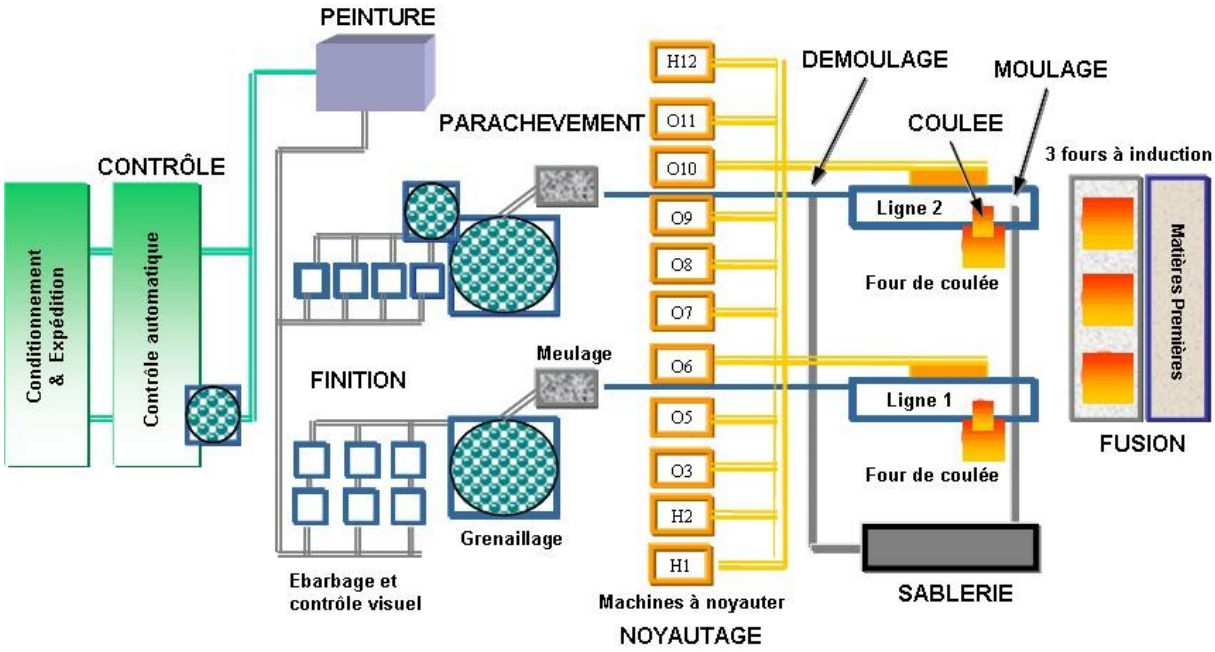
Le procédé de production des carters consiste à couler de la fonte liquide dans un moule afin d'obtenir, après refroidissement, une pièce ayant une forme intérieure (les cavités) et extérieure donnée. La technique employée à FPF est celle du moulage en sable, schématisée dans la figure suivante.



− Schématisation de la technique du moulage en sable : coulée du moule (gauche) et pièce finie (droite) −

La forme extérieure de la pièce est donnée par une motte à usage unique constituée d'un mélange de sable réfractaire (silice), d'argile, d'eau et d'additifs carbonés. Ce mélange, appelé sable à vert, est durci à froid par une compression mécanique. En revanche, la forme intérieure des pièces est obtenue par l'insertion d'éléments solides dans la motte. Ces éléments, appelés noyaux, sont constitués d'un mélange de sable et des résines qui est durci à froid par un procédé chimique utilisant un gaz catalyseur. Le moule, c'est-à-dire l'ensemble constitué de la motte et des noyaux, est réalisé selon le procédé Disamatic qui sera décrit dans la suite du rapport.

À FPF deux lignes de production existent. D'un point de vue fonctionnel, le procédé de fabrication peut être représenté par le schéma à la page suivante.



− Schéma représentant les différents secteurs de l'usine −

Les secteurs de l'usine qui contribuent à la transformation des matières premières en produits finis prêts au conditionnement et à l'expédition sont :

1. Fusion : les matières premières sont introduites dans les fours de fusion afin de préparer la fonte liquide qui est ensuite transférée au moulage.
2. Sablerie : le sable à vert est préparé par un procédé de « régénération » du sable des mottes et des noyaux récupérés après le démoulage, et il est transporté au moulage.
3. Noyautage : les différents éléments des noyaux sont produits, assemblés et convoyés au moulage.
4. Moulage: la motte est préparée à partir du sable à vert et les noyaux assemblés sont introduits dans la motte afin de former le moule.
5. Coulée : le moule est rempli de fonte liquide.
6. Démoulage : la fonte solidifiée est extraite du moule et la pièce est grenaillée afin d'enlever une grosse partie du sable à vert et des noyaux qui restent collé aux parois de la pièce.
7. Parachèvement : après un certain temps de refroidissement, la pièce est séparée des éléments formés à la coulée, meulée sur 4 des 6 surfaces et nettoyée définitivement par un dernier grenaillage.
8. Finition : les pièces sont ébarbées et contrôlées visuellement pour détecter des éventuels défauts.
9. Peinture : les parois externes des pièces sont peintes.
10. Contrôle : d'autres contrôles de qualité sont effectués par des systèmes automatisés.

Lorsque un défaut est détecté, la pièce est déclarée « rebut » et elle est stockée en attente d'être recyclée. En effet, entre 35% et 42% de la « matière première » utilisée aux fours de fusion est constituée des rebuts. Les carters qui passent tous les contrôles qualité prévus sont toutefois encore définis « bruts » puisque ils subiront d'autres opérations d'usinage chez les clients de FPF avant d'être assemblées avec les autres composants

des moteurs. Pendant ces opérations d'usinage d'autres défauts peuvent être découverts et les pièces déclarées rebuts « externes » (par opposition aux autres rebuts « internes » déclarés chez FPF). Il s'agit notamment des défauts qui ne peuvent pas être toujours détectés par une simple inspection de la surface du carter. De ce point de vue, le défaut de soufflure, qui fait l'objet du stage, est peut-être le plus difficile à détecter puisque il se présente comme une cavité ou porosité créée par l'occlusion de gaz lors de la solidification de la fonte. En 2012, les taux de rebut interne et externes atteignaient respectivement le 6% et 0.5% en moyenne, toute référence produit confondue. Des efforts sont constamment faits afin de réduire ces valeurs. Le coût de la non-qualité a en effet un impact non négligeable sur les résultats financiers de l'entreprise, sur son image auprès des clients et sur sa capacité à conquérir des nouveaux marchés.

Un des moyens utilisés à FPF pour la réduction du taux de rebut est l'analyse des paramètres procédé. Au fil des années une importante base de données a été alimentée en permanence avec des enregistrements réalisés tout au long du processus de production. Cette base, constituée de dizaines de milliers d'enregistrements, est utilisée en particulier par les services Procédé et Qualité pour la supervision du processus, le repérage des anomalies et la compréhension des mécanismes à l'origine des défauts constatés. Les responsables de ces services estiment toutefois de ne pas être capable de l'exploiter comme ils souhaiteraient. En particulier, à cause de l'énorme volume de données, ils rencontrent des difficultés lorsqu'ils essaient de prendre en considération simultanément plusieurs paramètres procédé (de l'ordre d'une dizaine, par exemple) afin de mettre en évidence et corriger (en temps réel si possible) des éventuelles « corrélations » entre la valeur de ces paramètres et l'apparition des défauts sur les produits finis. Ce genre d'opérations est effectué pour la plupart du temps « manuellement » et/ou « visuellement », en se basant sur des rapports sous forme de tableaux et/ou de graphiques, et en s'appuyant sur l'expertise métier des techniciens. Un système de Maîtrise Statistique des Procédés est couramment utilisé basé sur le logiciel commercial SESAME. Ceci permet en principe un contrôle du procédé en cours de production dans le but d'obtenir une production stable avec un minimum de produits non conformes aux spécifications. Toutefois il n'est pas déployé sur la totalité des secteurs ni intégré avec les autres logiciels, ce qui limite son utilisation comme outil de supervision et/ou monitorage du procédé. En outre, les outils d'analyse statistique disponibles dans SESAME sont très limités par rapport à d'autres logiciels applicatifs existants.

2.2 L'objet du stage et les missions confiées

Dans ce contexte, l'objet de mon stage a été l'analyse des paramètres procédé pour comprendre les défauts récurrents et identifier les améliorations à apporter. Les missions définies en début de stage étaient les suivantes : centraliser les données, faire un repérage des anomalies, étudier l'adéquation des outils utilisés à FPF et les résultats attendus, présenter une modélisation aidant à la compréhension des défauts, proposer des améliorations sur l'enregistrement actuel des paramètres.

Ce « cahier des charges » a été mieux précisé au fil du temps. Après une première semaine de stage dédiée à la découverte de l'entreprise (visite de tous les secteurs pour comprendre le fonctionnement du procédé), pendant la deuxième semaine j'ai pu travailler en autonomie afin de réviser et approfondir l'ensemble des notions apprises au cours de la phase de découverte, et de dresser un état de l'art. Cette activité m'a permit de mener une réflexion personnelle visant à mieux délimiter le champ d'intervention et mieux préciser la nature de ma mission à FPF vis-à-vis de mes compétences, du temps à disposition et des attentes de FPF. Ensuite, lors d'une réunion avec M. Franck Gaté, Directeur, M. Ludovic

Courcelles, Responsable Procédé, ainsi que mon tuteur, Jean-Yves Koster, Responsable Industrialisation, mes missions ont été clairement spécifiées. Les points suivants résument les problématiques à résoudre :

1. L'analyse s'est focalisée sur un seul type de défaut (appelé « soufflure ») et sur une seule famille de références produit fini (les carters cylindres « M9 »), cette famille étant celle qui présente le plus grand taux de défauts de ce type.
2. Le premier objectif était d'identifier quels paramètres procédé sont à l'origine de l'apparition d'un défaut de soufflure ou, du moins, peuvent être considéré comme des facteurs influents.
3. Le deuxième objectif était de déterminer quelles corrections apporter éventuellement aux valeurs nominales de ces paramètres (et aux fourchettes de conduite) afin de minimiser le taux d'apparition du défaut de soufflure.
4. J'étais responsable du choix et de la mise en place des méthodes d'analyse à appliquer. J'ai proposé de démarrer l'étude en utilisant la technique des Réseaux Bayésiens, sans pour autant exclure la possibilité de tester d'autres méthodes, en fonction de résultats obtenus « en cours d'œuvre ».

3 Méthodes et moyens mis en œuvre pour résoudre le problème

3.1 Organisation du projet

Les attentes de FPF vis-à-vis de mon stage concernaient donc la conception d'outils d'analyse de grandes masses de données et d'aide à la décision, permettant d'atteindre les objectifs décrits au paragraphe précédent. J'ai donc dirigé mon attention vers les techniques dites d'« exploration automatique des données » (« data mining » en anglais) en essayant d'abord de définir une bonne démarche de conduite du projet. L'approche que j'ai suivi correspond à celui proposé par la méthode CRISP-DM [1].

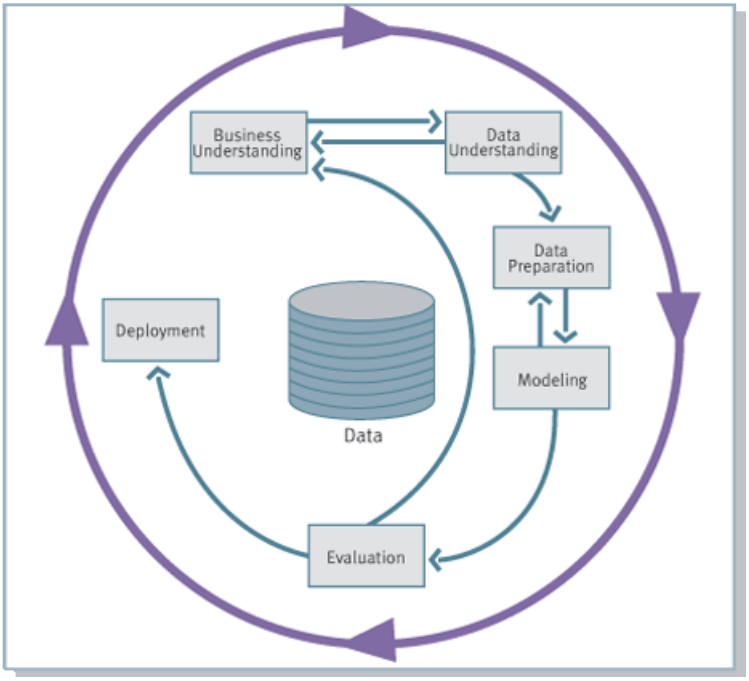
Le CRSIP-DM (Cross Industry Standard Process for Data Mining) est une méthode créée et mise au point jusqu'en 2006 par un consortium formé de compagnies privé (SPSS, Teradata, Daimler AG et OHRA) dans le cadre du projet ESPRIT (European Strategic Program on Research in Information Technology) financé par la Communauté Européenne. Il s'agit d'une méthode « neutre » par rapport aux métiers et aux outils, communément utilisée par les experts en analyse de données, qui décompose le cycle de vie d'un projet d'exploration de données en six phases principales :

1. Compréhension du Métier (Business Understanding)
2. Compréhension des Données (Data Understanding)
3. Préparation des Données (Data Preparation)
4. Modélisation (Modeling)
5. Évaluation (Evaluation)
6. Déploiement (Deployment)

Cette approche préconise de revenir cycliquement sur ces phases en suivant les trajectoires représentées dans la figure à la page suivante. L'analyste et son « client » doivent interagir afin d'atteindre une vision partagée : des objectifs à atteindre, des problématiques et des ressources disponibles (phase 1), de la nature et de la qualité des données à analyser (phase 2), de l'interprétation à donner aux résultats de l'analyse et des « corrections » à apporter (phase 5). Afin de se mettre constamment à jour sur l'avancement de la mission, effectuer des « brain stormings », discuter des résultats

partiels obtenus, etc., j'ai donc proposé aux experts métiers de FPF d'effectuer des réunions hebdomadaires de la durées d'une heure.
L'analyste est en revanche chargé de la sélection, de l'intégration et du prétraitement de données (phase 3), et du choix et de l'application de la méthode d'analyse (phase 4). J'étais totalement responsable de ces activités pendant le stage.
Enfin le déploiement (phase 6) consiste à appliquer « sur le terrain » les recommandations issues de l'analyse afin d'atteindre les objectifs préfixés. Cette dernière phase est essentiellement à la charge du « client ». L'analyste joue un rôle de superviseur de la mise à jour de la base de données et de la bonne utilisation des outils d'analyse mis en place. C'est l'adéquation des résultats obtenus dans la phase de déploiement avec les objectifs fixés qui, éventuellement, justifie le démarrage d'un nouveau cycle, c'est-à-dire la définition d'un nouveau projet d'analyse des données ou la redéfinition du précédent.

La dernière phase pourra démarrer seulement après la fin du stage. Dans les paragraphes suivants, je présenterai donc seulement les activités et les résultats obtenus dans les 5 premières phases prévues par la méthode CRISP-DM.



− Le cycle de vie d'un projet d'analyse de données selon la méthode CRISP-DM −

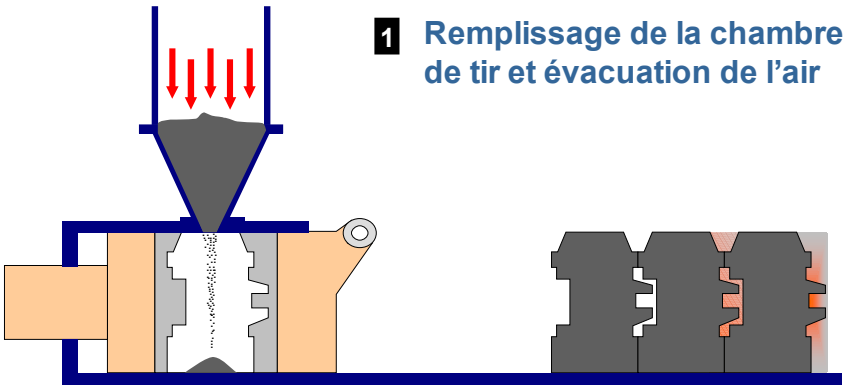
3.2 Compréhension du métier

3.2.1 Le moulage

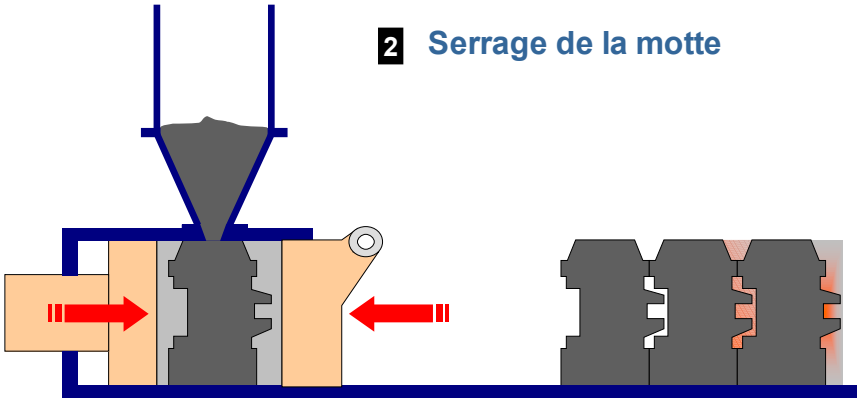
Le moulage est le secteur de la fonderie où les mottes sont préparées à partir du sable à vert et les noyaux assemblés sont introduits dans les mottes afin de former les moules. La coulée est le secteur où les moules sont transférés pour être remplis de fonte liquide.

À FPF le moulage s'effectue selon le procédé Disamatic. Ce procédé a été breveté en 1957 par Vagn Aage Jeppesen, professeur à l'Université Technique du Danemark. Depuis 1960, la société danoise Dansk Industri Syndikat (DISA) produit et commercialise des machines à mouler basées sur ce procédé. Plus de 1430 fonderies dans le monde ont installé ces machines pour la production d'un très large éventail de types et de dimensions

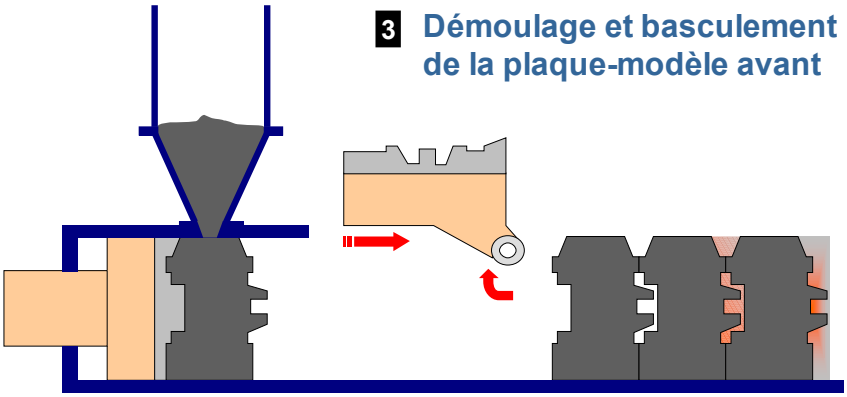
de pièces coulées. Le moulage Disamatic consiste en la fabrication en continu des moules présentant un plan de joint vertical. Les phases d'un cycle de fabrication sans noyaux sont représentée dans les figures suivantes.



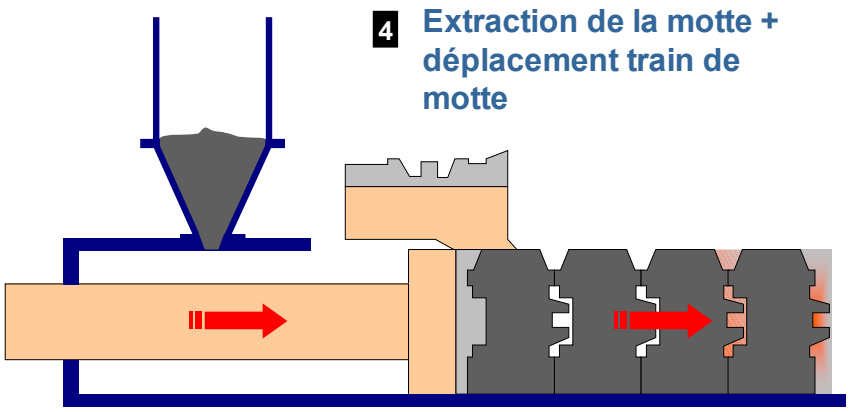
Le sable à vert est stocké en haut de la machine dans une trémie. Un tiroir entre la trémie et une chambre de tir sous-jacente s'ouvre pour laisser tomber le sable. La chambre de tir est remplie uniformément de sable à vert par un jet d'air comprimé.



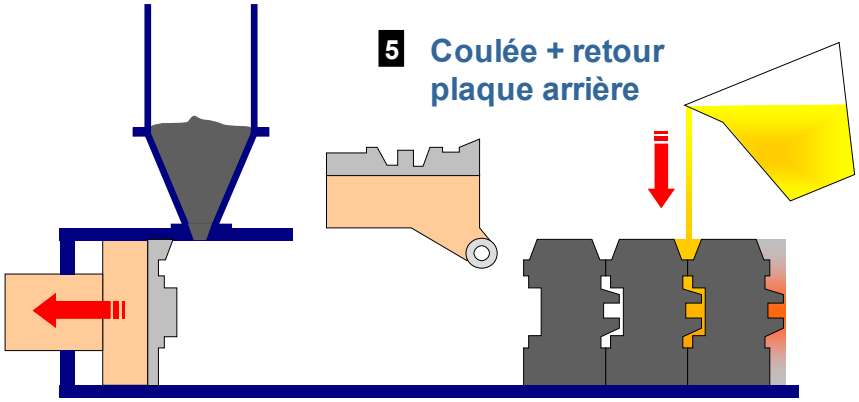
Le tiroir se referme et un piston, sur lequel est fixée la plaque-modèle arrière, comprime le sable contre la plaque-modèle avant, jusqu'à ce que la pression de serrage souhaitée soit atteinte. Chaque côté de la motte ainsi formée représente une face de la pièce à obtenir.



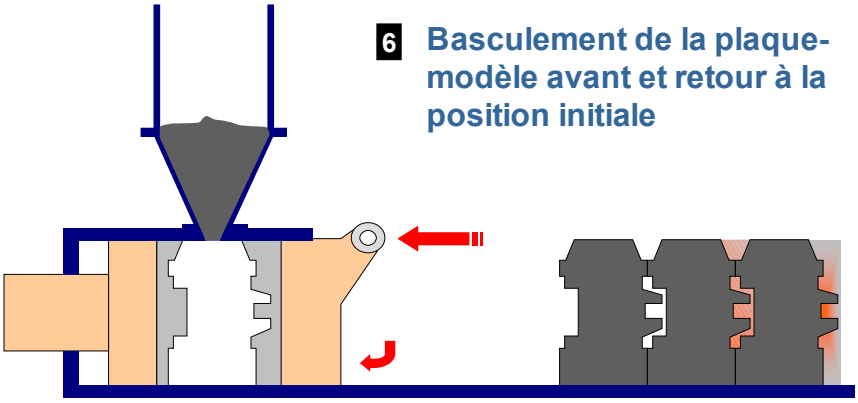
Le volet sur lequel est fixée la plaque-modèle avant se détache de la motte (opération appelée démoulage) et bascule, permettant au piston de reprendre sa course.



La motte est extraite de la chambre de tir poussée par le piston. La motte vient en contact avec la précédente, formant ainsi un moule serré. La poussée déplace d'un pas le train de moules déjà formés et qui se trouve sur un convoyeur à hauteur de la base de la chambre.



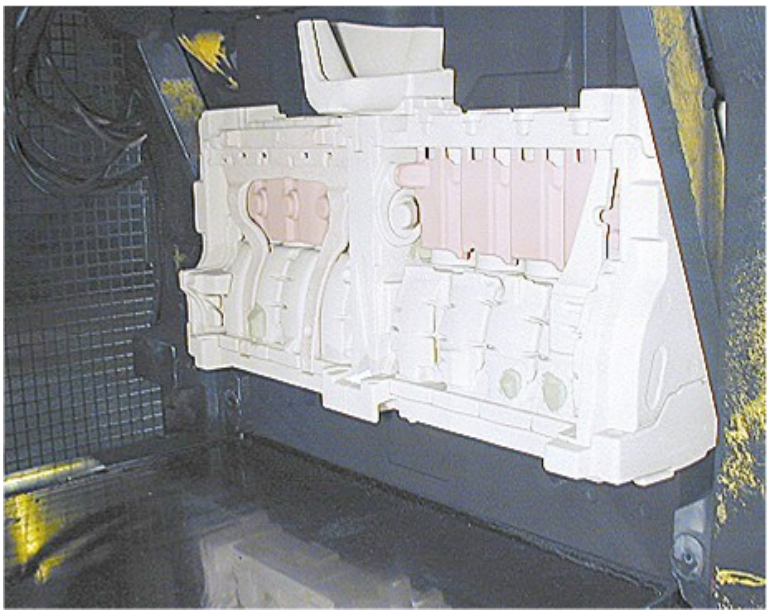
Le piston reprend sa place en arrière de la chambre. La coulée d'une des moules formé précédemment peut démarrer.



Le volet se referme et un autre cycle peut démarrer.

Le temps disponible pour la coulée d'un moule correspond au laps de temps entre le début de la phase 5 et la fin de la phase 3, pendant lequel le train de moules est à l'arrêt. Les pièces coulées refroidissent dans les moules pendant une période qui dépend de la cadence de production et de la longueur du convoyeur. En bout du convoyeur, à partir de la fin de la phase 4, une machine extractrice effectue le démoulage d'une pièce par cycle.

Lorsqu'un noyau doit être inséré entre deux mottes, le cycle comporte une phase supplémentaire, placée entre la fin de la phase 3 et le début de la phase 4. Un mécanisme dédié transporte le noyau en dessous du volet et le positionne partiellement dans la dernière motte ajoutée au train sur le convoyeur, comme montré dans la figure suivante. Pendant le serrage du moule, phase 4, la nouvelle motte recouvre entièrement le noyau, poussée par le piston.



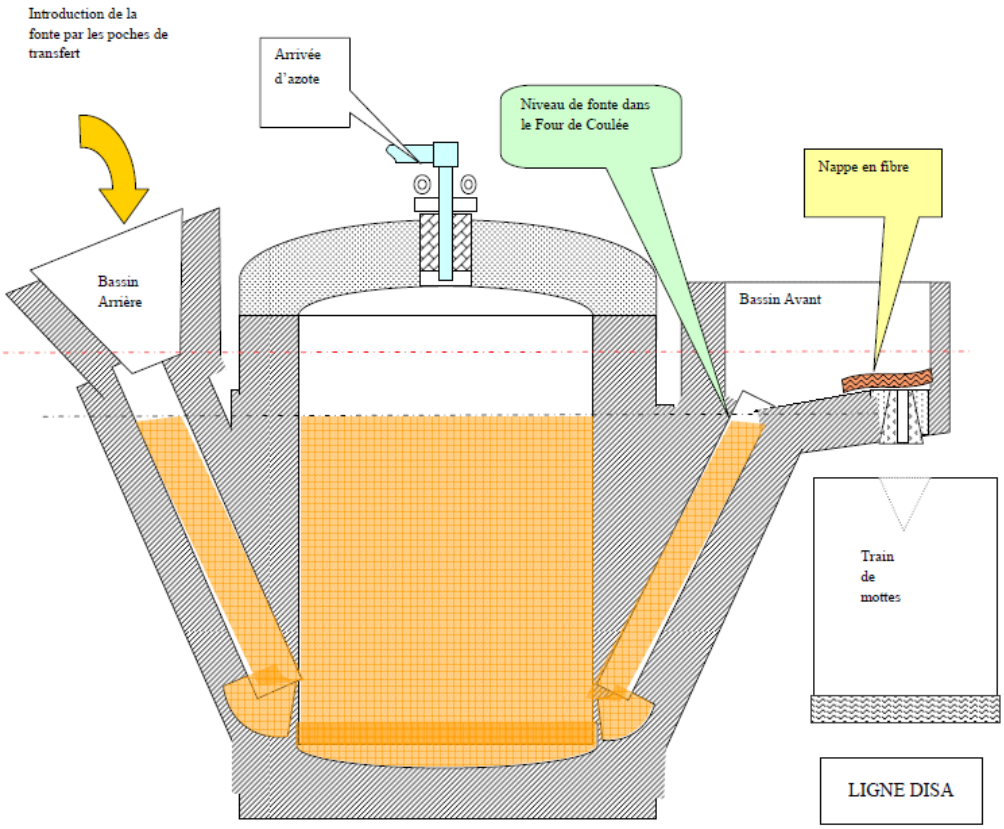
− Un noyau positionné dans une motte, en attente d'être renfermé dans son moule −

3.2.2 La coulée

La coulée est le secteur de la fonderie où les moules sont remplis de fonte liquide. La fonte coulée dans les moules ne provient pas directement des fours de fusion de 42 t. En revanche, des poches de 3.5 t sont prélevées périodiquement et transférées par des chariots vers les fours de coulée qui ont une capacité de 7 t. Ces derniers se trouvent très proches à la sortie des deux machines de moulage Disamatic, à coté des trains de moules. La figure à la page suivante représente en section un de ces fours.

L'insertion de la fonte se fait par le bassin arrière. Un orifice de remplissage en forme de siphon plonge au fond d'une cuve cylindrique, où la fonte est maintenue à la température de coulée souhaitée (entre 1350 °C et 1400 °C) par un système à induction électrique (dont la puissance est 200 KW). Un autre siphon relie le fond de la cuve au bassin avant. La partie supérieure du four est fermée hermétiquement par une voûte comportant une ouverture centrale.

Pendant la coulée, de l'azote est injectée par l'ouverture de la voûte. La pression exercée par l'azote fait diminuer le niveau de la fonte à l'intérieur de la cuve. Simultanément la fonte remonte par les siphons en remplissant le bassin avant. Le niveau de la fonte dans ce bassin est mesuré par une caméra et un détecteur laser, et il est réglé en boucle fermée. Le bassin, qui se trouve juste en dessus du train de moules, est pourvu d'une busette sur lequel va s'appliquer le tampon d'une quenouille. Le mouvement vertical de la quenouille, synchronisé avec le moulage Disamatic, provoque l'ouverture et la fermeture de la busette. La fonte est donc coulée par gravité dans le moule sans pression extérieure.



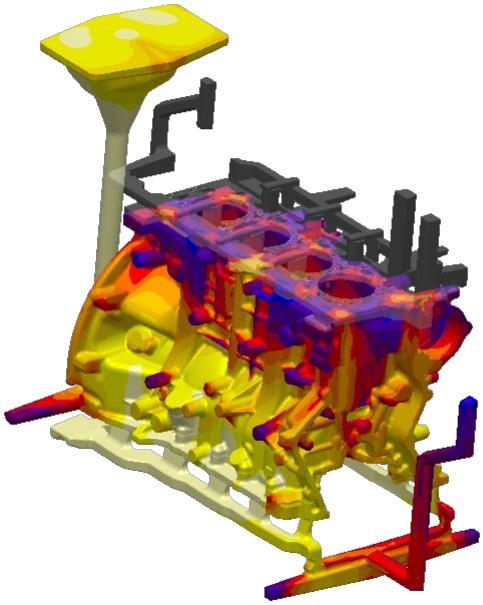
− Section d'un four de coulée −

Les deux figures suivantes montrent le four de coulée en activité et le bassin avant avec la quenouille.

− Four de coulée (gauche) et sa quenouille (droite) −

Afin de garantir un remplissage optimal du moule, des lois de coulée prédéfinies et mises au point par le chef du projet produit doivent être respectées. La durée du remplissage et les variations du flux de fonte pendant la coulée sont réglées automatiquement et surveillées par une caméra. De plus, le moule possède un ensemble d'évents et chenaux permettant aux gaz dégagés de s'échapper et d'obtenir une vitesse de remplissage régulière de la pièce. Des logiciels de simulation existent pour l'étude du remplissage ainsi que de la solidification. FPF utilise le logiciel MAGMA pour étudier des paramètres tels que

les flux et les vitesses de remplissage, la température de la fonte avant et après solidification, et prévoir les défauts potentiels (un exemple de simulation est montré dans la figure suivante). Toutefois aucun logiciel n'existe sur le marché permettant de prévoir l'apparition d'un défaut de soufflure.



− Exemple de simulation réalisée avec MAGMA (température de la fonte pendant le remplissage) −

Trois « ingrédients » sont donc utilisés pour la production d'une pièce : la fonte liquide, le sable à vert et les noyaux. Leur production est décrite dans les paragraphes suivants.

3.2.3 La fusion

La fusion est le secteur de la fonderie où les matières premières sont introduites dans les fours de fusion afin de préparer la fonte. La figure suivante montre le magasin de stockage des matières premières.

− Magasin de stockage des matières premières −

Le lit de fusion, c'est-à-dire les matières premières utilisées, est constitué en moyenne de

- 58% d'aciers (chutes d'emboutissage de l'industrie automobile),

- 38% de retours (rebuts et « mise au mille »[1]),
- 4% d'additions (C, CSi, FeSi, FeMn, FeCr, FeS, Sn).

Les additions sont ajoutées en quantité variable afin d'atteindre avec précision la composition chimique recherchée. Après la phase de démarrage d'un four et lorsque le premier lit de fusion à atteint la température nominale, les opérations se succèdent de la manière suivante :

1. Une poche de 3.5 t de fonte liquide, correspondant à 8% du poids du lit de fusion, est prélevée et transportée aux fours de coulée.
2. Une charge métallique de 3.5 t de matières premières est introduite dans des bennes et convoyée au four par un système entièrement automatisé. Cette charge va remplacer le prélèvement effectuée précédemment.
3. Les additions sont introduites dans le four par les fondeurs.

Un four n'est donc jamais vidé complètement, ce qui permet une stabilité de la composition chimique et de la température, et une vitesse de fusion optimisée. Les trois figures suivantes montrent des détails lors de l'exécution des trois opérations décrites.

− Prélèvement d'une poche (gauche), remplissage d'un four de fusion (centre) et introduction des additions (droite) −

3.2.4 Le noyautage

Le noyautage est le secteur de la fonderie où les différents éléments constituant les noyaux sont produits et assemblés avant d'être convoyés au moulage. Les phases de production des noyaux se succèdent comme suit :

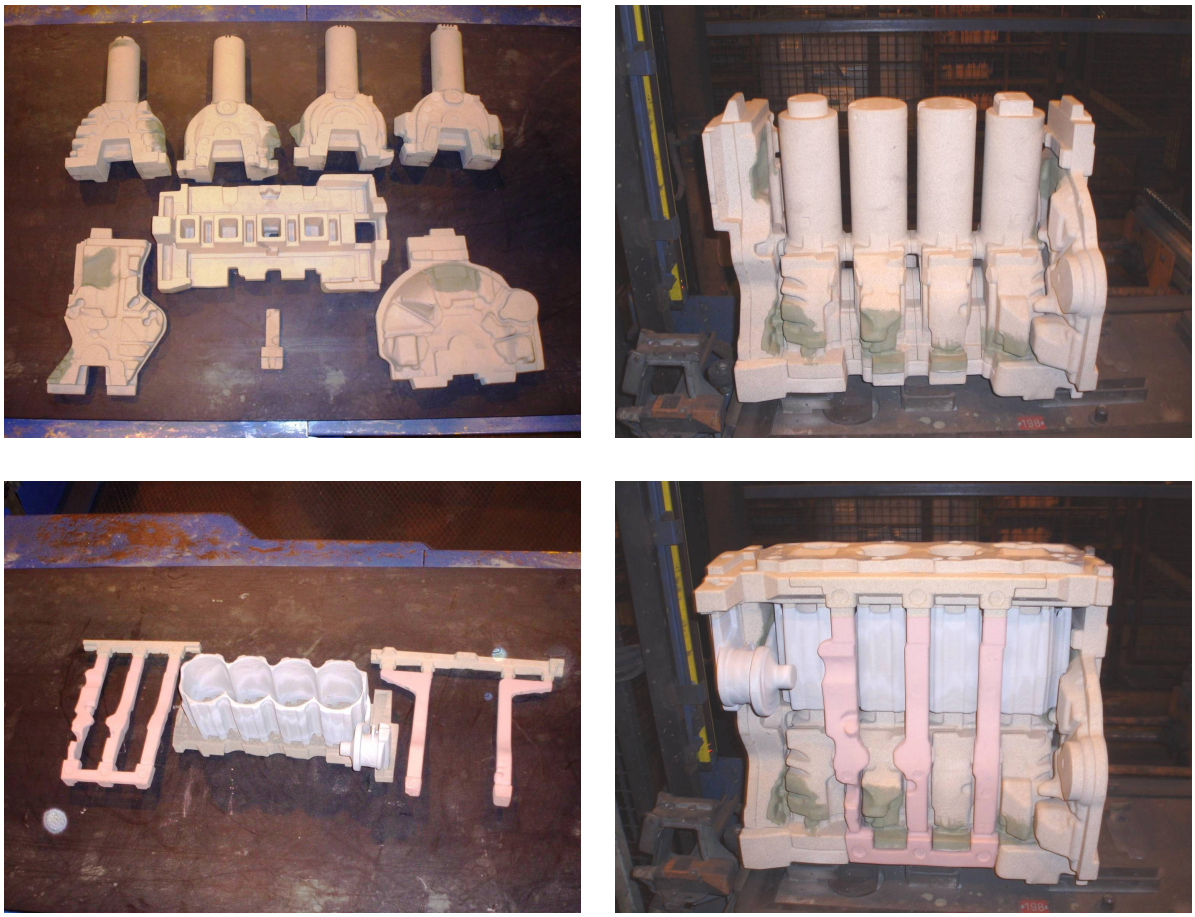
1. Du sable réfractaire est malaxé avec deux types de résines et d'autres additifs en proportions prédéfinies, afin d'obtenir des noyaux ayant les caractéristiques mécaniques exigées. Le sable préparé est stocké dans des trémies qui surmontent les différentes machines à noyauter.
2. Dans chaque machine, le sable préparé est tiré à l'aide d'air comprimé dans des boîtes à noyaux (moules reproduisant la forme des éléments).
3. Le durcissement du noyau est obtenu par le passage d'un gaz catalyseur à travers les porosités du sable. Une phase de rinçage par soufflage d'air permet d'évacuer le surplus de gaz.
4. Des opérations éventuelles de parachèvement (ébavurage, perçage, vissage) sont effectuées. Certains éléments subissent aussi un poteyage, qui consiste à les

[1] La mise au mille est le poids de la fonte nécessaire pour obtenir une pièce divisée par le poids d'une pièce. Cette quantité représente donc le poids d'un carter plus le poids de la fonte solidifiée dans le système de coulée.

enduire d'un produit réfractaire. Le poteyage protège les noyaux des chocs thermiques et donne un meilleur état de surface à la pièce.

5. Après séchage des enduits, les éléments sont contrôlés visuellement et les noyaux assemblés définitivement.

Les figures suivantes montrent comment un ensemble d'éléments sont assemblés pour obtenir un noyau.



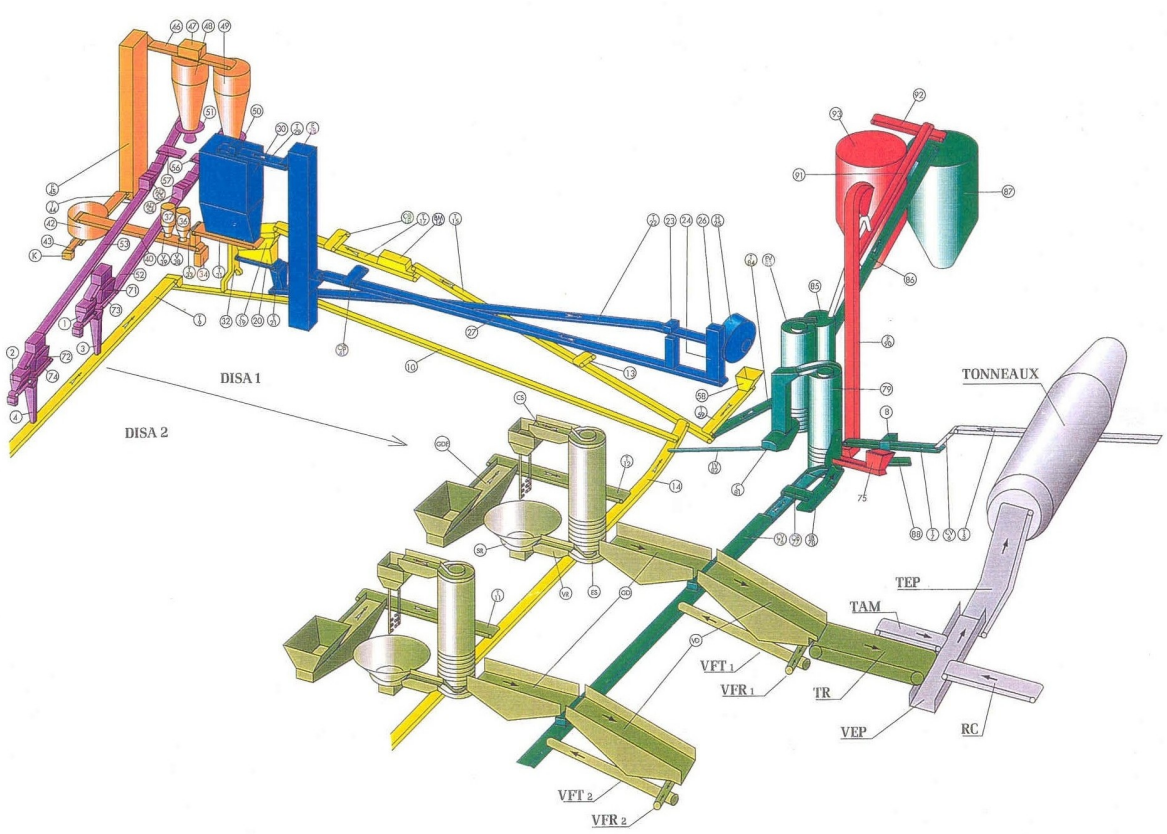
− Éléments constituant un noyau (photos à gauche) et leur assemblage (photos à droite) : fûts et chambre de bielles (haut), retours d'huile et chambre d'eau (bas) −

3.2.5 La sablerie

La sablerie est le secteur de la fonderie où le sable à vert est préparé. Il s'agit d'un ensemble d'équipements destinés à la récupération et à la régénération du sable à vert qui, après son utilisation au moulage et à la coulée, perd ses caractéristiques physiques et mécaniques exigées. Un schéma de la sablerie de FPF est représenté dans la figure à la page suivante. Les fonctions principales sont :

1. Le sable prêt à l'emploi est distribué vers les trémies qui alimentent les deux lignes de moulage (violet).
2. Après la coulée, deux installations de démoulage effectuent des opérations de décochage, débourrage et dessablage des pièces. La majeure partie du sable des mottes est récupéré et convoyé vers la régénération (jaune). Le sable des noyaux est récupéré séparément, conditionné et stocké (vert foncé). Ce sable, dit « neuf », est introduit en flux variable dans la sablerie afin d'en rééquilibrer le niveau qui diminue à cause des pertes au démoulage.

3. Une première étape de régénération (bleu) est effectuée en éliminant les particules fines par aspiration et en refroidissant le sable par humidification.
4. Pendant la deuxième étape de régénération (orange) de l'argile, des additifs carbonés et de l'eau sont additionnés et malaxés, dans des proportions qui permettent au sable à vert d'atteindre les propriétés physiques et mécaniques prédéterminées. Ensuite, le stockage du sable à vert prêt à l'emploi ferme cette boucle d'utilisation-régénération, qui est couramment appelée « circuit du sable ».

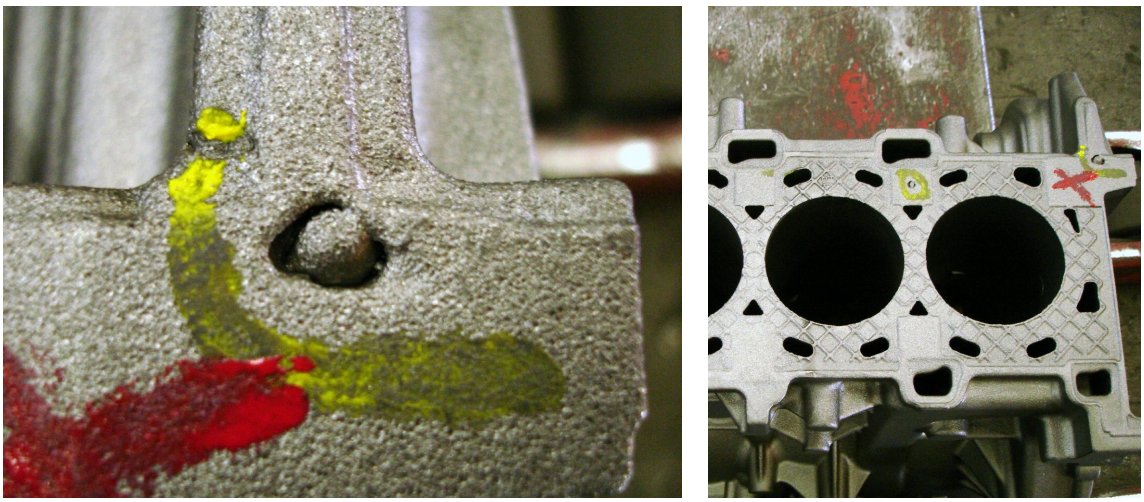


− Schéma de la Sablerie de FPF −

3.2.6 Le défaut de soufflure

Le défaut de soufflure est une porosité créée par l'occlusion de gaz endogènes ou exogènes[2], lors de la solidification de la fonte liquide dans le moule. Quand l'occlusion atteint la surface de la pièce, ce défaut peut être détecté pendant les contrôles visuels effectués en finition (les deux figures à la page suivante montrent un exemple de soufflure visible). En revanche, si l'occlusion est « emprisonnée » à l'intérieure de la pièce sa détection nécessite d'autres moyens de contrôle afin d'éviter que le défaut soit découvert pendant l'usinage chez le client. À FPF une station de contrôle par ultrasons existe. Cette station n'est pas automatisée : un opérateur effectue avec une sonde le contrôle d'un certain nombre de points « critiques » sur les carters. En outre, ce contrôle s'effectue sur la totalité des pièces sortant des lignes des production seulement si un seuil d'alerte est atteint en finition (c'est-à-dire seulement si un nombre minimum de soufflures ont déjà été détectées visuellement).

[2] On définit endogènes les gaz dégagés par le métal liquide, et exogènes les gaz qui ont une origine différente.



− Exemple de défaut de soufflure (gauche) et son positionnement sur un bloc cylindre (droite) −

Puisque une soufflure se produit lors de la solidification de la fonte, toutes les étapes du procédé de fabrication en amont sont potentiellement à l'origine de ce défaut. Les secteurs fusion, noyautage, sablerie, moulage et coulée sont donc concernés. Je résume ici quelques concept de base concernant le mode de formation de ce défaut.

Les causes du défaut de soufflure sont à rechercher à la fois dans un excès de gaz à l'intérieur du moule, et dans une mauvaise évacuation des gaz vers l'extérieur du moule. L'origine des gaz est multiple :

1. L'air présent dans les cavités du moule avant la coulée et qui peut être entrainé pendant le remplissage par des phénomènes de turbulence, ou qui n'est pas remplacé par du métal à cause d'un apport insuffisant. La dimension de la porosité résultante est proportionnelle au rapport de la température (ou de la pression) à l'instant où la fonte se solidifie et celle à l'instant où l'occlusion se produit.
2. La présence de gaz dissous dans le métal liquide, liée aux procédures de fusion, de transport et de coulée. Certains gaz, comme l'oxygène, l'hydrogène ou l'azote sont fortement solubles dans la fonte liquide. Leur solubilité est fonction de la température et de la pression.
3. L'air présent dans les porosités de la motte et du noyau avant la coulée, qui est réchauffé par le contact avec le métal liquide, et les gaz formés par des réactions chimiques à l'interface entre métal liquide et motte, ou métal liquide et noyau (comme par exemple l'évaporation d'eau ou la combustion des résines). Lorsque la pression de ces gaz est localement supérieure à la hauteur métallo-statique du liquide[3], ces gaz se dissolvent ou diffusent (sous forme de bulles) dans le liquide plutôt que de s'échapper à traverser la motte ou le noyau.

Lors de la conception du carter, les solutions qui peuvent être apportées dans ces trois cas pour contrecarrer l'apparition du défaut de soufflure sont respectivement :

1. Modifier la géométrie du carter et du système de coulée (chenaux et évents), afin d'améliorer le comportement dynamique des fluides (liquide et gaz) lors de la coulée. Le but est de garantir des flux le plus possible laminaires et un remplissage régulier du carter.
2. Optimiser le système de coulée d'un point de vue du comportement thermodynamique de la fonte lors du remplissage et du refroidissement du carter.

[3] Concept qui s'apparente à la pression hydrostatique d'un liquide en fonction de la profondeur.

En particulier, il est important d'éviter des gradients thermiques excessifs.

3. Modifier la géométrie du carter et du système de coulée (évents), ainsi que la conception des éléments du noyau, afin de minimiser les zones à risque de surpression. Ces zones correspondent, par exemple, aux points sur les interfaces métal-motte excessivement distants d'un évent ou de la surface extérieure du moule, ou les points à l'intérieur d'éléments du noyaux présentant un rapport surface exposée au métal / section trop élevé.

L'objectif de mon stage était d'identifier et d'évaluer les solutions qui peuvent être apportées au niveau du procédé de production. Par rapport aux trois sources de gaz mentionnées précédemment, il est essentiel de :

1. Surveiller l'état des machines et des outillages afin de garantir la précision des opérations de moulage (alignement des mottes et des noyaux, serrage optimale du moule) et de coulée (respect des lois de remplissage).
2. Surveiller la composition et la qualité métallurgique de la fonte, ainsi que les conditions de refroidissement après la coulée.
3. Surveiller la composition et les caractéristiques mécaniques des mottes et des noyaux, afin de minimiser la production de gaz et de garantir une perméabilité optimale du moule.

J'ai effectué rapidement des recherches bibliographiques sur ces sujets. J'ai pu trouver principalement des articles concernant des études effectuées en laboratoire (voir la bibliographie dans [2]). Les connaissances issues de ces recherches peuvent certainement orienter les techniciens d'une fonderie vers l'identification des causes d'un défaut de soufflure. Toutefois la « réalité » des usines est extrêmement plus complexe par rapport aux conditions idéales que l'on peut reproduire en laboratoire. Le nombre des paramètres procédé à considérer est nettement plus élevé, et la valeur de la plupart de ce paramètres ne peut pas être maintenue constante mais varie dans des fourchettes de travail plus ou moins étendues, en fonction de la maîtrise que les techniciens ont des différents aspects du procédé. Par conséquent, en cours de production il est extrêmement difficile de garantir des conditions expérimentales idéales (permettant de faire varier la valeur d'un paramètre, tout en gardant constantes les autres, et d'évaluer les conséquences de la variation sur le défaut). En outre, les caractéristiques des fonderies (équipement, outillage, etc.) diffèrent d'une usine à l'autre, même si leurs procédés sont similaires. Les solutions concrètes à apporter pour résoudre un problème de soufflure restent donc à rechercher au cas par cas.

Ces considérations expliquent la difficulté du problème à résoudre pendant mon stage et justifient le recours à des analyses de type statistiques sur de gros volumes des données, issus d'observations du procédé sur des longues périodes, plutôt qu'à la mise en place d'essais ciblés et de courte durée. Le paragraphe suivant fourni des informations sur les paramètres procédé auxquels j'ai eu accès à FPF.

3.3 Compréhension des données

Les analyses « statistiques » des données procédé peuvent être considérées complémentaires aux analyses basées sur des modélisations « physiques » du problème de soufflure. Comme décrit au paragraphe précédent, la complexité des phénomènes physiques est telle que plusieurs hypothèses peuvent être formulées sur les causes du soufflure. Mais souvent, comme j'ai pu le vérifier pendant le stage, les avis des experts ne convergent vers aucune explication claire et univoque permettant de définir une stratégie de résolution du problème. Dans ces cas, les données, surtout si elles sont abondantes,

peuvent « cacher » des informations précieuses qui peuvent être révélées d'une manière objective par une méthodologie d'analyse adaptée. J'ai donc démarré le travail de définition d'une telle méthodologie par un « brain storming » avec les responsables et techniciens des différents secteurs de l'usine, visant à déterminer quels étaient les paramètres à prendre en considération.

Au total, 67 paramètres procédé (ou groupes de paramètres) ont été recensés, parmi lesquels 39 sont actuellement enregistrées sur un support informatique ou papier. Je liste rapidement quelques-unes des variables mesurées en les groupant par secteur :

1. Fusion : proportion des matières premières constituant le lit de fusion, composition chimique de la fonte[4], analyse thermique de la fonte[5], etc.
2. Noyautage : granulométrie du sable, viscosité des résines et pourcentage d'utilisation dans le mélange, humidité et concentration d'acide dans le gaz catalyseur, résistance à la flexion des noyaux, viscosité et humidité résiduelle des enduits réfractaires, perte au feu[6], etc.
3. Sablerie : caractéristiques mécaniques du sable à vert (perméabilité, aptitude au serrage[7], résistance au cisaillement), caractéristiques physiques du sable à vert (humidité, température, teneur en carbone, teneur en argile active[8], teneur en particules fines), etc.
4. Moulage : référence des outillages utilisés dans chaque campagne de production, cadence de moulage, dureté de la motte, aptitude au serrage[9] du sable à vert, etc.
5. Coulée : température de coulée, diamètre du jet, hauteur métallo-statique dans le « bol » de coulée, temps de « stockage » des carters avant leur démoulage, etc.

À cette liste non exhaustive s'ajoute aussi le « facteur humain » : il a été reconnu que les compétences, le niveau d'attention et la motivation des équipes peuvent jouer un rôle non négligeable sur la qualité finale des produits.

Après avoir dressé la liste des paramètres candidats à être analysés, j'ai commencé à recenser les procédures de mesure et d'enregistrement. J'ai pu remarquer les faits suivants :

1. Les mesures des paramètres sont effectuées de façon asynchrone et avec des fréquences d'échantillonnage très différentes allant de quelques dizaines de secondes (mesures effectuées automatiquement par des capteurs placés sur le processus), à quelques minutes, heures, voir une journée entière (mesures réalisé manuellement par les opérateurs en laboratoire). Dans le cas des mesures réalisées manuellement, les périodes d'échantillonnage peuvent fluctuer considérablement, et des délais parfois importants et incertains peuvent être

[4] La composition chimique est obtenue par l'analyse au spectromètre d'échantillons de métal.

[5] L'analyse thermique consiste à mesurer l'évolution dans le temps de la température d'un échantillon de métal liquide lors de son refroidissement. L'analyse de la courbe de température en fonction du temps permet de déterminer quelle est la structure microscopique de la fonte. L'analyse thermique peut donc être considérée comme un indicateur de qualité métallurgique.

[6] La perte au feu mesure le teneur en matières volatiles d'un échantillon soumis à un séjour dans un four à 900 °C en atmosphère confinée.

[7] L'aptitude au serrage (exprimée en pourcentage) correspond à la diminution de hauteur d'un échantillon de sable à vert placé dans une éprouvette standard et soumis à une pression prédéterminée.

[8] L'argile est dite « active » lorsque se constituants, organisés dans des structures ayant la forme de feuillets, se lient à l'eau grâce à l'action du malaxeur. À des températures supérieures à 450 °C, cette liaison est perdue irréversiblement : les feuillets d'argile ayant subi cette transformations ne peuvent plus être « réactivés ».

[9] L'aptitude au serrage mesuré au moulage correspond à l'aptitude au serrage de la motte entière, contrairement à l'aptitude au serrage mesuré sur des échantillons standards prélevés en sablerie.

observés entre l'instant où les échantillons sont prélevés et l'instant où les mesures sont effectuées.
Par exemple, l'analyse chimique de la fonte à la coulée est effectuée 2÷4 fois par heure sur des prélèvements qui sont envoyés au laboratoire. Les mesures du spectromètre sont disponibles avec un délai variable normalement entre 10 et 20 min. La mesure des la plupart des caractéristiques du sable à vert se fait en revanche toutes les 2÷3 heures et dure entre 30 et 45 minutes.

2. Chaque pièce est marquée avec un code permettant de savoir dans quelle « tranche horaire » elle a été coulée. La tolérance sur les heures nominales de début et de fin de chaque tranche horaire est de ±10 min. L'horodateur prévu à cet effet est actualisé manuellement par les conducteurs des machines. Les procédures en vigueur leur imposent de respecter cette intervalle de tolérance. Sous l'hypothèse que l'ordre d'arrivée de pièces à la finition respecte l'ordre de coulée, il serait possible d'exploiter le système de supervision des machines à mouler et à couler (qui enregistre l'heure exacte de moulage et de coulée de chaque pièce) afin de remonter avec plus de précision à l'heure de coulée des pièces présentant des défauts. Toutefois, un tel système de traçabilité n'a pas été mis en place. La raison est que certains événements imprévus (par exemple une panne nécessitant le stockages temporaire de pièces en amont du parachèvement) peuvent interrompre ou modifier le flux des pièces : l'ordre dans lequel les pièces sont contrôlées n'est pas forcement l'ordre dans lequel elles ont été coulées.
3. Pour certains paramètres les procédures de traçabilité ne sont pas appliquées correctement dans l'usine, avec pour conséquence l'impossibilité de savoir précisément à quelle tranche horaire doivent être affectées certaines mesures. Par exemple, juste après avoir été produits certains éléments des noyaux sont accrochés sur des charriots et transportés sur la ligne d'assemblage. Toutefois, l'heure de sortie des charriots de l'atelier des production n'est pas toujours renseigné et les charriots ne sont pas toujours vidés en respectant l'ordre d'arrivée. Dans ces cas, les mesures des paramètres effectuées au noyautage ne peuvent pas être assignées à une tranche horaire précise.
4. À la finition, lorsque les pièces sont contrôlées afin de détecter des éventuels défauts, l'examen est toujours arrêté au premier défaut trouvé. Dans ce cas la pièce est déclarée rebut et le défaut enregistré dans une base de données dédiée à cet effet. Puisque plusieurs défauts peuvent coexister sur la même pièce, le nombre de défauts de soufflure est systématiquement sous-estimé.
5. La même base de données utilisée en finition enregistre aussi les déclarations de productions. Chaque déclaration indique le nombre des pièces coulées depuis la dernière déclaration. Étant effectuées par le conducteurs des machines à mouler et couler avec une cadence irrégulière, la quantification du nombre de pièces produites dans chaque tranche horaire peut contenir des imprécisions.
6. L'organisation des opérations de mesure, ainsi que le support et le format des enregistrements et même les unités de mesure, ont été définis pour répondre à des besoins et à des contraintes qui peuvent être différents d'un secteur à l'autre, et qui ont subi des évolutions au fil du temps. Un système de gestion de bases de données, permettant la centralisation et facilitant l'extraction et le traitement des données procédé, n'a pas été implémentée.
Les fichiers Excel (remplis manuellement ou issus des journaux des automates) représentent quasiment la norme, quelques bases de données alimentées par des logiciels propriétaires ou commerciaux sont utilisées, mais des fiches en papier sont encore présentes. Le format de certains tableaux rend très complexe l'extraction

automatique d'informations. Par exemple, par inspection des fichiers de suivi du secteur fusion, je pouvais déterminer à quels moments un des trois fours avait servi une des deux ligne de production, mais je n'ai pas pu mettre en place une requête permettant d'extraire automatiquement un historique sur une période donnée.

En raison de ces faits, il est impossible d'obtenir une « traçabilité » exacte des valeurs des paramètres pour chaque pièce. La modification des procédures de mesure et d'enregistrement n'était pas possible pendant mon stage pour des raisons organisationnelles évidentes, mais aussi de coût (exception faite pour la mise en place d'essais ciblés et de courte durée). En effet, les bénéfices apportés par les analyses statistiques ne pouvaient pas être évalués précisément avant leur démarrage. Le risque de « gaspillage » de ressources (humaines et matérielles) était à mon avis non négligeable.

J'ai donc décidé de me tourner vers des solutions simples permettant d'exploiter la masse des données collectées dans le passé selon les procédures courantes à FPF. Ces solutions seront présentées dans le paragraphe suivant. En ce qui concerne les procédures de mesure et les formats d'enregistrement, je me suis simplement limité à formuler quelques propositions de modifications qui, à mon avis, permettraient d'exploiter d'une manière plus efficace la base de données procédé. J'ai communiqué ces propositions lors des réunions d'avancement et je les ai ensuite synthétisées dans les comptes rendus des réunions.

3.4 Préparation des données

Le travail de préparation des données comporte généralement deux phases :

1. Les données « brutes » sont centralisées dans une base de données conçue pour faciliter les prétraitements.
2. Les données subissent des prétraitements afin de les mettre sous la forme exigée par les algorithmes d'analyse.

En ce qui concerne la centralisation, il aurait été judicieux de mettre en place une base dédiée aux données procédé, qui à FPF n'existe pas. Des systèmes de gestion professionnels existent, soit commerciaux (Oracle, par exemple), soit open source (MySQL et PostgreSQL, par exemple). Toutefois, mes compétences en ce domaine ne m'auraient pas permis d'atteindre des résultats significatifs avant la fin du stage. Pour cette raison, j'ai réalisé ce que je définirais juste un « brouillon » de base de données en utilisant simplement les fonctionnalités offertes par les logiciels Excel et Microsoft Query (avec lequel on peut créer des requêtes en langage SQL). La structure de cette base est décrite dans la suite du paragraphe.

J'ai fait d'abord l'hypothèse de tranches horaires « parfaites », autrement dit que la mise à jour de l'horodateur se fait à l'heure nominale de changement de tranche horaire. J'ai donc défini un système de numérotation séquentielle des tranches horaires, dans le quel la première tranche prise en considération (numéro 1) correspond à intervalle de 00:00 à 01:00 du 01 janvier 2010[10].
J'ai créé ensuite plusieurs tableaux à partir des sources de données disponibles à FPF. Certaines de ces données, disponibles exclusivement sur support papier, ont nécessité d'une transcription préliminaire sur support informatique. Pour d'autres, dont la

[10] Nous avons décidé de ne pas remonter plus loin dans le temps. Puisque l'étude s'est focalisée sur le carter « M9 », le risque aurait été de prendre en considération de données concernant des versions anciennes de ce carter, différentes de l'actuelle, et donc de fausser dans une certaine mesure les résultats des analyses.

récupération était laborieuse, j'ai demandé l'aide du Service Informatique. Le premier de ces tableaux est censé contenir principalement les données de production, c'est-à-dire la quantité Q de pièces produites pour chaque référence dans chaque tranche horaire, plus d'autres informations complémentaires (date, heure de début de la tranche, etc.). Le format de ce tableau est similaire à l'exemple suivant :

PRD			
N° T.H.	**Réf.**	**Q**	**...**
1	M9T-469R	150	...
2	M9T-469R	80	...
2	M9R-510	60	...
3	M9R-510	170	
...	...	...	...

Le deuxième tableau contient le nombre des défauts constatés en Finition (rebuts « internes ») ou chez le client (rebuts « externes ») pour chaque référence et tranche horaire et classés par type de défaut (Y_1, Y_2,...), plus éventuellement d'autres informations complémentaires (position du défaut sur le carter, etc.) :

FNT						
N° T.H.	**Réf.**	**Int./Ext.**	**Y_1**	**Y_2**	**...**	**...**
1	M9T-469R	Interne	2	0	...	...
2	M9T-469R	Interne	0	1	...	...
2	M9R-510	Interne	1	5	...	...
2	M9R-510	Externe	0	3	...	...
3	M9R-510	Externe	1	0	...	...
...	...	...	...	...	...	...

Dans la suite, j'indiquerai le nombre de défauts de soufflure avec Y_S.

Les autres tableaux contiennent les données procédé (X_1, X_2,...). Il y a au moins un tableau par secteur, dans lequel sont renseignées : la date et l'heure de prélèvement des échantillons sur lesquels les mesures sont effectuées, la valeur des mesures, plus éventuellement d'autres informations complémentaires. Le numéro de tranche horaire dans laquelle chaque mesure a été effectuée est calculé automatiquement à partir de la date et de l'heure :

FSN, NYT, SBL, MLG ou CLE						
N° T.H.	**Date**	**Heure**	**X_1**	**X_2**	**...**	**...**
1	01/01/10	00:13	3,15	0,18	...	...
2	01/01/10	01:27	3,25	--	...	...
2	01/01/10	01:49	3,45	--	...	...
3	01/01/10	02:05	3,15	0,12	...	...
...	...	...	...	...	...	...

Cette structure de la base m'a permis d'effectuer d'une manière plus efficiente : la sélection des paramètres procédé et des défauts que je voulais soumettre aux différentes analyses; le prétraitement des mesures; l'exportation des données vers les logiciels d'analyse. Grâce à des opérations de jointure, groupage et filtrage des ces tableaux, il est possible d'obtenir facilement des tableaux destinés à l'analyse ayant le format suivant :

Données à analyser										
N° Obs.	N° T.H.	Paramètres procédé				Nombre pièces	Nombre défauts			
		X_1	X_2	...	X_M	Q	Y_1	Y_2	...	Y_P
1	1	3,15	0,18	...	...	150	2	0	...	...
2	2	3,35	0,15	...	...	140	1	9	...	...
3	3	3,15	0,12	...	...	170	1	0	...	...
...	...	...	...	...	...	...	...	...	...	...

Ce format est celui demandé par les algorithmes d'analyse que j'ai utilisé. Chaque ligne de ce tableau, que j'appelle observation, correspond à une tranche horaire. À chacun des M paramètres procédé sélectionnés (X_1, X_2,..., X_M) est assigné une et une seule valeur pour chaque tranche horaire. Cela correspond à supposer que le paramétré reste constant à l'intérieur d'une tranche. En ce qui concerne les tranches horaires où il n'y avait pas de mesure disponible les solutions possibles étaient :

1. Ne pas prendre en considération les observations où il y avait des données manquantes.
2. Ne pas prendre en considération les paramètres mesurés avec des périodes trop longues relativement aux autres.
3. Estimer les données manquantes par interpolation (comme montré dans l'exemple précédent pour X_2, dont la valeur sur la tranche numéro 2 à été calculée par une interpolation linéaire).

En pratique, le choix a toujours été porté sur les deux premières solutions. Lorsque pour certains paramètres j'évoquais la troisième possibilité, les techniciens avec lesquels j'ai pu discuter se sont toujours montré sceptiques, puisque ils considéraient que l'hypothèse d'une dérive lente de ces paramètres, nécessaire pour utiliser la méthode d'interpolation, ne pouvait pas être justifiée en pratique. Par exemple, il n'était pas raisonnable d'effectuer d'analyses où la granulométrie du sable à vert, qui est mesurée une fois par mois, était considérée conjointement à d'autres paramètres Sablerie mesurés toutes le 2÷3 heures.

En revanche, en ce qui concerne les tranches horaires où il avait deux mesures où plus du même paramètre, il était nécessaire de définir une quelque fonction permettant d'agréger ces mesures. La fonction retenue dans tous les cas a été la moyenne arithmétique (comme montré dans l'exemple précédent pour X_1, dont la moyenne des mesures a été calculée sur la tranche horaire numéro 2). Dans ces cas, l'hypothèse de dérive lente du paramètre pouvait toujours être justifiée. Par exemple, lorsque la fonte est produite par le même four de fusion, il est raisonnable de supposer que la composition chimique des « poches » de métal liquide (qui sont transférées approximativement tous les 10 min aux fours de coulée) ne présente pas de variations brusques d'une poche à l'autre.

Les analyses que j'ai conduit par la suite sur les données de ces tableaux avaient pour objectif de répondre aux questions suivantes :

1. supposant d'avoir assignées les valeurs des paramètres X_1, X_2,..., X_M, quel sera vraisemblablement le taux constaté du défaut de soufflure (qui par définition est égal à Y_S / Q) ?
2. quelles sont les valeurs à assigner aux paramètres X_1, X_2,..., X_M afin de minimiser le taux du défaut de soufflure qui sera vraisemblablement constaté avec ces valeurs ?

La méthodologie d'analyse que j'ai conçu est présentée dans les paragraphes suivants.

3.5 Modélisation

3.5.1 État de l'art

Plusieurs techniques existent pour l'analyse statistique (ou plus généralement pour l'« exploration ») des données [3]. Leur but commun est l'extraction d'un savoir utile à une entreprise à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. Elles construisent des modèles des données en découvrant des structures sous-jacentes souvent invisibles à des inspections « visuelles ». Ces techniques répondent à trois besoins :

1. Description : générer un modèle capable de représenter d'une manière concise les données associées à un ensemble de paramètres, dans le but de faire des bilans, voir plus clair, etc.
2. Prédiction : générer un modèle capable de prédire la valeur d'une quantité cible en fonction de la valeur d'autres paramètres, dans le but d'aider à la décision, de maîtriser les risques, etc.
3. Planification : générer un modèle capable de proposer la valeur à assigner à un ensemble de paramètres afin d'optimiser la valeur d'une quantité cible, dans le but de définir des plans d'expériences, redéfinir des plages nominales de travail, etc.

En simplifiant en peu, il est possible de distinguer trois étapes dans un travail de modélisation :

1. Une classe de modèles est choisie, parmi le nombreuses classes existantes. Il s'agit de la classe qui est mieux censée représenter les relations entre les variables présentes dans les données.
2. Des algorithmes d'optimisation sont appliqués afin d'identifier le modèle, appartenant à la classe choisie, qui mieux décrit les données.
3. Le modèle identifié est utilisé selon les besoins (prédictions, prise de décisions, planification, etc).

Mon travail a eu pour but d'étudier et de sélectionner une technique de modélisation (une classe de modèles et les algorithmes d'identification) me permettant d'apporter des réponses aux questions suivantes :

1. Quel sera le nombre des défauts de soufflure sur un lot de $\bar{q}$ pièces si M paramètres procédé, X_1, X_2,..., X_M, sont maintenus à des valeurs $\bar{x}_1$, $\bar{x}_2$,..., $\bar{x}_M$ prédéterminées ? La solution de ce problème consiste à trouver une manière de prédire l'inconnue Y_S en se basant sur les informations qui peuvent être extraites à partir des N observations « passées », comme schématisé dans l'exemple suivant :

Le problème de la prédiction					
N° Obs.	N° T.H.	Paramètres procédé		Nombre pièces	Nombre défauts
		X_1	X_2	Q	Y_S
1	1	3,15	0,18	150	0
2	2	3,35	0,15	140	9
3	3	3,15	0,12	170	0
		$X_1 = \bar{x}_1$	$X_2 = \bar{x}_2$	$Q = \bar{q}$	Y_S = ?

2. Quelles sont les valeur à assigner à ces M paramètres procédé afin de ne pas dépasser un seuil prédéterminé de défauts de soufflure, $\bar{y}$, sur un lot de $\bar{q}$ pièces ?

La solution de ce problème consiste à trouver une manière de planifier les valeurs des inconnues X_1, X_2,..., X_M en se basant sur les informations qui peuvent être extraites à partir des N observations « passées », comme schématisé dans l'exemple suivant :

Le problème de la planification					
N° Obs.	N° T.H.	Paramètres procédé		Nombre pièces	Nombre défauts
		X_1	X_2	Q	Y_S
1	1	3,15	0,18	150	0
2	2	3,35	0,15	140	9
3	3	3,15	0,12	170	0
		X_1 = ?	X_2 = ?	Q = $\bar{q}$	$Y_S < \bar{y}$

3. Quels sont, parmi la totalité des paramètres procédé, les M paramètres X_1, X_2,..., X_M plus significatifs, c'est-à-dire essentiels et suffisants pour résoudre correctement les problèmes de la prédiction et de la planification ?

Une recherche bibliographique m'a permis de repérer une série d'articles (apparus à partir de 2008 et écrits par une équipe de chercheurs en collaboration avec deux fonderies aux Pays Basques) où ce genre de questions ont été abordées [4][5]. Dans ces articles est présentée une application de la technique des réseaux bayésiens, grâce à laquelle il a été possible de générer un modèle capable de prévoir l'apparition d'un particulier type de défaut (la micro retassure) en se basant sur la mesure d'un nombre réduit de paramètres procédé (24 à la place des 50 initialement prévus). Ce modèle a été utilisé pour réduire le nombre des contrôles non-destructifs (aux ultrasons et aux rayons X) appliqués aux pièces finies, en réduisant ainsi le coût de cette opération. La même équipe a testé aussi d'autres techniques pour résoudre le même problème [6][7][8] et, dans d'autres travaux, a aussi proposé des solutions pour estimer les propriétés mécaniques des pièces finies à partir des paramètres procédé, en réduisant ainsi le recours à des contrôles destructifs [9].

3.5.2 Sélection des techniques de modélisation

L'étude de l'état de l'art a initialement orienté mon choix vers la classe de modèles appelés réseaux bayésiens. Après avoir appris les fondements de cette technique de modélisation et avoir familiarisé avec les logiciels implémentant quelques algorithmes d'identification, j'ai décidé de l'utiliser conjointement à une autre technique, les cartes auto-adaptatives. Je décrirai plus en détail ces deux techniques dans le paragraphe suivant. Ici je rappelle simplement que la méthodologie d'analyse que j'avais commencé à mettre en place prévoyait d'utiliser les réseaux bayésiens afin de sectionner les paramètres pouvant être considérés vraisemblablement comme les « causes » du défauts de soufflure, et les cartes auto-adaptatives afin de calculer les fourchettes de travail optimales pour ces paramètres.

J'ai démarré ensuite le travail d'analyse sur un jeu de données constitué des mesures de la composition chimique de la fonte à la coulée (920 observations de 15 paramètres et du taux de soufflure sur la période octobre 2011-mars 2012).
Les résultats que j'ai obtenus par cette série de tests préliminaires ont été encourageants, puisque ils ont confirmé quantitativement une hypothèse qui était déjà supposée vraie par les experts de FPF. C'est-à-dire que des taux relativement bas de carbone et de silicium peuvent être associés à des taux élevés de soufflure. Par conséquent, j'ai considéré comme « validée » la méthodologie d'analyse et j'ai décidé de continuer les analyses en

intégrant des données métallurgiques additionnelles (377 observations effectuées sur la période juillet 2011-septembre 2011) et de les tester aussi séparément sur les données issues du secteur sablerie (889 observations de 11 paramètres et du taux de soufflure sur la période janvier 2011-mars 2012).
Toutefois, les résultats de cette deuxième série de tests se sont démontrés moins « satisfaisants ». Par exemple, dans le réseau bayésien identifié à partir des nouvelles données, la relation entre le taux de soufflure et le taux de carbone, dont l'existence avait été confirmée par le premiers tests, était devenue moins significative, au profit d'autres relations avec le cuivre, le plomb, le zinc et le titane. Mais l'application des cartes auto-adaptatives ne me permettait pas de conclure avec certitude sur l'utilité de hausser ou de baisser les fourchettes de travail de ces éléments chimiques afin de diminuer le taux de soufflure. L'analyse des mesures effectuées en sablerie a aussi fourni des résultats difficiles à interpréter. Par exemple, j'ai effectué différents tests de modélisations avec les réseaux bayésiens, en faisant varier d'un test à l'autre le sous-ensemble des paramètres analysés. Les résultats obtenus dans certains tests montraient que un paramètre pouvait être considéré significativement influent sur le taux de soufflure. Les résultats obtenus dans d'autres tests, en revanche, montraient que le même paramètre n'était pas du tout influent.

J'ai donc jugé nécessaire d'effectuer une étude plus approfondie de ces techniques, en particulier des réseaux bayésiens, afin de comprendre les raisons de ces résultats. Je ne rentre pas sur les détails concernant le déroulement de ce travail qui, si on se réfère à la méthode CRISP-DM, a consisté à parcourir plusieurs boucles entre les phases de Compréhension du Métier, Compréhension des Données, Préparation des Données, Modélisation et Évaluation. Ici je me limite à rappeler que ce travail m'a amené à adopter une classe de modèles différente : les modèles de régression logistique. Je n'ai pas pour autant abandonné ni la technique des réseaux bayésiens ni celle des cartes auto-adaptatives, que j'ai intégrées à d'autres techniques dans une méthodologie d'analyse en peu plus complexe de celle que j'avais envisagée initialement. En effet, j'ai trouvé que chaque technique présentait des limites par rapport aux objectifs de l'analyse et aux caractéristiques des données disponibles, et que ces limites pouvaient être surmontées par une utilisation conjointe d'un ensemble de techniques.

Dans le prochain paragraphe je présenterai une à une ces techniques, en expliquant les raisons pour lesquelles je les ai sélectionnées et leurs limites. Par la suite je décrirai la méthodologie que j'ai adopté en utilisant comme exemple l'analyse que j'ai effectué sur les données du secteur sablerie.

3.5.3 Description des techniques employées

3.5.3.1 Les cartes auto-adaptatives

Les cartes auto-adaptatives[11] permettent d'étudier comment des données multidimensionnelles se repartissent dans l'espace (multidimensionnel) qui les contient. Il s'agit d'une technique de classification, c'est-à-dire d'une technique qui vise à déterminer automatiquement une partition des données telle que les objets d'une même classe[12] se ressemblent le plus possible (grande homogénéité de chaque classe) et que les objets de classes différentes soient le plus différents possible (bonne séparation des classes) [10]. Les cartes auto-adaptatives répondent principalement à un besoin de description des

[11] Self-Organising Maps (SOM), en anglais.
[12] Couramment appelée aussi *cluster*, de l'anglais.

données, puisque elles permettent d'en résumer les caractéristiques saillantes. Mais les résultats qu'elles fournissent peuvent être interprétés aussi comme une forme de prédiction et peuvent être utilisés pour la planification. Je présente dans la suite comment je les ai initialement utilisées dans le but d'identifier les classes dans lesquelles le taux de soufflure était le moins élevé et de calculer les fourchettes de valeurs optimales des paramètres.

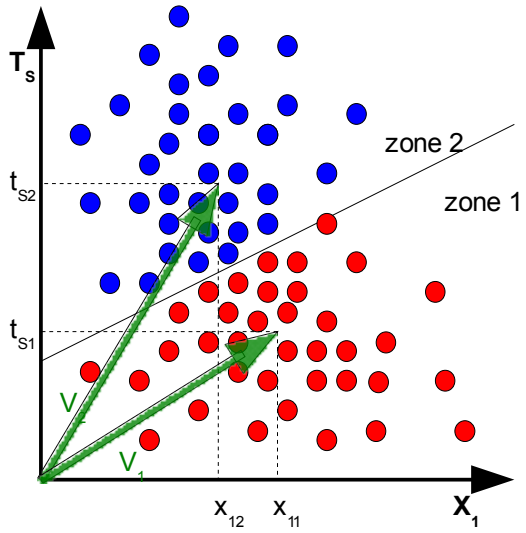
Supposons que les données d'entrée soient constituées de N observations (M+1)-dimensionnelles v(1), v(2),..., v(N) où

$$v(i) = [\ x_1(i)\ \ x_2(i)\ \ \ldots\ \ x_M(i)\ \ t_S(i)\ ]$$

est un vecteur qui contient les i-èmes mesures des paramètres X_1, X_2,..., X_M et du taux de soufflure, $T_S = Y_S/Q$. L'algorithme divise l'espace (M+1)-dimensionnel en P zones (où la valeur de P est un paramètre que l'on peut choisir librement), chacune contenant un sous-ensemble des N observations. Un vecteur référent, est assigné à chaque zone. Ces vecteurs, notés

$$V_j = [\ x_{1j}\ \ x_{2j}\ \ \ldots\ \ x_{Mj}\ \ t_{Sj}\ ]$$

où j est un index allant de 1 à P, représentent le sous-ensemble des observations v(i) qui appartiennent à la j-ème zone. Autrement dit, ils en résument les caractéristiques communes en minimisant une mesure de distance entre observations opportunément choisie. L'algorithme est initialisé en choisissant d'une manière aléatoire P vecteurs référents, et procède ensuite par itérations successives (par « raffinements ») jusqu'à convergence des vecteurs référents V_j (et donc des P zones) vers des valeurs stables. La figure suivante montre un exemple très simple de partitionnement.



− Exemple de partitionnement des données dans le cas où M=1 et P=2 −

L'interprétation intuitive du résultat de cette analyse des données est la suivante : lorsque les valeurs des paramètres étaient « proches » au référent $x_j = [\ x_{1j}\ x_{2j} \ldots x_{Mj}\ ]$ de la j-ème zone, le taux de soufflure observé était « proche » au référent t_{Sj} de la même zone. Le modèle ainsi obtenu peut donc être utilisé soit pour la prédiction, soit pour la planification :

1. Prédiction. Soient assignées aux M paramètres les valeurs appartenant au vecteur $\bar{x} = [\bar{x}_1\ \bar{x}_2 \ldots \bar{x}_M]$. Ce vecteur est comparé aux P référents x_j. L'objectif est de déterminer quel x_j se trouve à la distance minimale de $\bar{x}$. Le référent t_{Sj}

correspondant représente une prédiction du taux de soufflure.

2. Planification. Les valeurs des référents t_{Sj} sont ordonnées. Les référents x_j correspondant aux t_{Sj} plus bas représentent les valeurs optimales des paramètres, c'est-à-dire les valeurs autours desquelles centrer les fourchettes de travail pour minimiser le taux de soufflure. Ces fourchettes peuvent être déterminées en analysant la distribution des observations dans les zones dans lesquelles les taux de soufflure sont plus bas.

La particularité des cartes auto-adaptatives, par rapport à d'autres techniques de classification, est qu'elles permettent de cartographier sur une grille bidimensionnelle les P zones de la partition, tout en respectant les relations de « voisinage » entre les zones[13]. Lorsque M≥2, cette forme de visualisation des résultats peut, en principe, faciliter leur interprétation. Malgré cette intéressante propriété, les résultats que j'ai obtenu ne présentaient jamais des partitionnements où les « bonnes » et les « mauvaises » fourchettes de travail pouvaient être déterminées clairement, et grâce auxquels les paramètres significatifs pouvaient être repéré facilement. La figure à la page précédente met en évidence ce genre de difficultés dans un cas très simple : X_1 apparaît insignifiant pour la prédiction du taux de soufflure. Mais imaginons que les points bleus correspondent approximativement à des valeurs petites d'un autre paramètre, X_2, et que les points rouges à des valeurs élevées de X_2. On pourrait alors être tenté de conclure que c'est l'interaction entre X_1 et X_2 qui compte, et que des valeurs simultanément élevées de X_1 et X_2 favorisent la diminution du taux de défaut. Cependant, les cartes auto-adaptatives ne permettent pas d'évaluer le degré de confiance avec lequel on peut accepter cette hypothèse : d'autres techniques statistiques rigoureuses existent et doivent être utilisées à leur place.

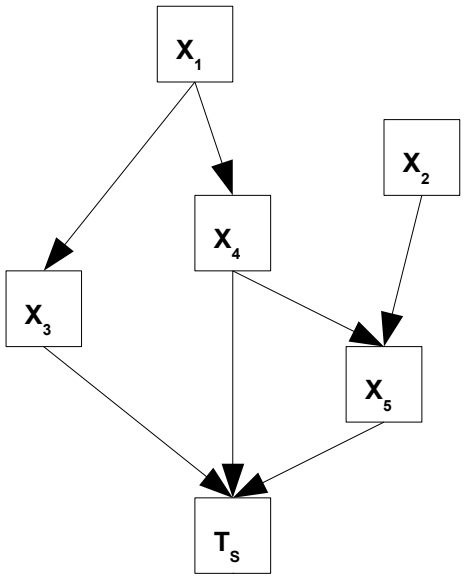
À ces problèmes s'ajoute le fait que l'application des cartes auto-adaptatives nécessite d'un choix préalable du nombre P de zones de la partition et d'une mesure de distance entre observations convenable. Pour effectuer ce choix je me suis appuyé sur des euristiques et sur une démarche par essais et erreurs (faute d'avoir trouvé d'outils appropriés). Mais, lorsque j'interprétais les résultats obtenus en fonction des différentes choix, j'ai observait souvent que des conclusions différentes pouvaient apparaître plausibles.

En définitive, l'interprétation des résultat obtenus avec les cartes auto-adaptatives était source d'introduction d'un degré non négligeable de subjectivité dans l'analyse. Afin de surmonter cette limite, je me suis donc tourné vers l'utilisation d'une autre technique d'analyse, les réseaux bayésiens, décrits dans le paragraphe suivant.

3.5.3.2 Les réseaux bayésiens

Les réseaux bayésiens sont des modèles de représentation des connaissances grâce auxquels les relations de cause à effet entre des quantités données sont décrites par un graphe acyclique : les nœuds du graphe représentent les quantités qui font l'objet de l'analyse ; les arcs liants les nœuds représentent les relations entre ces quantités [11]. La figure à la page suivante montre un exemple de réseau bayésien qui décrit X_3, X_4 et X_5 comme des causes directes du taux de soufflure T_S, X_1 comme une cause commune de X_3 et X_4, et X_5 comme étant causé par X_2 et X_4.

[13] Deux zones contiguës dans l'espace multidimensionnel des données sont représentées par deux nœuds adjacents sur la grille bidimensionnelle.



− Exemple de réseau bayésien −

L'intérêt d'utiliser cette technique dans le contexte de ma mission peut s'expliquer par les raisons suivantes :

1. Les réseaux bayésiens permettent de visualiser les relations de cause à effet entre les paramètres et le taux de soufflure sous une forme graphique, ce qui en facilite l'interprétation par des personnes n'ayant pas des compétences pointues en analyse des données.
2. Ils permettent de tenir compte des connaissances a priori des experts métier : il suffit d'ajouter des arcs orientés entre les paramètres qui sont réputés être liés par des relations de cause à effet, ou d'interdire la présence d'arcs entre les paramètres qui ne sont pas censés être en relation entre eux.
3. Des nombreux algorithmes d'identification permettent d'affiner la structure des relations définies par des experts métier en intégrant l'évidence expérimentale contenue dans les données : des arcs additionnels sont éventuellement ajoutés au réseau initial.
4. Plusieurs logiciels open source (dont l'utilisation n'implique aucun coût d'achat de licences pour FPF) sont disponibles, qui implémentent les nombreux algorithmes d'identification proposés dans la littérature statistique.

Les réseaux bayésiens ne sont des modèles déterministes. En revanche, les relations entre les quantités (soient elles renseignées à priori ou identifiées par les algorithmes) sont probabilisées. Considérons encore la figure précédente et supposons, par exemple, que les valeurs mesurées des paramètres procédé X_3, X_4 et X_5 sont égales respectivement à $\bar{x}_3$, $\bar{x}_4$ et $\bar{x}_5$. Le modèle ne fournit pas une seule valeur pour le taux de défaut T_S (ce qui constituerait une relation déterministe). En revanche il fournit la probabilité d'observer un taux de défaut égal à une valeur $\bar{t}_S$, pour toute valeur $\bar{t}_S$ arbitrairement choisie. Cette fonction de $\bar{t}_S$, $\bar{x}_3$, $\bar{x}_4$ et $\bar{x}_5$, ou loi de probabilité, est notée

$$\text{Prob}\left[T_S = \bar{t}_S \mid X_3 = \bar{x}_3,\ X_4 = \bar{x}_4,\ X_5 = \bar{x}_5 \right].$$

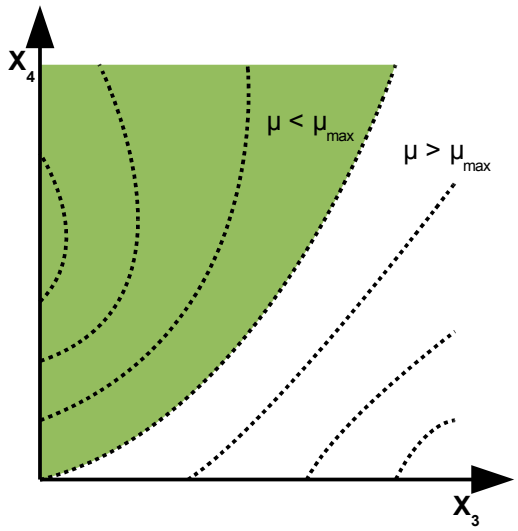
Un réseau bayésien représente donc un modèle pour des situations où l'observation des causes n'entraîne pas systématiquement les effets qui en dépendent, mais détermine seulement la probabilité d'observer ces effets. Autrement dit, même si les paramètres sont maintenus approximativement constants, le taux de défaut constaté peut varier d'une

tranche horaire à l'autre. Ces variations sont la conséquence de toute perturbation qui n'est pas observée ou mesurée, ou des variations des paramètres procédé qui ne sont pas pris en considération dans le modèle.

En résumant, la structure du réseau bayésien apporte une réponse au problème de repérer les paramètres significatifs, et la connaissance de la loi de probabilité indiquée plus haut permet de résoudre les problèmes de la prédiction et de la planification. Pour ces derniers l'approche plus répandu consiste en une transformation préliminaire de la loi de probabilité en un « indicateur », selon une fonction appelée souvent fonction de risque. Le choix de cette fonction est laissé à appréciation de l'analyste. Le choix plus simple est la moyenne, que j'indique avec la notation

$$\mu_{T_S}(\bar{x}_3, \bar{x}_4, \bar{x}_5),$$

mais d'autres solutions sont possibles. La prédiction fournie par le modèle est alors le taux de soufflure que l'on peut constater en moyenne si les valeurs des paramètres sont fixées à $\bar{x}_3$, $\bar{x}_4$, et $\bar{x}_5$. Et la planification cherche à déterminer les régions de l'espace tridimensionnel dans lesquelles la valeur de μ est inférieure à un seuil donné, comme montré dans la figure suivante.



− Exemple de résultat de planification : courbes de niveau de μ en fonction de X_3 et X_4 pour une valeur constante de X_5 et région où μ est inférieur au seuil μ_{max} −

Les réseaux bayésiens sont donc en principe des outils de modélisation très générales et flexibles. Cependant, la majorité des algorithmes qui, à partir des données, identifient la structure du réseau et estiment les lois de probabilité, ne sont justifiés que sous l'hypothèse de relations linéaires entre les quantités analysées. La structure représentée dans la figure précédente, par exemple, correspond à des équations ayant la forme suivante :

$$X_3 = a_{30} + a_{31}X_1 + E_3$$

$$X_4 = a_{40} + a_{41}X_1 + E_4$$

$$X_5 = a_{50} + a_{52}X2 + a_{54}X_4 + E_5$$

$$T_S = b_0 + b_3X_3 + b_4X_4 + b_5X_5 + E_0$$

où a_{ij} et b_i sont des coefficients à estimer à partir des données, et les E_i sont des termes

d'erreur (à moyenne nulle) qui rendent compte de la nature stochastique du modèle. Afin d'identifier la structure du réseau et d'estimer les coefficients du modèle j'ai utilisé l'algorithme PC-LiNGAM [12]. Contrairement à la majorité des autres algorithmes disponibles,PC-LiNGAM permet de s'affranchir de l'hypothèse additionnelle de distribution gaussienne des paramètres et des termes d'erreur. Cette hypothèse n'était pas toujours vérifiée dans les données que j'ai analysé. Les données d'entrée de cet algorithme sont constituées, comme pour les cartes auto-adaptatives, de N observations (M+1)-dimensionnelles v(1), v(2),..., v(N) où

$$v(i) = [\ x_1(i)\ \ x_2(i)\ \ ...\ \ x_M(i)\ \ t_S(i)\ ]$$

est un vecteur qui contient les i-èmes mesures des paramètres X_1, X_2,..., X_M et du taux de soufflure, $T_S = Y_S/Q$. J'ai pu constater les faits suivants :

1. Des relations non linéaires existent entre certains paramètres (documentées dans des rapports techniques que j'ai pu consulter), bien que les plages de variations des valeurs des paramètres ne soient pas généralement très larges à FPF (±10% au maximum par rapport aux valeurs nominales ciblée). L'hypothèse de linéarité s'est avérée trop restrictive dans ces cas.
2. Le taux de soufflure T_S est une quantité comprise entre 0 et 1. En revanche, une combinaison linéaire, telle que celle montrée plus haut ($b_0 + b_3X_3 + b_4X_4 + b_5X_5 + E_0$) n'est pas bornée pour toute valeur des des paramètre. Par conséquent, à la rigueur T_S n'est pas correctement modélisée. De plus, les analyses effectuées avec les cartes auto-adaptatives semblaient aussi montrer que cette relation n'est pas linéaire.
3. Les algorithmes d'identification des relations de cause à effet (la structure du réseau) se basent sur des procédures itératives permettant d'éviter le recours à des longues recherches exhaustives (lesquelles testent toutes les combinaisons possibles des quantités analysées). L'élimination ou l'introduction erronées d'une relation pendant une itération peut entrainer une accumulation d'erreurs qui peut rendre difficilement interprétable le résultat final. Par exemple, j'ai pu observer que des relations sans aucun fondement physique apparaissaient dans certains modèles simplement à cause de la présence de corrélations linéaires fortuites dans le jeu de données analysé. Afin de limiter l'occurrence de ces situations il est donc impératif d'intégrer toute connaissance a priori avant d'appliquer ces algorithmes. Cela a été un point manifestement difficile pour moi et les techniciens de FPF en raison de la complexité du procédé.

Ces considérations, et les résultats peu « satisfaisants » que j'ai déjà décrit plus haut, m'ont conduit à penser que les réseaux bayésiens n'étaient pas forcement la meilleure classe de modèles pour les données que je devais analyser. Je me suis donc tourné vers l'étude d'une solution alternative basée sur une technique bien connue en statistique : la régression logistique.

3.5.3.3 La régression logistique

La régression logistique (voir les Chapitres 4 et 5 de [13]) est une technique adaptée aux situations où la quantité à modéliser se présente à l'observateur seulement dans un nombre fini de modalités mutuellement exclusives. C'est la cas du défaut de soufflure, lequel peut être soit « présent » sur un carter, soit « absent ». L'approche de la régression logistique est alors de postuler que ces deux modalités peuvent être observées chacune avec une certaine probabilité, et d'assumer que ces probabilités dépendent des paramètres procédé. Autrement dit, si la probabilité de l'événement « défaut présent » est

indiquée par π, alors la probabilité de l'événement « défaut absent » est égal à 1-π, et la valeur de π varie en fonction des valeurs des paramètres. Le choix plus simple, consiste à utiliser une composition d'une fonction non linéaire et d'une combinaison linéaire des paramètres. Il est courant d'utiliser la fonction non linéaire définie par l'équation suivante, dite fonction logit :

$$\log(\pi/(1-\pi)) = b_1X_1 + b_2X_2 + ... + b_MX_M$$

où les b_i sont des coefficients à estimer à partir des données. De cette façon, la valeur de la fonction $\pi(X_1,X_2,...,X_M)$ est toujours comprise dans l'intervalle [0;1].

Dans l'hypothèse de maintenir constants les paramètres aux valeurs $\bar{x}_1, \bar{x}_2,..., \bar{x}_M$, la probabilité de constater un nombre des défauts égal à $\bar{y}_S$ dans une séquence de $\bar{q}$ observations est donnée par une loi binomiale. Cette probabilité s'écrit comme suit :

$$\text{Prob } [Y_S = \bar{y}_S \mid Q = \bar{q}, X_1 = \bar{x}_1, X_2 = \bar{x}_2 ... X_M = \bar{x}_M] = \binom{\bar{q}}{\bar{y}} \pi^{\bar{y}} (1-\pi)^{\bar{q}-\bar{y}}$$

où $\binom{\bar{q}}{\bar{y}}$ dénote le coefficient binomiale. La connaissance de cette loi permet de résoudre tout problème de prédiction et de planification, d'une manière similaire à celle que j'ai décrit dans le paragraphe sur les réseaux bayésiens, c'est-à-dire en calculant comme indicateur de risque le taux de soufflure moyen, par exemple,

$$\mu_{T_S}(\bar{x}_1, \bar{x}_2,..., \bar{x}_M).$$

Pour la loi binomiale, cette moyenne est exactement égal à $\pi(\bar{x}_1, \bar{x}_2,..., \bar{x}_M)$. Il est donc nécessaire de connaitre les coefficients b_i du modèle. Leur estimation est effectuée en maximisant une fonction $L(b_1,b_2,...,b_M)$, dite de vraisemblance. Je ne rentre pas trop dans les détails de la technique; il suffit de savoir que la fonction de vraisemblance est égale au produit pour i = 1, 2,..., N des lois binomiales suivantes

$$\text{Prob } [Y_S = y_S(i) \mid Q = q(i), X_1 = x_1(i), X_2 = x_2(i) ... X_M = x_M(i)],$$

définies à partir de N observations (M+2)-dimensionnelles v(1), v(2),..., v(N), où

$$v(i) = [x_1(i)\ \ x_2(i)\ \ ...\ \ x_M(i)\ \ q(i)\ \ y_S(i)]$$

est un vecteur qui contient les i-èmes mesures des paramètres $X_1, X_2,..., X_M$, de la quantité produite Q, et du nombre de défauts de soufflure Y_S. L'algorithme d'optimisation cherche donc les valeurs des coefficients pour lesquelles, selon le modèle, les observations sont plus probables.

Par rapport aux trois considérations faites à la fin du paragraphe dédié aux réseaux bayésiens, l'application de la régression logistique permet de modéliser le taux de soufflure plus correctement qu'avec les réseaux bayésiens, en prenant en considération le fait que ce taux est une quantité bornée entre 0 et 1. Mais la régression logistique telle que je l'ai décrite jusqu'à ici n'apporte aucune réponse aux deux autres problématique mentionnées : la présence de relations non-linéaires et la sélection des paramètres plus significatifs. La solution que j'ai trouvé est basé sur l'utilisation de la régression logistique conjointement à une autre technique, la méthode MARS, que je présenterai dans le paragraphe suivant.

3.5.3.4 La méthode MARS

La méthode MARS[14] (régression multivariée par spline adaptative [14]) généralise la

[14] De l'anglais Multivariate Adaptive Regression Splines

régression linéaire classique. La façon plus simple de la présenter est peut être à travers deux exemples très simples.
Supposons de vouloir prédire la valeur d'une quantité Z en fonction de la valeur d'un paramètre X_1. Sous l'hypothèse que Z et X_1 sont liés par la relation linéaire $Z = a_0 + a_1X_1 + E$, où E est un terme d'erreur, il est possible d'effectuer une régression linéaire afin d'estimer les coefficients qui minimisent la variance de l'erreur E. La méthode MARS généralise la régression linéaire classique en faisant l'hypothèse que la relation entre Z et X_1 est linéaire par morceaux, et elle permet de déterminer automatiquement une partition optimale du domaine de X_1, dom(X_1), en intervalles définis par des relations du type $\alpha_i \leq X_1 \leq \alpha_{i+1}$.
Supposons maintenant de vouloir prédire la valeur d'une quantité Z en fonction de la valeur de deux paramètres X_1 et X_2. Dans l'hypothèse que Z est lié à X_1 et X_2 par la relation $Z = a_0 + a_1X_1 + a_{12}X_1X_2 + a_2X_2 + E$, il est encore possible d'effectuer une régression linéaire afin d'estimer les coefficients à partir des données. La méthode MARS généralise ce modèle en faisant l'hypothèse que la relation précédente entre Z, X_1 et X_2 est valide par régions, et elle permet de déterminer automatiquement une partition optimale de dom(X_1) x dom(X_2) en régions « rectangulaires », c'est-à-dire définies par des relations du type $\alpha_i \leq X_1 \leq \alpha_{i+1}$ et $\beta_i \leq X_1 \leq \beta_{i+1}$.

La méthode peut ainsi être généralisée à un nombre de paramètres et à un degré d'interaction arbitraires. Un modèle à 3 paramètres avec degré d'interaction 2, par exemple, contient les termes X_1, X_2, X_3, X_1X_2, X_1X_3, et X_2X_3. La partition optimale résulte d'un compromis entre la précision du modèle (variance de l'erreur de prédiction) et la complexité du modèle (nombre de régions). Elle n'est donc pas fixée a priori mais adaptée automatiquement. Dans tous les cas, le modèle identifié peut toujours être écrit comme une combinaison linéaire de fonctions non linéaires des paramètres :

$$Z = \sum_i b_i f_i(X_1,X_2,...,X_M) + E,$$

et la méthode fournit aussi l'expression de ces fonctions $f_i(X_1,X_2,...,X_M)$, i=1,2,...,S. L'utilisateur doit choisir le degré d'interaction et, s'il le souhaite, peut fixer un nombre maximale S_{max} de termes. De plus, la méthode est capable d'ordonner les paramètres présents dans ce modèle en fonction de leur « importance ». L'importance d'un paramètre peut être définie, un peu grossièrement, comme une mesure de la contribution du paramètre à la réduction de la variance de l'erreur de prédiction E : plus cette contribution est grande, plus le paramètre est considéré important. Les critères exacts et la procédure de calcul sont toutefois plutôt compliqués et je ne les présenterai pas ici.

En raison de ces propriétés, j'ai estimé que la méthode MARS pouvait être utilisée conjointement à la régression logistique afin d'apporter une réponse aux deux problématiques qui restaient encore ouvertes : la présence de relations non-linéaires entre le taux de soufflure et les paramètres, et la sélection des paramètres plus significatifs. L'idée consiste à définir la fonction logit du modèle de régression logistique comme étant une fonction polynomiale par régions des paramètres X_1, X_2,..., X_M (plutôt qu'une simple combinaison linéaire de ces paramètres) :

$$\log(\pi/(1-\pi)) = \sum_i b_i f_i(X_1,X_2,...,X_M)$$

Les fonctions f_i peuvent alors être sélectionnées grâce à MARS, tandis que les coefficients b_i peuvent être estimés en maximisant la fonction de vraisemblance, comme expliqué au paragraphe sur la régression logistique. Il existe une manière optimale pour effectuer cette sorte d'« hybridation » des deux techniques [15]. Toutefois, pendant le stage j'ai appliqué un autre algorithme, suboptimal, pour lequel une implémentation était déjà disponible avec

le logiciel d'analyse que j'ai utilisé.

3.6 Évaluation

Afin de valider les modèles obtenus, je les ai soumis d'abord à deux types de tests :

1. Dans le premier, les données disponibles sont fractionnées en deux sous-ensembles appelés respectivement données d'apprentissage et données test. Les données d'apprentissage (80-90% des observations) sont utilisées pour l'identification du modèle, tandis que les données test (les observations restantes) servent à évaluer la qualité des prédictions calculées grâce au modèle. Le critère de qualité que j'ai défini est le suivant. À partir de la loi binomiale, je calcule pour toute observation test le seuil $y_{S,max}(i)$ tel que

   $$\text{Prob}\ [\ Y_S \le y_{S,max}(i) \mid Q = q(i),\ X_1 = x_1(i),\ X_2 = x_2(i) \ldots X_M = x_M(i)\ ] = C,$$

   où C est une valeur de probabilité fixée en avance (je l'ai fixé égale à 0.99). Cette relation traduit le fait que, selon le modèle, le nombre des défauts constatés, $y_S(i)$ doit être inférieur au seuil $y_{S,max}(i)$ avec une probabilité de 99%. J'indique avec $N_{correct}$ le nombre d'observations pour lesquelles la condition $y_S(i) \le y_{S,max}(i)$ est vérifiée, et avec N_{test} le nombre total d'observations test. Plus le rapport $N_{correct}/N_{test}$ approche 100%, plus la qualité du modèle est satisfaisante. Ce test est répété en sélectionnant par randomisation différents sous-ensembles d'apprentissage et de test.
2. Le deuxième critère consiste au calcul de la moyenne des erreurs de prédictions en utilisant les données test. Ces erreurs sont définis comme la différence entre le taux de soufflure constaté et le taux prédit par le modèle :

   $$e(i) = y_S(i)/q(i) - \mu_{T_S}(x_1(i),\ x_2(i),\ldots,\ x_M(i)) = y_S(i)/q(i) - \pi(x_1(i),\ x_2(i),\ldots,\ x_M(i))$$

   Idéalement la moyenne devrait être nulle, en traduisant le fait que le modèle n'introduit aucun erreur systématique.

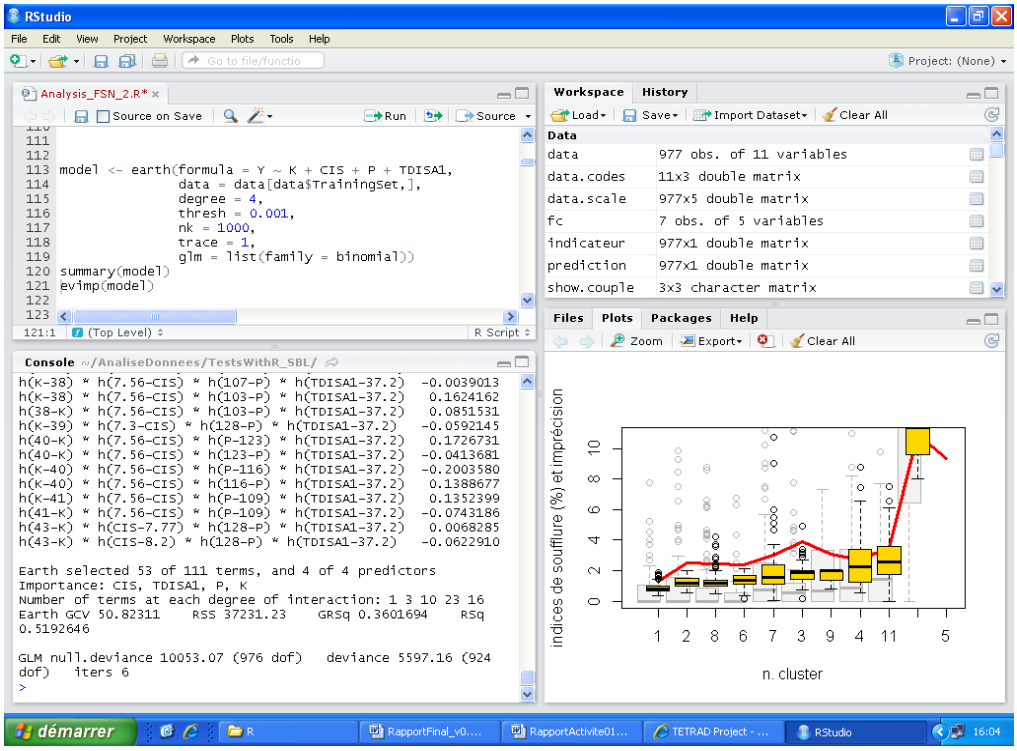
Ensuite, en collaboration avec les experts métier, j'ai toujours essayé de valider les modèles « statistiques » obtenus en cherchant une explication basée sur la « physique » du procédé. C'est surtout grâce à cette confrontation que j'ai affiné au fil du temps la méthodologie d'analyse.

3.7 Les logiciels utilisées

3.7.1 R

R est à la fois un langage de programmation et un environnement mathématique pour le traitement de données et l'analyse statistique [16]. R est distribué librement sous les termes de la GNU General Public Licence. Son développement est assuré par plusieurs statisticiens rassemblés dans le R Development Core Team. R est disponible sous plusieurs formes : le code (écrit principalement en C, C++ et certaines routines en Fortran et Java) surtout pour les machines Unix et Linux, ou des exécutables pré-compilés pour Windows, Linux et Macintosh. R est devenu ces derniers années un standard incontournable. Les plus grands logiciels commerciaux d'analyses statistiques, tels que Sas, Spss ou Statistica, proposent des interfaces pour intégrer des calculs et des graphiques réalisés avec R. Il est désormais utilisé par des sociétés informatiques (telles que Google, Microsoft et Facebook), pharmaceutiques (Johnson & Johnson, Merck et

Pfizer) et centaines d'autres.
R dispose de la plupart des fonctionnalités utiles pour la statistique de base et les graphiques, mais nombreux paquets (ou « extensions ») existent, mis librement à disposition, qui en augmentent considérablement les potentialités [17]. Pendant le stage j'ai utilisé R version 2.15.0 sous Windows XP et les paquets « earth » version 2.3-3 et « kohonen » version 2.0.9 qui implémentent respectivement la méthode MARS et les cartes auto-adaptatives. La régression logistique est implémentée dans la version de base de R. En outre, plusieurs interfaces graphiques et environnements de travail sont disponibles, qui facilitent le développement des programmes en R. La figure suivante montre l'environnement de travail RStudio que j'ai utilisé.

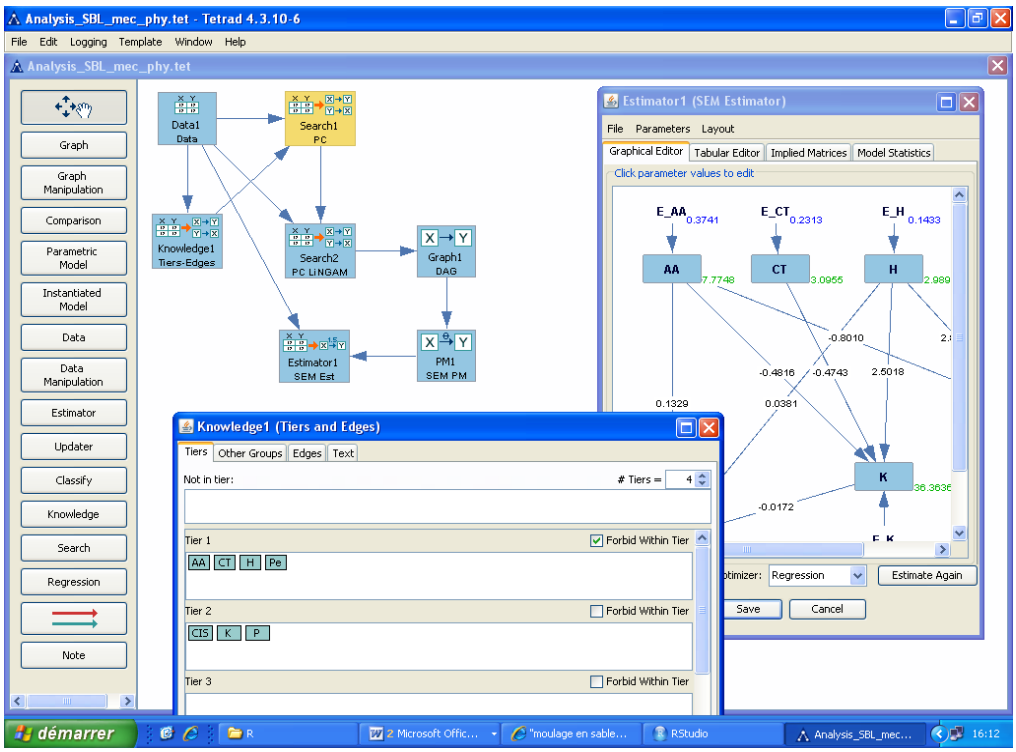


− Le logiciel R avec l'environnement de travail RStudio −

3.7.2 Tetrad IV

Tetrad IV est un logiciel qui implémente de nombreux algorithmes d'identification des réseaux bayésiens [18]. Il a été développé en Java par une équipe de chercheurs de la Carnegie Mellon University aux États Unis avec le support financier, entre autre, de la NASA[15] et du NSF[16]. Il est distribué gratuitement; de ce fait est une bonne alternative à des logiciels commerciaux tels que Netica, Hugin, LISREL, EQS, parmi les plus connus. L'environnement de travail proposé par le logiciel Tetrad est montré dans la figure suivante. Il a été conçu pour des utilisateurs n'ayant pas de connaissances particulières en programmation : toute concaténation d'opérations sur les données est représentée graphiquement par des liaisons entre blocs constituant les opérations élémentaires, et les graphes peuvent être facilement manipulés directement à l'écran.
Bien qu'il se présente comme un outil de modélisation extrêmement puissant et intuitif à utiliser, la documentation est toutefois très sommaire et, pour une utilisation correcte il est nécessaire de se référer à de la littérature statistique de niveau avancée.

[15] The National Aeronautics and Space Administration
[16] The National Science Foundation



– Le logiciel Tetrad IV –

3.8 Exemple d'application de la méthodologie d'analyse

Dans ce paragraphe je présente un exemple d'application des techniques décrites précédemment aux données du secteur sablerie. Pour information, je disposais de N=1001 observations réalisées sur la période de janvier 2011 au juin 2012. J'ai d'abord classé les paramètres dans les catégories suivantes :

1. Les mesures des flux des matières premières entrant dans le circuit du sable (argile, additifs carbonés, sable, eau). Ces quantités n'ont aucune relation avec le défaut de soufflure. Je les ai donc exclues de l'analyse.
2. Les mesures des propriétés physiques et chimiques du sable à vert, telles que sa composition, sa température, etc. Les paramètres que j'ai retenu pour l'analyse sont les suivants :

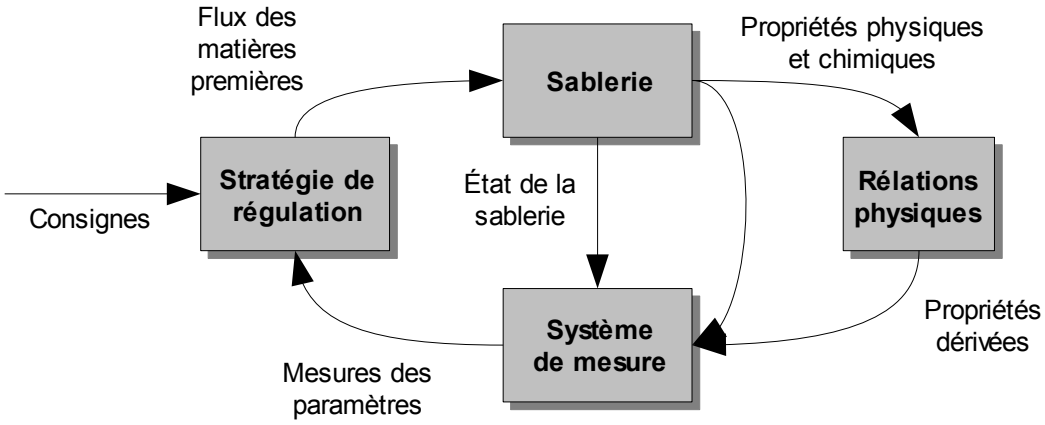
AA	teneur en argile active
CT	teneur en carbone
H	teneur en eau
Pe	teneur en silice
TDISA1	température du sable

   Je n'ai pas pu retenir les mesures de granulométrie en raison du fait qu'elles sont effectuées avec une cadence trop faible (une fois par mois) par rapport aux autres (toutes le 2-3 heures).
3. Les mesures d'autres propriétés du sable à vert qui dépendent des propriétés physiques et chimiques mentionnées au point précèdent, et que j'appelle propriétés « dérivées ». J'ai retenu les paramètres suivants :

CIS	résistance au cisaillement
K	aptitude au serrage
P	perméabilité

J'ai dû toutefois exclure les mesures de perte au feu (grâce auxquelles on peut estimer le teneur en matières volatiles) puisque elle ne sont pas effectuées régulièrement.

4. D'autres mesures de l'état du système de régénération, telles que le niveau du sable à vert dans les trémies de stockage et le rendement du malaxeur. Ces mesures peuvent influencer les propriétés du sable à vert. Le malaxeur est responsable, par exemple, de l'activation de l'argile et donc son rendement est lié au teneur en argile active. Toutefois il est mesuré avec une cadence trop faible (une fois par mois) et, par conséquent, je l'ai dû exclure de l'analyse. Le niveau du sable peut fournir des informations sur le délai de stockage avant utilisation, délai pendant lequel les propriétés du sable peuvent se dégrader. À cet effet il serait nécessaire d'analyser la série temporelle des observations et d'identifier un modèle dynamique en boucle fermée de la sablerie tel que montré dans le schéma suivant :



− Schéma de principe du système de régulation de la sablerie −

Ce genre d'étude dépassait toutefois le cadre des objectifs de mon stage. J'ai donc exclu aussi ce paramètre de l'analyse.

En me basant sur cette classification, j'ai procédé d'abord à l'identification d'un modèle permettant de prédire le taux de soufflure en fonction des paramètres dérivés CIS, K, et P, plus le paramètre physique TDISA1. Faire dépendre « directement » le taux de soufflure des paramètres chimiques, AA, CT, H et Pe, ne semblait pas correct. La liste suivante montre un aperçu des fonctions f_i sélectionnées[17] par la méthode MARS et les coefficients b_i estimées par la régression logistique :

```
(Intercept)                                  -4.4410206
h(TDISA1-37.3)                                0.4418897
h(TDISA1-44)                                  3.3728890
h(K-31) * h(TDISA1-44)                       -0.0617582
h(CIS-7.07) * h(TDISA1-37.3)                 -0.4565293
h(CIS-7.56) * h(TDISA1-37.3)                  0.2743029
h(104-P) * h(TDISA1-44)                      -1.1524882
h(K-35) * h(CIS-7.24) * h(TDISA1-44)         -7.3037333
h(K-40) * h(7.56-CIS) * h(TDISA1-37.3)       -0.5654309
...
```

Cette liste est non exhaustive puisque elle contient en réalité 38 termes jusqu'au degré d'interaction 4. Au delà du degré 4, les gains en précision du modèle sont relativement très modestes et contrebalancés par une augmentation sensible de la complexité du modèle

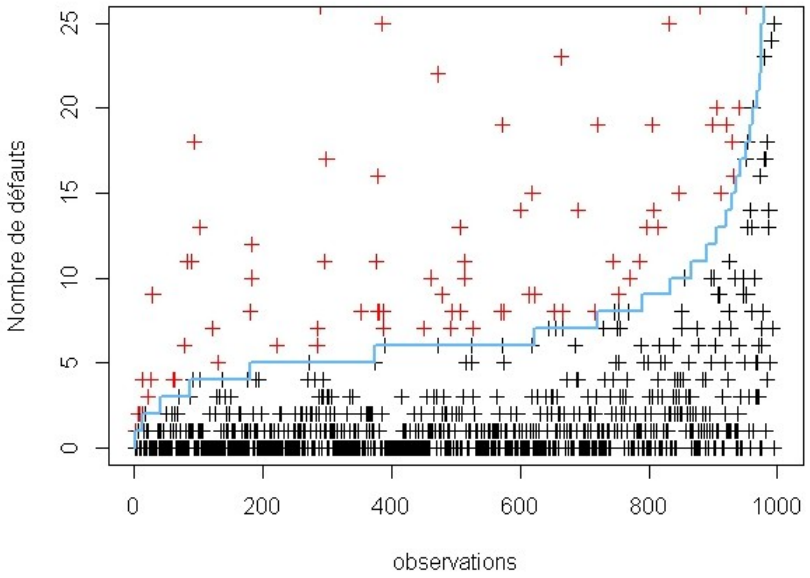
[17] MARS utilise la fonction h définie par : h(x)=0 si x<0 et h(x)=x si x≥0.

(nombre de coefficients à estimer trop élevé par rapport au nombre N d'observations). L'importance assignée aux paramètres est la suivante (en ordre décroissant) :

	nsubsets	gcv	rss
TDISA1	37	100.0	100.0
CIS	35	78.5	83.7
P	29	52.8	63.6
K	28	51.7	62.3

TDISA1 est le paramètre le plus significatif par rapport à tous les trois critères employés par la méthode ; K le moins significatif. Aucun paramètre peut être considéré négligeable.

Dans la figure suivante, le nombre des défauts constatés, $y_S(i)$, est comparé pour chaque observation à la limite supérieure $y_{S,max}(i)$ prédite par le modèle.

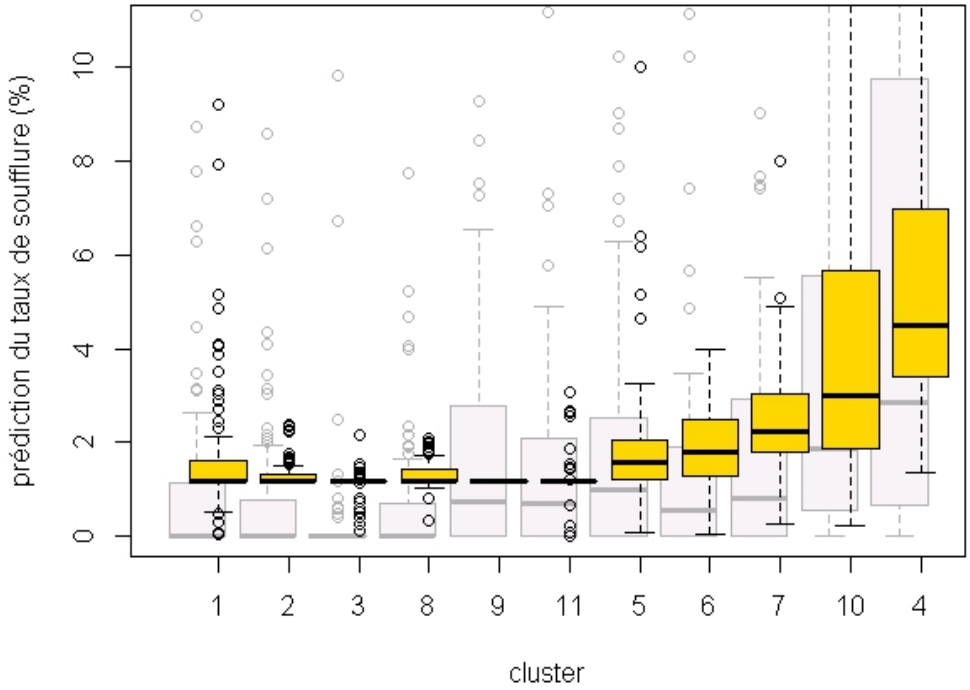


− *Nombre de défauts constaté dans chaque observation (croix) et limite supérieure prédite par le modèle (ligne bleue) en fonction des paramètres sablerie* −

Dans les différents tests de validation que j'ai effectué, le rapport $N_{correct}/N_{test}$ était égal approximativement à 90% et la moyenne des erreurs de prédiction égale à 1%. Ce modèle surestime donc systématiquement le nombre des défauts. La présence d'observations « extrêmes », représentées par les croix rouges, est la cause de ce biais. Ces taux élevés de soufflure ne semblent pas être imputables aux paramètres sablerie (puisque le biais reste approximativement constant à 1% pour toute valeur des paramètres), mais vraisemblablement à d'autres paramètres procédé ou bien à d'autres origines.

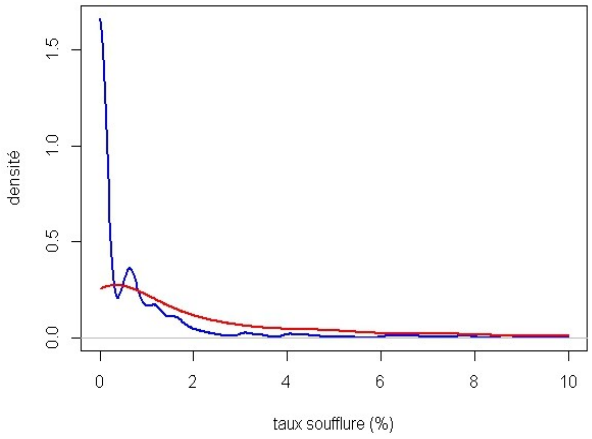
Afin de faciliter la tâche de planification, j'ai utilisé les cartes auto-adaptatives. La figure en haut de à la page suivante représente, avec un graphe « boites et moustaches »[18], une partition des observations en 11 classes.

[18] La « boîte » contient 50% des valeurs appartenant à la classe et comprises entre le premier et le troisième quartiles. Les quartiles sont calculés à partir de l'ensemble des valeurs de la classe. La valeur médiane est indiquée par la ligne épaisse au milieu de la boîte. La position des « moustaches » est calculée en fonction de la différence entre le troisième et premier quartiles. Les ronds représentent les valeurs extrêmes.



− Partition des observations : distribution des prédictions du taux de soufflure (jaune) et distribution des valeurs mesurées (gris clair) −

Par une comparaison visuelle ces classes, il est possible de conclure que les classes 1, 2, 3 et 8 sont celles où le taux de soufflure est vraisemblablement le plus bas (la médiane des taux constatés est nulle). Cette hypothèse peut être confirmé en séparant les mesures du taux de soufflure en deux populations : la première, P_1, regroupe les mesures appartenant aux classes 1, 2, 3 et 8 (439 mesures) ; la deuxième, P_2, celles appartenant aux autres classes (562 mesures). Les densités de probabilité[19] estimées de P_1 et de P_2 sont représentées dans la figure suivante.



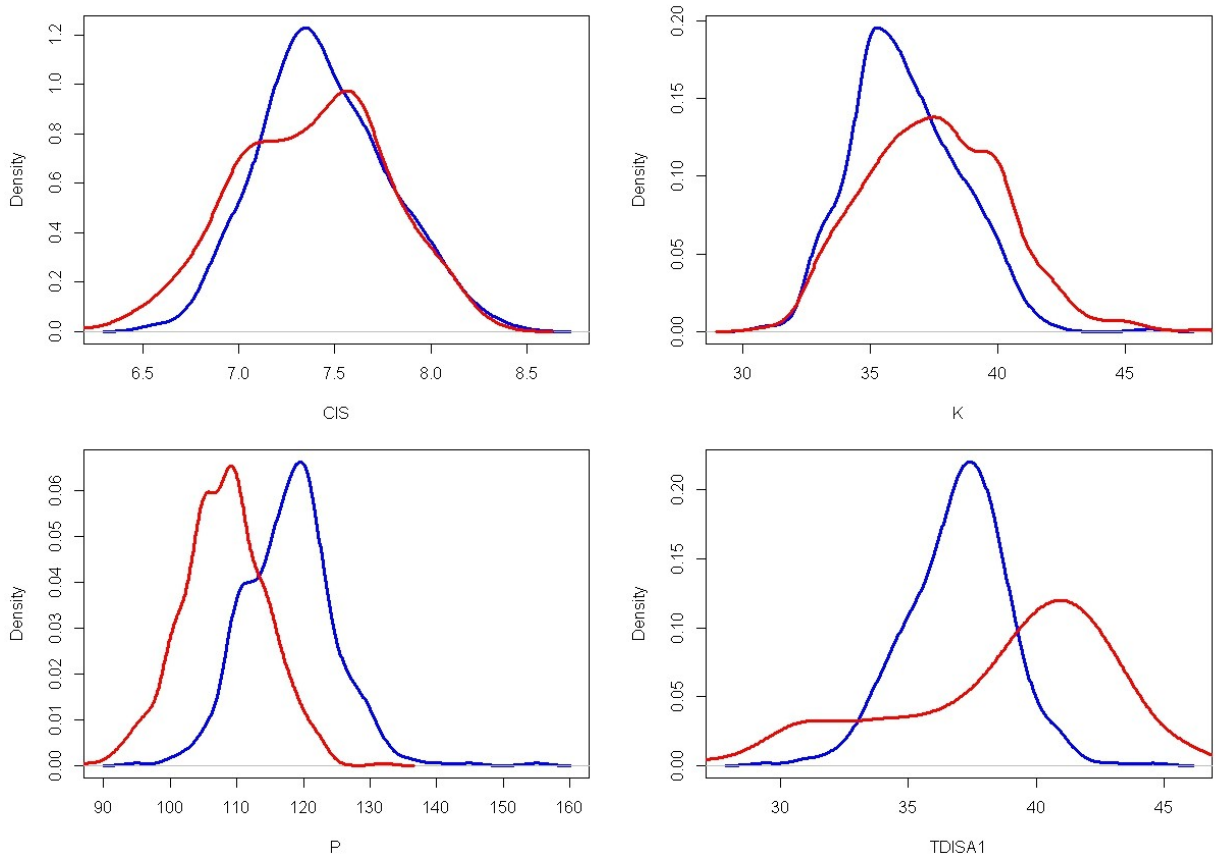
− Densités de probabilité du taux de soufflure : population P_1, classes de la partition présentant les taux plus bas (bleu), population P_2, classes présentant des taux plus élevés (rouge) −

Cette figure montre que les taux plus bas de soufflure sont plus probables dans les classes 1, 2, 3 et 8 (par exemple,75% des mesures sont inférieures à 0.7%) ; vice versa,

[19] Une densité de probabilité peut être assimilée grossièrement à un histogramme « lissé » d'un population de valeurs.

les taux plus élevés sont plus probables dans les autres classes (50% des mesures sont supérieures à 1%).
J'ai testé rigoureusement cette hypothèse en appliquant des tests statistiques opportuns, tels que le test des rangs signés [19]. Ces tests déterminent quantitativement le dégrée de confiance avec lequel ce type d'hypothèses peut être acceptée, mais je n'expliquerai pas ici ces procédures.

Le même genre de considérations peuvent être faites à propos des mesures des paramètres. Les quatre figures suivantes ont été générés en suivant la même procédure décrite pour le taux de soufflure.



− Densités de probabilité des paramètres : population P_1, classes de la partition présentant les taux plus bas (bleu), population P_2, classes présentant des taux plus élevés (rouge) −

Qualitativement, ces résultats montrent que les taux plus bas de soufflure (lignes bleues) s'obtiennent plus probablement lorsque simultanément : CIS n'est pas trop basse, K n'est pas trop élevée, P est élevée et TDISA1 est basse (relativement aux fourchettes des valeurs dans lesquelles ces paramètres varient). L'effet bénéfique d'une augmentation de P sur le taux de soufflure est évident, puisque la perméabilité favorise le passage des gaz à travers la motte. Des valeurs élevées de TDISA1 peuvent signaler une dégradation du rendement du malaxage du sable avec une diminution du teneur en argile active[20]. Des températures élevées du sable peuvent aussi favoriser le dépôt de vapeur d'eau sur la surface des noyaux, qui sont généralement plus froids que la motte. Dans les deux cas la

[20] L'argile active forme avec l'eau une structure en feuillets qui lui confère les propriétés de plasticité et cohésion bien connues. Cette liaison eau-argile est perdue irréversiblement à des températures supérieures à 450°C.

conséquence est une augmentation de la quantité de vapeur d'eau dégagée à la coulée[21]. En revanche, une explication claire des rôles de CIS et K sur l'apparition des soufflures est moins facile à trouver. Ces deux paramètres sont toutefois corrélés au teneur en argile active et au teneur en eau, comme montré dans la suite du paragraphe. Ils peuvent donc fournir indirectement des indications sur la quantité de vapeur d'eau dégagée à la coulée.

Sur la base de ces résultats, j'ai proposé alors la redéfinition des limites de conduite cibles, comme suit :

	Limites en vigueur	**Nouvelles limites**
CIS	6.8 – 7.5	7.3 –7.8
K	35 – 39	35 – 38
P	110 – 130	112 – 120
TDISA1	*pas définie*	35 – 38

La différence majeure concerne la résistance au cisaillement, CIS, qu'il faudrait augmenter sensiblement. En ce qui concerne la perméabilité, la fourchette en vigueur était déjà satisfaisante, mais les techniciens de FPF m'ont confirmé qu'il ne sont pas capable de stabiliser le procédé à des valeur supérieurs à 120, à cause vraisemblablement de la granulométrie du sable. En ce qui concerne la température, à FPF le sable est refroidi par une pré-humidification mais sa température n'est pas réglée finement en boucle fermée par un équipement dédié. Pour cette raison il n'y a pas une fourchette cible en vigueur pour le paramètre TDISA1.

J'ai ensuite déterminé, grâce aux réseaux bayésiens, quelle est la composition chimique du sable à respecter afin de garantir que les propriétés dérivés atteignent ces nouvelles limites. Le réseau est montré dans la figure à la page suivante. Ce modèle linéaire correspond au bloc appelé « Rélations physiques » dans le schéma de la sablerie montré précédemment. Les valeurs en vert représentent les moyennes des paramètres, les valeur en bleu les écarts type des termes d'erreur, et les valeurs en noir sur les arcs les coefficients définissant les relations linéaires entre les paramètres. Dans ce réseau TDISA1 n'apparaît pas. En effet, la température est mesurée directement à la sortie de la trémie de stockage. Par conséquent il ne me semblait pas correct de la corréler aux autres paramètres, dont les mesures sont effectuées en revanche sur un échantillon du sable transporté préalablement au laboratoire. Dans ce réseau j'ai imposé de nombreuses relations que j'ai trouvé documentées dans un rapport techniques [20]. Les coefficients estimés correspondent bien au courbes obtenues expérimentalement et mentionnées dans ce rapport. Deux relations (CT → K, AA → P) ont été identifiées par l'algorithme.

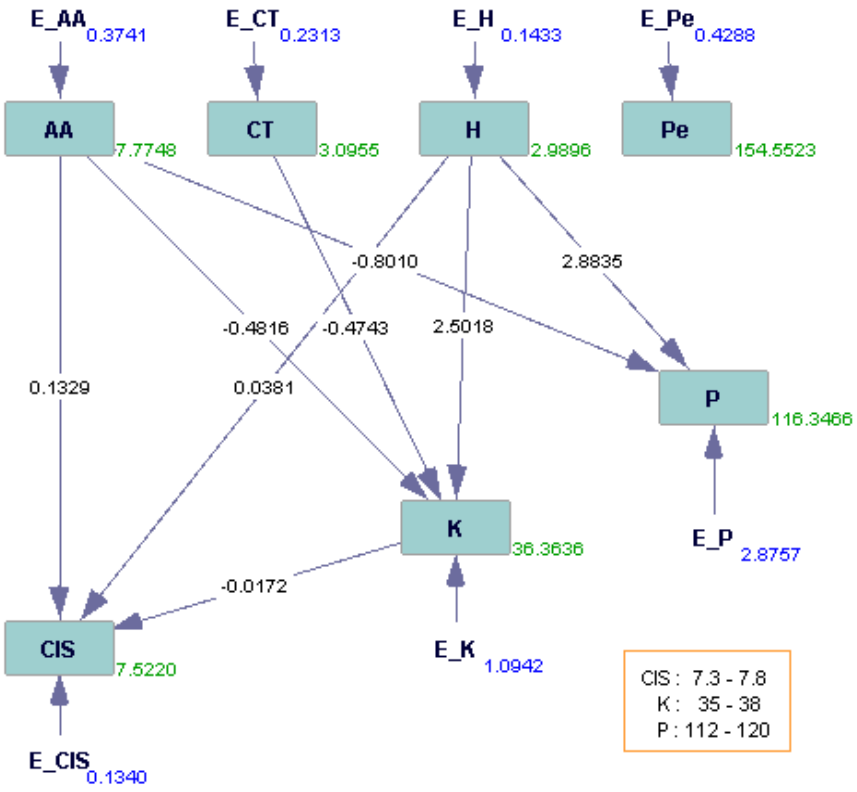
J'ai proposé la redéfinition des limites de conduite cible pour la composition chimique du sable comme suit :

	Fourchettes en vigueur	**Nouvelles fourchettes**
AA	7.3 – 7.7	7.4 – 8.15
CT	2.8 – 3.25	2.85 – 3.30
H	2.8 – 3.25	2.85 – 3.15
Pe	153 – 155	154 – 155

La différence majeure concerne le teneur en argile active, AA, qu'il faudrait augmenter sensiblement, soit en incrémentant le flux d'argile neuve entrant dans le circuit du sable

[21] La vapeur représente toutefois seulement une partie des matières volatiles, le reste provenant des additifs carbonées et des résidus de résines de noyautage.

(avec une augmentation des coûts en matières premières), soit en améliorant le malaxage qui est responsable de son activation (d'où la nécessité de diminuer TDISA1 par un système de refroidissement plus efficace).



− Réseau bayésien décrivant les relations entre les propriétés chimiques (AA, CT, H, Pe) et les propriétés dérivées (CIS, K, P) du sable à vert −

4 Résultats obtenus et perspectives

4.1 Analyses effectuées

Les chapitres précédents résument le travail que j'ai accompli pendant la période de mars à juillet 2012. J'ai dédié la majeure partie de ce temps à la compréhension du procédé, à l'apprentissage des techniques statistiques et de logiciels, et à la préparation des données. L'avantage de la méthodologie d'analyse que j'ai mis en place est le fait d'être suffisamment générale pour être appliquée indistinctement à tous les secteurs de la fonderie. Je considère abouties les deux analyses suivantes et à finaliser la troisième :

1. Sablerie : j'ai présenté dans le chapitre précédent les résultats de l'analyse effectuée sur les données de ce secteur.
2. Fusion (composition chimique) : l'objectif était de modéliser le taux de soufflure en fonction de la composition chimique de la fonte à la coulée. En résumé, l'analyse a montré que le taux de soufflure prédit par le modèle varie très faiblement en fonction de la composition chimique. Je n'ai pas pu justifier statistiquement la redéfinition des limites de conduites de ces paramètres. Ce résultat confirme peut-être ce que certains experts estiment, c'est-à-dire que la composition de la fonte peut être responsable de l'apparition du défaut de piqure (un autre type de défaut lié aux gaz), mais ne joue pas un rôle sur le défaut de soufflure.

3. Fusion (analyse thermique) : l'objectif est de modéliser le taux de soufflure en fonction de l'analyse thermique de la fonte, et d'identifier ensuite les relations entre la composition chimique et l'analyse thermique.

4.2 Analyses planifiées

En septembre, pendant le dernier mois du stage, je compte poursuivre les analyses concernant les autres secteurs :

1. Moulage : une procédure de prétraitement a été implémentée par le service informatique de FPF, grâce à laquelle je pourrai extraire de la base les données et les analyser ensuite rapidement. Les paramètres intéressants caractérisent certaines propriétés géométriques des moules, la cadence de production et la modalité de conduite de la machine à mouler (manuelle ou automatique).
2. Coulée : une procédure de prétraitement est nécessaire aussi pour les données issues de ce secteur. Je demanderai encore l'aide du service informatique pour la réaliser. Les paramètres intéressants caractérisent la dynamique du remplissage des moules.
3. Noyautage : l'audit que j'ai fait en début de stage m'a montré que le prétraitement des données de ce secteur présente des difficultés : la traçabilité de certains paramètres ne me semble pas être garantie avec une précision à la tranche horaire prés, et nombreuses informations sont encore stockées sur papier. Je devrai évaluer attentivement l'opportunité d'une analyse.

4.3 Améliorations possibles de la méthodologie d'analyse

Pour l'amélioration de la méthodologie, je préconise les actions suivantes :

1. Un seul modèle devrait être identifié, qui prend en considération simultanément les paramètres de tous les secteurs. De cette manière il serait possible de mettre en évidence des interactions entre des paramètres appartenant à des secteurs différents. Cela nécessiterait toutefois d'un volume de données plus important afin de pouvoir estimer un nombre plus large de coefficients inconnus.
2. Le modèle ne devrait pas prédire seulement le taux du défaut de soufflure mais aussi les taux d'autres types de défaut[22]. De cette manière, la planification des limites de conduite optimales du procédé pourrait être effectuée plus correctement. En effet, il n'est pas possible d'exclure des situations où certaines fourchettes sont optimales par rapport au défaut de soufflure, mais sous-optimales par rapport à d'autres défauts. Dans ces situations la détermination des limites de conduite serait le résultat d'un compromis. Ce genre de problématique de planification pourrait être formalisé en s'appuyant sur la théorie de l'optimisation multiobjectif.
3. En complément de la modélisation du procédé, il serait très intéressant de modéliser la variable coût. La modification des limites de conduite peut entrainer des coûts additionnels, par exemple à cause d'un besoin plus important en certaines matières premières. Il faudrait s'assurer que ces coûts ne dépassent pas les gains apportés par une réduction des taux de défaut.

4.4 Bilan et préconisations

Par rapport aux objectifs du stage, je peux dresser le bilan suivant :

[22] Cette généralisation de la régression logistique classique est dite « polytomique ». Des outils existent sous R pour effectuer ce genre d'analyses.

1. J'ai analysé 35% des paramètres procédé (sans inclure les analyses qui seront effectuées en septembre).
2. Les résultats obtenus ont pu être interprétés par les experts métier sur la base de leurs connaissances des mécanismes de formation des soufflures. Ces résultats ont donc fourni une confirmation quantitative de certaines hypothèses sur les causes de ce défaut.
3. Les résultats justifient sur des bases statistiques l'utilité d'une redéfinition des limites de conduite de certains paramètres procédé. Les responsables de FPF décideront s'il est pertinent de procéder à la validation expérimentale des nouvelles consignes. Mes connaissances actuelles du procédé ne me permettent pas d'évaluer ni la capacité de FPF à stabiliser le procédé dans les limites que j'ai proposé, ni l'impact que celles ci peuvent avoir sur d'autres types de défauts, ni les coûts engendrés.

Je rédigerai avant la fin du stage un rapport technique présentant en détail l'ensemble des analyses effectuées et les résultats obtenus.

L'utilisation des outils d'analyse statistique présentés dans le rapport peut être considéré comme une innovation à FPF. Afin d'améliorer dans le futur l'efficience et l'efficacité de ce genre d'études je préconise l'abandon du logiciel Excel et de tout support en papier, et la mise en place d'un système informatisé basée sur un logiciel de gestion de bases de données couplé à un logiciel d'analyse. Des solutions professionnelles et open source sont disponibles (déjà cités dans le rapport) si FPF souhaite réaliser ces études en interne au moindre coût. Toutefois, la redéfinition des consignes sur les limites de conduite ne résout pas entièrement le problème. Afin de respecter ces consignes, il est nécessaire de définir aussi une stratégie optimale de régulation du procédé. Les données pourraient être exploitées aussi pour la modélisation du procédé (je pense en particulier au secteur sablerie). Grâce à des logiciels de simulation, il serait possible de définir et d'évaluer des stratégies de régulation du procédé différentes de celles utilisées à ce jour, avant de les tester directement sur le procédé. Des logiciels professionnels et open source existent (par exemple l'environnement Scilab/Xcos).

5 Conclusion personnelle

Ce stage conclut une démarche personnelle de reprise d'études, que j'ai entrepris après plusieurs expériences professionnelles en tant que chercheur au sein de gros établissements de recherche, au début de ma carrière, et comme employé d'une société de conseil en matière de financement de l'innovation, dans les dernières années. J'avais à mon actif la participation à plusieurs projets de recherche et innovation. Je dois donc remercier d'abord le Directeur et les Responsables de FPF de m'avoir concédé une ample autonomie dans la gestion du projet et d'avoir fait confiance en mes compétences dès le début du stage.

Malgré mes expériences professionnelles passées, je sentais toutefois que le monde industriel m'était en un certain sens encore étranger. En tant que chercheur j'avais été impliqué dans des projets de recherche fondamentale à long terme, ce qui m'avait souvent permis de faire abstraction des problématiques plus concrètes de nature technique, organisationnelle et économique, liées aux systèmes de production industrielle. En tant que conseiller j'étais souvent en contact avec des entreprises de tailles et secteurs d'activité variés mais, vues de l'extérieur, je n'ai pu comprendre qu'une facette de la réalité quotidienne de ces entreprises. De ces points de vue le stage s'est avéré très enrichissant. Je veux citer en particulier deux expériences que j'ai pu faire.

La première. Le groupe Teksid, actionnaire unique de FPF, a décidé de remplacer dans ses fonderies les actuels systèmes de gestion par un système intégré basé sur le progiciel commercial SAP. En France ce système devra être opérationnel à partir de 2013. S'agissant d'un groupe italien, Teksid a fait appel à des consultants de IBM Italie pour assurer la mise en place de SAP et la formation du personnel concerné. Des ateliers de formation et discussion ont donc été organisés chez FPF en langue italienne, auxquels j'ai participé officiellement en tant qu'interprète. Cette activité, bien que n'étant pas corrélée directement avec ma mission, m'a néanmoins permit d'acquérir des connaissances à la fois sur l'organisation de la fonderie (en particulier sur les activités de maintenance, sur la gestion des magasins, des achats, de la production et sur la comptabilité) et sur le logiciel SAP, leader mondial sur le segment des progiciels de gestion. En outre, n'ayant pas des connaissances dans ces domaines, pour assurer dignement mon rôle d'interprète j'étais souvent obligé de participer activement aux discussions afin de bien comprendre les détails, et aussi, mais dans une moindre mesure, de jouer le rôle d'animateur.

La deuxième. La direction de FPF m'a invité à présenter mon travail lors de deux réunions qui se sont tenues chez Renault, l'un des deux clients de FPF. Le projet de stage s'inscrit, en effet, parmi les multiples démarches d'amélioration de la qualité des produits. J'ai donc fait partie de la délégation de personnes (le Directeur de FPF, les Responsables Industrialisation et Qualité, deux Chefs de Projet et un Assistant Qualité permanent chez le client) qui ont participé à ces réunions : la première, le 23 mai 2012, au Centre Technique Renault de Rueil en présence des Responsables Qualité/Achats et Métiers Fonderies de Renault ; la deuxième, le 30 mai 2012, à l'Usine d'Usinage et de Montage de Cléon en présence des Responsables Qualité Fournisseurs de Renault. Cette expérience m'a permis d'abord de tester mes capacités d'exposer d'une façon claire et concise un sujet technique en m'adaptant le plus possible aux compétences et aux attentes de l'auditoire. Mais surtout m'a donné la possibilité de découvrir comment se déroulent les relations client-fournisseurs dans un marché très concurrentiel comme celui de l'automobile.

Pour conclure, les connaissances techniques que ce stage m'a apporté et le « savoir être » que j'ai pu développer grâce aux contacts avec nombreux employés, me permettent de dresser un bilan plus que positif de mon stage à la Fonderie de Poitou Fonte.

6 Bibliographie

[**1**] C. Shearer, *The CRISP-DM model: the new blueprint for data mining*, Journal of Data Warehousing, no. 5, pp. 13-22, 2000
[**2**] R. Monroe, *Porosity in castings*, American Foundry Society Transactions, no. 05-245(04), 2005
[**3**] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009
[**4**] Y.K. Penya, P.G. Bringas, A. Zabala, *Advanced fault prediction in high-precision foundry production*, Proceedings of the 6th IEEE International Conference on Industrial Informatics, pp. 1673-1677, 2008
[**5**] Y.K. Penya, P.G. Bringas, A. Zabala, *Efficient failure-free foundry production*, Proceedings of the13th IEEE International Conference on Emerging Technologies and Factory Automation, pp. 237-240, 2008
[**6**] I. Santos, J. Nieves, Y.K. Penya, P.G. Bringas, *Optimising Machine-Learning-Based Fault Prediction in Foundry Production*, Lecture Notes in Computer Sciences, vol. 5518, pp. 53-56, 2009
[**7**] I. Santos, J. Nieves, Y K. Penya, P.G. Bringas, *Towards noise and error reduction on*

foundry data gathering processes, Proceedings of the International Symposium on Industrial Electronics, pp. 1765-1770, 2010
[**8**] I. Santos, J. Nieves, C. Laorden, B. Sanz, P.G. Bringas, *Collective Classification for the Prediction of Microshrinkages in Foundry Production*, International Journal of Computer Systems Science & Engineering, to be published, 2012
[**9**] J. Nieves, I. Santos, Y.K. Penya, S. Rojas, M. Salazar, P.G. Bringas, *Mechanical Properties Prediction in High-Precision Foundry Production*, Proceedings of the 7th IEEE International Conference on Industrial Informatics, pp. 31-36, 2009
[**10**] T. Kohonen, *Self-Organizing Maps*, Springer Verlag, 1995
[**11**] P. Spirtes, C. Glymour, R. Scheines, *Causation, Prediction, and Search*, MIT Press, 2000
[**12**] P.O. Hoyer, A. Hyvarinen, R. Scheines, P. Spirtes, J. Ramsey, G. Lacerda, S. Shimizu, *Causal discovery of linear acyclic models with arbitrary distributions*, Proceedings of the 24th Conference on Uncertainty in Articial Intelligence, , 2008
[**13**] P. McCullagh, J.A. Nelder, *Generalized Linear Models*, Chapman & Hall/CRC, 1989
[**14**] J. Friedman, *Multivariate adaptive regression splines*, Annals of Statistics, vol. 19, no. 1, pp. 1-141, 1991
[**15**] C. Stone, M. Hansen, C. Kooperberg, Y. Truong, *Polynomial splines and their tensor products*, Annals of Statistics, vol. 25, no. 4, pp. 1371-1470, 1997
[**16**] R Core Team, *R: A Language and Environment for Statistical Computing*, 2012, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org
[**17**] J. Adler, *R, L'essentiel*, Pearson Education France, 2011
[**18**] J.A. Landsheer, *The specification of causal models with TETRAD IV: A review*, Structural Equation Modeling, vol. 17, pp. 630-640, 2010
[**19**] J.Dickinson Gibbons, S. Chakraborti, *Nonparametric Statistical Inference*, Taylor & Francis, 2003
[**20**] Alain Colbaut, *Les sables de moulage à vert*, Formation spécifique A3F, Association de Formation Forge Fonderie, 2012

Résumé

Ce rapport est une synthèse des travaux que j'ai effectués pendant un stage de six mois à la Fonderie de Poitou Fonte (FPF). Cette usine produit plusieurs références de blocs cylindres à destination de l'industrie automobile. La problématique posée était d'analyser les données procédé afin d'identifier les causes d'un défaut récurrent sur les produits finis : le défaut de soufflure. Afin d'aboutir à une méthodologie d'analyse statistique qui répond le mieux possible au besoins de FPF, j'ai mis en place une gestion du projet basée sur une approche très répandu parmi les professionnels : le CRISP-DM. Cette approche préconise d'affiner dans le temps et de manière itérative la méthodologie d'analyse, en évaluant les résultats partiels sur la base d'une compréhension toujours meilleure du métier du client, de sa problématique et des données disponibles. La méthodologie que j'ai développé intègre quatre techniques différentes : les cartes auto-adaptatives, les réseaux bayésiens, la régression logistique et la régression multivariée par spline adaptative. Son application a permis de justifier la redéfinition des limites de conduite de certains paramètres procédé afin de réduire le taux de soufflure moyen constaté.

Abstract

This report summarize the work I done during a six months internship at the Fonderie de Poitou Fonte. This foundry produces several types of cylinder blocs for the automotive industry. The problem I was ask to solve was to analyze the available process data in order to identify the causes of a defect frequently found on the finished pieces : the blow holes. In order to develop a statistical data analysis methodology which best satisfies FPF needs, I followed a project management approach well known among the practitioners: the CRSIP-DM. This approach recommends to improve iteratively the analysis methodology by evaluating the partial results against better and better business and data understanding. The methodology I developed incorporates four statistical techniques: the self organizing maps, the bayesian networks, the logistic regression and the multivariate adaptive regression splines. Based on its application, the reference bounds on variation of some process parameters have been redefined in order reduce the mean defect rate.