INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN NORMANDIE

BARBOSA
Antoine
MS ESD

Thèse professionnelle pour l'obtention du mastère spécialisé « Expert en Science des Données »

Construction d'un modèle statistique pour l'optimisation d'un processus de production en milieu industriel

Aptar pharma

2017/2018

Table des matières

Introduction

La notion de performance dans le secteur industriel repose sur deux objectifs principaux : la maximisation de la productivité des machines, et la qualité optimale du produit. L'évaluation de la qualité d'un produit est faite en cours ou en sortie de ligne de production. Elle est définie par des procédures spécifiques de contrôle qualité, avec leur propre protocole, et qui permettent de valider ou non la conformité du produit en fonction des résultats obtenus.

C'est dans une démarche de développement qualité que s'inscrit Aptar Val-de-Reuil, site Aptar Pharma de production de joints en caoutchouc. Un joint est le résultat d'un processus de production incluant le traitement de matières premières et des procédés complexes de liaisons chimiques. A cet effet, et dans le but de satisfaire les exigences client, une importante procédure de contrôle qualité est appliquée sur chacune des étapes du processus.

Aptar Val-de-Reuil souhaite développer la performance qualité d'une des composantes de son processus de production : la ligne de vulcanisation. L'influence actuelle limitée des conducteurs de ligne sur cette phase, et les facteurs inconnus, propres à l'action chimique exercée sur la matière, sont des éléments difficilement contrôlables impactant la variabilité des résultats qualités. Ces conditions amènent à réfléchir, d'une part à la configuration optimale de la ligne, et d'autre part à l'impact des caractéristiques de la matière avant vulcanisation, afin d'obtenir des résultats qualités valides sur la durée.

Dans ce contexte, la thèse professionnelle tentera, par une démarche de modélisation statistique, et à partir de données brutes, de répondre à la question suivante : quels facteurs impactent significativement les résultats qualités mesurés en sortie de ligne de vulcanisation ? L'approche de modélisation devra notamment prendre en compte les liens d'influences connus entre les mesures qualités.

Il s'agira dans un premier temps de définir formellement la problématique métier à partir des connaissances du processus de production. Dans un second temps, la problématique métier sera transcrite en problématique statistique, et nous décrirons l'approche de modélisation choisie à cet effet. La dernière partie consistera à appliquer cette démarche à partir des données brutes, puis d'analyser et comparer les résultats obtenus.

Les environnements de programmation python et R ont été utilisés pour le nettoyage de données, et la construction du modèle.

I. Contexte industriel et enjeux métier

A. Activité du site Aptar Val-de-Reuil

Le site Aptar du Val-de-Reuil est un site de production d'Aptar Pharma, filiale d'AptarGroup©. Son activité consiste à produire 3 types de joints en caoutchouc : des joints de col, flottants et fonctionnels. Ces joints se distinguent par leurs dimensions, leur aspect, mais également leurs propriétés chimiques. Ils sont destinés à l'assemblage de dispositifs d'administration pharmaceutiques et cosmétiques tels que les sprays, ventolines, et sont un élément central d'une valve doseuse dans les différentes variétés de pompe. La fonctionnalité principale d'un joint consiste, de par son étanchéité, à sécuriser l'afflux du produit vers l'extérieur au moment de la pression exercée par l'utilisateur sur la pompe.

B. Description du « process » et protocole de production

La production d'un joint combine un mélange complexe de matières premières, un processus de transformation chimique spécifique, et une méthode de découpe. Ces 3 phases sont intégrées dans le processus de production du Val-de-Reuil, organisé de la manière suivante :

- **Etape mélange** : composée de deux mélangeurs interne et externe. L'objectif de cette étape est de mixer plusieurs composants (gommes de caoutchouc, poudres, etc.) afin d'atteindre la formation requise, caractérisée par des bandes d'élastomères crues obtenues en sortie.
- **Etape vulcanisation** : consiste en une ligne de vulcanisation (LVC), prenant en entrée le mélange cru pour le transformer en bandes de caoutchouc vulcanisées, aux propriétés mécaniques (élastiques) et d'épaisseurs appropriées.
- **Etape découpe** : phase de découpe des bandes vulcanisées via l'usage de presses, pour obtenir en sortie des joints aux dimensions appropriées.

Chaque étape du process fonctionne selon un programme prédéfini de production de lots. Un lot correspond à une série d'unités produites en continu sur une même machine, à partir d'un ou plusieurs lots issus de l'étape précédente du process. Une unité, ou « carton », prend plusieurs formes selon l'étape : une charge de matières premières (mélange), une bobine de bandes vulcanisées (vulcanisation), ou un filet de joints découpés (découpe). La configuration de production d'un lot (paramétrage des automates, etc.) dépend de la matière traitée (catégorie normalisée d'élastomère).

La gestion de la production d'un lot est soumise à un cadre précis de contrôle qualité. La qualité d'une production est définie par un ensemble d'opérations de contrôle effectués selon une méthodologie propre à la matière traitée. Un lot est validé et libéré si les résultats obtenus sur l'ensemble des opérations sont conformes. En atelier de production, il existe deux types de contrôle : les contrôles d'aspects (effectués à l'œil nu) et dimensionnels relatifs à l'allure extérieur du produit, et les contrôles physiques liés aux propriétés élastiques de la matière. En excluant l'aspect, ces contrôles consistent à mesurer différents indicateurs numériques, dont les valeurs obtenues doivent respecter des limites haute et basse de spécification. Dans le cas contraire, le carton est considéré « hors-spécification » et ne peut être validé. Ces valeurs seuil sont définies en accord avec le client, et dépendant également de la matière produite.

Les contrôles physiques, effectués en phases de mélange et vulcanisation, sont des attributs qualités critiques qui déterminent la performance globale d'un processus de production sur le long-terme, et sa capacité à maximiser sa productivité en limitant le taux de rebut.

C. Problématique métier

La vulcanisation est un mécanisme chimique complexe, dont le déroulement est conditionné par un paramétrage précis de la LVC. Par conséquent, c'est une étape qui génère de l'incertitude sur les résultats qualités obtenus, de par la connaissance actuelle relative du process, et le peu de contrôle exercé par le conducteur de ligne sur la production. Le comportement chimique de l'élastomère, au contact de la chaleur et de la pression de vulcanisation dépend également de l'historique de production à l'étape de mélange, caractérisée par les mesures rhéomètriques obtenues en sortie de mélangeur. Ces mesures sont les suivantes :

- **Ts2** (en minutes) : temps de fixation. C'est le temps disponible pour la mise en œuvre du mélange.
- **T90** (en minutes) : temps optimum de vulcanisation.
- **ML** (en lbf-in) : couple de force minimum atteint lors du test rhéomètrique.
- **MH** (en lbf-in) : couple de force maximum atteint après vulcanisation.

Ces indicateurs rhéomètriques définissent les caractéristiques de la future vulcanisation, et permettent d'anticiper la probable réaction chimique du mélange, une fois consommée en LVC.

L'enjeu métier de ce projet consiste à analyser l'impact des paramètres process et des caractéristiques qualités du mélange sur 4 mesures qualités observées en sortie de LVC :

- **Résistance à l'allongement** (en %) : pourcentage d'allongement atteint au moment de la rupture de la bande prélevée pour le contrôle.
- **Résistance à la rupture** (en megapascal – MPa) : force de traction nécessaire pour rompre la bande prélevée pour contrôle.
- **Module à 100%** (en MPa) : force de traction nécessaire pour allonger de 100% la bande prélevée pour contrôle.
- **Dureté** (en point shore A) : mesuré à l'aide d'un duromètre.

Ces indicateurs sont des mesures physiques et définissent les propriétés mécaniques (PMC) de la bande vulcanisée. Ils quantifient le niveau d'élasticité global atteint après vulcanisation, et sont essentiels pour la validation d'une production de bandes.

Il existe un lien d'influence entre ces PMC, qui réside dans la gestion de la ligne de la vulcanisation. Un changement provoqué sur certains paramètres process impacte simultanément les valeurs futures des PMC.

Les consignes de production et les spécifications des PMC diffèrent selon la matière. Dans notre cas, l'étude sera axée uniquement sur la matière « 403E ». Les données utilisées sont des données historisées portant sur 6 mois de production.

L'objectif de la thèse professionnelle sera donc d'apporter des éléments de réponse à la problématique métier suivante : **en intégrant l'ensemble des paramètres de production de la LVC, les indicateurs rhéomètriques obtenus sur le mélange consommé, et le temps de maturation de ce même mélange entre la sortie du mélangeur et son chargement sur LVC, quels critères impactent significativement les 4 propriétés mécaniques** ? **Précisément, en tenant compte des liens de corrélation connus entre ces 4 attributs qualités, existe-il une approche de modélisation permettant une prédiction simultanée des PMC** ?

II. Problématique statistique et démarche de modélisation

La formulation de la problématique statistique résulte d'une observation des données disponibles et de leurs caractéristiques. D'autre part, en considérant la problématique métier et ses enjeux, l'idée est de déterminer le type de modèle statistique pouvant répondre à la problématique.

A. Définition de la problématique statistique

1. Données disponibles

Tout d'abord, les 4 propriétés mécaniques mesurées sur chaque bobine sont des données observées. Il s'agit d'une série de statistiques descriptives, incluant la moyenne, l'écart-type, le minimum, le maximum, qui résument les résultats des contrôles effectués sur les éprouvettes de caoutchouc prélevées. Dans le cadre du projet, les moyennes des PMC seront les variables à prédire.

Les données process de 28 capteurs postés sur les zones d'influence des paramètres de la LVC sont observées, et caractérisent des températures, vitesses, niveaux de pression et des mesures de distance entre différents composants de la LVC (notamment l'écart des cylindres pour l'ajustement de l'épaisseur d'une bande). De même, les mesures rhéomètriques pour un lot de mélanges sont connues : valeurs discrètes pour le ts2 et t90, valeurs continues pour les mesures ML et MH.

Enfin, le temps de maturation du lot de mélange consommé, en tant qu'indicateur influent à ajouter au modèle, n'est pas une donnée observée. Elle devra être créée à partir des métadonnées disponibles sur le lot de mélange.

2. Problématique statistique

La connaissance générale des données permet de définir le cadre d'apprentissage statistique et la structure du modèle recherché. Les moyennes observées des propriétés mécaniques sont les variables dépendantes et sont de distribution continue. Il s'agit d'un modèle d'apprentissage supervisé pour une tâche de régression. Les données d'entrée sont caractérisées par vecteur de variables continues et discrètes (ts2, t90, temps de maturation), aux échelles de valeurs différentes. De plus, le principe de prédiction simultanée des PMC, énoncé dans la problématique métier, implique une démarche de modélisation multivariée, qui consiste à prédire une matrice de variables dépendantes $\boldsymbol{Y}$, à partir d'une matrices $\boldsymbol{X}$ de prédicteurs.

Le modèle statistique devra apporter des éléments de réponse à la problématique suivante : **quelles sont les variables explicatives agissant significativement sur la variabilité de la matrice $\boldsymbol{Y}$** ? En incluant cet objectif exploratoire, l'objectif est de **définir et construire un modèle multivarié pour la prédiction simultanée de 4 variables numériques présentant des signes de corrélation, à partir d'une matrice $\boldsymbol{X}$ de variables aléatoires**.

B. Démarche de modélisation adoptée

La méthode usuelle pour la prédiction simultanée de plusieurs variables dépendantes est la construction de modèles séparés, ajustés sur chaque $Y_j, j = 1 \dots q$. Cependant, cette approche ne tient pas compte du lien de corrélation entre les variables dépendantes, notamment pour mesurer l'effet des variables explicatives sur la variabilité globale de la matrice $\boldsymbol{Y}$.

Dans le but de répondre à la problématique statistique posée, une démarche reposant sur 2 types de modélisation a été adoptée. En premier lieu, l'idée est d'étudier le lien linéaire entre les deux ensembles de variables, par l'implémentation d'un modèle linéaire multivarié. Le modèle multivarié est basé sur l'apprentissage de q modèles de régression multiple, et une procédure inférentielle MANOVA pour la significativité des variables explicatives. Dans un second temps, 4 modèles random forest ont été construits, dans une volonté de comparaison des performances prédictives avec le modèle linéaire multivarié.

1. Modèle linéaire multivarié et MANOVA

La régression linéaire multivariée (« Multivariate Multiple Regression » - MMR) est une généralisation de la régression linéaire multiple (MLR) pour la prédiction d'un vecteur aléatoire de variables dépendantes $\boldsymbol{Y} = (Y_1 \dots Y_q), q > 1$, en fonction d'un vecteur aléatoire de variables explicatives $\boldsymbol{X} = (X_1 \dots X_p), p \geq 1$ (Fox & Weisberg, An R Companion to Applied Regression (Second Edition), 2011). La formulation matricielle du modèle est la suivante :

$$\boldsymbol{Y}_{n\times q} = \boldsymbol{X}_{n\times(p+1)}\ \boldsymbol{\beta}_{(p+1)\times q} + \boldsymbol{\epsilon}_{n\times q}$$

Où :

- $\boldsymbol{\beta}$ est la matrice des coefficients de régression, avec $\boldsymbol{\beta}_j$ le vecteur colonne des $(p+1)$ coefficients associés à la $j^{ème}$ variable dépendante, $j = 1 \dots q$.
- $\boldsymbol{\epsilon}$ est la matrice des erreurs non-observées et de même dimension que $\boldsymbol{Y}$.

Les hypothèses du modèle MMR sur la matrice $\boldsymbol{\epsilon}$ sont analogues à celles d'un modèle MLR : les erreurs sont indépendantes, telles que $\epsilon_i \perp \epsilon_{i'}, i \neq i'$ et $(i, i') \in [1 \dots n]$, et normalement distribuées telles que $\epsilon_i \sim N_q(\boldsymbol{0}, \boldsymbol{\Sigma})$, où $\boldsymbol{\Sigma}_{q\times q}$ est la matrice de variance-covariance des erreurs, constante pour tout i.

L'estimateur $\widehat{\boldsymbol{\beta}}$ de la matrice $\boldsymbol{\beta}$ est obtenu par une succession de modèles de régression ajustés sur chaque Y_j par méthode des moindres carrés ordinaires :

$$\begin{cases} \widehat{\boldsymbol{\beta}}^{ols} = (\hat{\beta}_1^{ols}, \dots, \hat{\beta}_q^{ols}) \\ \hat{\beta}_j^{ols} = (\boldsymbol{X^T X})\boldsymbol{X Y}_j \end{cases}$$

La construction des q modèles MLR implique également une estimation indépendante des résidus $\hat{\boldsymbol{\epsilon}}_{n\times q}$, où $\hat{\boldsymbol{\epsilon}}_j$ est la colonne des résidus obtenus après l'ajustement du modèle sur la $j^{ème}$ variable dépendante.

La présence de corrélation entre les variables dépendantes impacte $\widehat{\boldsymbol{\beta}}^{ols}$ et la matrice de variance-covariance des résidus $\widehat{\boldsymbol{\Sigma}}_{res}$ (Carey, 2013). Cette dernière affiche des valeurs de covariance $Cov(\hat{\epsilon}_j, \hat{\epsilon}_{j'})$ non-nulles. De plus certains coefficients estimés d'un même prédicteur sont corrélés si ce dernier influe simultanément sur les variables dépendantes. Ces liens doivent être pris en compte dans l'optique de tests inférentiels pour la significativité des variables explicatives sur le modèle multivarié.

Parallèlement à la procédure ANOVA (Analysis of Variance), utilisée dans le cas d'un MLR pour la comparaison de modèles emboités (test de Fisher), il existe un cadre inférentiel pour la comparaison de modèles multivariés, appelé MANOVA (Multivariate Analysis of Variance). La procédure

MANOVA est également fondée sur le principe de décomposition de la variance totale de $\boldsymbol{Y}$, mais dans un cadre matriciel, via un calcul des matrices des sommes des carrées et produits croisés (SSCP). Une matrice SSCP est symétrique. Son expression est utilisée pour l'estimation des matrices de variance-covariance et de corrélation (Maitra, 2012) :

$$\boldsymbol{SSCP_{tot}} = (\widehat{\boldsymbol{Y}}^{\boldsymbol{T}}\widehat{\boldsymbol{Y}} - n\bar{\boldsymbol{y}}\boldsymbol{y}) + \hat{\boldsymbol{\epsilon}}^{\boldsymbol{T}}\hat{\boldsymbol{\epsilon}}\text{ , où } \begin{cases}\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} \\ \widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{ols}\end{cases}$$
$$\boldsymbol{SSCP_{tot}} = \boldsymbol{SSCP_{reg}} + \boldsymbol{SSCP_{res}}$$

Où :

- $\boldsymbol{SSCP_{reg}}$ est la matrice caractérisant la variance totale de $\boldsymbol{Y}$ captée par le modèle multivariée.
- $\boldsymbol{SSCP_{res}}$ est la matrice résiduelle captant la variance totale de $\boldsymbol{Y}$ non-expliquée par le modèle.

Une procédure MANOVA est une généralisation du test de Fisher d'une ANOVA pour le cas d'une matrice $\boldsymbol{Y}$ multidimensionnelle. Elle permet en outre de répondre à la question suivante : existe-il un lien statistique entre variables dépendantes et variables indépendantes ?

Dans le cas multivarié, le test d'hypothèse pour la comparaison de deux modèles emboités peut-être littéralement formulé de la manière suivante (Fox, Friendly, & Weisberg, Hypothesis Tests for Multivariate Linear Models Using the R "car" package, 2013) :

$$\begin{cases} H_0 : \text{"le modèle restreint suffit-il à expliquer la variance totale de } \mathbf{Y} \text{ ?"} \\ H_1 : \text{"}\textit{le modèle complet explique davantage la variance totale de } \boldsymbol{Y} \text{ ?"} \end{cases}$$

Le modèle multivarié restreint sous H_0 intègre l'ensemble des variables explicatives, sauf la variable dont on souhaite tester l'influence, tandis qu'à H_1 est associé le modèle multivarié complet.

La démarche consiste ensuite à obtenir l'expression de la matrice $\boldsymbol{SSCP_{reg|H0}}$ sous H_0, puis de calculer la différence matricielle avec la matrice $\boldsymbol{SSCP_{reg}}$ du modèle complet. Le test de significativité est effectué via la statistique de test de Pillai-Bartlett ($T_{PB} \in [0,1]$), dont la loi de distribution est déterminée par une approximation de la loi de Fisher. Une forte augmentation de T_{PB} implique un rejet de H_0 au niveau α.

Le modèle multivarié est fondé sur la construction parallèle régressions linéaires multiples, et sur une procédure de tests multivariés pour la comparaison et l'optimisation de ces modèles en gardant les variables d'influences significatives sur $\boldsymbol{Y}$.

2. Random Forest appliqué à la régression

Un random forest (RF) est une classe de méthodes ensemblistes appliquée à des tâches de classification et de régression. Il consiste en un ensemble d'arbres de décisions entrainés sur des échantillons bootstrap tirés de l'ensemble d'apprentissage. La prédiction $\hat{y}$ est obtenue par agrégation des prédictions individuelles des arbres (moyenne des prédictions dans le cas de la régression). Un modèle random forest combine donc un ensemble de modèles instables afin d'améliorer les performances de généralisation, et limiter le risque de sur-apprentissage propre aux arbres de décision.

La spécificité d'un modèle random forest réside dans la manière dont sont construits les arbres de décision en phase d'apprentissage. Tandis que pour un arbre, le meilleur split est défini sur l'ensemble des variables explicatives, un RF sélectionne aléatoirement un sous-ensemble de variables, et le split

optimal est achevé à partir de ce sous-ensemble. Cette méthode ajoute un caractère aléatoire au processus d'apprentissage des arbres de décision, et réduit la variance des prédictions. En régression le meilleur split est obtenu par minimisation de l'erreur moyenne quadratique (MSE).

Un modèle RF comporte trois hyperparamètres principaux, à déterminer par cross-validation : le nombre de variables candidates à sélectionner pour la recherche du split optimal, la profondeur de l'arbre (nombre maximal de splits), et le nombre d'arbres de décisions.

Par ailleurs, un modèle RF fournit, à l'issue de l'étape d'apprentissage, un classement des variables explicatives selon leur importance, caractérisée par une mesure d'augmentation du MSE sur les données out-of-bags (OOB). Dans le cas d'une régression, une variable explicative est classée en fonction de son impact sur la croissance du MSE, avant et après permutation de ses valeurs, à partir des données OOB, et sur tous les arbres.

Les caractéristiques ensemblistes d'un random forest sont une alternative à un modèle de régression linéaire standard. Il s'agit d'un agrégat de fonctions par morceaux permettant d'approximer une relation statistique plus complexe entre $\boldsymbol{X}$ et $\boldsymbol{Y}_j, j = 1..q$. De plus, ses résultats exploratoires par le biais de l'indicateur mean decrease impurity, donnent des indications sur les variables explicatives les plus influentes.

III. Construction du modèle vulcanisation

A. Nettoyage de données

La définition de la problématique statistique apporte des informations sur le format et la structure du jeu de données final souhaité pour l'apprentissage du modèle. A partir de fichiers de données brutes, l'objectif de cette étape de nettoyage de données est de construire l'ensemble $D = \{(x_i, y_i) \in R^{n \times p} \times R^{n \times q}\}_{i=1\ldots n}$, où n est le nombre de bobines de matière 403E produites entre Janvier et Juin 2017, p est le nombre de prédicteurs, et $q = 4$ le nombre de variables dépendantes.

1. Manipulation de tables de données brutes

Les données proviennent de différents secteurs opérationnels (production, qualité). Elles sont stockées et gérées par plusieurs systèmes d'information, à partir desquels des démarches d'extractions de données sont effectuées. Dans notre cas, les données brutes sont restituées sous la forme de fichiers excel (fichiers .csv, classeurs). Il s'agit de table brutes issus de base données relationnelles, et de granularités différentes. Nous présentons brièvement les tables disponibles stockant les données d'intérêts pour D :

- Table de suivi de production : contient les métadonnées relatives à un carton de mélange, vulcanisation ou découpe. Elle fournit les différents identifiants des lots et cartons, utilisables comme clés de jointures avec d'autres tables.
- Tables des résultats qualités (mélange et vulcanisation) : tables de granularités différentes, conservant les résultats de toutes les opérations qualités effectuées, ainsi que les métadonnées de ces opérations. Par exemple, la table des valeurs individuelles donne les résultats des contrôles sur les échantillons d'un carton, la table des caractéristiques, les statistiques descriptives calculés à partir des échantillons.

- Tables de données process : disposant d'une table par capteur, celle-ci comporte deux champs : l'horodatage (heure unix), et la valeur numérique du paramètre. Le nombre de lignes de ces tables varient en fonction de la fréquence d'enregistrement des valeurs.

17-juin-17 16:52:05	19,39999962
17-juin-17 16:52:10	22,70000076
17-juin-17 16:52:15	23,70000076
17-juin-17 16:52:40	20,5

Exemple : données process d'un capteur

- Table des résultats rhéomètriques : donne les 4 indicateurs rhéomètriques mesurés à l'aide d'un rhéomètre sur les charges d'un lot de mélange, et l'identifiant du lot.

Les tables de données à disposition présentent des caractéristiques différentes en termes d'identification de lignes, de types de données et de dimensions. Une première tâche de nettoyage de données consistait à manipuler et agréger ces tables par des procédures de base d'algèbre relationnel (jointures, concaténations, etc.), en gardant comme clé de repère l'identifiant des bobines produites entre Janvier et Juin 2017.

2. Création de la variable « temps de maturation » et décomposition (moyenne, écart-type, minimum, maximum).

Parallèlement à cette phase de traitement des tables, une stratégie de création de variables a été définie pour la matrice $\boldsymbol{X}$, en fonction des données brutes à disposition. Dans un premier temps, le format des données process et rhéomètriques ne permettent pas d'obtenir une ligne de données pour chaque bobine.

Il s'agit d'abord de définir l'historique de production d'une bobine, à partir des données horodatées pour chaque capteur.

Le temps de production d'une bobine est compris entre une 1h et 1h30, selon les évènements pouvant ralentir la production et la longueur exigée de la bobine. De plus, la cadence de ligne (en mètres/h) dépend de la matière produite. Elle est paramétrée selon une consigne spécifique renseignée dans les tables de données. En utilisant l'heure de début de production d'un lot de vulcanisation et la cadence standard de la ligne, on obtient un horaire présumé de fin de production. Cette méthode fut appliquée pour chaque table de données process, de manière à obtenir une séquence de valeurs par capteur. Ces séries de données définissent le suivi de production d'une bobine. Pour décrire numériquement ce suivi, chaque séquence sera résumée par un vecteur de statistiques descriptives : moyenne, écart-type, minimum, maximum. Ce choix suit l'idée de représenter statistiquement la variabilité du paramètre process pendant le temps de production.

Ainsi, pour un paramètre donné, cette variabilité sera illustrée par le vecteur [« Param_avg », « Param_stdev », « Param_min », « Param_max].

La même approche sera utilisée pour les indicateurs rhéomètriques d'un lot mélange. Ceux-ci étant mesurés sur chaque charge d'un lot. Le résultat qualité global du lot sera donc décrit par le vecteur [« rheo_avg », « rheo_stdev », « rheo_min », « rheo_max »].

Par ailleurs, le temps de maturation d'un lot de mélange n'est pas une donnée disponible. Cette variable (mesurée en jours) a été créée en calculant le delta entre deux dates clés : la date de fin de production du lot (avant sa mise en magasin), et celle de chargement dans la ligne LVC. En phase de vulcanisation, un seul lot de mélange doit être consommé pour la production d'un lot de bandes vulcanisées, mais ce lot est consommable pour plusieurs lots de vulcanisation.

3. Spécificité des tables process et ajout de données

Les tables de données process sont de dimensions variables. Le nombre d'enregistrements pour une période de 6 mois peut aller du millier pour un capteur, jusqu'à 5 millions pour un autre capteur. Cette différence s'explique par la fréquence de stockage des valeurs, différente selon le capteur.

Pour un paramètre process affichant une valeur constante sur une longue période, seule la première valeur sera enregistrée et stockée sur la base de données, jusqu'au prochain changement effectué sur le paramètre. Certaines périodes présentent donc des trous dans les tables de données process. Ce type de stockage de données rend impossible le calcul des indicateurs statistiques pour certaines capteurs et bobines produites.

La méthode adoptée pour remédier à ce problème consiste à combler les vides par l'ajout fréquentiel (toute les 5 secondes) de valeurs égales à celle enregistrée lors de la dernière modification du paramètre.

maturation_time	ML_mean	ML_std	ML_min	ML_max	MH_mean	MH_std	MH_min	MH_max	TS2_mean	TS2_std	TS2_min	TS2_max	T90_mean	T90_std	T90_min	T90_max	M_Caland_B
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3255215
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3255215
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3255215
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3231102
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3231102
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3231102
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3233294
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3233294
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3233294
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3255215
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3255215
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3255215
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3255215
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3255215
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3255215
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3255215
26	1,8427451	0,0683612	1,68	2,07	25,185373	0,4601438	23,67	26,96	0,4635686	0,0283848	0,41	0,51	3,1743529	0,026969	3,11	3,28	0,3255215

Exemple : temps de maturation et variables
rhéomètriques du jeu de données final *D*

Les différents points évoqués dans cette section constituaient les principaux défis de cette phase de nettoyage de données. A l'issue de cette étape, nous disposons d'un jeu de données D comportant $n = 1698$ observations, $p = 129$ variables explicatives et $q = 4$ variables dépendantes.

M_Extrud_PressureBeforeFilter_max	M_Extrud_ScrewSpeed_mean	M_Extrud_ScrewSpeed_stdev
109,709	17,17712984	0,06487201
109,709	17,17712984	0,06487201
109,709	17,17718602	0,064835924
110,195	17,17720361	0,060606703
110,195	17,17714433	0,060563822
110,195	17,17735907	0,060389457
112,196	17,173005	0,063422886
112,196	17,173005	0,063422886
112,196	17,17303125	0,063161878
112,103	17,1757212	0,0643847
112,103	17,1757212	0,0643847

Exemple : variables statistiques pour deux
paramètres process issues du jeu de
données final *D*

B. Statistiques descriptives et compréhension des données

Une étape d'analyse exploratoire des données, à travers l'étude des statistiques descriptives, doit nous permettre de comprendre et cibler les caractéristiques du jeu de données D. Par ailleurs, l'objectif de cette section est de révéler d'éventuelles propriétés statistiques chez certaines variables, en vue des modèles de régression sélectionnés et du respect de leurs hypothèses ou conditions optimales d'implémentation.

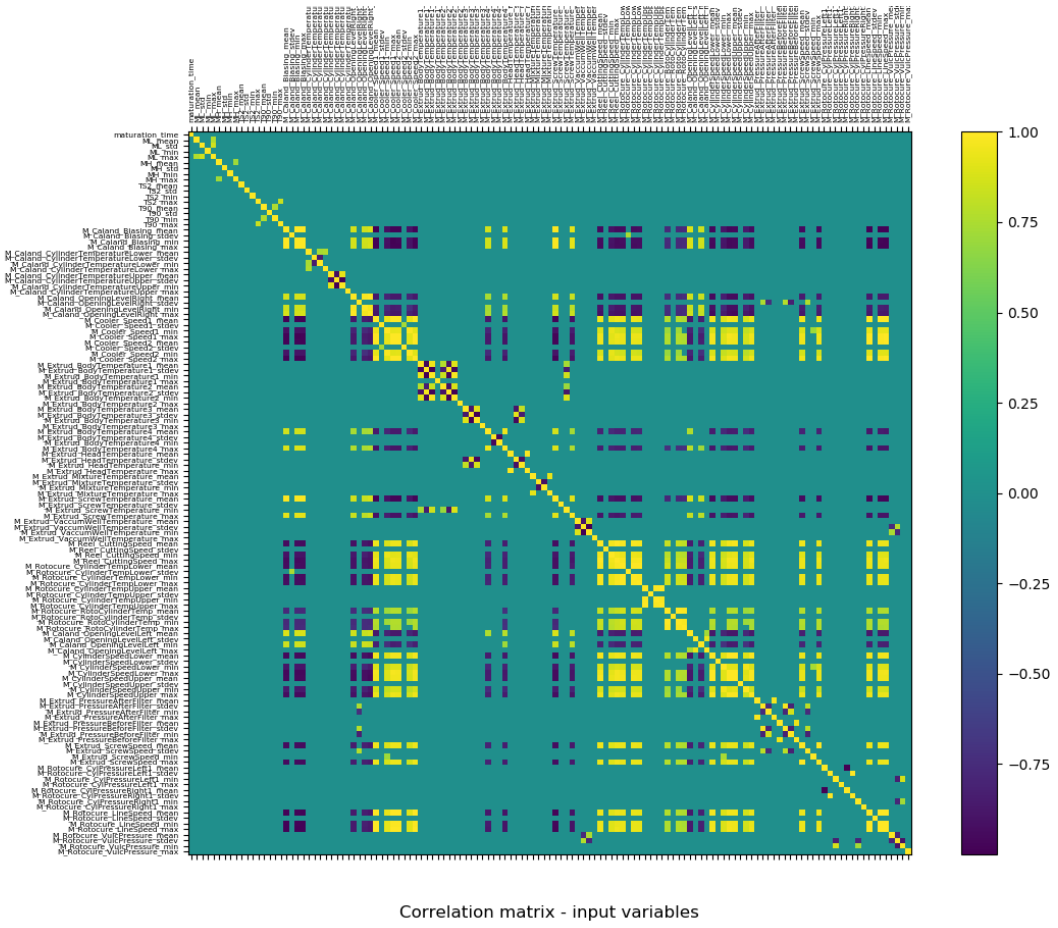
Un modèle de régression linéaire est altéré par la présence de multicollinéarité entre les prédicteurs, et se caractérise notamment par une erreur standard élevée des coefficients (associés aux prédicteurs

corrélés). Un modèle linéaire ajusté sur des variables corrélées est caractérisé par une faible performance de généralisation, et des coefficients difficilement interprétables.

Le classement des variables importantes propre au modèle random forest est également influencé par la multicollinéarité. Ce facteur est dû au choix arbitraire du split optimal lorsque plusieurs variables explicatives corrélées sont sélectionnées comme variables candidates.

1. Evaluation de la multicollinéarité et filtrage des variables

La démarche sera donc dans un premier temps d'étudier la matrice de corrélation (coefficient de pearson) associée à $\boldsymbol{X}$.

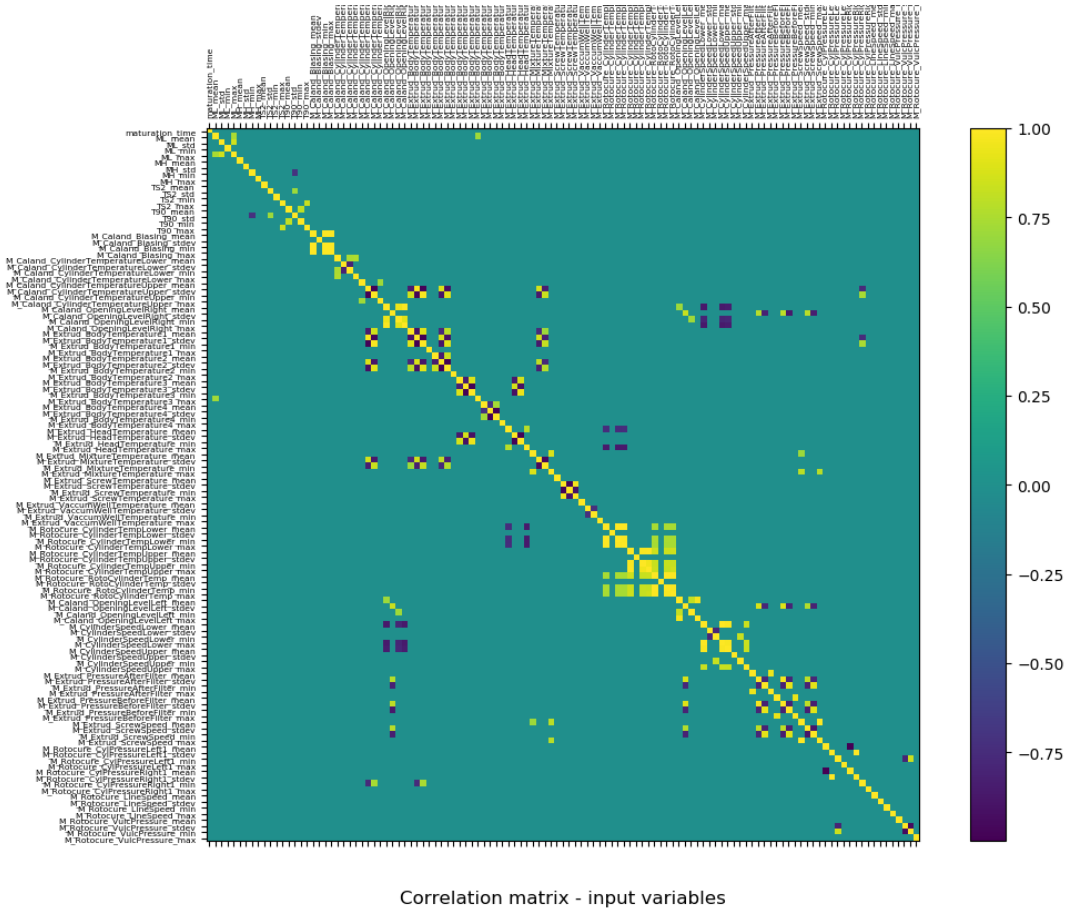


Graphique : matrice des corrélations calculées sur X

On constate la présence d'une corrélation groupée entre deux ensembles de variables : celles associées au positionnement des deux cylindres de la calandre (Caland_Biasing, Caland_OpeningLevelRight, M_Caland_Opening_LevelLeft), et celles associées aux paramètres de vitesse de la ligne de vulcanisation (vitesse des cylindres de la calandre, au niveau des refroidisseurs, de la machine de réfonte, vitesse des cylindres de la rotocure, vitesse de vis de l'extrudeuse). En effet, l'écart (en mm) entre les deux cylindres sur le côté droit de la calandre est corrélée positivement avec le biais cylindrique (en mm) et l'écart à gauche. Ce même paramètre est, à l'inverse corrélé négativement avec les paramètres de vitesse de ligne.

Cette relation statistique trouve son origine dans l'épaisseur de bande souhaité pour un certain lot (qui est également une mesure qualité contrôlée). La configuration de la LVC pour ces paramètres change entre des bandes d'épaisseur 1 mm, et 1.5 mm. Une bande fine implique un rapprochement des

cylindres de la calandre, et une augmentation de la vitesse globale de la ligne, inversement pour une bande plus épaisse de 1.5 mm.



Graphique : matrice des corrélations calculées sur X sans les bobines de 1.5mm et les 3 paramètres non-influents.

Ce constat fait, les bobines d'épaisseurs 1.5mm ont par la suite été retirée de l'analyse, et pour la suite de cette section. Le focus sera mis sur les bobines de 1 mm, qui constituent la majorité de la production en matière 403E. De plus, 3 paramètres process, liés à la vitesse de ligne dans les zones des refroidisseurs et de la machine de refonte de la bande caoutchouc, ont aussi été supprimés de l'étude. Ce choix a été fait par les experts métiers au court d'une réunion de présentation du projet, et des contours du modèle à construire.

L'exclusion des bobines d'épaisseur 1.5 mm, ainsi que des trois paramètres d'importance négligeable sur les PMC, a résolu en partie la forte corrélation observée dans un premier temps. Il reste néanmoins plusieurs groupes de variables corrélées, correspondant majoritairement à des indicateurs statistiques d'un même paramètre process. Ce constat s'applique aux paramètres de l'extrudeuse, et de température des cylindres de la rotocure.

Pour supprimer l'effet de ces groupes corrélés, une procédure de sélection de variable par calcul récursif du coefficient VIF (« Variance Inflation Factor ») a été utilisée. Cette mesure de multicollinéarité est calculée dans le cadre d'une régression linéaire, et repose sur l'ajustement d'un modèle de régression, par méthode des moindres carrés ordinaires, sur chaque prédicteur en fonction des $(p-1)$ variables restantes. L'idée de cette récursive est de prendre en compte les fluctuations des VIF comme conséquence de l'exclusion de la variable associée au VIF le plus élevé. La procédure est

répétée jusqu'à ce que le VIF maximal soit inférieur à un seuil spécifié en amont. Dans notre cas, un seuil $VIF = 20$ a été choisi, et 48 variables ont été exclues à la suite de cette routine.

Le nouveau jeu de données obtenu est de dimension $(n = 1260, p = 69, q = 4)$.

2. Distributions des variables dépendantes

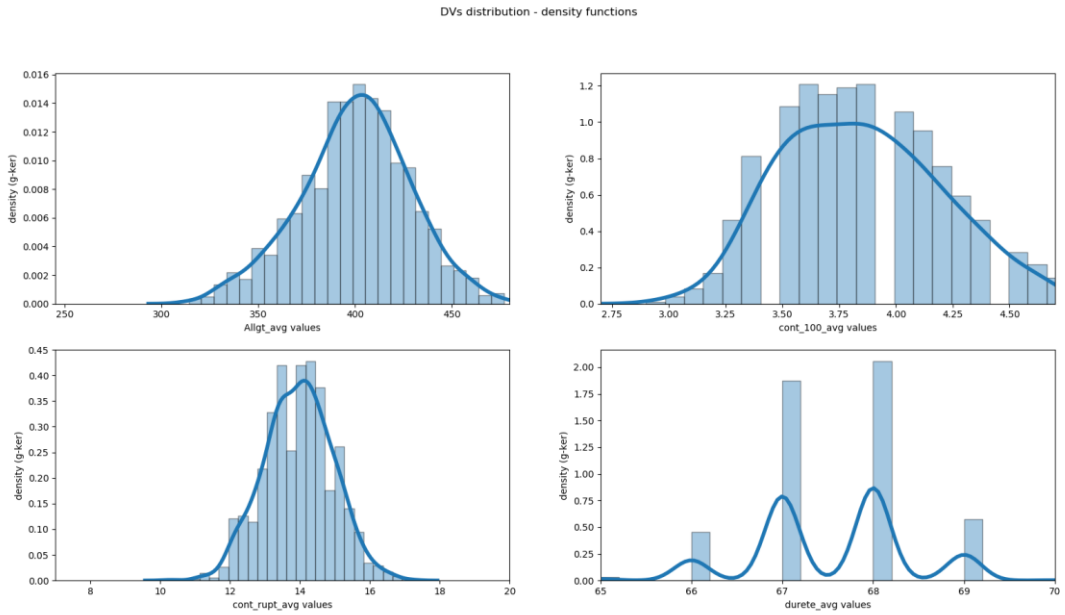
A la lecture des statistiques descriptives et des histogrammes combinés aux densités estimées (noyau gaussien), les distributions des variables dépendantes « Allgt_avg » et « cont_rupt_avg » semblent se rapprocher d'une loi normale, centrées sur leur moyenne respective (400 % et 13.96 MpA).

	Moyenne	Ecart-type	Minimum	Percentile 25%	Percentile 50%	Percentile 75%	Maximum
Allgt_avg (%)	400	28.5	314	382	402	419	477
Cont_100_avg (MPa)	3.85	0.35	2.9	3.6	3.8	4.1	5
Cont_rupt_avg (MPa)	13.96	0.98	10	13.3	14	14.6	17.2
Durete_avg (Point shore A)	68	0.84	65	67	68	68	70

Table : statistiques descriptives des 4 variables dépendantes

Parallèlement, les tests de shapiro-wilk effectués sur ces deux variables ont abouti à un rejet de l'hypothèse nulle de normalité pour l'allongement moyen, et un non-rejet pour la contrainte à la rupture moyenne, au niveau $\alpha = 5\%$. Ces deux variables présentent une répartition relativement étalée de leurs valeurs sur leur intervalle de spécification.

Les variables « cont_100_avg » et « durete_avg » affichent des distributions spécifiques. Avec un écart-type plus faible que la contrainte à la rupture (même unité de mesure), le module à 100% possède une distribution plus répartie autour de sa moyenne, et vers ses limites de spécification. La dureté moyenne affiche une faible amplitude de valeurs, avec un couple minimum/maximum respectif à 65 psA et 70 psA. De plus, 50% des observations prennent les valeurs 67 psA et 68 psA.



Graphique : histogramme et densité estimée par noyau gaussien des 4 variables dépendantes – les limites des abscisses correspondent aux spécifications respectives.

Les distributions des variables « Allgt_avg » et « cont_rupt » se distinguent donc des autres variables dépendantes. Cette caractéristique s'explique notamment par l'étendue des spécifications pour chaque propriété mécanique. Les spécifications min / max plus souples pour l'allongement et la contrainte à la rupture impliquent une marge d'erreur plus élevée par rapport à la valeur cible. A l'inverse, les spécifications très réduites pour la dureté ont pour conséquence des mesures majoritairement centrées sur quelques valeurs seulement.

3. Niveau de corrélation entre les 4 PMC.

Le lien d'influence entre les 4 PMC est un enjeu important dans l'approche de modélisation choisie. Il s'agit de vérifier statistiquement cette dépendance par la matrice de corrélation (coefficient de pearson).

	Allgt_avg	Cont_100_avg	Cont_rupt_avg	Durete_avg
Allgt_avg	**1**	-0.58	0.4	-0.49
Cont_100_avg		**1**	0.3	0.48
Cont_rupt_avg			**1**	-0.01
Durete_avg				**1**

Table : matrice de corrélation calculée sur Y

L'analyse de la matrice met en lumière trois liens pour lesquels $r \geq 0.5$ et $r \leq -0.5$.

L'allongement semble corrélé négativement avec le module à 100%, et dans une moindre mesure avec la dureté. D'autre part le module à 100% est faiblement corrélé avec la dureté.

Dans un premier temps, cette étape de compréhension des données a permis de distinguer les conditions de production entre deux types de bobines de matière 403E, pour privilégier par la suite les bobines d'épaisseur 1 mm. Un algorithme récursif de filtrage des variables explicatives, basé sur le coefficient de multicollinéarité VIF a ensuite été appliqué. Dans un second temps, l'étude des statistiques descriptives sur $\boldsymbol{Y}$ a mis en lumière la faible répartition des données pour les variables « durete_avg » et « cont_100_avg ». Les valeurs obtenues sur la matrice des corrélations nous permettent d'envisager la prise en compte des liens entre PMC pour la construction des modèles.

C. Régression linéaire multivariée

L'implémentation du modèle linéaire multivarié (MMR) et de la procédure MANOVA associée s'est faite sur logiciel R via le package de statistiques inférentielles *{car}* et la fonction de R base *lm( )* pour la construction d'un modèle linéaire.

On rappelle que le jeu de données obtenu à l'issue de l'étape de pre-processing consiste en un ensemble $D = \{(x_i, y_i) \in R^{n \times p} \times R^{n \times q}\}_{i=1 \dots n}, n = 1260, p = 69, q = 4$.

Celui-ci est découpé en un couple (D_{app}, D_{test}) selon un rapport 80/20.

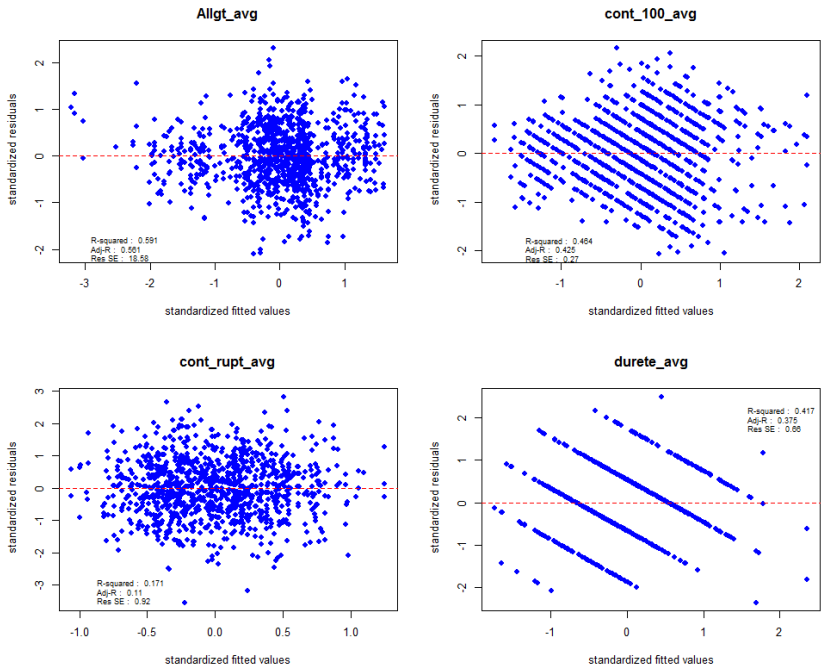
Les ensembles (D_{app}, D_{test}) sont dans un premier temps standardisés à l'aide de la fonction *preProcess( )* du package *{caret}* permettant notamment de réutiliser sur D_{test} les paramètres de standardisation calculés sur D_{app}.

Un modèle linéaire multivarié est ensuite ajusté sur D_{app} via la fonction *lm( )*, tout en spécifiant le caractère multidimensionnel de Y_{app} (fonction *cbind( )*).

```
mmr.1 <- lm(cbind(Allgt_avg, cont_100_avg, cont_rupt_avg, durete_avg) ~ ., data = Dapp_stdz)
```

Code R : apprentissage du modèle multivarié

La solution $\hat{\beta}^{ols}$ d'un modèle multivarié est la concaténation des vecteurs de coefficients $\hat{\beta}_j^{ols}, j = 1 \dots q$, obtenus par ajustement d'une régression linéaire multiple (MLR) sur chaque variable dépendante. La validité du modèle MMR dépend dans un premier temps de la qualité d'ajustement des q modèles MLR.



Graphiques R : résidus contre valeurs ajustées pour les 4 MLR

En inspectant le graphique des résidus contre les valeurs ajustées, le modèle 'Allgt_avg' semble être le plus robuste en terme d'ajustement aux données ($R^2_{adj} = 0.561$) et de respect des hypothèses sur les erreurs (compromis entre l'indépendance des erreurs et la variance constante). Le modèle MLR ajusté sur la variable 'cont_rupt_avg' affiche également un nuage de points relativement dispersé autour de la valeur résiduelle nulle, mais affiche un faible coefficient de détermination ajusté ($R^2_{adj} = 0.171$). Les modèles 'durete_avg' et 'cont_100_avg' sont caractérisés par des résidus corrélés négativement avec les valeurs ajustées, indiquant la présence d'hétéroscédasticité, et affichent des coefficients R^2_{adj}de valeurs 0.425 et 0.375 respectivement. Un test de Shapiro-Wilk atteste de la normalité des résidus pour les modèles 'Allgt_avg' et 'cont_rupt_avg' (non-rejet de l'hypothèse nulle au niveau $\alpha = 5\%$) ainsi que celle des modèles 'cont_100_avg' et 'durete_avg' au niveau $\alpha = 1\%$.

La fiabilité relative de ces deux derniers modèles concernant le respect des hypothèses des erreurs est due à la distribution quasi-discrète des variables dépendantes et notamment leur faible amplitude de valeurs.

Les résultats des tests de significativité de Student pour chaque modèle MLR confirment et appuient les valeurs des coefficients de détermination obtenus. Notamment, le modèle 'cont_rupt_avg' affiche seulement 3 variables significatives au niveau $\alpha = 5\%$, et se distingue des 3 autres modèles pour lesquels davantage de variables explicatives impactent les 4 variables à prédire.

L'évaluation du modèle multivarié ajusté sur l'ensemble de test donne les RMSE suivants :

	RMSE
Allgt_avg	19.29
Cont_100_avg	0.29
Cont_rupt_avg	0.98
Durete_avg	0.68

Tableau : performance de test du modèle MMR

Enfin, évaluons l'impact des liens de corrélations, entre les 4 variables à expliquer exposés précédemment, sur la matrice de corrélation (coefficient de pearson) des résidus.

	Allgt_avg	Cont_100_avg	Cont_rupt_avg	Durete_avg
Allgt_avg	**1**	-0.35	0.6	-0.2
Cont_100_avg		**1**	0.37	0.23
Cont_rupt_avg			**1**	-0.01
Durete_avg				**1**

Tableau : matrice de corrélation des résidus, obtenue à l'issu du modèle MMR

Mis à part le lien quasi-nul entre les résidus des modèle « cont_rupt_avg » et « durete_avg », les coefficients observés nous permettent d'envisager des liens d'influence entre les 4 PMC. Précisément, on distingue un coefficient $r = 0.6$ entre les résidus des modèles ajustés sur l'allongement et la contrainte à la rupture. Les coefficients de pearson restants varient entre -0.3 et 0.37.

Il s'agit dans un second temps d'appliquer une MANOVA au modèle multivarié, pour évaluer l'effet individuel de chaque variable explicative sur la variance totale de Y. La fonction *Manova( )* du package *{car}* est utilisée dans ce sens.

```
> manova.mmr <- Manova(mmr.1)
> print(manova.mmr)

Type II MANOVA Tests: Pillai test statistic
                            Df test stat approx F num Df den Df    Pr(>F)
maturation_time              1  0.012020   2.8439      4    935 0.0232044 *
ML_std                       1  0.001465   0.3430      4    935 0.8489719
ML_min                       1  0.014594   3.4618      4    935 0.0081000 **
MH_std                       1  0.048314  11.8668      4    935 2.082e-09 ***
MH_min                       1  0.056315  13.9491      4    935 4.659e-11 ***
MH_max                       1  0.018901   4.5032      4    935 0.0013141 **
TS2_mean                     1  0.056140  13.9033      4    935 5.065e-11 ***
TS2_std                      1  0.035790   8.6765      4    935 7.062e-07 ***
TS2_min                      1  0.013241   3.1367      4    935 0.0141381 *
TS2_max                      1  0.033043   7.9876      4    935 2.478e-06 ***
T90_min                      1  0.056896  14.1017      4    935 3.528e-11 ***
```

Sortie R : résultats d'une MANOVA à un facteur sur le modèle multivarié – 11 premières variables

La sortie R obtenue est une table MANOVA affichant les résultats des tests multivariés à un facteur (« one-way MANOVA ») via un calcul des sommes des carrés de type II. Cette approche (à distinguer des types I et III) est liée à la formulation du test d'hypothèse. Plus précisément, l'hypothèse nulle caractérise un modèle réduit intégrant toute les variables explicatives sauf celle dont nous souhaitons tester l'effet sur la variance totale de Y. L'hypothèse alternative désigne le modèle complet avec toute les variables. Les matrices SSCP sont ensuite calculées sous H_0. La trace de Pillai-Bartlett est choisie par défaut comme statistique de test, et calculée pour le résultat du test au niveau 5% et la p-value associée.

L'analyse des résultats permet de constater l'effet significatif de plusieurs groupes de variables explicatives. Tout d'abord, l'ensemble des variables associées aux résultats rhéométriques des lots de mélange consommés, i.e les mesures de ML, MH, ts2 et t90, impactent significativement la variance totale de Y. Ce résultat inférentiel se révèle cohérent avec le lien fonctionnel existant entre les phases de mélange et de vulcanisation. En effet, les experts métiers semblent alignés sur le fait qu'une durée optimale de vulcanisation (t90) ou un MH élevés (par rapport aux spécifications) impactera le processus de vulcanisation du mélange, une fois le lot chargée en LVC. On distingue par ailleurs 5 paramètres process dont les indicateurs statistiques se sont révélés fortement significatifs : M_Caland_OpeningLevelRight, M_Extrud_BodyTemperature3, M_Extrud_HeadTemperature, M_Extrud_ScrewTemperature, M_Rotocure_CylinderTempUpper et M_Rotocure_RotoCylinderTemp.

Afin d'améliorer la clarté du modèle et sélectionner les variables explicatives influentes, une solution existante serait d'utiliser la procédure MANOVA pour la comparaison de modèles emboités. En se servant des résultats des tests à un facteur, l'idée serait de tester l'hypothèse nulle d'un modèle réduit excluant les variables non-significatives, contre l'hypothèse alternative du modèle complet. La liste des paramètres process et de leurs indicateurs statistiques associés (au nombre de 18) dont on souhaite tester l'effet est la suivante :

- La température des cylindres de la calandre : M_Caland_CylinderTemperatureLower (min/max), M_Caland_CylinderTemperatureUpper (mean/max).
- L'écart entre les deux cylindres à gauche de la calandre : M_Caland_OpeningLevelLeft (stdev/max).
- La pression avant-filtre de l'extrudeuse : M_Extrud_PressureBeforeFilter (mean).
- La température du corps n°2 de l'extrudeuse : M_Extrud_BodyTemperature2 (mean/max).

- La température du cylindre inférieur de la rotocure : M_Rotocure_CylinderTempLower (stdev).
- La pression de vulcanisation de la rotocure : M_Rotocure_VulcPressure (mean/stdev/min/max).
- La pression exercée côté droit du cylindre tendeur : M_Rotocure_CylPressureRight (stdev/min/max).

Le test multivarié est formulé à l'aide de la fonction *linearHypothesis( )* du package *{car}*. L'argument *hypothesis.matrix* de cette fonction permet de spécifier la forme de l'hypothèse nulle, soit par une matrice de contraste, soit par une chaine de caractère de type « Var = 0 ».

```
> lh.out_10 <- linearHypothesis(mmr.1, hypothesis.matrix = var_tests)
> lh.out_10

Sum of squares and products for the hypothesis:
              Allgt_avg cont_100_avg cont_rupt_avg durete_avg
Allgt_avg      8.668851    -2.942789      7.170840  -1.328555
cont_100_avg  -2.942789     6.744145      1.962094   1.936253
cont_rupt_avg  7.170840     1.962094     15.442523  -2.905263
durete_avg    -1.328555     1.936253     -2.905263  16.142197

Sum of squares and products for error:
              Allgt_avg cont_100_avg cont_rupt_avg  durete_avg
Allgt_avg      411.5153    -166.9748    351.913858 -101.160904
cont_100_avg  -166.9748     539.5910    252.876892  129.006826
cont_rupt_avg  351.9139     252.8769    834.377699   -9.629506
durete_avg    -101.1609     129.0068     -9.629506  586.715793

Multivariate Tests:
                Df test stat approx F num Df   den Df    Pr(>F)
Pillai          17 0.0906023 1.278743     68 3752.000 0.0625753 .
Wilks           17 0.9122482 1.278544     68 3671.524 0.0627410 .
Hotelling-Lawley 17 0.0931103 1.278213     68 3734.000 0.0629053 .
Roy             17 0.0399565 2.204661     17  938.000 0.0033211 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sortie R : résultat du test MANOVA pour la comparaison du modèle réduit à 51 variable avec le modèle complet

La sortie R renvoie les expressions de la matrice $SSCP_{Reg|H0}$ calculée sous H_0 et de $SSCP_{Res}$, la matrice de dispersion des résidus associée au modèle complet. Le résultat du test multivarié est un tableau affichant les valeurs des 4 statistiques de test (trace de Pillai, lambda Wilk, Hotelling-Lawley et Roy) calculées pour le test d'hypothèses, ainsi que leur approximation de Fisher.

L'hypothèse nulle n'est pas rejetée pour 3 statistiques de test sur 4 au niveau 5%. Seule la plus grande racine de Roy amène au rejet du modèle contraint.

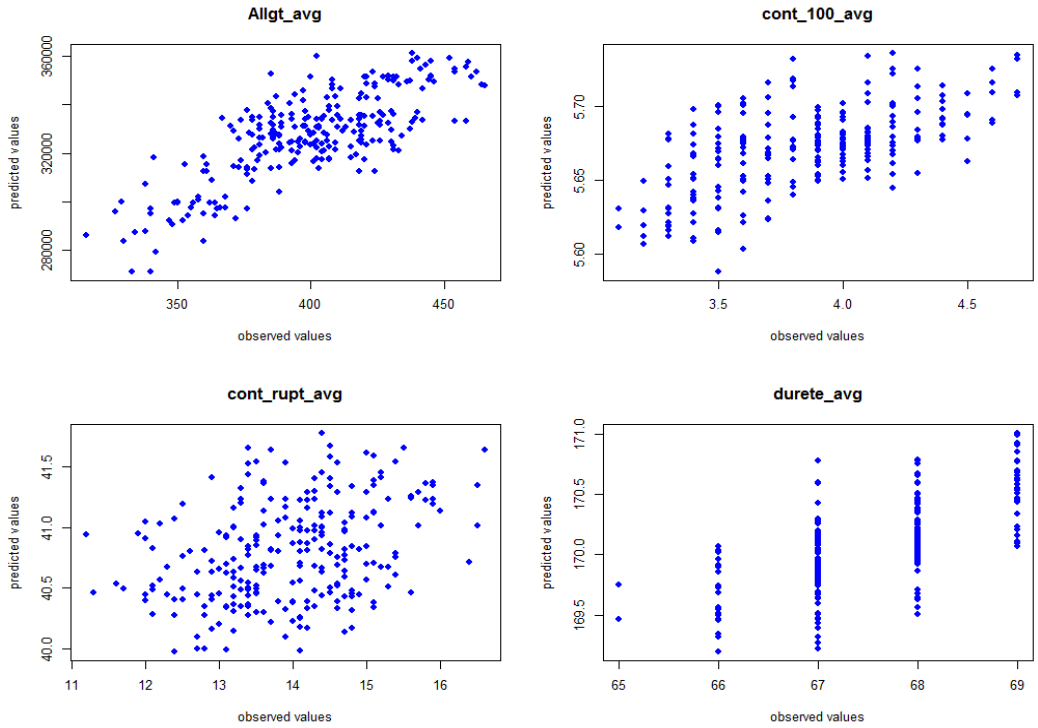
En suivant les résultats des tests pour les statistiques de Pillai-Bartlett, Wilk et Hotteling-Lawley, il parait judicieux d'explorer le modèle réduit à 51 variables, celui-ci s'étant révélé informatif en terme de variabilité de Y expliquée.

L'étape finale consiste à construire ce dernier modèle (à l'aide de la fonction *update( )*), évaluer sa performance sur l'ensemble D_{test}, et déterminer par des tests MANOVA à un facteur, l'impact individuel des variables explicatives restantes.

	RMSE
Allgt_avg	19.03
Cont_100_avg	0.29
Cont_rupt_avg	0.96
Durete_avg	0.67

Tableau : performance du nouveau modèle réduit sur D_{test}

Les RMSE obtenus sont proches de ceux calculés sur le modèle complet, avec une faible amélioration de 0.2 pour le modèle « Allgt_avg ».



Graphiques R : valeurs observées contre valeurs prédites sur D_{test}

L'analyse des graphiques des valeurs prédites contre les valeurs observées nous permet de constater que le modèle final « Allgt_avg » modélise efficacement la relation entre $\boldsymbol{X}$ et $\boldsymbol{Y}$, avec un nuage linéaire de points. A l'inverse, le modèle final « cont_rupt_avg » affiche une forte dispersion entre prédictions et valeurs observées.

Les résultats des tests MANOVA ont notamment montré une contribution significative de la température du cylindre central de la rotocure et des mesures rhéomètriques sur la variabilité des 4 PMC. A l'inverse, la vitesse de ligne mesurée sur la rotocure, les températures des cylindres de la calandre, ou encore la pression avant-filtre de l'extrudeuse n'influent pas sur la variance de Y.

Les résultats obtenus sont à pondérer avec la qualité d'ajustement observée sur les 4 modèles MLR indépendants. On peut envisager l'influence plus grande des modèles 'Allgt_avg' et 'durete_avg' sur les résultats inférentiels, du à leur coefficient R^2 élevés par rapport aux deux autres modèles.

D. Random forests

L'objectif de cette seconde approche de modélisation est d'évaluer le lien non-linéaire entre X et Y à l'aide d'un modèle de régression *random forest* (RF). En construisant cette fois-ci 4 modèles indépendants, l'idée est de mettre à profit les propriétés ensemblistes d'un *random forest* (agrégation d'arbres de décisions sur-ajustés) pour modéliser une relation fonctionnelle plus complexe entre X et Y, en comparaison aux modèles de régression linéaire usuels (régression linéaire multiple, Ridge, Lasso, etc). Bien que cette démarche multi-modèles ne prenne pas en compte la corrélation constatée entre les 4 variables dépendantes, le classement des variables explicatives influentes propres aux RF apporte des éléments de réponse supplémentaires concernant l'impact de certains paramètres sur les 4 propriétés mécaniques. Il s'agira également de comparer les performances des 4 modèles et leur RMSE sur l'ensemble de test avec les modèles de régression multivariés.

A la différence du modèle de régression multivarié, le jeu de données initial $D(n = 1260, p = 69, q = 4)$ est décomposé en 4 ensembles $(D_1, \dots, D_4)$, avec X inchangé et y un vecteur de taille $(n \times 1)$ comme variable à expliquer pour chaque D.

La même stratégie de découpage $(D_{app})_k$ / $(D_{test})_k, k = 1 \dots 4$, est appliquée selon un rapport 80 / 20. Toute les démarches d'implémentation, sélection de modèles et évaluations seront effectuées sous python via le module machine learning *scikit-learn*.

1.1. Sélection de modèles sur D_{app} par grille de recherche

Une première série de 4 modèles random forest est construite par une stratégie *grid-search* sur les 3 hyperparamètres évoqués lors de la description du modèle.

Hyperparamètre (dénomination scikit - klearn)	Valeurs candidates (p = 69)
max_features	p, $\log_2$(p), $\sqrt{p}$, 25, 50
max_depth	5, 10, 15, 25
n_estimators	10, 100, 500, 1000, 5000

Table : valeurs candidates choisies pour 3 hyperparamètres d'un random forest

Toute les combinaisons possibles de ces hyperparamètres sont testées via une procédure 5-*fold cross-validation* sur l'ensemble $(D_{app})_k, k = 1 \dots 4$. La meilleure combinaison est celle maximisant le coefficient de détermination R^2 moyen calculé sur les 5 blocs de validation. Le modèle random forest ainsi que l'approche *grid-search* sont implémentées via les classes d'objets *RandomForestRegressor( )* et *GridSearchCV( )*.

Random forest optimal	R^2
Allgt_avg	**60**
Cont_100_avg	**39.2**
Cont_rupt_avg	7.2
Durete_avg	**49**

Table : valeurs candidates choisies pour 3 hyperparamètres d'un RF

A l'image des résultats obtenus pour la MMR, le coefficient de détermination le plus élevé correspond au modèle 'Allgt_avg', avec un nombre d'arbres de décision égal à 100, un nombre de variables candidates sélectionnées pour le split à $\log_2(p) \approx 6$, et une profondeur maximal d'arbres à 10 splits.

Le modèle 'cont_rupt_avg' affiche le plus faible coefficient de détermination (7.2).

1.2. Constructions et évaluations des modèles finals par sélection de variables

Disposant d'une série de 4 modèles RF construits à partir de $p = 69$ prédicteurs, l'étape suivante consiste à réapprendre 4 modèles sur les mêmes jeux d'apprentissage en adoptant une démarche post-hoc de sélection de variables.

En s'appuyant sur les tables d'importance des variables propres à chaque modèle, le but est de construire 4 nouveaux modèles à partir des p' premières variables importantes, $p' \in N^*$.

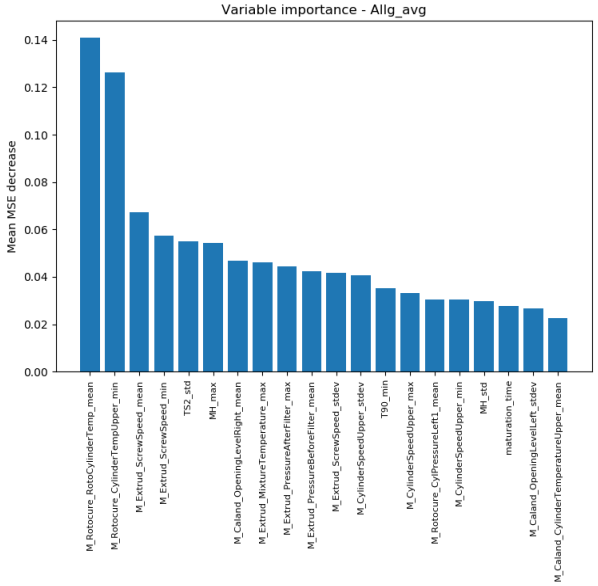
Le nouveau modèle optimal est associé au coefficient de détermination maximal parmi les p' obtenus. Enfin, le RMSE est calculé sur l'ensemble de test $(D_{test})_k$.

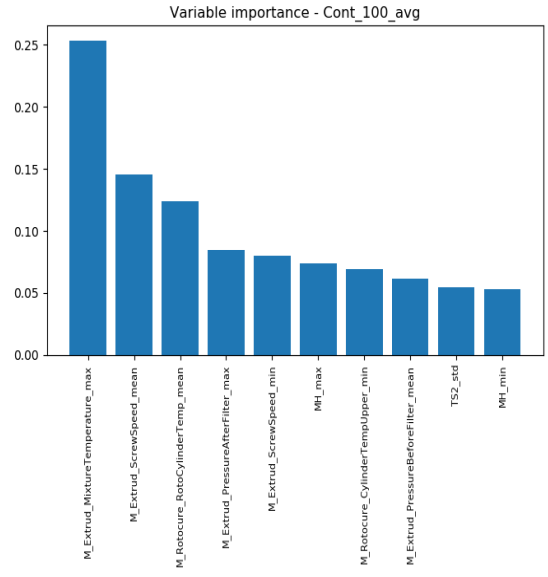
RF réduit optimal	p'	R^2 (apprentissage)	RMSE (test)
Allgt_avg	**20**	**60.1**	**16.65**
Cont_100_avg	**10**	**41.1**	**0.277**
Cont_rupt_avg	15	7.1	0.963
Durete_avg	**15**	**48.8**	**0.56**

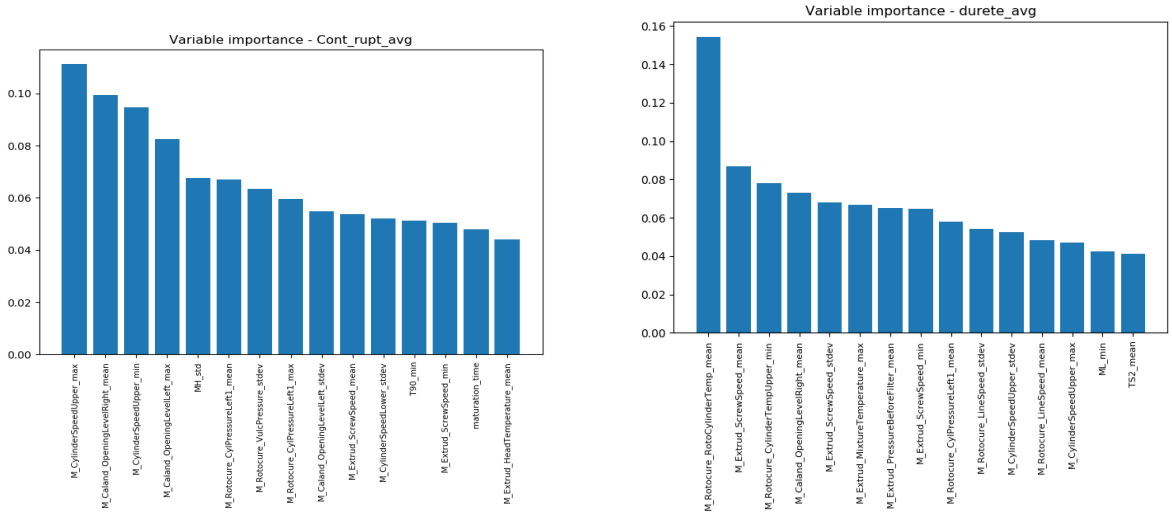
Table : performances des 4 modèles RF réduits obtenus à partir de p' variables explicatives

En comparant avec les résultats obtenus pour les modèles de régression, on constate une amélioration des performances sur le jeu de test pour 3 modèles sur 4. Seul le modèle 'cont_rupt_avg' affiche une valeur de RMSE comparable.

En outre, les diagrammes d'importance des variables permettent de confirmer certains constats faits à l'issue des modèles de régression. La température du cylindre central de la rotocure (M_Rotocure_RotoCylinderTemp) apparait significative sur la construction des modèles 'Allgt_avg' et 'durete_avg'.







Graphiques: diagramme d'importance des variables issus des modèles restreints finals

En d'autres termes, ce paramètre impacte significativement la structure du random forest et sa capacité à délivrer des feuilles minimisant le MSE, et donc la variabilité des deux propriétés mécaniques. D'autre part, deux paramètres *process* liés à l'extrudeuse impactent la moyenne du module à 100 %., la température du mélange passé dans l'extrudeuse ('M_Extrud_MixtureTemperature') et la vitesse de vis ('M_Extrud_ScrewSpeed'). On retrouve également le paramètre 'M_Rotocure_RotoCylinderTemp' comme troisième variable importante. Enfin, l'interprétation du diagramme associé au modèle 'cont_rupt_avg' est fortement conditionnée par la faible valeur du R^2 obtenue. On distingue cependant 4 variables dominantes sur la décroissance du MSE, toutes liées à la calandre de la LVC : l'écart entre les deux cylindres ('M_Caland_OpeningLevelRight', 'M_Caland_OpeningLevelLeft') et la vitesse du cylindre du haut de la calandre ('M_CylinderSpeedUpper').

Conclusion

Les deux types de modélisation choisies ont abouti à des résultats exploratoires relativement similaires, mais des performances de généralisation différentes.

D'un point de vue exploratoire, la constante observée tout au long de l'étape de modélisation est l'impact significatif de la température des cylindres de la rotocure, et plus particulièrement le cylindre central, sur la variabilité des 4 propriétés mécaniques. D'autre part, le modèle de régression multivarié met davantage en lumière la forte influence des critères rhéomètriques sur les PMC. Le modèle random forest pour le module à 100% cible le paramétrage de l'extrudeuse comme élément influant, et plus particulièrement la température de la matière et la vitesse de vis.

L'implémentation du modèle linéaire multivarié et sa procédure de tests MANOVA, ont permis de sélectionner efficacement un modèle réduit optimal, et d'exclure des paramètres non-influents sur les 4 propriétés mécaniques : la température des cylindres de la calandre, certaines composantes de la rotocure telles que la pression de vulcanisation, la pression du cylindre tendeur et la température du cylindre inférieur. La performance de généralisation obtenue sur le jeu de test est néanmoins plus faible que celle des 4 modèles random forest, notamment sur les modèles « Allgt_avg » et « durete_avg ».

Toutefois, du point de vue de la qualité d'ajustement au données, les résultats différents obtenus selon les propriétés mécaniques pour les deux modèles sont à prendre compte. Les modèles ajustés sur l'allongement, la dureté et le module à 100% moyens ont donné des coefficients de détermination entre 0.3 et 0.6, tandis que le coefficient lié à la prédiction de la résistance à la rupture moyenne ne dépasse pas 0.1. Dans ce sens, le modèle « Allgt_avg » est le modèle le plus robuste en termes de résultats exploratoires apportés.

Une autre stratégie de représentation du signal d'un paramètre de production peut être adoptée. Au lieu d'une décomposition statistique (moyenne, écart-type, minimum, maximum) calculée pour chaque paramètre, une analyse davantage orientée sur le traitement du signal, et plus précisément la décomposition des séries de données horodatées en plusieurs composantes temporelles, est une possibilité, afin de mettre en lumière leur comportement tendanciel. Dans cette optique, les décompositions fréquentielles de Fourier, les méthodes d'estimation de densité spectrales sont des méthodes possibles.

Enfin, d'autres pistes de modélisation peuvent être envisagées pour ce type de problématique. Dans le cadre du modèle multivarié, l'introduction de termes polynomiaux pour certains paramètres process pourrait mettre en avant des relations non-linéaires jusque-là inconnues entre les prédicteurs et $\boldsymbol{Y}$. L'apprentissage multi-tâches (« Multi-Task Learning » - MTL) est une autre approche existante de modélisation multivariée. Celle-ci est appliquée dans le cas d'une optimisation jointe de plusieurs fonctions de coût, une pour chaque tâche de régression. L'expression de la fonction objective d'un problème MTL, varie selon le modèle de base choisi (LASSO, Elastic-Net, etc.).

Glossaire

Définitions des termes industriels et qualité utilisés pour la rédaction de la thèse professionnelle :

- **Processus de production** : série d'opérations de production reliées entre elles. Un processus de production prend en entrée un ensemble de matières premières, et renvoie en sortie un produit fini (ou semi-fini). Abréviation courante : « process ».

- **Mélange** : première étape du process qui consiste à incorporer et disperser de façon homogène différents produits chimiques pulvérulents, pâteux ou liquides dans un élastomère selon un ordre déterminé. Phase répartie sur une machine avec deux fonctions différentes : 1) Mélangeur interne (incorporation des matières premières) – 2) Mélangeur externe (homogénéisation du mélange).

- **Vulcanisation** : processus de transformation (réticulation chimique) de bandes d'élastomères crues (issues du mélange après maturation), en bandes vulcanisées possédant les propriétés mécaniques souhaitées. Cette transformation s'opère sur la ligne de vulcanisation.

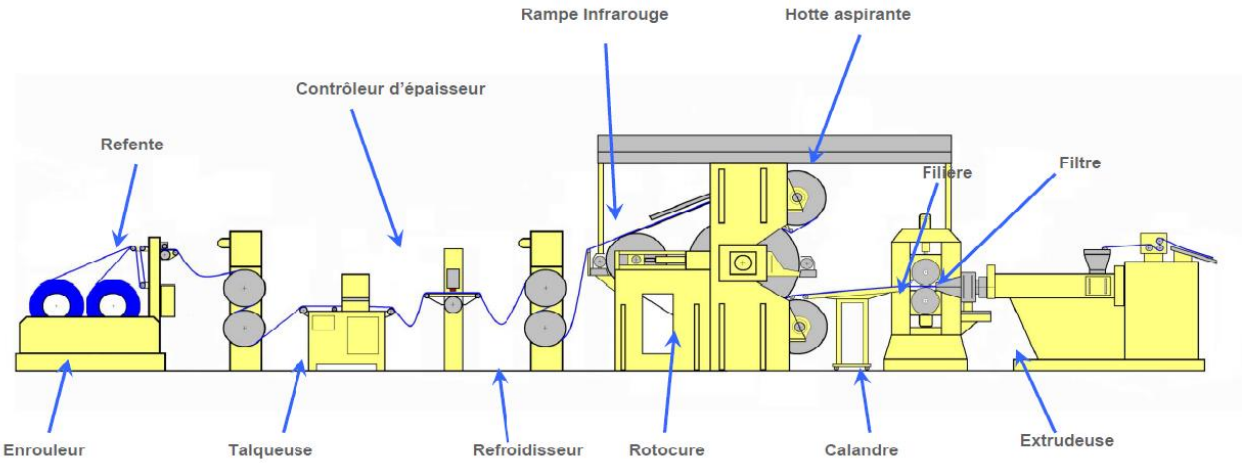


Schéma : ligne de vulcanisation et ses éléments constitutifs

Fonctions des composantes de la ligne :

- Extrudeuse : cisaillement de la matière (par une vis chauffante) et préformage de la bande.
- Calandre : réalise l'épaisseur de la bande par deux cylindres superposés
- Rotocure : assure la vulcanisation de la bande via une combinaison pression / température exercées par 4 cylindres dont un cylindre tendeur (en fin de cycle) et trois cylindres chauffants superposés.

- **LVC** : abréviation du terme « ligne de vulcanisation en continu ».

- **Lot** : une série de cartons produits en continu sur une même machine, à partir d'un ou plusieurs lots issus de l'étape précédente du process.

- **Carton** : terme générique pour une unité de production. Un carton est une quantité fixe produite, et prends plusieurs formes en fonction de l'étape du process :
    - Mélange : charge de matières premières mélangées
    - Vulcanisation : bobine de bandes vulcanisées
    - Découpe : filet de joints découpés.

- **Spécifications** : intervalle définissant les tolérances inférieure et supérieure d'un indicateur qualité. Ces intervalles varient en fonction de la matière produite. Un carton est déclaré **conforme** pour un indicateur donné si les résultats obtenus respectent les spécifications (i.e tombent dans l'intervalle défini).

- **Maturation** : durée minimale à respecter avant utilisation d'un lot de mélange. C'est la période repos pour rendre le mélange consommable en phase de vulcanisation.

- **Propriétés mécaniques** (PMC) : 4 indicateurs physiques qui décrivent le niveau d'élasticité atteint par la bande vulcanisée en sortie de ligne : résistance à l'allongement (en %), module à 100% (MpA), résistance à la rupture (MpA) et dureté (shore A). La conformité d'une bobine sur ces 4 critères sont indispensables pour la validation et la libération d'un lot de vulcanisation.

- **Rhéométrie** : mesure des propriétés viscoélastiques d'une matière.

- **lbf – in** ("pound-force per inch") : unité de pression anglo-saxonne.

Bibliographie / Références

Carey, G. (2013). *General Linear Models (GLM) : Multiple dependant variables.* University of Colorado.

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression (Second Edition).* Sage Publications.

Fox, J., Friendly, M., & Weisberg, S. (2013). Hypothesis Tests for Multivariate Linear Models Using the R "car" package. *The R Journal*.

Hastie, T., Jerome, F. H., & Tibshirani, R. (2009). *The Elements of Statistical Learning (Second Edition).* Springer.

Maitra, R. (2012). *Multivariate Linear Regression Models.* Iowa State University.

https://data.library.virginia.edu/getting-started-with-multivariate-multiple-regression/

https://cran.r-project.org/web/packages/car/car.pdf

http://scikit-learn.org/stable/