Centrale Méditerranée | STMicroelectronics

Rapport de stage assistant ingénieur

Application de la Data Science et du machine learning dans l'optimisation du processus de fabrication des semiconducteurs

Manufacturing Data Science (MDS)

STMicroelectronics Crolles

Determining and Optimizing Routes in Semiconductor Wafer Fabrication and Production

Réalisé par:
Akdi Oussama

Supervisé par:
M. Joao Paulo Sales Brum
M. Loïc Rolland

Tuteur de l'école:
M. Salah Bourennane

Année universitaire 2023-2024

Contents

Introduction

Le marché des semiconducteurs est en constante évolution, alimenté par l'innovation technologique et la demande croissante pour des appareils électroniques plus performants et plus efficaces. Les semiconducteurs sont au cœur de nombreux dispositifs modernes, des smartphones aux voitures autonomes, en passant par les équipements médicaux et les systèmes de communication. Ils entrent aussi dans la fabrication d'autres composants électroniques tels que les processeurs, les microcontrôleurs, la mémoire et les capteurs. Ainsi, cette industrie particulière est un atout et un maillon indispensable pour les chaînes industrielles en général et principalement celle des appareils électroniques.

Le marché des semiconducteurs tend vers une réduction continue de la taille des composants dans le but d'augmenter la performance, de réduire la consommation énergétique et les coûts de fabrication. Il y a également une tendance à s'adapter aux fortes demandes pour les puces capables de traiter les algorithmes d'intelligence artificielle et les puces capables de traiter des masses volumineuses de données.

Chez STMicroelectronics, la fabrication Front-End est continue 24h/24 et toute la semaine, dans les deux salles blanches. Des petites plaques de matière semiconductrice brute passent par un ensemble d'opérations traversant une multitude de machines dans un ordre bien précis, et ce jusqu'aux derniers tests électriques faits sur les produits fabriqués.

Le sujet de mon stage, intitulé "La recherche du golden route dans le process de fabrication", s'inscrit dans le cadre de l'optimisation du processus. Il avait pour but d'élaborer une solution basée sur les outils de machine learning et de la statistique inférentielle, de développer un outil de classement des machines intervenant dans le processus, afin de pouvoir choisir parmi ces machines celles qui composeront le meilleur chemin, celui qui permettra une amélioration du rendement du processus.

Présentation de STMicroelectronics

STMicroelectronics (ST) est une entreprise multinationale spécialisée dans la conception, le développement, la fabrication et la commercialisation de solutions de semiconducteurs. Fondée en 1987, STMicroelectronics est le résultat de la fusion de deux entreprises pionnières dans le domaine des semiconducteurs : SGS Microelettronica en Italie et Thomson Semiconducteurs en France. Aujourd'hui, ST est un acteur majeur dans l'industrie des semiconducteurs, offrant une gamme diversifiée de produits pour divers marchés, notamment l'automobile, l'industriel, l'électronique grand public, et les communications.

Figure 2.0.1: Carte des sites STMicroelectronics

STMicroelectronics est présente sur plusieurs sites de production répartis en Europe, en Asie et en Afrique du Nord. Le siège social de la société est situé à Genève, en Suisse. Avec un effectif global de 46 000 employés (2020), en France, l'entreprise compte 10 300 employés répartis sur 9 sites, dont 2 800 se consacrent à la recherche et au développement. L'entreprise dessert plus de 100 000 clients dans le monde et a réalisé un chiffre d'affaires de 10,2 milliards de dollars en 2020.

Site de Crolles

Le site de Crolles, situé près de Grenoble en France, est l'un des centres de production les plus avancés de STMicroelectronics. Ce site est spécialisé dans la fabrication de plaquettes de silicium de 200 mm et 300 mm, et est également un centre de recherche et développement.

Figure 2.0.2: Les deux fabs Crolles200 et Crolles300 du site STMicroelectronics Crolles

La salle blanche

Figure 2.0.3: Salle blanche

Le site de Crolles est le plus grand site de production «Front-End» en France, composé de deux usines de fabrication : Crolles 200 et Crolles 300. Crolles 200 utilise des plaques de silicium de 200 mm de diamètre, tandis que Crolles 300 utilise des plaques de 300 mm. Crolles 200 et Crolles 300 ont respectivement des capacités de production allant jusqu'à 120 nm et 28 nm. La surface du site STMicroelectronics à Crolles est d'environ 40 hectares avec une surface au sol de 60 000 m^2. Les deux unités de production fonctionnent en continu, 24 heures sur 24 et 7 jours sur 7.

Les conditions environnementales dans les deux usines sont strictement contrôlées pour minimiser la contamination des plaquettes de silicium. Ces conditions incluent une filtration de l'air, un contrôle de la température et de l'humidité, et des protocoles stricts de propreté.

La MDS

Mon stage a été effectué dans le service MDS (Manufacturing Data Science), un service interne et intégré de STMicroelectronics. Il s'agit du pôle de Data Science qui développe des solutions pour le processus de fabrication. Chez STMicroelectronics, la MDS joue un rôle crucial dans l'amélioration de l'efficacité de la production, la réduction des coûts, et l'amélioration de la qualité des produits. La MDS a aussi un rôle de support pour les autres sites et centralise les solutions qui démarrent localement et peuvent être étendues à d'autres usines. Les données collectées au cours et à la fin du processus de fabrication sont analysées pour identifier des opportunités d'amélioration et pour prédire et détecter les problèmes potentiels liés aux puces fabriquées, en passant le plus souvent par des tests de contrôle de qualité et ensuite l'analyse des résultats de ces tests EWS (Electrical Wafer Sort) et PT (Parametric Test).

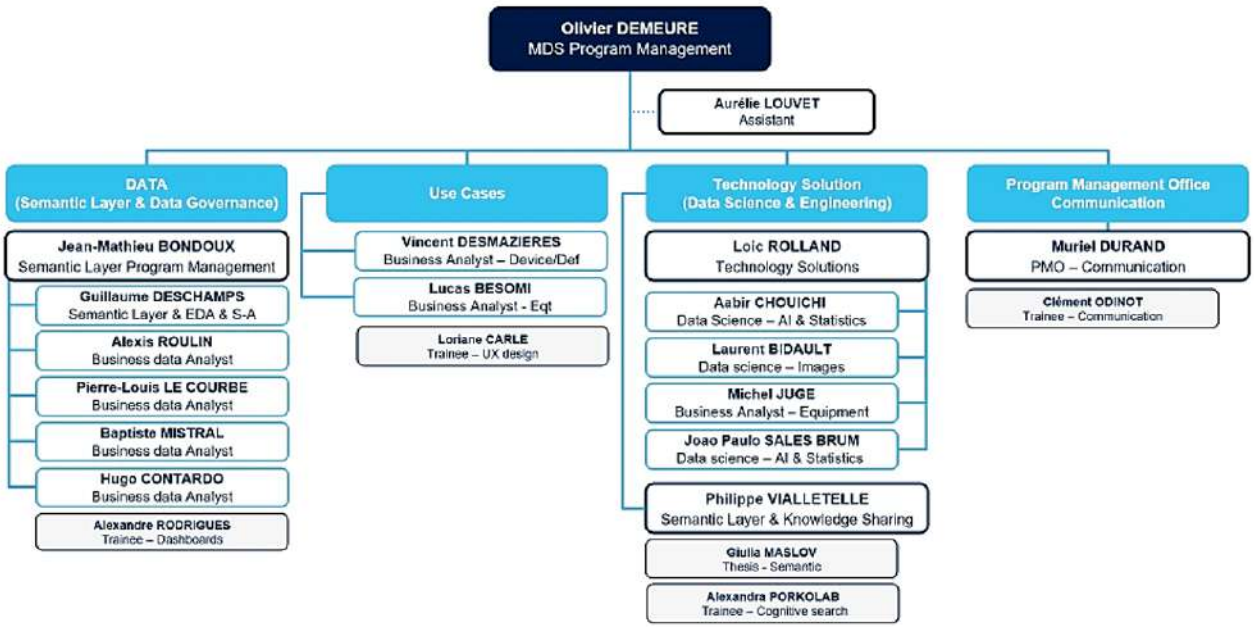


Figure 2.0.4: Organigramme de la MDS

Classement des Routes dans la Fabrication des Semi-conducteurs

3.1 Processus de Fabrication des Wafers

Un wafer est le résultat d'un processus de fabrication de puces électroniques chez STMicroelectronics. Il s'agit d'une plaque ronde implantée avec un très grand nombre de puces (transistor, diode, mémoire, capteur, etc.). Les wafers se présentent en deux catégories selon leur diamètre :
Wafer200: Ayant un diamètre de 200 mm, ils sont fabriqués dans la chaîne de fabrication nommée Crolles200, qui est la première ligne de fabrication de semiconducteurs sur le site de Crolles.
Wafer300: Ayant un diamètre de 300 mm, ils sont fabriqués dans la chaîne de fabrication nommée Crolles300. Cette fabrication est plus efficace que celle de Crolles200 grâce aux dernières technologies. Plusieurs étapes du processus Crolles200 ont été automatisées en Crolles300.

Le passage des semiconducteurs aux puces électroniques se fait en deux étapes : Frontend et Backend.

Le Frontend désigne toutes les étapes de fabrication de circuits intégrés tant qu'ils sont encore implantés sur le wafer. Ces micro-dispositifs se fabriquent en parallèle et suivent le même cheminement que leur wafer porteur, en passant par différents ateliers de traitement de matière : gravure, nettoyage, implantation, photolithographie, contrôle, etc. À la fin de sa fabrication, le wafer subit un ensemble de tests et un contrôle de qualité, ce qui marque la fin de la fabrication Frontend. Le wafer est alors prêt pour aller dans les ateliers de Backend.

Le Backend regroupe les étapes de découpage des wafers et la séparation des circuits intégrés initialement logés sur le wafer, leur mise en boîtier et les derniers tests de qualité de ces circuits avant leur arrivée sur le marché.

Géographiquement, les sites Backend et Frontend sont séparés, et les lots de wafers doivent être transmis des sites Frontend vers les sites dédiés au Backend. Dans les deux usines de Crolles, des lots de wafers sont fabriqués sous les yeux des experts qui contrôlent en permanence les étapes du processus

et assurent les transitions des lots de wafers entre ces étapes. Le processus est long et dure plusieurs semaines ; le wafer passe par un grand nombre de machines et subit beaucoup de transformations avant d'être prêt pour être envoyé au Backend. Cependant, le nombre de tests effectués sur le wafer au cours du processus est faible ; la majorité des tests se passent en fin de fabrication, car le processus est compliqué et minutieusement chronométré, et les wafers sont très sensibles et fragiles. Ainsi, une interruption du processus pour mettre en place des tests intermédiaires peut ne pas être envisageable, par exemple entre deux étapes qui requièrent une isolation totale du wafer.

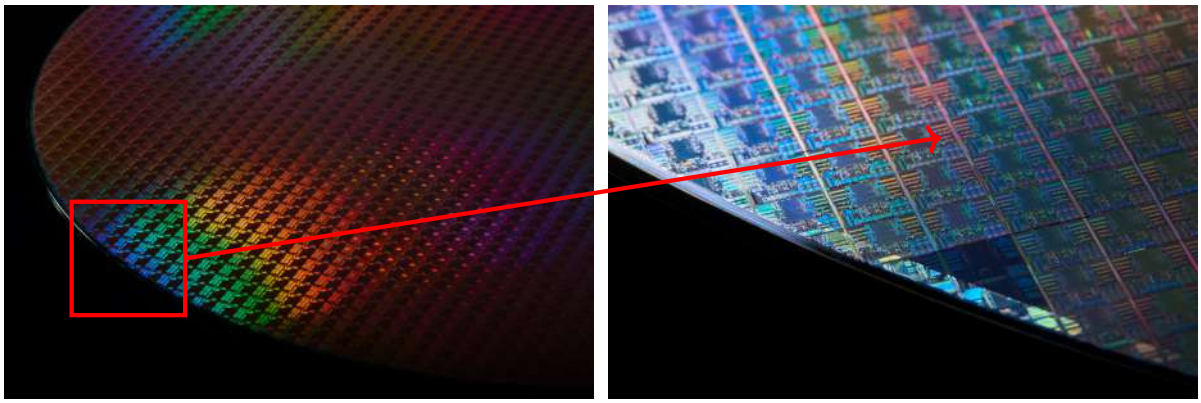


Figure 3.1.1: A gauche l'image d'un wafer et à droite un zoom sur une le bord du wafer montrant un nombre de circuits imprimés

Sur un wafer, on appelle rendement la proportion des puces qui réussissent un test donné parmi l'ensemble des tests EWS* et PT*. Donc, pour un même wafer, on peut parler de plusieurs rendements, autant qu'il y a de différents tests (binning). Un test mesure la proportion de circuits intégrés qui vérifient le critère du test : cela peut correspondre, par exemple, à la mesure du gain du circuit sous une configuration de courant ou de tension donnée. Si ce gain n'atteint pas le seuil imposé pour la puce (test failed), celle-ci sera marquée comme présentant un bin X donné.

*PT (Test paramétrique) : Test de validation des caractéristiques électriques des C.I., se fait au cours du process

*EWS (Electrical wafer sorting) : On test la fonctionnalité électronique de chaque C.I. juste avant de l'envoyer au Backend.

3.2 Problématique

Au fil du temps, les résultats des tests finaux sont enregistrés dans de grandes bases de données, mises à disposition de l'équipe MDS (Manufacturing Data Science), le pôle de la Data Science chez STMicroelectronics. Moi, en tant

que stagiaire dans ce service, j'avais accès à deux grandes tables de données qui m'ont servi tout au long de ma période de stage.

Le sujet de mon stage s'inscrit dans le cadre de l'enjeu du **Golden Process Route**, une théorie avec des applications surtout dans les domaines du manufacturing, qui vise à identifier les meilleurs chemins et les pires chemins le long d'une chaîne de fabrication.

Il faut savoir qu'un wafer passe par un ensemble d'opérations successives et dans un ordre bien défini. À chaque opération, le wafer passe par un certain nombre d'étapes successives et dans un ordre bien précis. À chaque étape, le wafer passe par une machine. Pour certaines étapes, il n'y a qu'une seule machine à disposition et donc le passage par celle-ci est certain. Mais pour d'autres étapes, le wafer peut avoir le choix entre plusieurs machines qui font la même tâche. Pourtant, elles n'auront pas le même rendement, car la performance des machines diffère d'une machine à une autre. Cela peut dépendre de plusieurs facteurs comme l'âge de la machine, sa génération, sa fréquence d'utilisation, etc.

Un chemin est une suite finie d'opérations-étapes-machines qui, pour un wafer, sera l'ensemble des machines par lesquelles il a dû passer lors de sa fabrication.

Exemple : OP-1 ET-1 M-2, OP-1 ET-2 M-3, . . . , OP-n ET-m M-l

Ici dans cet exemple, OP-1 ET-1 M-2 veut dire que le wafer a passé par la machine 2 à l'étape une de la première opération. Et n, m, l représentent respectivement les indices de la dernière opération, étape, machine.

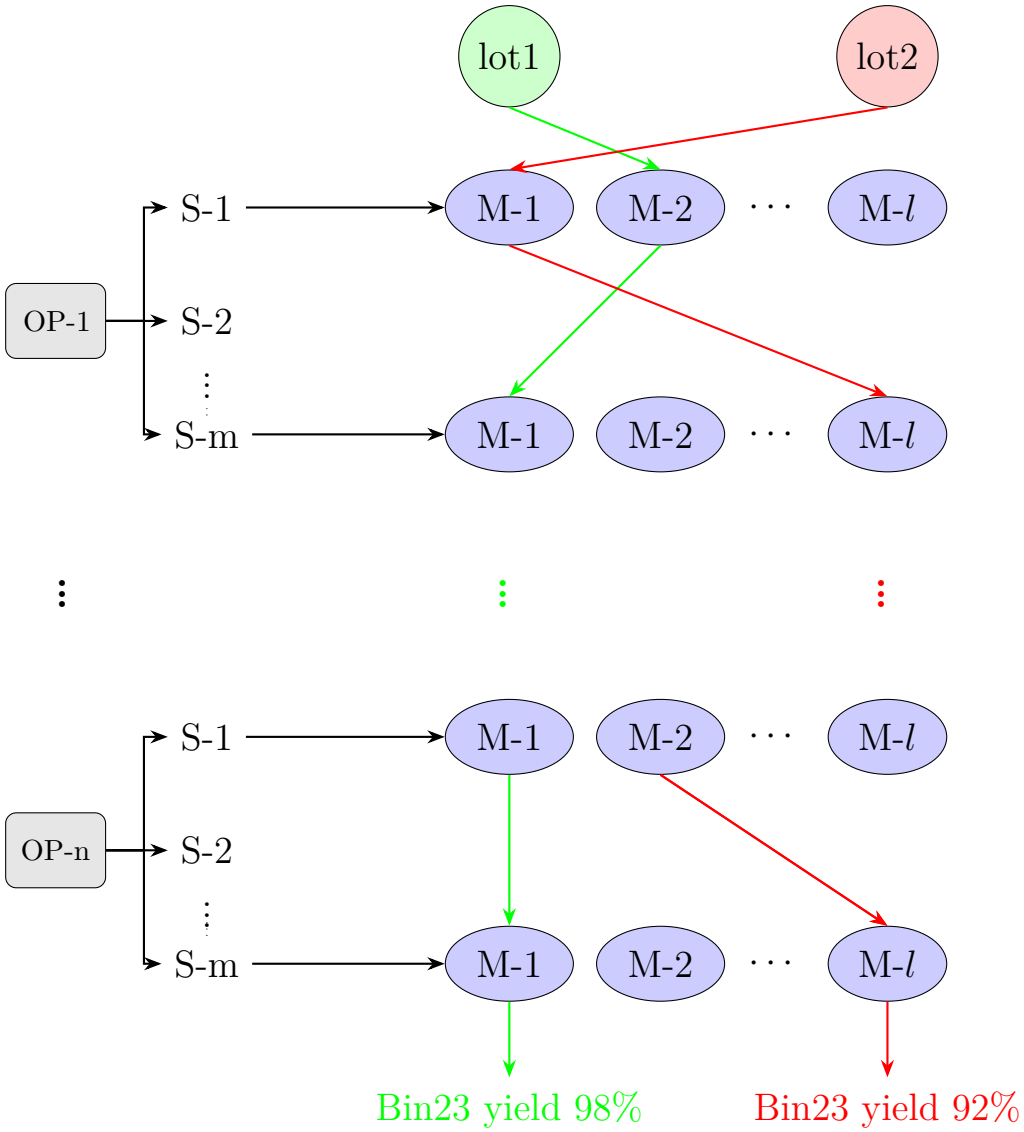
Le fait qu'à une étape donnée on ait le choix entre plusieurs machines si elles sont disponibles implique que le wafer peut suivre différents chemins. Si on fait le choix de donner le wafer à une machine qui est fatiguée ou qui se trompe à cause d'un problème technique, on risque d'augmenter le nombre de défauts et donc de nuire au rendement final du wafer.

Le passage par toutes les machines de haute qualité n'est pas toujours possible à cause de leur disponibilité à un moment donné et du fait que le processus est en flux continu. On ne peut pas simplement attendre que la meilleure machine soit libre. Mais parfois, on ne sait même pas quelle est la meilleure machine. Notons bien que l'influence d'une machine sur le rendement n'est pas la même pour tous les types de bins. Ainsi, un chemin peut être favorisé si on veut contrôler un paramètre (un bin) mais peut être désavantageux avec d'autres types de défauts.

La recherche du Golden Route se fait par type d'erreur. Dans la suite du rapport, l'étude sera faite sur le bin23, dont le test associé est évalué via un paramètre beta qui représente le gain du transistor.

3.3 Modélisation

On peut modéliser le problème du Golden Route Process par le schéma ci-dessous. On voit qu'il y a n opérations, pour chaque opération il y a m étapes, et à chaque étape il y a l machines possibles.



La procédure de la recherche du meilleur chemin commence par la quantification de l'influence de chaque étape sur le rendement final, définir un critère et en déduire une classification des machines du processus. Via des études statistiques et des validations au cours du temps, on peut générer un classement (ranking) pour chaque équipement du processus.

Avec le classement mentionné précédemment, il sera possible d'identifier le golden route qui maximise le rendement final selon le critère de la méthode statistique adoptée. Mais aussi d'identifier les bons chemins pour générer une distribution d'équipements maximisant le rendement et d'identifier les

mauvais chemins afin de les éviter.

3.4 Les Défis du Problème

Le problème de la recherche du golden route tel qu'il est défini précédemment souffre de nombreuses limitations et défis, tels que :

Le fait qu'il n'y ait pas de tests intermédiaires pour le bin23 rend la tâche difficile, car le rendement final du wafer est informatif quant à la qualité globale du chemin suivi, mais ne permet pas de se renseigner sur les effets de chaque équipement sur le rendement.

Les données récoltées sur 261 jours de fabrication montrent que l'hypothèse selon laquelle le rendement est uniquement fonction de la performance des équipements du processus peut ne pas être vraie en cas de crises par exemple, d'accidents ou à la suite d'une mauvaise gestion des transitions entre les étapes du processus. D'autres sources d'aléatoire peuvent intervenir en dehors des équipements, ces sources étant considérées comme des variables manquantes dans notre modélisation, ce qui peut ajouter un biais aux inférences.

Avec une centaine d'opérations et une dizaine d'étapes à chaque opération, le nombre de chemins possibles est énorme (de l'ordre d'un billion). Trouver un algorithme de recherche global sur un tel ensemble avec une complexité raisonnable peut ne pas être possible.

3.5 Prise en Main des Données

La manipulation des données va se faire sur la plateforme Databrics. Il s'agit d'une plateforme de gestion de données et d'analyse basée sur le cloud. Fondée par les créateurs d'Apache Spark, elle offre une infrastructure unifiée pour le traitement des données compatible avec plusieurs langages de programmations orientés données.

Databricks permet aux équipes de Data Science de travailler ensemble de manière plus efficace grâce à des fonctionnalités telles que les notebooks collaboratifs, l'intégration avec divers outils de données et une gestion simplifiée des clusters.

Sur cette plateforme j'ai pu importer deux tables qui feront la matière première de mon stage.

La première table « flow_normalized_bamb10 » est un historique du processus de fabrication du produit **BAMBOO** enregistré dans la salle blanche de Crolles200, les enregistrements ont été effectués entre : 2023-09-28 13:01:28 et 2024-06-16 02:34:21, donc 261 jours de fabrication. Dans cette table on

compte 48M enregistrement.

La table comporte 67 variables de natures différentes : Variables temporelles, variables catégoriques (ex : identifiant d'une opération). Un enregistrement correspond à une transaction: information sur un lot de wafers (groupement de 25 Wafer). Cela peut être par exemple la date de passage du lot à travers une machine donnée, la date de fin d'une opération sur un lot etc...

Vu la taille de cette base de données, son traitement se fera via l'outil Spark (Apache Spark),c'est est un moteur de traitement de données open-source conçu pour la vitesse et la facilité d'utilisation. Spark est capable de traiter des vastes quantités de données en mémoire, ce qui le rend beaucoup plus rapide et avantageux qu'un traitement traditionnel via la librairie Pandas sous Python. Il peut aussi interagir avec divers langages de programmation, notamment Scala, Java, Python et R, et offre des bibliothèques intégrées pour le SQL.

La deuxième table « flow_bin_die_bamb10 » contient les enregistrements des résultats des tests faits sur les wafers de chaque lot à la fin du process. Toujours pour le même produit **BAMBOO** et le résultat consiste au comptage du nombre de défauts présentes sur le wafer. Comme précisé avant, le type de défaut considéré le long de cette étude est le bin23.

stdf_mes_lot_name	**stdf_sublot_id**	**stdf_prr_hard_bin_number**	**count_dies**
J408VWW	25	23	2241
J408VWW	19	23	2491
J408VWW	15	23	383
J408VWW	14	23	358
J408VWW	11	23	1009
J408VWW	05	23	983
J343ZAP	23	23	198
J343ZAP	21	23	21
J343ZAP	20	23	37
J343ZAP	16	23	28
J343ZAP	01	23	202
J408VWW	20	23	4

Table 3.5.1: Table flow_bin_die_bamb10

L'attribut stdf_mes_lot_name est la variable en commun entre les deux tables et c'est bien cette colonne qui jouera après le rôle de clé de jointure des deux tables. Comme chaque lot est formé de 25 wafers ; stdf_sublot_id est alors une variable qui va de 1 à 25 pour chaque valeur de stdf_mes_lot_name.

Avec quelques transformations sur lesquelles nous reviendrons plus tard avec plus de détails, on peut passer de ces deux tables à une seule qui a la

forme d'une table pivot, une matrice remplie de 0 et de 1. Dans le contexte de notre étude, les lignes de cette table représenteront les chemins parcourus par les différents lots. Nous allons nous limiter aux chemins déduits de nos bases de données et ne traiterons pas toutes les combinaisons possibles pour des raisons expliquées au paragraphe : défis du projet. L'étude va se focaliser sur la recherche du meilleur chemin possible, qu'il ait été parcouru ou non, mais en se basant seulement sur les chemins parcourus et nos connaissances sur ces chemins.

La table pivot soigneusement choisie peut être vue comme la matrice X en machine learning. Le choix de cette forme sera justifié lors de la partie suivante qui va présenter la méthodologie adoptée pour la recherche du Golden Route.

Lot_id	3322-3130-TELO1B	3322-3130-TEL32	3322-3130-TEL33	3322-3130-TEL41	3322-3200-ASM01B	...
J344CSK	0	0	0	0	0	...
J344KHT	0	0	0	0	0	...
J344PWZ	0	0	0	0	0	...
J345MNF	0	0	0	0	0	...
J345NPT	0	0	0	0	0	...
...	...	...	...	...	...	...
J412TLG	0	0	0	0	0	...
J412TRB	0	0	0	0	0	...
J412TTP	0	0	0	0	0	...
J413ANV	0	0	0	0	0	...
J413EVF	0	0	0	0	1	...

Table 3.5.2: Table pivot

La variable OST est synthétisée à partir de la concaténation de trois variables: Operation, Step et Tool. Donc, 3322-3130-TEL01B signifie simplement la machine TEL01B de l'étape 3130 de l'opération 3322.

L'ensemble des colonnes de cette table comprend toutes les combinaisons apparues au moins une fois dans la table « flow_normalized_bamb10 » et sont au total 239 OST. Les autres possibilités sont concrètement envisageables mais ne seront pas prises en compte dans la modélisation car nous avons aucune mesure du bin23 sur les lots correspondants.

Chaque ligne de la table pivot représente un wafer différent. Il faut savoir que les 48 millions d'enregistrements ont abouti, après prétraitement, à seulement 914 lots différents parce qu'un grand nombre de transactions correspondent au même lot et en plus il y a beaucoup de valeurs manquantes.

Les 1 et 0 indiquent un passage ou pas de passage ; par exemple le 1 qui se trouve à la position (J413EVF,3322-3200-ASM01B) **Table 3.5.2** veut dire que le lot matriculé sous J413EVF a fait un passage à travers la machine ASM01B à l'étape 3200 de l'opération 3322. De la même manière les 0 affirment qu'il n'avait aucune correspondance entre le lot et l'OST.

La table pivot est un résumé statique (sans prendre en compte les variables du temps) des informations importantes de la table « flow_normalized_bamb10 » qui seront utilisées dans les modèles de Golden ranking routes.

Classement Basé sur la Régression de Comptage

On sait jusqu'ici que le problème est paramétré par un nombre d'Operation-Step-Tool, ou simplement dit OST, 239 dans l'exemple précédent, mais cela peut varier selon le modèle. Par exemple, Lasso peut éliminer un nombre de variables non significatives grâce à la régularisation. Ou bien, on peut prendre la décision de supprimer quelques variables corrélées si leur corrélation dépasse un certain seuil fixé.
Remarque : Rien n'empêche le lot de passer par la même machine deux ou plusieurs fois pendant différentes opérations et étapes

La métrique qui nous intéresse est une fonction de la moyenne du comptage (count) du nombre de défaut bin23 que présente le wafer, c'est une variable entière non négative. L'algorithme que l'on va développer se base sur une régression de $g(\mu)$ où μ est la moyenne du comptage de défauts et g est une fonction qui dépend de la nature du modèle utilisé, appelée aussi Link function.

Considérons un ensemble de N combinaisons O-S-T ($N = 239$ dans l'exemple précédent).Alors, l'équation de régression peut se mettre sous la forme de :

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_N X_N \tag{4.1}$$

Où X_i dummy variable (variable muette)

$$X_i = \begin{cases} 1 & \text{si il y a un passage correspondant avec la i-ème OST} \\ 0 & \text{sinon} \end{cases}$$

$\beta_1, \beta_2, \ldots, \beta_n$: les coefficients de la régression qui dépendent du modèle choisi et de la table pivot et β_0 est le biais *the intercept*.

La recherche des coefficients β_i qui optimisent l'ajustement des variables X_i avec la target permettra de quantifier l'effet relatif de chaque variable X_i sur la moyenne des erreurs. Cela revient à déterminer une estimation de la contribution individuelle des équipements (Tools) et des opérations-étapes correspondantes sur $g(\mu)$.

Prétraitement des Données.

Dans cette partie on va résumer les principales étapes du prétraitement des données.

Tout d'abord, nous avons effectué un filtrage des étapes qui n'interfèrent pas avec le bin23 et donc n'impactent pas le rendement. Ensuite, la première table contenait des enregistrements sur les chemins des lots (groupement de 25 wafers), tandis que la seconde table englobait des mesures du bin23 sur les wafers pour chaque lot. Pour pouvoir joindre les deux tables selon la clé "id_du_lot", nous avons procédé à un groupement des 25 wafers de chaque lot de la deuxième table, en sommant leur bin23 total, ce qui a donné une mesure du bin23 par lot.

Nous avons ensuite généré une table pivot : chaque ligne représentait un lot différent, et chaque colonne était une combinaison opération-étape-machine envisageable par le processus. Nous avons mis un 1 dans une position (lot, OST) si le lot présent avait passé l'étape S de l'opération O par la machine T, sinon nous avons mis un 0. Une ligne de la table pivot est une suite de 0 et de 1 traduisant le chemin parcouru par le lot de cette ligne.

La table pivot a été ensuite jointe avec la table du comptage du nombre de bin23 par lot obtenue à la suite du groupement effectué pendant l'étape précédente. Nous avons sélectionné 62 caractéristiques parmi les 239 variables à l'aide de la méthode de régression Lasso. La régression Lasso est particulièrement utile pour réduire la dimensionnalité et éliminer les variables non significatives, ce qui est essentiel pour garantir que les modèles ne deviennent pas non significatifs ou ne convergent pas en raison de la présence de trop de variables.

En outre, la multicolinéarité, qui peut entraîner une perte d'information et rendre les modèles instables, a posé un défi majeur durant cette étude. La régression Lasso a aidé à atténuer ce problème en sélectionnant un sous-ensemble de variables pertinentes, ce qui assure que les modèles restent robustes et pertinents.

Enfin, l'élimination des variables avec peu de mesures est une étape essentielle dans le prétraitement des données. Il est crucial d'identifier et de supprimer ces variables, car elles peuvent introduire du bruit et ne pas

apporter d'information pertinente à l'analyse. En effet, les variables avec peu de mesures peuvent fausser les résultats et diminuer la robustesse des modèles prédictifs.

Méthodes d'évaluation des Modèles

6.1 Estimation des Coefficients

Dans les modèles de régression, les coefficients β_i représentent l'effet de chaque variable indépendante X_i sur la variable dépendante target Y. L'estimation de ces coefficients se fait généralement soit par la méthode des moindres carrés ordinaires MCO, Cette méthode minimise la somme des carrés des résidus: différences entre les valeurs observées et les valeurs prédites par le modèle, soit par la méthode du maximum de vraisemblance MLE.

6.2 Écart-Types Standard des Coefficients

L'écart-type standard σ_i ou erreur standard d'un coefficient β_i est une mesure de la variabilité de l'estimation de ce coefficient, permet d'évaluer la précision des coefficients. Un écart-type standard faible indique que l'estimation du coefficient est précise, tandis qu'un écart-type standard élevé indique plus d'incertitude.

6.3 Test de Significativité du Modèle

Estimer tous les coefficients avec les plus petits écarts types n'est pas facilement atteignable et n'est pas toujours possible. Mais même si nous y parvenons, cela ne signifie pas que le travail est terminé. Le modèle peut ne pas être statistiquement significatif, même si l'estimation est bonne. Cela peut se produire si le modèle est mal spécifié, c'est-à-dire si le choix de la famille du modèle n'est pas approprié. Dans ce cas, l'ajustement (fitting) est impossible. Pour vérifier la significativité du modèle choisi, nous procéderons à un test d'hypothèse portant sur les coefficients du modèle.

Dans notre cas d'étude, tous nos modèles sont des modèles de régression. Le test d'hypothèse sera donc le même.

L'hypothèse nulle

Le test en question évalue l'hypothèse nulle selon laquelle tous les coefficients

du modèle sont nuls

$$H_0 : \beta_0 = \beta_1 = \cdots = \beta_N = 0 \tag{6.1}$$

Cette hypothèse signifie que le modèle renvoie toujours la même valeur, qui est zéro, indépendamment de l'entrée. En d'autres termes, le modèle n'a rien appris sur les données. Cela implique que le modèle ne fournit aucune information utile sur les données et que les prédictions du modèle ne varient pas en fonction des variables indépendantes.

Statistique F pour les Modèles de régression linéaire

La statistique F est utilisée dans le cadre de ce test d'hypothèse pour évaluer la signification globale d'un modèle de régression linéaire. Elle compare la variabilité expliquée par le modèle à la variabilité non expliquée (ou résiduelle). Voici une explication détaillée : La statistique F est calculée comme suit :

$$F = \frac{\text{Variabilité expliquée par le modèle/Nombre de paramètres du modèle}}{\text{Variabilité résiduelle/Degrés de liberté résiduels}}$$
$$= \frac{SSR/p}{SSE/n-p-1}$$

- SSR : est la somme des carrés des régressions (Sum of Squares for Regression), qui mesure la variabilité expliquée par le modèle.
- SSE : est la somme des carrés des erreurs (Sum of Squares for Error), qui mesure la variabilité non expliquée par le modèle.
- p : est le nombre de paramètres du modèle (y compris l'intercept).
- n : est le nombre total d'observations.
- $n - p - 1$:sont les degrés de liberté résiduels.

Calcul des Composantes de F

Somme des Carrés Totale (SST) :

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

où $\bar{y}$ est la moyenne des valeurs observées y_i.
Somme des Carrés des Régressions (SSR) :

$$\text{SSR} = \sum_{k=1}^{n} (\hat{y}_k - \bar{y})^2$$

où $\hat{y}_k$ est la valeur prédite par le modèle pour l'observation k.
Somme des Carrés des Erreurs (SSE) :

$$\text{SSE} = \sum_{k=1}^{n} (\hat{y}_k - y_k)^2$$

Relation entre les Sommes des Carrés :

$$\text{SST} = \text{SSR} + \text{SSE}$$

Interprétation de la Statistique F

Sous l'hypothèse nulle, les valeurs prédites sont toutes nulles, donc la variabilité expliquée est également nulle ($F = 0$) ce qui indique que le modèle n'est pas significatif.

Une valeur élevée de la statistique F indique que la variabilité expliquée par le modèle est grande par rapport à la variabilité non expliquée, ce qui suggère que le modèle est significatif. Sachant que F peut prendre d'autres formes quand on change de modèle, les calculs faits correspondent à un cas particulier d'un régresseur linéaire.

$\mathbf{P}_{\text{value}}$

La $\mathbf{P}_{\text{value}}$ associée au test F indique la probabilité d'observer les données si H_0 est vraie. Une $\mathbf{P}_{\text{value}}$ faible suggère que le modèle global est significatif. Généralement, on compare la $\mathbf{P}_{\text{value}}$ avec 0,05 (Accepter un risque d'erreur de première espèce de 5%) selon le critère suivant :
Si $\mathbf{P}_{\text{value}} \leq 0{,}05$:

On rejette l'hypothèse nulle et donc le modèle est significatif.

Si $\mathbf{P}_{\text{value}} > 0{,}05$:

On ne rejette pas l'hypothèse nulle et le modèle n'est pas significatif.

Intervalle de Confiance

L'intervalle de confiance à 95% pour un coefficient donne une plage de valeurs dans laquelle le vrai coefficient est susceptible de se trouver avec une probabilité de 95%. Si cet intervalle ne contient pas zéro, cela renforce l'idée que le coefficient est significatif.

Conclusion

La statistique F est un outil puissant pour évaluer la signification globale d'un modèle de régression. Une valeur élevée de F, accompagnée d'une $\mathbf{P_{value}}$ faible, suggère que le modèle explique une part significative de la variabilité des données,indiquant que le modèle est pertinent.

6.4 Test de Significativité des Coefficients

Le test décrit précédemment permet de juger de la qualité globale du modèle et d'avoir une idée de la pertinence des estimateurs des coefficients en même temps. On peut descendre à l'échelle des coefficients et tester individuellement la pertinence de chaque variable. Cela permet de voir l'effet de la variabilité de cette variable sur la variance de la target et de déterminer si cette variable ajoute une information utile au modèle ou non.

Pour cela on procède à des tests d'hypothèse sur les coefficients individuels, exactement comme on l'avait fait avant en testant pour chaque coefficient β_i l'hypothèse nulle suivant :

$$H_0 : \beta_i = 0 \tag{6.2}$$

La statistique utilisée ici est :

$$t - Student = \frac{\hat{\beta}_k}{\sigma_{\hat{\beta}_k}} \tag{6.3}$$

La P_{value} associé s'interprétent exectement de la même manière que dans la partie précédente mais ici pour une seule variable et non pas pour l'ensemble du modèle.

6.5 L'indicateur R-squared

Il s'agit d'une métrique qui renseigne sur la performance globale du modèle. Il indique la proportion de la variance de la target qui est expliquée par les variables du modèle, calculé comme suit :

$$\begin{aligned} R^2 &= 1 - \frac{SS_{res}}{SS_{tot}} \\ &= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \end{aligned}$$

- SS_{res} Est la somme des carrés des résidus (ou erreurs)

- SS_{tot} Est la somme des carrés totaux

Le R^2 varie entre 0 et 1, Un R^2 élevé indique que le modèle explique bien la variance des données. Par contre un R^2 faible indique que la relation entre les variables indépendantes et la variable dépendante est faible ou non linéaire.

Choix du Modèle

La méthodologie adoptée consiste à tester un nombre de modèles et garder le meilleur au sens des critères d'évaluation détaillés avant. Une fois ce modèle trouvé, le ranking des opérations-étapes-équipements va se baser sur les coefficients de ce modèle. Plus le coefficient est grand plus l'OST correspondante génère le plus d'erreurs et vice-versa. Cela entraîne une classification des OST, qui va aboutir par la suite au meilleur chemin selon la règle suivante : Pour chaque couple opération & étape on préfèrera le passage par l'équipement générant le moins d'erreurs potentielles.

7.1 Régression Linéaire

Dans le contexte de la régression linéaire la fonction g sera égale à la fonction identité

$$g(\mu) = \mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_N X_N \tag{7.1}$$

Résultats de l'application de ce modèle :

OLS Regression Results

Dep. Variable:	sum(count_dies)	**R-squared:**	0.144
Model:	OLS	**Adj. R-squared:**	0.060
Method:	Least Squares	**F-statistic:**	109.1
Date:	Wed, 10 Jul 2024	**Prob (F-statistic):**	0.00
Time:	14:48:00	**Log-Likelihood:**	-7491.0
No. Observations:	848	**AIC:**	1.514e+04
Df Residuals:	771	**BIC:**	1.550e+04
Df Model:	76		
Covariance Type:	HC0		

Le R^2 est faible : 14,4 % de la variance de la Target est expliquée par les variables La statistique F et la P_{value} montrent que le modèle est significatif en globalité. Mais cela ne suffit pas, vu qu'il y a peu de variables significatives **Table 7.1.1** P>|z| est grande pour la majorité des coefs.

	coef	std err	z	P> |z|	[0.025	0.975]
const	786.1000	295.795	2.658	0.008	206.353	1365.847
3322-3130-TEL41	-463.7312	408.817	-1.134	0.257	-1264.998	337.536
3331-3000-EHE1	283.7314	133.396	2.127	0.033	22.280	545.183
3331-3000-VHE1	245.0739	188.977	1.305	0.192	-123.516	613.662
3332-3000-VHE1	-66.0685	161.313	-0.410	0.682	-382.236	250.099
3334-3000-MCU12	-12.2831	203.439	-0.060	0.952	-466.315	441.749
3334-3000-MCU5	89.1639	231.807	0.385	0.700	-365.169	543.497
3334-3500-FSIF	-166.7592	165.096	-1.010	0.312	-490.342	156.823
3334-3500-FSIH	-2051.1582	562.847	-3.644	0.000	-3154.398	-947.918
3334-3500-FSIK	220.0466	227.197	0.969	0.333	-225.251	665.344
3334-3500-FSIS	-303.6494	149.175	-2.036	0.042	-623.028	-38.271
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 7.1.1: lLinear Regression Coefficients

Les variables **3331-3000-EHE1**, **3334-3500-FSIH** et **3334-3500-FSIS** sont celles qui apportent le plus d'informations au modèle, il y a d'autres qui ne sont pas affichées dans cette table !
Les valeurs de *std err* sont importantes vise à vise des valeurs des coeff; cela suggère que l'estimation n'est pas bonne.
A partir de ces résultats on voit que ce modèle particulier n'est pas efficace. *Cela veut-t-il dire que la famille des modèles de regression linéaire n'est pas adaptée à nos données ou bien juste que ce modèle en particulier est mauvais et il faut le réajuster d'avantage pour améliorer la performance?*
Pour répondre on va tester la validité de quatre hypothèses cruciales sur les données, si ces hypothèses ne sont pas valides alors la régression linéaire est à exclure vu que la relation entre les variables indépendantes et la target n'est pas linéaire.

Hypothèse de linéarité

Si le modèle est bien spécifié (relation linéaire entre les variables et la target) on s'attendait à avoir un nuage de point aléatoirement distribué autour de la ligne rouge. Aucun pattern ne doit apparaitre.

Mais dans la figure **Figure 7.1.1**(voire page suivante) les points ont tendance à se situer au-dessus d'une droite oblique , cela suggère la présence d'un lien non linéaire dans les données.

Hypothèse de normalité des résidus

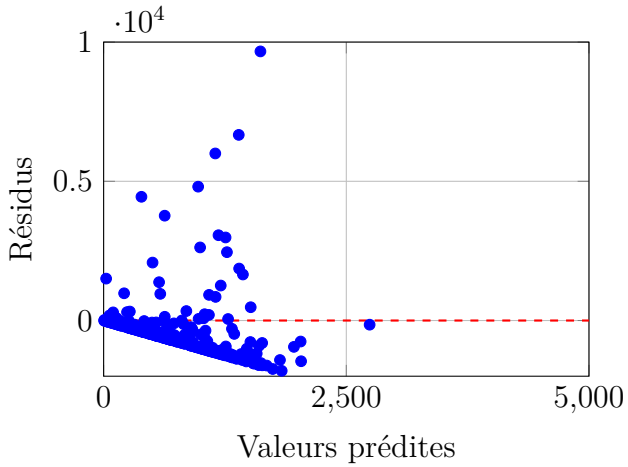


Figure 7.1.1: Nuage de résidus en fonctions des valeurs prédites

Cette hypothèse dit que les résidus suivent une loi normale si la relation entre X et target est linéaire, afin de tester si les résidus sont normalement distribués on trace un graphique PP **Figure 7.1.2** (probabilité-probabilité). En représentant les centiles de la distribution des résidus en fonction des centiles de la distribution théorique. La droite rouge permet de comparer ; si le tracé est sur cette ligne alors les résidus sont bien normales s'il y'a un écart les résidus ne suivent pas une loi normale.

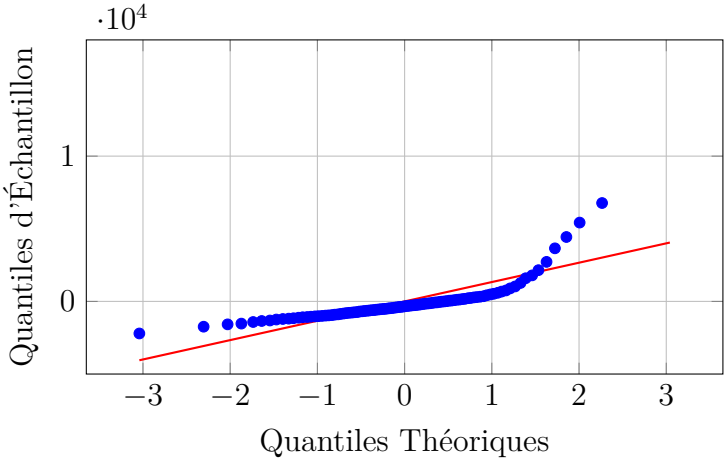


Figure 7.1.2: QQ plot des résidus

On voit bien, sur la **Figure 7.1.2**, un écart entre la courbe bleue et la ligne rouge. Cela montre que les résidus ne sont pas normalement distribués.

Hypothèse d'homoscédacité

L'hypothèse d'homoscédasticité stipule que la variance des erreurs (résidus) est constante lorsque X augmente. En d'autres termes, les résidus doivent avoir une variance uniforme à travers toutes les valeurs prédictives. Si cette hypothèse est violée, cela signifie que les erreurs varient systématiquement avec les valeurs des variables indépendantes, ce qui peut biaiser les résultats des tests statistiques et les intervalles de confiance. Pour cela on utilise le test d'hypothèse de **Breusch Pagan** défini comme suit:

Hypothèse nulle H0: Les variances d'erreur sont égales.
Hypothèse alternative: Les variances d'erreur ne sont pas égales.

Si la P_{value} est inférieure à 0,05 nous rejetons l'hypothèse nulle. Ainsi, nous rejetons l'hypothèse selon laquelle les variances d'erreur sont égales.
Résultats obtenus:

Statistique de test de Breusch-Pagan : 60.149036089170274

P_{value} du test de Breusch-Pagan : 0.90868587798434
Conclusion:
On ne rejette pas $H0$, la variance des erreurs est bien constante en augmentant X.

Hypothèse de multicolinéarité

La multicolinéarité se produit lorsque les variables indépendantes sont corrélées et que la variation de l'une entraîne la variation de l'autre dans le même sens. La multicolinéarité nuit à la significativité des paramètres estimés par le modèle. Si beaucoup de variables sont corrélées, les résultats de la statistique de Student issus de l'estimation de ces variables ne sont pas pertinents.

Pour cette hypothèse, on peut calculer la matrice des corrélations. Généralement, on définit un seuil, par exemple 0.7, et on élimine une composante des paires dont la corrélation dépasse ce seuil.

Conclusion 1:

Vu que la majorité des hypothèses de la regression linéaire sont violées. La régression linéaire sera exclue, elle n'est pas adaptée aux données traitées. Elle pose un problème de spécification du modèle, dans la suite on va continuer de tester une autre famille de modèle tout en faisant des tests de spécification du modèle.

7.2 Régression de Poisson

Ce qui nous'a guidé vers le choix de cette regression est la distribution de la target qui plus ou moins s'identifie à une distribution de poisson.

La modélisation du nombre moyen d'erreurs dans une machine par une loi de Poisson est tout à fait appropriée dans de nombreux contextes. La loi de Poisson est une distribution de probabilité discrète qui décrit le nombre d'événements se produisant dans un intervalle de temps ou d'espace fixe T, selon un taux r lorsque ces événements se produisent avec une moyenne constante $E[Y] = rT$ et indépendamment les uns des autres.

Dans le contexte d'une régression de Poisson, les nombres négatives n'ont pas de sens et la régression donnera toujours des valeurs positives.
La fonction g sera la fonction logarithme et μ sera la rate r:

$$g(r) = \ln r = \beta_0 + \beta_1 X_1 + \cdots + \beta_N X_N \tag{7.2}$$

Ce que l'on modélise est le log de la rate, donc une fois les coefficients sont trouvés l'estimation de la rate sera égale :

$$\hat{r} = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_N X_N} \tag{7.3}$$

Lorsque X_i varie de 1 le $\ln r$ varie de β_i et donc, r varie de e_i^{β} qui est le taux relatif de changement de r dû à la $i - ème$ machine.
Résultats de l'application du modèle:

Poisson Regression Model Results			
Dep. Variable:	sum(count_dies)	**No. Observations:**	848
Model:	GLM	**Df Residuals:**	771
Model Family:	Poisson	**Df Model:**	76
Link Function:	Log	**Scale:**	1.0000
Method:	basinhopping	**Log-Likelihood:**	-6.5801e+05
Date:	Fri, 12 Jul 2024	**Deviance:**	1.3122e+06
Time:	11:26:29	**Pearson chi2:**	2.41e+06
No. Iterations:	100	**Pseudo R-squ. (CS):**	1.000
Covariance Type:	HC3		

Bien que le modèle présente un excellent ajustement global avec un pseudo R^2 de 1, indiquant une très bonne qualité d'ajustement, il est crucial de noter que la variance des données vaut 3225358 est beaucoup plus élevée que la moyenne 636, alors que dans le cas d'une vraie distribution de poisson ces deux quantités sont égales ce qui suggère que le modèle de Poisson simple pourrait ne pas être approprié et que la valeur du R^2 soit erronée.

	coef	std err	z	P> |z|	[0.025	0.975]
const	6.1386	0.483	12.712	0.000	5.192	7.085
3322-3130-TEL41	-5.0792	0.836	-6.116	0.000	-7.892	-4.066
3331-3000-EHE1	0.3944	0.202	1.889	0.059	-0.015	0.803
3331-3000-VHE1	0.4593	0.352	1.280	0.201	-0.231	1.413
3332-3000-VHE1	-0.0993	0.250	-0.397	0.691	-0.550	0.391
3334-3000-MCU12	0.0551	0.291	0.189	0.850	-0.516	0.629
3334-3000-MCU5	0.0998	0.343	0.291	0.771	-0.573	0.773
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 7.2.2: Poisson Regression Coefficients

La valeurs élevée de de Pearson chi2 par rapport au nombre de degrés de liberté Df Residuals suggère également une surdispersion.

Cette surdispersion peut imposer l'utilisation d'un modèle plus complexe, tel qu'une distribution négative binomiale, ou quasi-poisson, la surdispersion observée doit être adressée pour améliorer la pertinence et la précision du modèle.

Il y a deux autres modèles qui sont la Négative Binomial Regression et la Quasi-Poisson Regression qui peuvent être utilisés lorsque la condition de moyenne et variance sont égales n'est pas vérifiée. La sélection se fait selon la règle expliquée dans le tableau suivant :

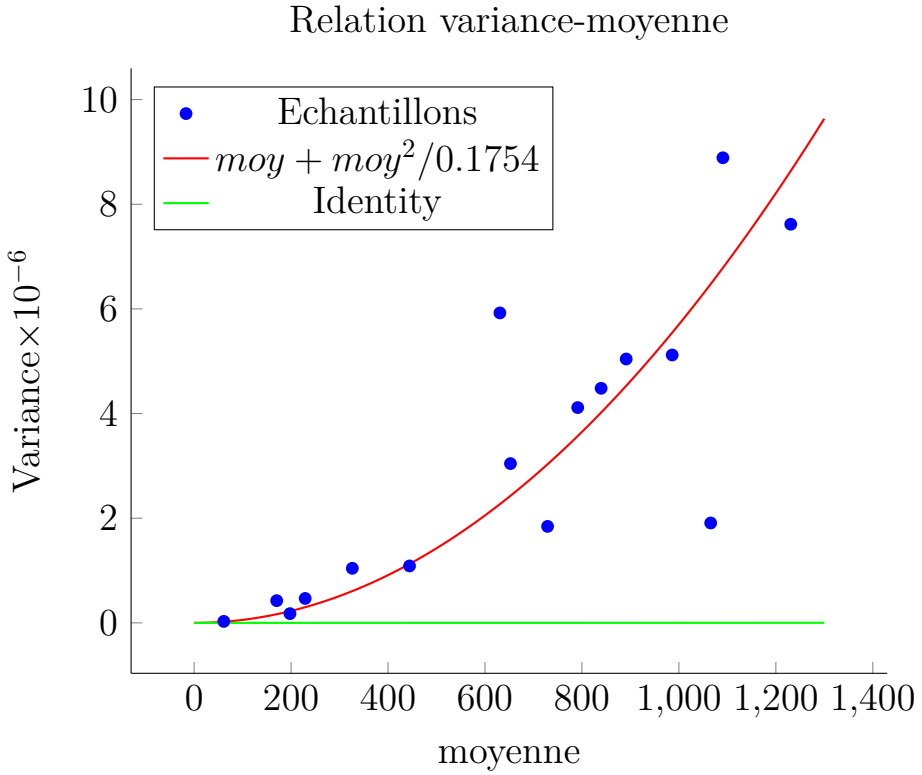
Sous-dispersion	**Moyenne=Variance**	**Surdispersion**
Quasi-poisson	**Poisson**	**Quasi-poisson/ Négative Binomial**
Variance = $K \times$ Moyenne Avec ($K < 1$) Utiliser le modèle Quasi-Poisson dans le cas d'une sous-dispersion lorsque la variance est proportionnelle à la moyenne avec un facteur $K < 1$.	Utiliser le modèle de Poisson quand la moyenne et la variance sont égales	Deux cas se présentent : **Une Surdispersion linéaire** Variance = $K \times$ Moyenne Avec ($K > 1$) dans ce cas utiliser un modèle quasi-poisson comme la sous-dispersion. **Une surdispersion quadratique** Variance = moy + $\frac{\text{moy}^2}{k}$ Dans ce cas on utilise une négative binomiale régression

Table 7.2.3: Tableau comparatif des modèles selon l'état de dispersion

Nous nous positionnons bien dans la 3-ème colonne (variance=3225358 et moyenne=636), mais dans quel type de surdispersion! *s'agit-t-il d'une*

surdispersion linéaire ou bien quadratique ?

Pour cela on va partitionner nos données Y de la target en sous-parties de même taille. Pour chaque partie, nous allons évaluer la moyenne et la variance. Ensuite un scatter plot permettra d'identifier la relation entre les deux quantités. Voici la courbe obtenue :



La première chose qui attire l'attention est la pente de l'identité, qui montre à quel point les variances dépassent les moyennes. Si la surdispersion était linéaire la pente de la courbe rouge ne devrait pas beaucoup s'écarter de l'identité. Pourtant on voit que les points bleus s'ajustent bien via la courbe en rouge, qui est quadratique, avec un facteur de surdispersion $k = 0.1754$. Ce dernier a été trouvé par l'optimiseur de la régression binomiale négative, qui sera abordée par la suite.

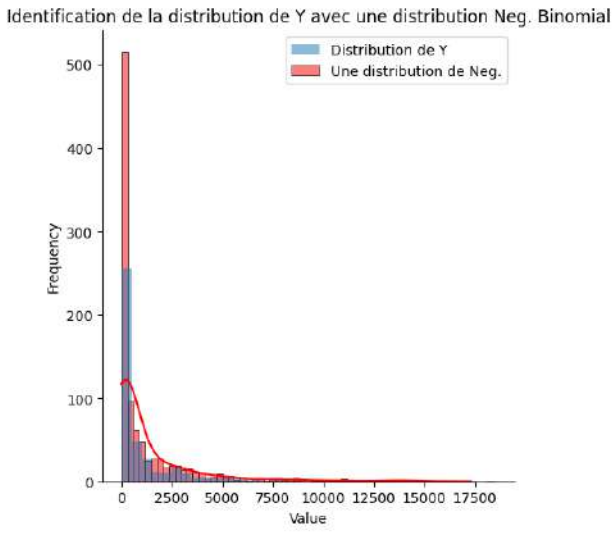
Conclusion 2

Cette deuxième étude a développé la régression de Poisson, une famille de modèles plus adaptée qu'une régression linéaire lorsqu'on traite des données de comptage. Dans le cas étudié, un simple modèle de Poisson ne s'applique pas correctement en raison du problème de surdispersion quadratique. L'étude réalisée nous oriente vers l'application du modèle de régression binomiale négative.

7.3 Régression Binomiale Négative

En raison de l'impossibilité d'ajuster un modèle de Poisson avec les données disponibles, en raison de la non-validité de l'hypothèse variance = moyenne, le recours à la régression binomiale négative permet un ajustement des données lorsque la relation variance-moyenne est quadratique, de la forme :

$$variance = moyenne + \frac{moyenne^2}{k} \tag{7.4}$$



Une variable aléatoire de distribution binomiale négative modélise le nombre d'échecs avant d'atteindre un certain nombre de succès dans une série d'essais de Bernoulli indépendants. Les paramètres de cette distribution sont :

- n : Le nombre de succès à atteindre (paramètre de forme).
- p : La probabilité de succès lors de chaque essai.

La figure ci-dessus, permet de visualiser la distribution de la target Y en même temps que celle d'une vrai variable aléatoire suivante une loi de négative binomiale, les paramètres n et p de cette loi sont déduites à partir des valeurs de la target pour être le plus proche de celle-ci.

L'équation d'une negative binomial régression est la même que celle d'une régression de poisson. La seule différence est la prise en compte de la nouvelle relation quadratique entre mean et variance.

Résultats d'application de la régression binomiale négative :

Les résultats de la régression binomiale négative montrent que plusieurs variables explicatives ont un impact significatif sur le nombre moyen de défauts **Table 7.3.4**. Par exemple, la variable **3331-3000-EHE1** a un coefficient égale à 0.5618 et une $P_{value} = 0.008$, son utilisation implique donc une décroissance du rendement du bin23. Pour mieux comprendre cet impact, nous pouvons comparer deux groupes : le groupe 0 (Baseline), des lots dont la fabrication

Negative Binomial Regression Results			
Dep. Variable:	sum(count_dies)	**No. Observations:**	848
Model:	Negative Binomial	**Df Residuals:**	786
Method:	MLE	**Df Model:**	61
Date:	Tue, 16 Jul 2024	**Pseudo R-squ.:**	0.01910
Time:	14:47:00	**Log-Likelihood:**	-4503.3
converged:	True	**LL-Null:**	-4591.0
Covariance Type:	HC3	**LLR p-value:**	5.253e-13

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	6.5223	0.468	13.932	0.000	5.605	7.440
3331-3000-EHE1	0.5618	0.213	2.638	0.008	0.144	0.979
3331-3000-VHE1	0.6780	0.254	2.670	0.008	0.180	1.176
3332-3000-VHE1	-0.2611	0.222	-1.177	0.239	-0.696	0.174
3334-3000-MCU12	-0.1007	0.274	-0.367	0.714	-0.438	0.639
3334-3000-MCU5	0.3806	0.336	1.130	0.258	-0.479	1.040
3334-3500-FSIF	-0.4300	0.302	-1.423	0.155	-1.022	0.162
3334-3500-FSIK	0.8794	0.411	2.142	0.032	0.075	1.684
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 7.3.4: Résultats de la régression binomiale négative

utilise la machine **EHE1** à l'étape **3000** de l'opération **3331**, et le groupe 1 (Treatment), des lots dont la fabrication n'utilise pas cette machine à cet endroit. En moyenne, l'utilisation de la machine augmente le nombre moyen de bins23 de $e^{0.5618}$ fois par rapport au treatment. Cela revient à dire que l'utilisation de cette machine à cet endroit augmente en moyenne 1.75 fois ($e^{0.5618=} = 1.75$) le nombre moyen de bins23 par rapport au treatment.

Le modèle a un log-vraisemblance de -4503.3, et le test de log-vraisemblance (LLR) est significatif avec une P_{value} de 5.253e-13, indiquant que le modèle est globalement significatif.

Les P_{values} associées aux coefficients montrent que certaines variables sont statistiquement significatives, tandis que d'autres ne le sont pas. Cela est dû aux limitations détaillées dans le paragraphe « Les défis du problème » que l'on ne peut pas résoudre. Cela étant dit, les résultats du modèle retenu sont considérés comme des approximations, et les machines reconnues comme défectueuses ne sont mauvaises que statistiquement. Leur comportement peut évoluer ou dépendre d'autres variables manquantes, ce qui requiert une validation des résultats par les experts du processus de fabrication.

Conclusion 3

Parmi les trois modèles entraînés : régression linéaire OLS (ordinary least squares), régression de Poisson et régression binomiale négative, c'est ce dernier qui s'ajuste le mieux aux données. Il sera retenu par la suite et le classement des chemins se fera sur la base de ses résultats.

Golden route

Pour les 61 variables du modèle (à l'exception de la constante qui ne correspond à aucune machine réelle et qui ne fera donc pas partie du Golden Route), seront calculés : une estimation du coefficient de la régression, son écart type, la $P - value$, et les bornes de l'intervalle de confiance.

Selon la hiérarchie des opérations, des étapes et des équipements, certaines étapes n'ont qu'une seule machine possible, rendant le passage du lot par cette machine inévitable, quelle que soit la valeur du coefficient associé. Cependant, d'autres étapes permettent aux lots de choisir parmi plusieurs machines disponibles. Dans ce cas, et dans le cadre de la recherche du Golden Route, on préférera celle avec le coefficient le plus petit possible pour l'opération-étape considérée.

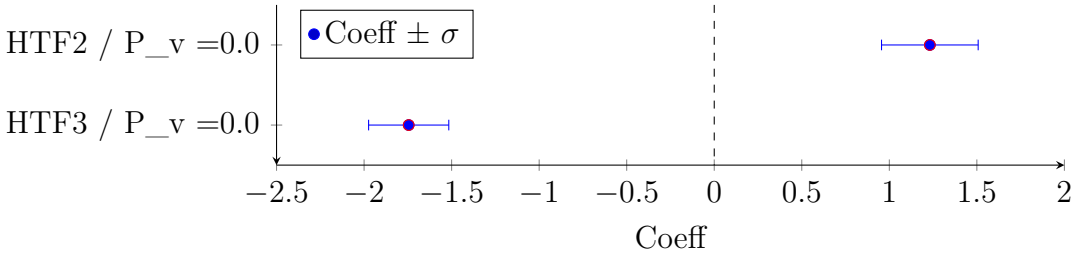


Figure 8.0.1: Coefficient estimation for machines of OP-STEP 3418-3000 : Depot SiGe

Ici **Figure 8.0.1**, en bleu est la valeur du coefficient, qui servira pour classer les machines: plus le Coeff est petit plus la machine est bonne. Elle est représentée aussi la plage des variations potentielles(écart-type) du coefficient. Devant le nom de chaque machine est calculée la P_{value} qui renseigne sur la significativité de la variable pour le modèle. Les résultats pour cet exemple montrent qu'à l'étape 3000 du dépôt SiGe de l'opération 3418, il sera mieux vise à vise du rendement de bin23 de prendre la HTF3, vu de la significativité des deux machines, et du fait que les valeurs des Coeffs avec leurs marges d'erreurs sont bien séparées.

Une autre figure **Figure 8.0.2** (voir page suivante) consiste au tracé sur une échelle temporelle, du nombre de bin23 total à la fois pour les lots qui ont traversé la HTF2 et les lots qui ont traversé la HTF3 à l'étape 3000 du dépôt SiGe de l'opération 3418. Il faut noter que ces valeurs de bin23 ne sont

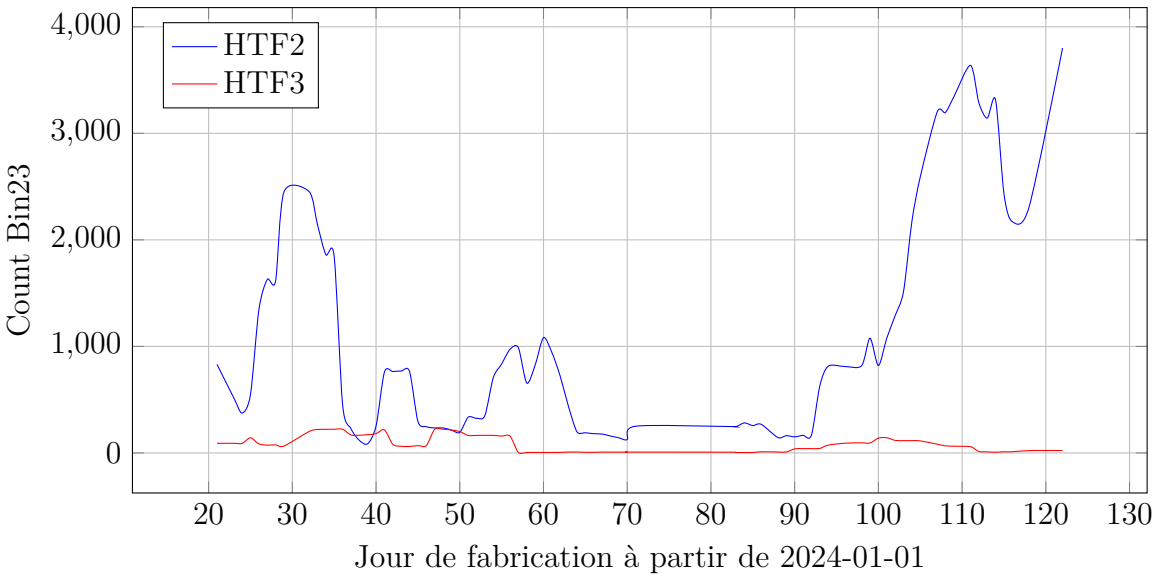


Figure 8.0.2: Count dies per day and per tool for OP_Step: 3418-3000: Depot SiGe

pas forcément directement générées au cours de cette étape. Ces valeurs sont calculées tout à la fin du processus. Cependant, cela permet de constater que les lots avec le plus de bin23 sont le plus souvent ceux qui sont passés par la machine HTF2 à l'étape 3000 de l'opération 3418. D'après les résultats du modèle, c'est bien la machine HTF2 qu'il faut éviter.
Malheureusement, cette distinction n'est pas toujours possible avec certitude pour plusieurs raisons : lorsque les performances des machines sont similaires, lorsque les marges d'erreur se chevauchent trop, ou encore lorsque les machines ne sont pas significatives pour le modèle.

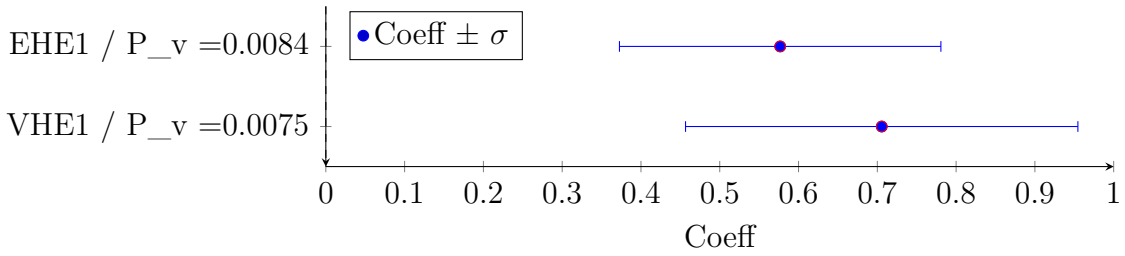


Figure 8.0.3: Coefficient estimation for machines of OP-STEP 3331-3000 : Implant.Phos

Ici, par exemple, **Figure 8.0.3** les deux coefficients sont très proches et, en prenant en compte les écarts types des estimateurs des coefficients, les deux intervalles rouges se superposent et se confondent quasiment.
Même les séries temporelles associées **Figure 8.0.4** (voir page suivante) ne peuvent servir à séparer ou classer ces deux machines selon leur performance.

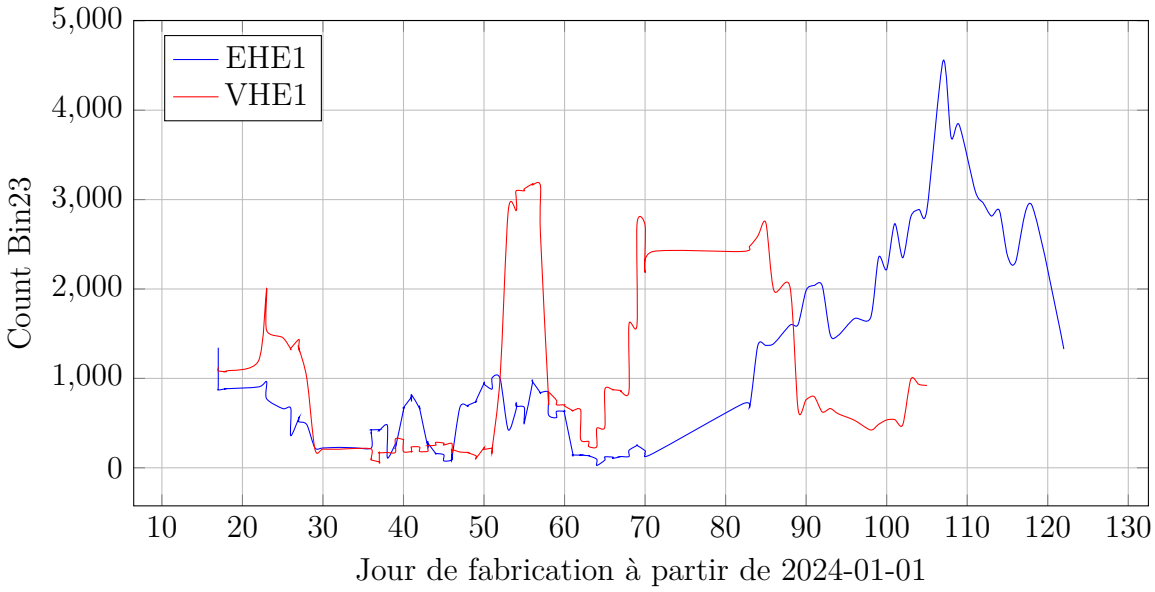
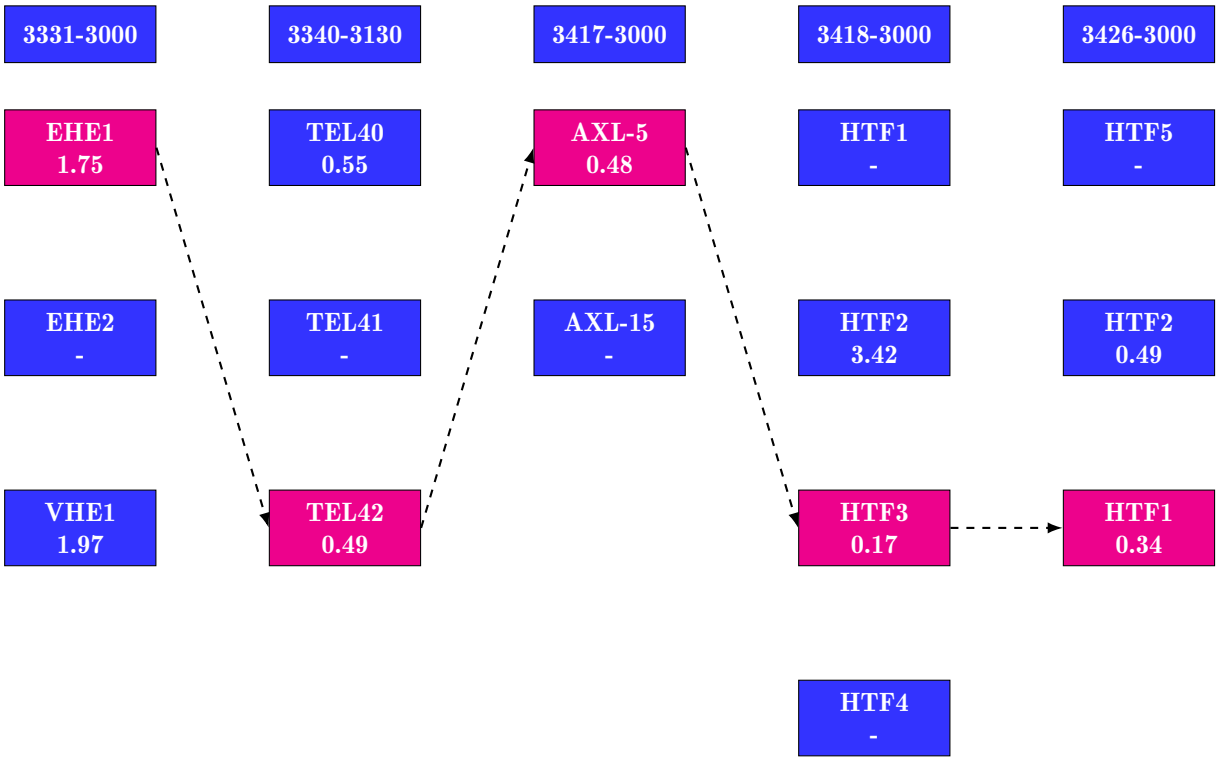


Figure 8.0.4: Count dies per day and per tool for OP_Step: 3418-3000: Depot SiGe

A la fin de notre analyse on est capable de déduire un meilleur chemin en récupérant la meilleure machine de chaque étape.



Les opérations-étapes sont classées dans l'ordre chronologique du processus. En bas de chaque opération-étape est affiché l'ensemble des machines possibles, la machine à prendre en rouge et l'exponentiel du coefficient pour quantifier l'effet de cette machine sur la moyenne des erreurs. Un tiret signifie l'absence

du coefficient dans le cas où la machine a été éliminée durant les étapes du prétraitement des données.
Cette méthode de la recherche du Golden Route présente des limitations :

Les variables du modèle ne sont pas toutes significatives. Certaines variables peuvent ne pas avoir un impact statistiquement significatif sur le rendement, ce qui complique l'interprétation des résultats et la prise de décision basée sur ces variables. Cela signifie que certaines machines ou étapes peuvent ne pas être clairement identifiées comme bonnes ou mauvaises en termes de performance.

Pour certaines opérations-étapes, les machines montrent des performances similaires, rendant le choix difficile. Lorsque les coefficients associés à différentes machines sont proches et que leurs intervalles de confiance se chevauchent, il devient compliqué de déterminer laquelle est la meilleure ou la pire. Cette similitude dans les performances peut rendre la sélection de la machine optimale moins évidente.

La séquence des machines trouvée représente une partie de l'ensemble des machines formant le meilleur chemin. Le modèle peut identifier une série de machines qui semblent optimiser le rendement, mais cette séquence n'est qu'une partie du chemin global. Il est important de considérer que d'autres machines et étapes, non incluses dans cette séquence, peuvent également jouer un rôle crucial dans la performance globale du processus.

Conclusion

Cette étude a permis d'explorer et de modéliser le processus de fabrication des wafers chez STMicroelectronics, en se concentrant sur l'identification du Golden Route, c'est-à-dire le chemin optimal à travers les différentes machines et étapes du processus.

Nous avons commencé par une phase de prétraitement des données. Ensuite, nous avons évalué trois modèles de régression: la régression linéaire OLS, la régression de Poisson et la régression binomiale négative, pour finalement retenir cette dernière en raison de sa meilleure adéquation aux données.

Les résultats obtenus ont mis en évidence les machines et étapes ayant un impact significatif sur le rendement final des wafers, notamment en termes de bin23. Nous avons pu identifier les machines à éviter, comme la machine HTF2 à l'étape 3000 de l'opération 3418, et celles à privilégier en fonction de leurs coefficients de régression. Cependant, cette distinction n'a pas toujours été possible avec certitude, en raison de performances similaires entre certaines machines et de la présence de variables non significatives.

En prenant du recul sur la méthodologie, il est important de noter que la complexité du processus de fabrication et la présence de multicolinéarité ont posé des défis majeurs. La régression binomiale négative s'est avérée être un outil précieux pour traiter la surdispersion des données, mais les limitations inhérentes à la collecte et au traitement des données, ainsi que la non-significativité de certaines variables, ont affecté la précision des résultats. De plus, l'absence de tests intermédiaires a rendu difficile l'évaluation de l'impact de chaque machine sur le rendement final.

Pour conclure, cette étude ouvre la voie à de futures recherches et améliorations. De plus, l'exploration de techniques de machine learning plus avancées pourrait offrir de nouvelles perspectives pour optimiser le processus de fabrication. Enfin, une collaboration avec les experts du processus de fabrication est

essentielle pour valider les résultats et affiner les modèles, afin de garantir une amélioration continue de la pertinence des résultats.

Références Bibliographiques

[1]:Lecture 9: *Modelling Counts* https://personalpages.manchester.ac.uk/staff/mark.lunt/R_course/9_counts/practical.html. Accessed 30 July 2024.

[2]:Gupta, S., & Hasenbein, J. J. (2024). *Ranking Routes in Semiconductor Wafer Fabs*. Operations Research & Industrial Engineering Graduate Program, The University of Texas at Austin.

[3]:Lee, C.-H., Lee, D.-H., Bae, Y.-M., & Kim, K.-J. (2017). *Determining Golden Process Routes in Semiconductor Manufacturing Process for Yield Management*. Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, Korea; Division of Interdisciplinary Industrial Studies, Hanyang University, Seoul, Korea.

[4]:Jiang, Dan, et al. *A Novel Framework for Semiconductor Manufacturing Final Test Yield Classification Using Machine Learning Techniques*. IEEE Access, vol. 8, 2020, pp. 197885–95. DOI.org (Crossref), https://doi.org/10.1109/ACCESS.2020.3034680.

[5]:Jiang, Dan, et al. *A Gaussian Mixture Model Clustering Ensemble Regressor for Semiconductor Manufacturing Final Test Yield Prediction*. IEEE Access, vol. 9, 2021, pp. 22253–63. DOI.org (Crossref), https://doi.org/10.1109/ACCESS.2021.3055433.