



Année scolaire 2016 - 2017

PROJET STATISTIQUE DE 2^e ANNÉE

Sujet n° 3

Analyse exploratoire de l'absentéisme et des arrêts de travail dans
une population de conducteurs

Réalisé par :

Patrick Boucly

Patrick Jabot

Régis Relland

Arnaud Wilczynski

Sous la direction de :

Denis Coutrot

Directeur du département Data Science de Transdev

Remerciements

Nous tenons à remercier notre tuteur, M. Denis Coutrot, pour nous avoir accompagnés et conseillés tout au long de ce projet. Nous avons également apprécié le recul par rapport à la technique statistique apporté par notre coach en communication, M^{me} Dominique Legoubey.

Enfin, nous adressons nos remerciements à l'équipe enseignante de l'Ensaï et tout particulièrement à MM. Guillaume Chauvet et François Coquet ainsi qu'à M^{me} Brigitte Gelein pour leur disponibilité et leur appui technique.

Abstract

The aim of this study is to analyze the absenteeism in Transdev, a French bus company. We rely on a database that contains information on bus drivers work schedule between January 2015 and December 2016. To carry out this study, we need a clear definition of the notion of absenteeism. We rely on a definition by the Anact¹ definition which tries to make a link between working conditions and absenteeism. That's why some causes of work absence will not be considered as absenteeism (training programmes, day of strike action, parental leave, ...) in this survey. After a database cleansing work, we use descriptive statistics to figure out which element or variable characterizes most clearly absenteeism of bus driver. This first approach gives us some interesting results. Absenteeism rate in the database is 9.6 %. Men, who represent 80 % of the 1 567 bus driver of our database, have a lower absenteeism rate than women. Absenteeism rates by agency are clearly different. Sunday is the day with the higher absenteeism rate, more than 30 %.

Next, we use three types of regression to try to understand the relationship between the explanatory variables and absenteeism. On the first one, we work on a logistic regression. The response variable is binomial : the driver was never missing versus was missing at least one day. The aggregate database used for this regression contains one observation per driver. Even if this dichotomous analysis might seem too simple, it confirms some information found on descriptive statistics : all else being equal, being a woman, seniority in the enterprise, or a high frequency of worked on Sundays increases the probability of missing work at least once.

On the second regression, we transform the response variable on a categorical variable with 3 values (depending of the absenteeism frequency : the driver was never missing, was missing but less than 5 % of the time or more than 5 %). This transformation gives a more nuanced picture of the consequences of the explanatory variables on absenteeism.

The third regression we apply is a GEE (Generalized Estimating Equation) regression. We regroup schedule information on a period of time we define. Here, the response variable is binary like in the first model but one observation matches with one period for one driver. Information is much less aggregated than it is on the other models. This approach provides a possibility of a better understanding of the relationship between work schedule and absenteeism. One of the main lesson learnt from this model is that a large modification of working hours increases the probability of being absent.

1. **Agence nationale pour l'amélioration des conditions de travail**, national agency for improvement of working conditions

Table des matières

Remerciements	1
Abstract	2
Introduction	7
Sous le capot de la base de données	8
La base de données initiale	8
La mise en forme et l'apurement des données	8
L'imputation en cas de non réponse partielle	9
Les différentes bases constituées	9
Tableau de bord de l'absentéisme	11
Saisir deux concepts mouvants par des indicateurs chiffrés	11
Proposer une définition partagée de l'absentéisme	11
Qualifier les « conditions de travail »	11
Construire des indicateurs à partir de la base de données	12
Décrire la population de conducteurs	12
Les hommes embauchés récemment représentent 40,1 % des agents	13
Des « conditions de travail » marquées par les contraintes de service	13
Un taux d'absentéisme global de 9,6 % sur les deux années	17
Un taux d'absentéisme supérieur à 30 % pour l'activité le dimanche	19
Travailler le week-end, principal frein à l'assiduité	21
Comment mesurer l'effet d'une variable sur l'absentéisme ?	21
Le travail le week-end comme facteur d'absentéisme	22
Les plannings, un levier pour atténuer l'absentéisme	28
Un modèle qui nuance les résultats précédents	28
Les fluctuations de planning accroissent l'absentéisme	32
Choix de la période de référence optimale	32
Les variations de planning aggravent l'absentéisme	35
Pistes d'amélioration et à explorer	37
Voies d'amélioration	37
Pistes non développées	38
Conclusion	40
Bibliographie	42

Annexes		43
1	Taux d'absentéisme par variable d'intérêt	43
2	Création de la base binaire	45
3	Classification CART	46
4	Méthodes de discrétisation supervisées utilisées	46
5	Corrélations entre les variables explicatives	47
6	Note méthodologique : test du χ^2 d'indépendance entre l'ancienneté et l'amplitude horaire	48
7	Validité du modèle binaire	49
8	Test de Hosmer-Lemeshow	50
9	Modèle GEE	50

Table des figures

1	Bases de données constituées pour l'analyse exploratoire et les régressions .	10
2	Les causes potentielles de l'absentéisme selon l'Anact	12
3	Dispersion de l'amplitude moyenne d'une journée de travail par conducteur par agence Transdev	14
4	Proportion des journées de travail selon leur heure de début	15
5	Proportion des journées de travail selon leur heure de fin	16
6	Taux d'absentéisme par mois entre janvier 2015 et décembre 2016	18
7	Odds Ratios, modèle binaire	24
8	Courbe ROC, modèle binaire	26
9	Classification CART selon la modalité d'absence dans la base binaire	27
10	Odds Ratios, modèle polytomique	29
11	Classification CART selon la modalité d'absence dans la base polytomique .	31
12	Odds Ratios, modèle GEE7	33
13	Comparaison des courbes ROC	36
14	Durée d'absence par épisode	39
15	4 profils selon 4 variables d'intérêt	39
16	Points de vigilance pour un planning optimisé	40
17	Nombre de jours de présence des agents durant la période étudiée	46
18	Exemple de discrétisation CART sur l'ancienneté	47
19	Matrice de corrélation des variables explicatives	48

Liste des tableaux

1	Ancienneté dans l'entreprise en années par sexe du conducteur	13
2	Taux d'absentéisme selon le sexe de l'agent	19
3	Taux d'absentéisme selon l'amplitude moyenne de la journée de l'agent . . .	19
4	Taux d'absentéisme selon la journée de la semaine	20
5	Légende, modèle binaire	24
6	Odds ratios, modèle binaire	25
7	Matrice de confusion, modèle binaire	26
8	Matrice de confusion, modèle polytomique	30
9	Comparaison de l'efficacité prédictive des modèles GEE selon la période de référence	32
10	Légende, modèle GEE7	33
11	Odds ratios, modèle GEE7	34
12	Taux d'absence selon le taux de travail le samedi, modèle GEE à 90 jours .	35
13	Taux d'absence selon la variation d'amplitude horaire, modèle GEE à 7 jours	36
14	Matrice de confusion, modèle GEE à 7 jours	36
15	Taux d'absentéisme selon la classe d'âge de l'agent	43
16	Taux d'absentéisme selon l'agence territoriale Transdev de l'agent	43
17	Taux d'absentéisme selon la proportion de travail matinal de l'agent	43
18	Taux d'absentéisme selon le taux de travail dominical de l'agent	44
19	Taux d'absentéisme selon la journée de la semaine	44
20	Taux d'absentéisme selon l'ancienneté en classes de l'agent	44
21	Taux d'absentéisme selon le régime horaire de l'agent	44
22	Taux d'absentéisme selon la durée moyenne de travail effectif par jour de l'agent	45
23	Taux d'absentéisme selon la période de vacances scolaires	45
24	Tableau des effectifs observées	48
25	Tableau des effectifs théoriques	48
26	Test de significativité de chaque variable, modèle binaire	50

Introduction

En mars 2015, la société de transports Transdev a présenté à ses cadres un plan d'action volontariste contre l'absentéisme. Structuré autour de quarante actions issues des bonnes pratiques observées localement, ce projet identifie quatre leviers de progression : la culture managériale, le poste de travail, l'organisation du travail et l'environnement socio-économique du collaborateur. L'absentéisme constitue en effet un défi majeur dans les services aux personnes et notamment dans les transports collectifs. Véronique Subileau, DRH du groupe Transdev, estimait ainsi en 2015, dans un entretien au magazine *Liaisons sociales*, que « *le coût [du phénomène] s'élevait à 50 millions d'euros par an, ce qui représente 250 bus à l'arrêt* ».

Au-delà même de ces coûts directs, des éléments moins tangibles, comme la dégradation du climat social ou de l'image de l'entreprise, renforcent la nécessité de lutter contre l'absentéisme. Celui-ci ne constitue pas seulement, d'ailleurs, une problématique pour l'entreprise mais pour la société toute entière puisque la plupart des transports urbains sont financés par la collectivité, dans un contexte économique contraint, comme le rappelait la Cour des Comptes dans un rapport de 2015.

Ce constat dressé et partagé, encore faut-il saisir les racines de ce comportement. Car l'absentéisme est évidemment un phénomène complexe, qui mêle des composantes personnelles et professionnelles. En première approche, il est probablement plus simple, pour une entreprise, de peser sur les facteurs d'organisation du travail. Mais faut-il se limiter aux recettes qui marchent, déjà retenues par le plan d'action de 2015 ? Denis Coutrot, directeur du département Data Science de Transdev, suggère une démarche complémentaire avec ce sujet en proposant de scruter la base de données décrivant les horaires de travail des conducteurs franciliens. Avec l'idée que certaines conditions de travail, plus contraignantes, pourraient probablement être associées statistiquement à un absentéisme plus marqué.

Autrement dit, les conditions de travail d'un conducteur peuvent-elles être précisément qualifiées à partir de son planning sur une période de deux années ? Et comment mesurer l'influence qu'elles exercent sur l'absentéisme ?

Explorer ce fichier, c'est donc tout d'abord se doter des outils conceptuels et techniques qui permettent de comprendre et de mesurer l'absentéisme et les contraintes exercées par les conditions de travail d'un service de transports urbains. **Décrire cette population** est une seconde étape qui devrait suggérer des liens, encore lâches, entre des formes de travail et des absences. **Modéliser ces relations** à travers les techniques de régression, c'est vérifier la robustesse de ces liens et tenter de leur donner un sens. La voie est tracée, si vous voulez bien nous suivre : en voiture...

Sous le capot de la base de données

L'analyse de l'absentéisme au sein de la société Transdev repose sur une base de données fournie par notre tuteur. Ces fichiers étaient initialement destinés à des traitements internes de gestion des ressources humaines. Des opérations de mise en forme, d'apurement, d'imputation et d'agrégation des informations ont donc été nécessaires pour mener à bien cette étude.

La base de données initiale

La base de données initiale, fournie par le commanditaire, est constituée de deux fichiers de format bureautique. Un fichier intitulé « *Idf Ouest_absenteism.csv* » comporte 986 774 observations caractérisées par 18 variables. Ces premières données quantitatives détaillent, pour chaque jour et pour les 2 063 agents en poste, le statut et les « conditions de travail » entre le 1^{er} janvier 2015 et le 29 décembre 2016. Elles permettent donc, dans une certaine mesure, de qualifier une journée de travail, de repos ou d'absence pour les employés des entités Transdev de l'ouest de l'Île-de-France.

En revanche, la seconde source, le fichier « *Agents.xls* », propose, sous une forme très succincte et anonymisée, des informations individuelles (date de naissance, dates d'entrée et de sortie de l'entreprise, sexe) pour 9 920 employés. Au total, la fusion des deux tables autorise, d'une part, un suivi des journées de travail des agents et, d'autre part, une analyse des caractéristiques des employés.

La mise en forme et l'apurement des données

À la demande du commanditaire, la création et l'analyse des bases de données ont essentiellement été réalisées avec le logiciel libre R. Ce travail a ainsi été l'occasion de découvrir les packages de manipulation de données (dplyr) et de création de graphiques (ggplot2).

La **mise en forme** comprend l'ensemble des opérations qui rendent la base utilisable sous un format brut sans tenir compte de la cohérence interne des données. Pour cette étude, la mise en forme s'est réduite au formatage des variables existantes et à la création de nouvelles variables. Le *formatage des variables* a essentiellement consisté à transformer les variables qualitatives en facteurs (codage en variables catégorielles) et à prendre en compte les formats spécifiques pour certaines données (les dates notamment).

La *création de nouvelles variables* a été, en revanche, plus intense, notamment pour préciser les conditions de travail et de l'absence. Dans un premier temps, il a fallu déplier l'information inscrite dans les horaires. À partir d'une heure de début et de fin de poste, des indicatrices de travail matinal, vespéral et nocturne ont été créées. Par ailleurs, les paramètres continus d'amplitude et de durée de travail effectif ont nécessité la constitution de classes pour densifier l'information disponible et permettre des croisements efficaces entre les données. Un fichier externe a par ailleurs été constitué précisant, pour chaque date, le jour correspondant (lundi, mardi...) et l'existence d'un férié officiel ou de vacances scolaires.

Fusionnées aux données Transdev, ces données ont permis de caractériser le travail en fin de semaine, les jours fériés et pendant les vacances scolaires. Enfin, des variables indicatrices des différentes

situations d'absence ont été élaborées pour bien identifier le phénomène. Au final, après ces traitements préliminaires, la base initiale a été enrichie d'une quarantaine de variables.

L'apurement des données désigne l'ensemble des opérations qui s'attachent à la pertinence des données au regard d'une problématique fixée.

Dans un premier temps, la définition du champ de l'étude (suivi de l'absentéisme dans une population de conducteurs) a conduit à supprimer tous les agents Transdev qui n'étaient pas des conducteurs dans la base fusionnée des horaires et des agents. Cette seule opération a conduit à la suppression de 170 000 lignes et 487 agents. Cette première base comportait alors 818 711 lignes correspondant aux horaires de travail de 1 576 conducteurs de Transdev franciliens. Trois colonnes inexploitable (convention collective) ou redondantes (code de l'entité managériale ou statut de conducteur) ont également été supprimées.

L'apurement plus fin des données pouvait débiter. Un contrôle de cohérence de calcul a tout d'abord été réalisé sur une variable (amplitude horaire) sans identifier d'erreurs. Des échanges avec le commanditaire nous ont permis d'identifier des valeurs incohérentes sur certaines variables saisies manuellement dans la base qui ont fait l'objet d'une annulation (mise en données manquantes pour des années de naissance erronées) ou d'un écretement (amplitude horaire et durée de temps de travail effectif respectivement bornées à 18 et 12 heures). La création d'une variable dédiée à l'apurement, stockant dans une liste les ou l'ancienne(s) information(s) avant modification, permet d'estimer le nombre de corrections de la base de données à ce stade à environ 20 000 (soit 2 % des observations).

L'imputation en cas de non réponse partielle

À l'issue des traitements de mise en forme des données, la qualité de la base peut notamment être évaluée selon son exhaustivité c'est-à-dire par le nombre d'observations caractérisables à partir de l'ensemble des variables. Certaines techniques statistiques (régression logistique ...) nécessitent d'ailleurs que les observations soient intégralement renseignées, tout au moins pour les variables d'intérêt, sans quoi elles sont tout simplement écartées.

À ce stade, notre base de données ne comporte pas de non réponse totale mais certaines variables ne sont pas systématiquement saisies. En cas de non réponse partielle, une méthode courante est l'imputation d'une valeur plausible. Une première étape a été menée de concert avec les essais de régression logistique. Il s'agissait d'identifier les variables devant prioritairement faire l'objet d'une imputation. À l'issue de cette phase, huit variables ont été sélectionnées pour leur importance même relative dans les modèles pressentis. Les données manquantes correspondant à ces caractères ne représentent que 4,5 % des observations, essentiellement du fait de la variable précisant l'âge du conducteur. Le choix d'imputation a été d'affecter une valeur plausible à la variable âge (qui représente 82,7 % des observations manquantes) et d'omettre les autres observations manquantes, notamment l'absence.

La méthode retenue est déterministe, utilisant les variables sexe, ancienneté et agence pour constituer des cellules de « comportement homogène ». Ainsi, l'âge affecté à une individu correspond à l'âge moyen des individus de sexe, ancienneté et agence d'appartenance identiques.

L'imputation réalisée, une suppression des observations comportant encore des valeurs manquantes a été opérée, finalisant l'apurement. La base obtenue n'est toutefois pas encore exploitable pour notre étude, elle décrit toujours des journées de travail alors que nous devons caractériser l'absentéisme des agents.

Les différentes bases constituées

Constituer une base utilisable a essentiellement consisté à agréger les données accessibles pour chaque journée de planning en informations disponibles pour chaque conducteur. Cette première opé-

ration réalisée, une base de données dédiée a été créée en entrée des régressions avec des sélections d'observations distinctes. La figure 1 décrit les différents fichiers construits et leurs principales caractéristiques.

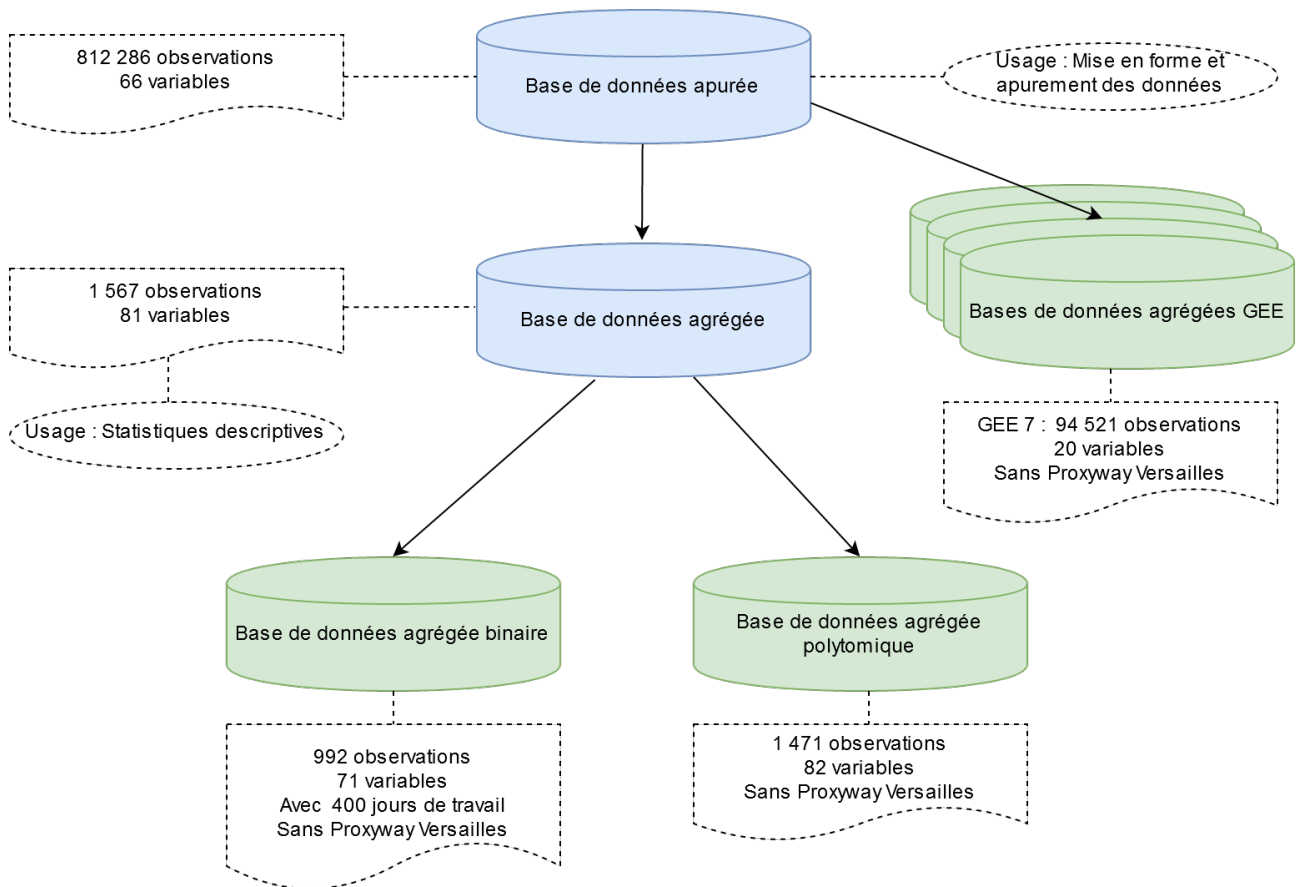


Figure 1 – Bases de données constituées pour l'analyse exploratoire et les régressions

Pourquoi sélectionner les agents ayant travaillé 400 jours dans la base binaire ?

La base destinée à la régression logistique binaire ne comporte que des individus durablement « installés » dans l'entreprise. Les statistiques produites dans cette base ne portent en effet que sur les **400 premiers jours** d'activité dans l'entreprise des agents ayant travaillé au moins 400 jours pendant la période. Cette restriction est notamment nécessaire pour saisir l'absence des conducteurs sur une durée comparable. Une justification plus détaillée du choix du seuil de 400 jours est proposée dans l'annexe 2.

Pourquoi écarter les agents de ProxyWay Versailles des bases des modèles ?

Toutes les bases utilisées par les régressions omettent par ailleurs les agents de l'entité **ProxyWay Versailles**. Cette suppression s'est imposée en pratique dès lors que les spécificités des conducteurs de cette agence ont apparu. Il s'agit d'une entité de petite taille, employant uniquement des hommes avec peu d'ancienneté. À 400 jours d'activité, seuls six agents pourraient être intégrés à la base binaire et ces seules observations influencent fortement les résultats de la régression binaire. Ils ne sont donc pas intégrés aux modèles mis en oeuvre mais quelques statistiques générales sont disponibles sur cette agence dans le Tableau de bord de l'absentéisme et en Annexes.

Tableau de bord de l'absentéisme

Avant même de proposer un modèle analysant les liens entre l'absentéisme et les « conditions de travail », l'exploration de la base de données dresse un premier tableau de la population des conducteurs de Transdev. Cette découverte des fichiers est guidée par une caractérisation du sujet, compatible avec le cadre posé par les données et la demande du commanditaire.

Saisir deux concepts mouvants par des indicateurs chiffrés

Évidente en apparence, la question de l'absentéisme dans l'entreprise, et son lien avec les « conditions de travail », se heurte à une définition plus précise des termes, nécessaire à la quantification. Ces deux notions peuvent en effet être décrites de manière très variable alors même que la base de données restreint le champ de perception de ces phénomènes.

Proposer une définition partagée de l'absentéisme

L'absentéisme recouvre dans le langage courant « *le fait d'être absent du lieu du travail ou de tout lieu où la présence est obligatoire*² ». Cette définition très générale présente notamment le défaut de placer toutes les absences sur le même plan. Or, certaines absences au travail sont nettement plus préjudiciables au fonctionnement quotidien de l'entreprise et au service qu'elle dispense aux clients ou aux usagers.

L'Agence nationale pour l'amélioration des conditions de travail (Anact) a proposé en 2016 une description plus opérationnelle qui met l'accent sur certaines absences en lien avec les conditions de travail. « *L'absentéisme caractérise toute absence qui aurait pu être évitée par une prévention suffisamment précoce des facteurs de dégradations de travail entendus au sens large*³ ». Certaines absences sont ainsi exclues de la définition (formations, grève, maternité et paternité...) qui est centrée sur quatre catégories d'événements : *les absences pour maladie, les accidents de travail (et de trajet), les maladies professionnelles et les absences injustifiées*.

Ce cadre de travail, assez rapidement partagé avec notre commanditaire, implique en revanche de pouvoir caractériser assez finement les conditions de travail.

Qualifier les « conditions de travail »

Le terme de « conditions de travail » est également très vaste puisqu'il peut tout à la fois désigner l'environnement, l'organisation ou même le contenu du travail que les effets qu'il produit sur la santé et l'épanouissement personnel et professionnel du salarié. Schématiquement, les conditions de travail recouvrent donc des paramètres organisationnels, normatifs, managériaux ou techniques qui contribuent à favoriser l'engagement qu'un employé peut déployer dans un emploi.

L'étude étant fondée sur une base de données brutes détaillant essentiellement les horaires quotidiens des conducteurs de bus ainsi que leurs repos et absences, les conditions de travail seront ici principalement saisies sous leur forme organisationnelle. Les éléments plus subjectifs, relevant du management ou des moyens techniques ou humains, ne pourront être perçus, en revanche, à travers le

2. *Encyclopédie Larousse en ligne*, <http://www.larousse.fr/encyclopedia/divers/absentéisme/17921>

3. *Anact*, L'absentéisme : Outils et Méthodes pour agir, 2016, p.3

fichier dont nous disposons. Nous n'aborderons donc directement qu'une seule des causes potentielles de l'absentéisme représentées dans la figure ci-dessous.



Figure 2 – Les causes potentielles de l'absentéisme selon l'Anact

Cette restriction peut être relativisée si l'on retient que les données disponibles permettent de qualifier assez finement l'organisation du travail. Les horaires d'activité quotidienne autorisent notamment une approche du travail en termes d'amplitude, de morcellement (nombre d'interruptions d'un service) ou de contraintes spécifiques (poste de matin, de soir...). De la même façon, le planning peut être analysé à la lumière des obligations (travail le week-end, lors des vacances scolaires) tandis que quelques informations plus succinctes enrichissent l'examen d'informations sur le contrat de travail (type de régime horaire). Bien plus, la mention du contenu d'une journée (travail, repos, absence) ainsi que la justification des absences peuvent concourir à caractériser le rythme de travail d'un employé (alternance d'activité et de repos par exemple).

Au final, l'organisation du travail est donc envisagée de manière assez globale pour espérer dégager une caractérisation de l'absentéisme par ces paramètres. On s'appliquera toutefois à construire dans un premier temps des indicateurs d'absentéisme opérationnels.

Construire des indicateurs à partir de la base de données

Il n'existe pas de consensus sur les indicateurs d'absentéisme à utiliser pour mesurer ce phénomène. Toutefois, l'Anact présente un indicateur⁴ simple à comprendre et analyser qu'il est possible de construire à partir de notre base de données.

$$\text{Taux d'absentéisme} = \frac{\text{Nombre de jours d'absences pendant une période}}{\text{Nombre de jours théoriques pendant la même période}}$$

Les outils de mesure étant précisés, nous disposons des moyens nécessaires pour mieux appréhender la base de données mise à notre disposition.

Décrire la population de conducteurs

Après apurement et sélection des conducteurs dans la base de données, la population regroupe 1 567 agents employés par différents établissements de la société Transdev. Ils ont notamment exercé une

4. Anact, L'absentéisme : Outils et Méthodes pour agir, 2016, p.6

activité, quelquefois très ponctuelle, entre le 1^{er} janvier 2015 et le 29 décembre 2016 sur un territoire correspondant grossièrement à l'ouest de l'Île-de-France. Une description succincte des agents permet de mieux les connaître avant d'aborder la caractérisation de l'absentéisme.

Les hommes embauchés récemment représentent 40,1 % des agents

La base de données ne comporte qu'une poignée de variables sociodémographiques pour esquisser une première description de cette population. Mais le genre, l'âge ou l'ancienneté dans l'entreprise constituent déjà des éléments d'information sur les agents d'autant qu'il est possible de décliner ces données selon les agences territoriales qui les emploient.

Sans surprise, les conducteurs de bus de Transdev dans l'ouest de l'Île-de-France sont essentiellement des hommes (80,8 %) et la prédominance masculine est même encore plus marquée parmi les individus embauchés dans les trois dernières années (87,0 %). Une agence se distingue par une parité de conducteurs et de conductrices (Rambouillet) mais la plupart des entités franciliennes de Transdev examinées dans cette étude affichent un net déséquilibre entre les hommes et les femmes. Plusieurs agences comptent d'ailleurs moins de 10,0 % de personnels féminins parmi les conducteurs.

L'âge moyen d'un chauffeur dans notre base est légèrement supérieur à 42 ans même si un cinquième de la population est âgé de plus de 52 ans. Les femmes sont sensiblement plus âgées que les hommes notamment parce qu'elles sont moins nombreuses parmi les agents plus jeunes.

Les conducteurs recrutés récemment (entre 2013 et 2016) représentent 38,1 % de l'effectif et 64,3 % des agents ont au plus sept ans d'ancienneté dans la société. Ici aussi, les agences territoriales se distinguent par des âges et anciennetés moyens très variables. Pour ne citer que les extrêmes : l'agence de Rambouillet emploie ainsi des agents en moyenne plus âgés (45 ans) et plus installés dans l'entreprise (plus de 9 années d'ancienneté) tandis que les conducteurs de l'entité Autocars Tourneux sont en moyenne plus jeunes (39 ans), avec moins d'ancienneté (6 années).

	Ancien. [0,3]	Ancien. (3,7]	Ancien. (7,12]	Ancien. (12,41]
Hommes	629 (40,1 %)	232 (14,8 %)	195 (12,4 %)	210 (13,4 %)
Femmes	86 (5,5 %)	61 (3,9 %)	73 (4,7 %)	81 (5,2 %)

Tableau 1 – Ancienneté dans l'entreprise en années par sexe du conducteur

Note de lecture : 629 conducteurs travaillant dans l'entreprise Transdev depuis moins de 3 ans sont des hommes. Ils représentent 40,1 % des conducteurs de notre base de données.

Source : Base Absentéisme Transdev, Base agrégée par matricule, 1 567 agents

Au final, le profil le plus courant du conducteur Transdev des agences de l'ouest de l'Île-de-France est un homme, récemment embauché dans la société (moins de trois années d'ancienneté dans l'entreprise), avec un âge moyen de 37 ans.

Des « conditions de travail » marquées par les contraintes de service

Les « conditions de travail » sont essentiellement accessibles dans la base de données à travers le déroulement des journées et des semaines de travail. Dans le secteur des transports, les contraintes de service aux usagers impriment un rythme spécifique à l'activité quotidienne mais elles s'avèrent également exigeantes en termes de disponibilité pendant les périodes d'inactivité habituelles (week-end, fériés et vacances scolaires).

Une très forte proportion (93,8 %) d'agents travaille à temps complet. Seuls une cinquantaine sont employés à temps partiel, notamment les salariés travaillant comme « conducteur en période scolaire » (CPS), un dispositif spécifique pour les employés chargés des transports scolaires. Deux tiers de ces conducteurs scolaires sont employés par une seule entité (TD Houdan) avec une proportion de conductrices insolite (43,6 %). Ces salariés se démarquent également par un âge plus élevé (48,7 % de ces

chauffeurs scolaires ont plus de 52 ans) mais paradoxalement une très courte ancienneté chez Transdev (74,4 % ont moins de trois ans d'expérience dans l'entreprise).

L'amplitude de la journée de travail excède largement le temps de travail effectif, du fait d'interruptions du service. La moitié des conducteurs de notre population réalisent ainsi des journées de travail d'une durée de 9 à 11 heures alors que leur temps de travail effectif est plus classique, entre 7 et 8 heures. Cette différence conséquente s'explique notamment par des interruptions, légales ou de service, qui découpent la journée en vacations. En réalité, plus de 75 % des conducteurs ne réalisent en moyenne qu'une à deux vacations, une configuration somme toute assez proche d'un emploi de bureau. Le morcellement régulier des journées de travail ne concerne d'ailleurs qu'une dizaine de conducteurs travaillant dans une seule agence (TD Conflans) avec un nombre moyen de vacations compris entre 4 et 12. Même peu nombreuses, les pauses entre des vacations peuvent cependant être longues. Ainsi, 22,0 % des agents dont l'amplitude d'activité est supérieure en moyenne à 11 heures travaillent en temps effectif moins de 7 heures. Ces écarts extrêmes ne concernent toutefois qu'une part modeste de l'effectif total (5,3 %).

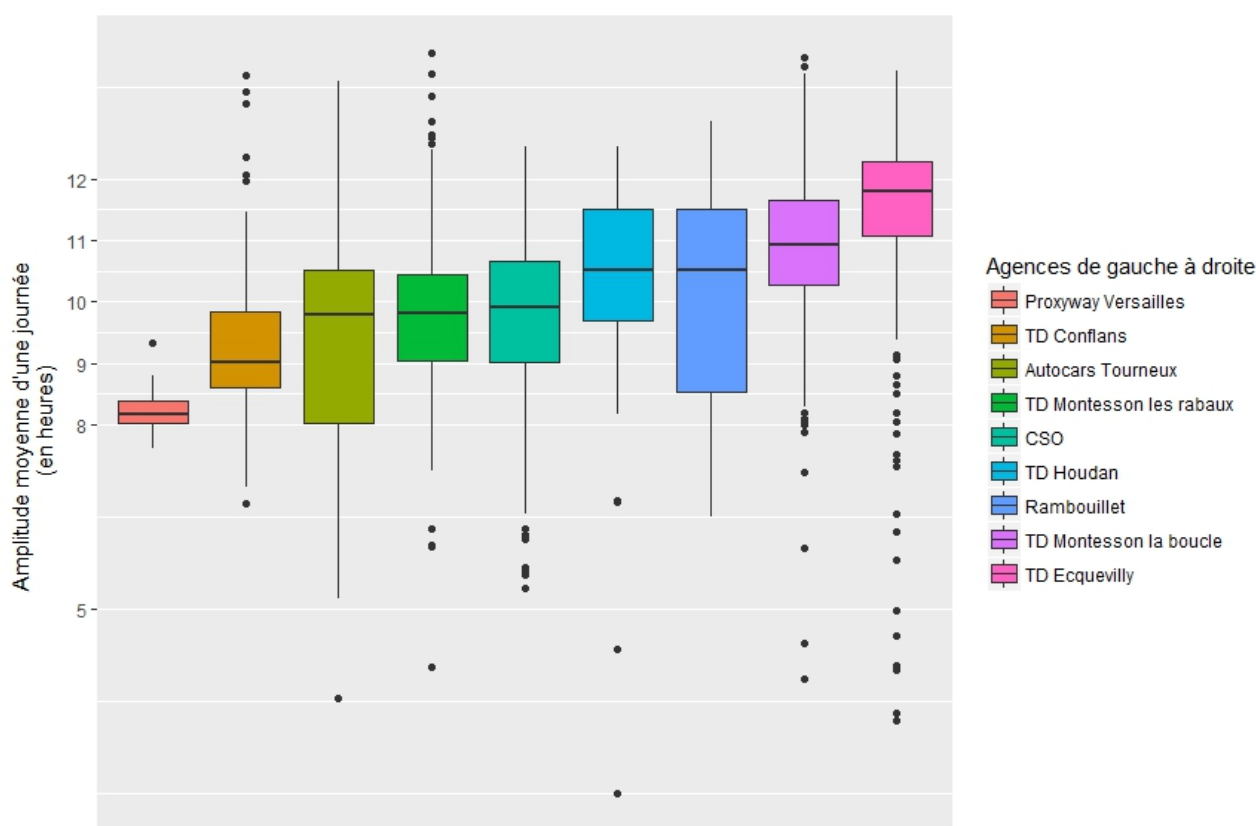


Figure 3 – Dispersion de l'amplitude moyenne d'une journée de travail par conducteur par agence Transdev

Note de lecture : L'amplitude moyenne d'une journée de travail est resserrée autour de 8 heures pour les conducteurs de l'agence de Versailles.

Source : Base Absentéisme Transdev, Base agrégée par matricule, 1 567 agents

L'amplitude de travail n'est pas seulement conséquente, elle est également relativement instable. Sur l'ensemble des journées travaillées, elle peut varier en moyenne de 2 heures et la durée d'activité effective, hors interruption, d'une heure. Les agents soumis à des fluctuations plus marquées ne se distinguent ici pas tant par le sexe, l'âge ou l'ancienneté mais par l'entité Transdev à laquelle ils appartiennent. Les conducteurs de l'agence Proxyway de Versailles réalisent ainsi des journées d'am-

plitude et de durée effective plutôt régulières, avec des différences inférieures à 2 heures en moyenne. En revanche, les agents de l'entité Autocars Tourneux et quelques employés de la société CSO sont soumis à des horaires beaucoup plus hétérogènes (Figure 3).

Les employés du secteur des transports doivent également composer avec des horaires décalés (matinaux, tardifs voire nocturnes), notamment en région parisienne. Cette contrainte quotidienne s'avère extrêmement forte puisque seule une poignée d'agents, de l'ordre de la dizaine, bénéficient au moins un jour sur deux d'horaires qu'on pourrait qualifier « de bureau » (début après 7 heures 30 et fin avant 18 heures 30). Les horaires décalés sont donc monnaie courante pour l'essentiel des agents de notre population. Pour 90 % d'entre eux, un travail de journée « classique » n'intervient au plus qu'une fois par mois en moyenne.

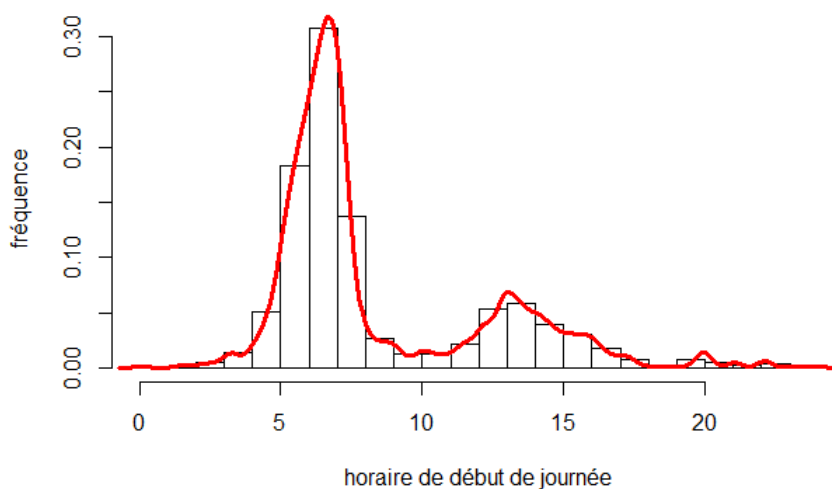


Figure 4 – Proportion des journées de travail selon leur heure de début

Note de lecture : Plus de 30 % des journées de travail débutent entre 6 heures et 7 heures du matin.

Source : Base Absentéisme Transdev, Base mise en forme et apurée, 812 286 observations

Premier type d'horaire décalé, l'entrée en poste matinale. Utilisée pour distinguer les agents notamment en termes d'absentéisme, elle a été définie dans cette étude de manière extrême. Il s'agit en effet d'une journée de travail qui débute avant 5 heures 15 du matin. Même dans cette acception, deux tiers des agents sont concernés au moins une fois sur les deux années. En retenant un seuil plus usuel (prise de poste avant 7 heures), la proportion de conducteurs concernés au moins une fois sur la période atteint 90,4 %. Ces chiffres recouvrent toutefois au moins partiellement une population de travailleurs matinaux occasionnels. En ne conservant que les conducteurs dont les journées commencées tôt représentent plus de 20 % de leur planning (au moins un jour par semaine en moyenne), on peut centrer l'analyse sur les travailleurs matinaux habituels qui constituent 11,2 % de l'effectif total. Essentiellement composée d'hommes (82,4 %) et concentrée dans quelques agences (19,8 % des effectifs de l'agence CSO et 14,2 % dans l'entité TD Ecquevilly), cette population se distingue également par une ancienneté moyenne dans l'entreprise supérieure à 9 ans.

Deuxième mode d'horaire atypique, les horaires tardifs de fin de poste. La définition retenue ici (journée terminée après 21 heures 45) est plus proche de l'acception commune. Cette contrainte de planning est un peu moins fréquente que la précédente puisque 26,3 % des agents ne l'ont jamais subie entre début 2015 et fin 2016. En conservant le seuil de 20 % d'horaires atypiques pour les distinguer des travailleurs tardifs occasionnels, les chauffeurs travaillant régulièrement le soir représentent 12,3 %

des effectifs de conducteurs. Essentiellement composée d'hommes (83,3 %) et mieux répartie entre les différentes entités Transdev franciliennes, cette population est nettement plus jeune, disposant d'une ancienneté moindre que celle des travailleurs du matin.

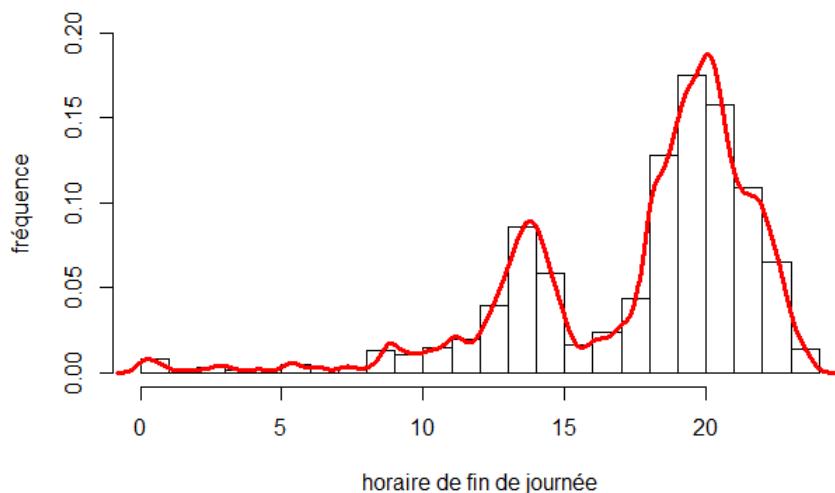


Figure 5 – Proportion des journées de travail selon leur heure de fin

Note de lecture : Plus de 15 % des journées de travail s'achèvent entre 19 heures et 20 heures le soir.

Source : Base Absentéisme Transdev, Base mise en forme et apurée, 812 286 observations

Dernière variante d'horaires décalés, sans doute la plus exigeante et qui recouvre au moins partiellement la précédente, le travail de nuit. À la définition légale, assez complexe et soumise à des variations selon les secteurs, on a préféré celle de l'enquête « Conditions de travail » plus facile à construire à partir de la base de données⁵. Relève alors du travail de nuit, une période de 9 heures consécutives au moins, incluant les heures comprises entre 24 heures et 5 heures et se situant dans l'intervalle 21 heures - 7 heures. Seuls 95 conducteurs (6,1 %) ont travaillé au moins une fois de nuit sur la période. Les chauffeurs réguliers de nuit (au moins un jour par semaine en moyenne) ne sont guère plus d'une vingtaine (1,4 %) presque tous employés par l'agence CSO.

Peu d'agents ne semblent finalement dédiés qu'à un type d'horaire atypique. Il est donc possible que certains agents cumulent certaines de ces contraintes, ce roulement constituant en soi une difficulté supplémentaire. La combinaison régulière de travail tôt le matin et tard le soir est ainsi réalisée par un trentaine de conducteurs (2,2 %). L'association répétée de travail le matin, le soir et la nuit est encore plus rare (8 agents) et uniquement pratiquée par l'agence CSO.

Ces contraintes ne s'expriment pas seulement sur le rythme quotidien, elles peuvent également concerner la fin de semaine, les jours fériés et les vacances scolaires. Travailler le week-end est somme toute assez courant : seuls 4,5 % des conducteurs de Transdev dans l'ouest de l'Île-de-France n'ont jamais travaillé en fin de semaine entre janvier 2015 et fin 2016. À l'inverse, 22,0 % des agents de notre base ont assuré leur service au moins un week-end par mois en moyenne sur la même période. Dans certaines agences territoriales, la proportion de conducteurs travaillant ainsi fréquemment en fin de semaine est substantielle : 41,8 % des agents de l'entité TD Montesson les rabaux et 30,3 % des chauffeurs de Proxyway Versailles assument ces plages d'activité au moins une fois par mois. Un service

5. *Direccte Nord-Pas-de-Calais, Mission Synthèse, Le travail en horaires décalés, Estimation des effectifs salariés concernés en région Nord-Pas-de-Calais, p.6, 2012*

de week-end bimensuel est plus rare puisqu'il ne concerne qu'une cinquantaine d'agents, essentiellement employés par les agences CSO et TD Ecquevilly. Ces agents sont probablement spécialement affectés à une grille de week-end puisque plus de 30 % de leur temps de travail est concentré sur la fin de semaine.

L'analyse de la fin de semaine comme un bloc noie toutefois les spécificités de chaque journée. En effet, travailler le week-end, c'est bien souvent travailler le samedi alors que l'activité le dimanche est loin d'être aussi systématique. 95,2 % des salariés de Transdev ont au moins assuré leur service une fois le samedi contre 71,3 % le dimanche. Le statut du samedi le rapproche d'ailleurs d'un jour « normal » : 55,5 % des conducteurs exercent leur activité en moyenne au moins un samedi par mois et 10,3 % travaillent un samedi sur deux. Le dimanche dispose en revanche d'une aura spécifique puisqu'il est moins fréquemment travaillé que les jours fériés. Prendre son service régulièrement le dimanche est d'ailleurs assez rare : 10,4 % des conducteurs de Transdev dans l'ouest de l'Île-de-France sont employés au moins un dimanche par mois et seulement 3,0 % assurent un service bimensuel le dimanche.

Travailler les jours fériés est en effet une pratique habituelle dans les secteurs de service au public : seuls 7,6 % des conducteurs n'ont pas subi cette contrainte entre janvier 2015 et décembre 2016. En moyenne, les agents ont travaillé 12 fériés sur les 26 jours fériés officiels possibles pour la période. Mais cette moyenne lisse des comportements très différents qui coupent la population en deux blocs. Un gros tiers des agents a travaillé moins de 15 jours fériés sur les deux années tandis qu'un deuxième tiers a travaillé entre 15 et 18 jours fériés, soit 7 jours fériés sur 10 sur cette période. C'est le cas notamment d'un agent sur deux de l'agence de Rambouillet et d'Autocars Tourneux.

Dernière contrainte éventuelle de planning : les vacances scolaires. Il s'agit ici de saisir dans quelle mesure les obligations de service pendant ces vacances pourraient impacter la vie des agents. Ce point est délicat car tous les salariés travaillent nécessairement assez fréquemment pendant les vacances scolaires qui couvrent grossièrement quatre mois de l'année. En moyenne, les agents de notre population ont d'ailleurs travaillé un jour sur deux pendant les vacances scolaires sur les deux années. Une stratégie peut dès lors consister à cibler les vacances de Noël, synonymes de congés en famille ou avec des proches, et les agents les plus jeunes. Premier constat : 13,2 % des conducteurs n'ont pas travaillé une seule journée pendant les vacances de Noël sur la période observée et parmi eux 49,8 % sont âgés de 35 ans au plus. En moyenne, les agents travaillent une des deux semaines pendant les vacances de Noël, des contraintes comparables à celles supportées par l'ensemble des salariés. Une autre perspective peut être adoptée en se concentrant sur les vacances d'été. En 2004, les Français de 35 à 39 ans prenaient en moyenne 40,0 % de leurs jours de vacances en août⁶. En comparaison, le mois d'août ne représente que 24,8 % des jours de vacances des agents de la même tranche d'âge de notre population. En considérant ces deux cribles très grossiers, il paraît bien difficile de trancher sur la contrainte exercée par le service aux usagers sur l'activité lors des vacances scolaires.

Au terme de cet examen de l'organisation du travail, les « conditions de travail » semblent bien marquées par les contraintes de service aux usagers. Ces exigences s'expriment notamment par des horaires décalés ou par une fréquence plus élevée de travail en fin de semaine ou les jours fériés. Des liens peuvent désormais être tissés entre les « conditions de travail » et les indicateurs d'absence. Des statistiques croisant les variables d'organisation du planning et l'absentéisme constaté pourraient ainsi permettre une première perception des relations entre les deux phénomènes.

Un taux d'absentéisme global de 9,6 % sur les deux années

Tous agents confondus, le taux d'absentéisme calculé selon la définition retenue s'élève à 9,6 % sur l'ensemble de la période. Il doit être interprété comme la proportion de jours d'absence pour certaines raisons (*les absences pour maladie, les accidents de travail et de trajet, les maladies professionnelles et*

6. *Le tourisme en France*, Les vacances des Français depuis 40 ans, Laurence Dauphin, Marie-Anne Le Garrec et Frédéric Tardieu, Insee, 2008, p.35

les absences injustifiées) parmi l'ensemble des jours devant être travaillés, exprimée en pourcentage. Mais il est tentant de comparer ce taux avec les chiffres disponibles sur l'absentéisme au niveau national.

Il se situe par exemple à un niveau nettement supérieur à celui proposé par l'Union des Transports Publics et ferroviaires (UTP) dans son « *Bilan social des transports urbains* » pour l'année 2015⁷. L'indicateur calculé par l'organisation syndicale s'élève à 7,4 %⁸ alors même qu'il n'intègre pas les absences injustifiées. Pour la même année, on obtient avec notre mode de calcul, à définition de l'absence identique, un taux de 9,0 %. Cette différence conséquente s'explique notamment par le choix du dénominateur : nous rapportons en effet le nombre de jours d'absence au nombre de jours normalement travaillés (travail et absence) en omettant les jours de repos (qui sont intégrés par l'UTP).

Une dernière statistique permettrait de situer les résultats obtenus pour notre population de conducteurs dans une perspective plus large. Il s'agit du nombre de jours d'absentéisme par agent pour une année, c'est-à-dire « *la moyenne des jours calendaires d'absence dans l'année pour l'ensemble des salariés des entreprises du panel* »⁹. Selon l'UTP, un salarié d'une entreprise de transport public urbain a été absent 29 jours en 2015, tous types d'absence confondus. Le même indicateur calculé à partir de notre population est de 27 jours d'absence par agent.

Il est également intéressant de saisir les composantes de cet indicateur. Les différents motifs d'absence ne présentent en effet pas le même poids dans le phénomène : sur la période entière, le taux d'absentéisme pour cause d'arrêts maladie (maladies professionnelles comprises) s'élève à 6,2 %. Les arrêts maladie représentent donc une part majoritaire des absences qui seront prises en compte dans ce travail, face aux absences pour accidents du travail (3,0 %) et aux absences injustifiées (0,4 %), nettement moins nombreuses.

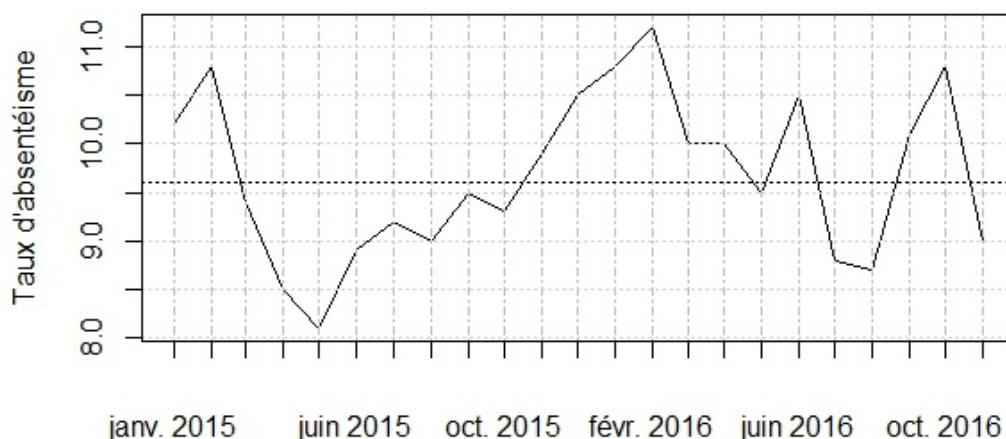


Figure 6 – Taux d'absentéisme par mois entre janvier 2015 et décembre 2016

Source : Base Absentéisme Transdev, Base mise en forme et apurée, 812 286 observations

Remarque : Par ailleurs, la base de données ne comporte pas toutes les observations du mois de décembre 2016 qui n'est donc pas représenté.

Déplions encore un peu plus notre indicateur en le déclinant selon des temporalités différentes. Tout d'abord, ce taux d'absentéisme, de 9,6 % sur une période qui s'étend de janvier 2015 à décembre 2016, est d'une relative stabilité si l'on considère chacune des années. Il atteint 9,4 % en 2015 et 9,7 % en 2016. La figure 6 fait toutefois apparaître des pics d'absentéisme en début et fin d'année (janvier et février notamment) tandis que la période estivale correspond à des taux nettement plus faibles.

Ce cadrage général sur l'absentéisme doit être complété par une mise en regard du phénomène avec

7. *Bilan social des transports urbains 2015*, UTP, Édition 2016, 56 pages

8. *Bilan social des transports urbains 2015*, UTP, Édition 2016, pp. 31-32

9. *Bilan social des transports urbains 2015*, UTP, Édition 2016, p. 31

les « conditions du travail » qui pourrait suggérer des liens entre une organisation atypique de l'activité et des absences plus nombreuses.

Un taux d'absentéisme supérieur à 30 % pour l'activité le dimanche

Le taux d'absentéisme semble varier, de façon quelquefois conséquente, selon les caractéristiques des agents mais également selon les horaires et contraintes d'activité qu'ils subissent.

	Taux d'absentéisme
<i>Sexe</i>	
Hommes	8,6 %
Femmes	12,9 %

Tableau 2 – Taux d'absentéisme selon le sexe de l'agent

Note de lecture : Selon la définition retenue, le taux d'absentéisme pour les hommes s'élève à 8,6 %

Source : Base Absentéisme Transdev, Base agrégée et apurée par matricule, 1 567 agents.

En première approche, les agents ne présentent pas les mêmes indicateurs d'absentéisme selon leur sexe, leur âge et l'agence de rattachement. D'une part, les femmes sont nettement plus absentes que les hommes et, d'autre part, le taux d'absentéisme décroît, de manière moins sensible, avec l'âge des agents. Les niveaux d'absentéisme entre les agences Transdev sont également très différents, avec des ordres de grandeur de un à quatre entre certaines entités¹⁰.

Les « conditions de travail » devraient également avoir une incidence sur le taux d'absentéisme des agents. Une première intuition consisterait à associer une forte amplitude de la journée de travail avec un absentéisme plus élevé. Cette relation n'est pas vérifiée : l'absentéisme décroît au contraire jusqu'à une amplitude de journée comprise entre 10,2 et 11,3 heures puis se redresse à 8,4 % pour les agents avec les journées en moyenne les plus longues.

	Taux d'absentéisme
<i>Amplitude moyenne journée</i>	
2 à 8,9 heures	10,7 %
8,9 à 10,2 heures	8,1 %
10,2 à 11,3 heures	5,2 %
11,3 à 14,1 heures	8,4 %

Tableau 3 – Taux d'absentéisme selon l'amplitude moyenne de la journée de l'agent

Note de lecture : Selon la définition retenue, le taux d'absentéisme pour les conducteurs dont la durée de la journée de travail est comprise en moyenne entre 2 et 8 heures 55 s'élève à 10,7 %

Source : Base Absentéisme Transdev, Base agrégée et apurée par matricule, 1 567 agents.

Dans le même ordre d'idée, on pourrait s'attendre à ce que l'absentéisme augmente à mesure que des horaires atypiques sont plus fréquemment réalisés. Les taux d'absentéisme pour le travail matinal et vespéral ne confirment pas cette analyse, ouvrant la voie à d'autres hypothèses comme celle, déjà évoquée, de l'importance du rythme régulier dans les horaires. Les agents répétant fréquemment un horaire atypique, même exigeant comme un poste matinal, pourraient ainsi être moins sujets à l'absentéisme.

Les indicateurs pointent, en revanche, de manière assez claire, la journée du dimanche. Non seulement, un horaire de dimanche sur trois est l'objet d'une absence mais même les agents travaillant

10. Des taux d'absentéisme pour certaines variables d'intérêt sont consultables en Annexes

	Taux d'absentéisme
<i>Journée</i>	
Lundi	8,2 %
Mardi	8,3 %
Mercredi	8,4 %
Jeudi	8,4 %
Vendredi	8,6 %
Samedi	14,8 %
Dimanche	35,8 %

Tableau 4 – **Taux d'absentéisme selon la journée de la semaine**

Note de lecture : Selon la définition retenue, le taux d'absentéisme s'élève à 8,2 % pour la journée du lundi en tenant compte de l'ensemble des agents

Source : Base Absentéisme Transdev, Base mise en forme et apurée, 812 286 observations

fréquemment ce jour-là sont plus régulièrement absents. En revanche, les vacances scolaires ne se distinguent pas vraiment en termes d'absentéisme : seules les vacances de février et de Toussaint présentent un taux d'absentéisme supérieur à celui constaté en moyenne sur la période.

En définitive, plusieurs caractéristiques des conducteurs mais surtout certains modes d'organisation de l'activité paraissent en mesure d'influencer le taux d'absentéisme. Mais un examen systématique des variables, les unes après les autres, ne constitue pas une garantie de la présence d'une relation avérée.

C'est tout l'intérêt des techniques qui seront mises en oeuvre par la suite de dégager la réalité et l'intensité de la relation qui unit, toutes choses égales par ailleurs, certaines de ses variables supposées explicatives avec l'absentéisme.

Travailler le week-end, un frein à l'assiduité

Les descriptions statistiques précédentes ont mis en lumière certaines caractéristiques de l'absentéisme des conducteurs de bus. Il reste maintenant à mettre en évidence les facteurs qui l'influencent. L'absence étant par définition binaire (chaque jour, un agent est soit présent, soit absent), nous avons choisi d'utiliser l'outil de la régression logistique pour objectiver l'influence de plusieurs variables que nous supposons explicatives du phénomène. Il va de soi que nous mesurerons uniquement l'influence des variables dont nous disposons et que nous ne pourrions étudier l'influence d'autres facteurs qui pourraient être pertinents (par exemple la situation familiale des agents, les lignes de bus utilisées, etc.) mais qui sont inaccessibles.

Comment mesurer l'effet d'une variable sur l'absentéisme ?

Choix de modélisation

Notre modélisation a pour objectif l'analyse de la variable à expliquer « Absence » par toute une liste de variables explicatives potentielles.

Dans ce premier modèle, l'**absence** ne compte que deux modalités :

- 0 si l'agent n'a jamais été absent pendant la période d'étude ;
- 1 si l'agent a été absent au moins une fois.

Les variables explicatives testées sont de natures différentes :

- Liées à l'individu : **âge**, **sexe** ;
- Liées à l'entreprise : **agence**, **type de régime horaire** ;
- Liées aux conditions de travail : **amplitude horaire** , **temps de travail effectif** , **nombre de vacances**, **horaire de début de journée**, **horaire de fin de journée**. Pour toutes ces variables, nous avons deux indicateurs : la **moyenne** et l'**écart type** sur la période. Enfin, nous avons ajouté le **taux de travail le samedi** (nombre de samedi au planning / nombre de samedi), le **dimanche** (nombre de dimanche au planning / nombre de dimanche), pendant les **vacances** (nombre de jours de vacances au planning / nombre de jours de vacances), afin de vérifier si le fait de travailler hors des périodes habituelles influait sur l'absentéisme.

Afin d'assurer une bonne comparaison entre les individus, nous avons dans ce premier modèle réduit le champ de l'étude aux **400 premiers jours** d'environ 1000 agents. L'annexe 2 détaille le processus de création de la base de donnée pour le modèle binaire ainsi que le choix du nombre de jour optimal. Enfin, avant d'entamer notre étude, nous nous sommes assurés que les variables explicatives n'étaient pas corrélées entre elles. La matrice de corrélation figure en annexe 5.

Découpage en classes

Nous avons choisi de découper les variables quantitatives en classes car :

- La relation entre la variable quantitative et l'absence n'est pas toujours linéaire,
- Les modèles obtenus sont plus souples car ils comportent davantage de paramètres (un pour chaque modalité nonobstant les contraintes d'identification) : les interprétations sont plus aisées et plus claires,
- Les prédictions sont grandement améliorées car le découpage se fait selon la variable cible d'absence, voir les annexes 3 et 4 pour le détail des méthodes de discrétisation supervisée utilisées ainsi qu'un exemple sur une variable.

Remarque : Avec la méthode CART, les variables sont discrétisées une par une avec des choix notamment de nombre maximum d'intervalles pour une lecture aisée de la sortie des modèles. Ainsi, les classifications CART présentées dans les encadrés pages 27 et 31 peuvent donner un découpage différent car optimisé pour cette classification et la lisibilité de celle-ci uniquement.

Sélection des variables

À ce stade, il est apparu indispensable de sélectionner les variables du modèle. En effet, toutes les variables initiales n'apportent pas d'information par rapport au phénomène étudié. Ensuite, l'interprétation est plus aisée lorsqu'il y a un nombre réduit de paramètres. Donc, après avoir supprimé les variables peu significatives d'après le test de Wald (p-value > 30 %), nous avons utilisé des outils basés sur le critère AIC¹¹ (stepwise, backward, forward) ce qui nous a conduit à comparer plusieurs modèles potentiels au vu de leur capacité prédictive. Nous avons conservé les variables **sexe** et **agence** afin d'observer leurs effets malgré le peu de significativité dans les modèles. Finalement, nous avons retenu le modèle suivant :

logit(absence) = sexe + agence + temps de travail effectif moyen + amplitude horaire moyenne + nb moyen vacations + écart-type du temps de travail effectif + écart-type de l'amplitude horaire + ancienneté + âge + taux travail samedi + taux travail dimanche

Le travail le week-end comme facteur d'absentéisme

La régression multivariée nous apporte des informations précieuses sur l'influence de chacune des variables de notre modèle (Figure 7 et Tableau 6). La relation entre l'âge des agents et l'absence est précisée : les agents les moins âgés sont davantage absents, la différence étant surtout marquée entre les moins de 30 ans et les plus de 30 ans. De plus, les conducteurs les plus récemment engagés sont moins sujets à l'absentéisme, tandis que ceux dont l'ancienneté est comprise entre 3.5 et 19.5 ans sont les plus absents (le rapport de cotes¹² est le double de celui des plus récemment engagés). Le modèle logistique binaire confirme aussi le fait que les femmes sont un peu plus absentes que les hommes (Odd-ratio (homme vs femme) = 0.7).

Le travail le week-end semble être un facteur essentiel pour caractériser l'absentéisme (les p-valeurs sont les plus petites et les odds-ratios très élevés, quel que soit le modèle testé). Nous avons vu précédemment que les taux d'absence sont particulièrement élevés le samedi et le dimanche. Le modèle

11. *Pratique de la régression logistique*, Ricco Rakotomala, p.117

12. Le rapport de cote (ou Odd-ratio) calcule le rapport $\frac{\frac{p}{1-p}}{\frac{q}{1-q}}$ avec p la probabilité d'absence des individus de la modalité 1 et q la probabilité d'absence des individus de la modalité 2 (dite de référence). Un odd-ratio supérieur à 1 indique que les individus de la modalité 1 sont davantage absents que ceux de la modalité 2.

logistique confirme ce constat : toutes choses égales par ailleurs, les agents qui travaillent le week-end ont une probabilité d'absence très fortement augmentée par rapport à ceux qui ne travaillent pas ces jours-ci (Odd-ratio de 2.3 pour le samedi et 3.1 pour le dimanche). Malheureusement, les données ne nous permettent pas de savoir si ces journées de travail le week-end sont contraintes ou choisies.

Le temps de travail effectif et l'amplitude horaire ont des influences similaires, linéaires ou quasi-linéaires : les agents qui travaillent beaucoup chaque jour ont une probabilité d'absence plus faible. C'est un résultat assez surprenant à première vue. Il existe en fait un lien statistique (établi par un test du Khi2, voir annexe 6 pour le détail des calculs) entre l'amplitude horaire journalière et l'ancienneté ainsi qu'avec l'agence (pas de lien en revanche avec le sexe ou l'âge). Ainsi, ce résultat est peut être dû à l'interaction d'autres variables. En outre, nous avons émis l'hypothèse que les agents avec une forte amplitude horaire journalière répartissaient leurs heures de travail sur un plus petit nombre de jours, mais en fait il n'en est rien. Sur les 400 jours de notre étude, les conducteurs avec une forte amplitude travaillent en moyenne 294 jours et ceux avec une faible amplitude travaillent en moyenne 281 jours.

Les agents qui ont en moyenne peu de coupures chaque jour (entre 1.74 et 1.94 vacations journalières moyennes) sont moins sujets à l'absentéisme que les autres. Cependant, les agents qui effectuent en moyenne le moins de vacations (moins de 1.74 vacations par jour) ont des probabilités prédites d'absence plus élevées que l'intuition le laisserait supposer à priori. Un test du Khi2 montre que le nombre moyen de vacations n'est pas indépendant de l'ancienneté, de l'âge, du sexe ou de l'agence.

Les deux derniers indicateurs, « écart-type de l'amplitude » et « écart-type du temps de travail effectif » ont des effets contrastés qui s'apparentent assez à l'effet de l'ancienneté. Un test statistique du Khi2 confirme la non-indépendance entre l'ancienneté et ces deux variables. Il est également possible que ce modèle ne soit pas pertinent pour mesurer les variations de planning, ce qui nous a conduit à envisager un modèle qui appréhende mieux les fluctuations temporelles.

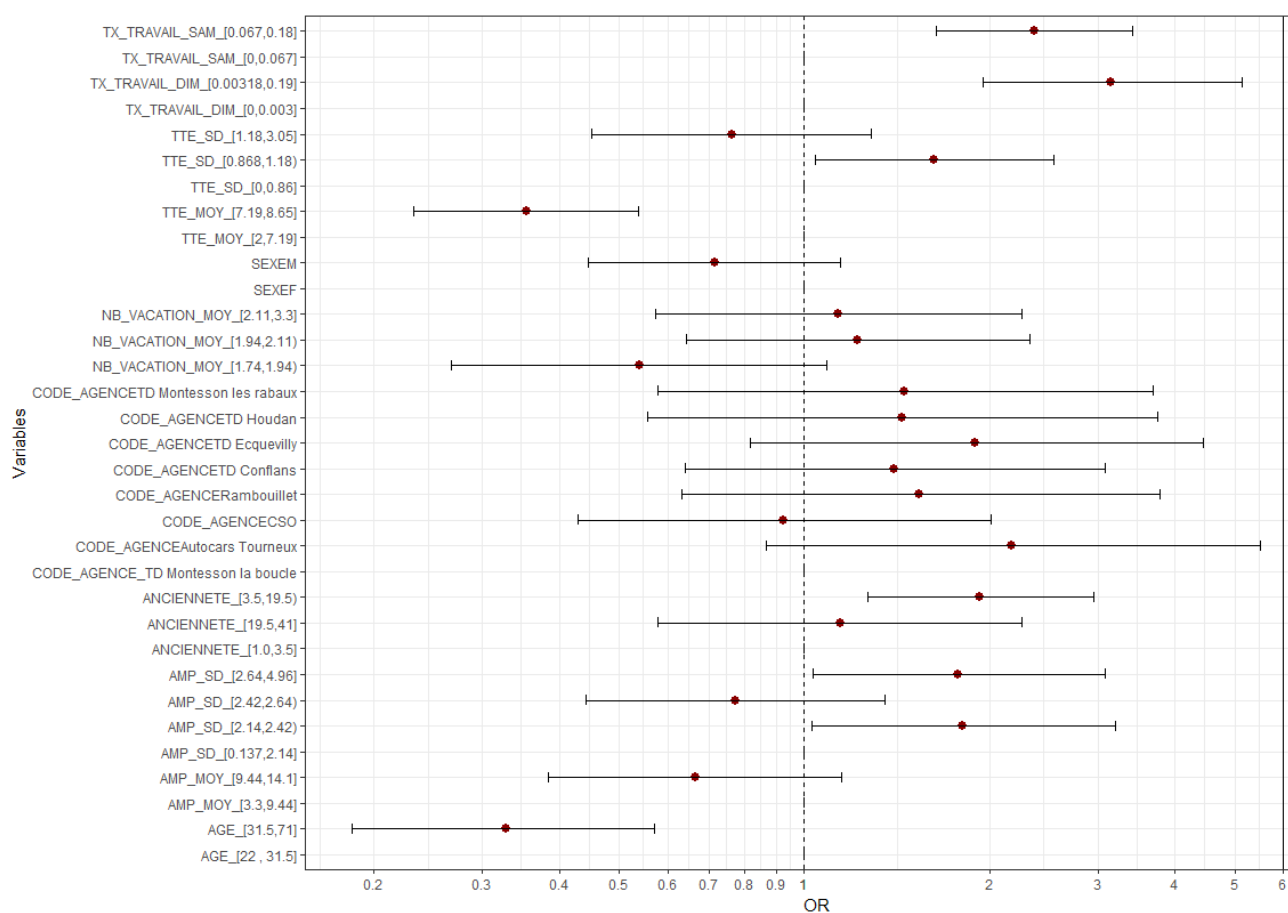


Figure 7 – Odds Ratios, modèle binaire
Source : Base Absentéisme Transdev, Base modèle binaire.

Variable	Libelle	Variable	Libelle
TX TRAVAIL SAM	taux de travail le samedi	AMP SD	écart-type de l'amplitude horaire
TX TRAVAIL DIM	taux de travail le dimanche	AMP MOY	moyenne de l'amplitude horaire
TTE SD	écart-type du temps de travail effectif	ANCIENNETE	ancienneté
TTE MOY	moyenne du temps de travail effectif	AGENCE	agence
SEXE	sexe	AGE	âge
NB VACATION MOY	moyenne du nombre de vacations		

Tableau 5 – Légende, modèle binaire

Variable	Odds Ratios	Intervalle de confiance
Constante	0.66	[0.20, 2.17]
ANC[1, 3.5]	ref	ref
ANC [3.5,19.5)	1.93 ***	[1.27, 2.95]
ANC [19.5,41]	1.15	[0.58, 2.25]
AGE [31.5,71]	0.328***	[0.18, 0.57]
AGE [22, 31.5]	ref	ref
Agence Autocars Tourneux	2.17	[0.867, 5.495]
Agence CSO	0.925	[0.43, 2.01]
Agence TD Conflans	1.4	[0.64, 3.08]
Agence TD Houdan	1.44	[0.56, 3.75]
Agence TD Montesson les rabaux	1.45	[0.58, 3.68]
Agence Ecquevilly	1.89	[0.818, 4.448]
Agence Rambouillet	1.54	[0.63, 3.77]
Agence TD Montesson la boucle	ref	ref
SexeF	ref	ref
SexeM	0.716	[0.45, 1.14]
TTE MOY [2, 7.19]	ref	ref
TTE MOY [7.19 , 8.65]	0.355	[0.23, 0.54]
TTE SD [0.19, 0.87]	ref	ref
TTE SD [0.87,1.18)	1.62	[1.04, 2.54]
TTE SD [1.18,3.05]	0.762	[0.45, 1.28]
AMP MOY [3.3, 9.44]	ref	ref
AMP MOY [9.44 , 14.1]	0.665***	[0.38, 1.15]
AMP SD [0.14, 2.14]	ref	ref
AMP SD [2.14,2.42)	1.8**	1.03, 3.2]
AMP SD [2.42,2.64)	0.774	[0.44, 1.35]
AMP SD [2.64,4.96]	1.78**	[1.03, 3.08]
VAC MOY	[1,1.74]ref	ref
VAC MOY [1.74,1.94)	0.541*	[0.27,1.09]
VAC MOY [1.94,2.11)	1.22	[0.64, 2.32]
VAC MOY [2.11,3.3]	1.14	[0.57, 2.25]
SAM [0, 0.067]	ref	ref
SAM [0.067,0.18]	2.36***	[1.64, 3.48]
DIM [0, 0.0032)	ref	ref
DIM [0.0032,0.19]	3.14***	[1.95, 5.15]

Tableau 6 – Odds ratios, modèle binaire

* p-value <10 %, ** p-value <5 %, *** p-value< 1 %

Source : Base Absentéisme Transdev, Base modèle binaire.

Note de lecture : l'odd-ratio d'une modalité correspond au rapport de l'odd ($\frac{p}{1-p}$) de la modalité divisé par l'odd de la modalité de référence.

Validité du modèle

Trois tests usuels ont été réalisés pour vérifier la validité globale de ce premier modèle. Les tests de Wald, de la déviance et du score rejettent tous trois l'hypothèse de nullité des coefficients du modèle, ce qui signifie que le modèle apporte suffisamment d'information. L'annexe 7 détaille les calculs.

Adéquation du modèle

Afin de vérifier l'efficacité de notre modèle, nous avons scindé notre échantillon initial (environ 1 000 agents) en deux groupes : un échantillon d'apprentissage et un échantillon de validation.

Lors de l'analyse descriptive, nous avons constaté des disparités de taux d'absences selon le sexe et l'agence des agents. Ceci nous a conduit à réaliser un échantillonnage stratifié pour scinder notre population en deux échantillons afin de limiter la variabilité des résultats liée à une sélection purement aléatoire. Pour chaque croisement des variables sexe et agence, nous sélectionnons de manière aléatoire 70 % des individus de cette strate que nous plaçons dans l'échantillon d'apprentissage, les 30 % restant sont placés dans l'échantillon test.

Nous avons donc bâti notre modèle sur la base de l'échantillon d'apprentissage ; ensuite nous avons utilisé ce modèle pour prédire l'absence des individus présents dans l'échantillon de validation afin de comparer les prédictions à la réalité. La matrice de confusion obtenue permet de mesurer l'efficacité du modèle à l'aide de plusieurs indicateurs.

	Absence prédite	Présence prédite
Absence réelle	103	51
Présence réelle	40	112

Tableau 7 – **Matrice de confusion, modèle binaire**

Source : Base Absentéisme Transdev, Base binaire.

- Taux d'erreur : 29.7 %
- Taux de bien classés : 70.3 %
- Spécificité : 73.7 % (indicateur du taux de présents détectés, soit : vrais présents/présents réels)
- Sensibilité : 66.9 % (indicateur de la capacité du modèle à détecter les absents : vrais absents/absents)

La courbe ROC est une illustration graphique du pouvoir discriminant du modèle envisagé. Elle sert à ajuster le seuil afin d'optimiser la sensibilité et la spécificité.

En effet, en fonction des coefficients déterminés par le modèle, pour chaque individu, on calcule une probabilité prédite d'absence (qui est donc comprise entre 0 et 1). On choisit ensuite un seuil de décision (par exemple $s=0.5$). Si la valeur calculée est supérieure au seuil, on affecte l'individu à la modalité absence=1, sinon à la modalité absence=0. Ces prédictions sont comparées à la réalité, ce qui permet de calculer une spécificité et une sensibilité. En réitérant l'opération pour différents seuils, on obtient plusieurs valeurs de spécificité et de sensibilité qui permettent de tracer la courbe.

La courbe ROC obtenue montre une bonne discrimination du modèle binaire (Auc¹³=0.738)

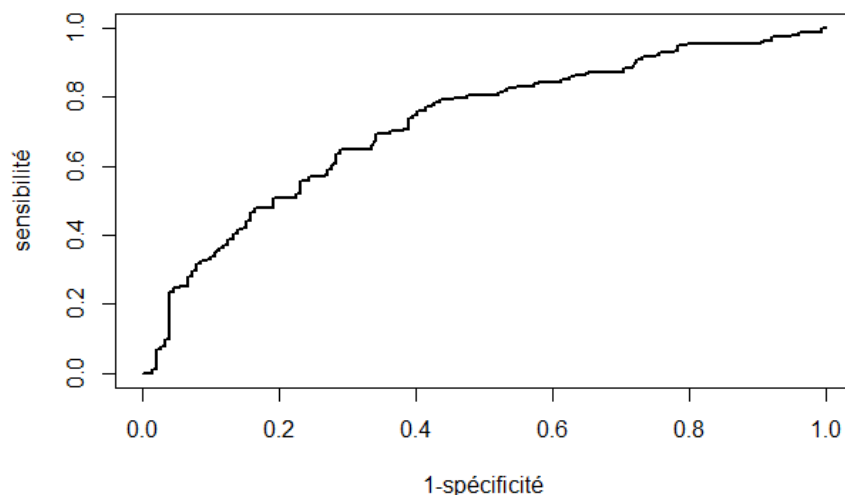


Figure 8 – **Courbe ROC, modèle binaire**

Source : Base Absentéisme Transdev, Base modèle binaire.

La matrice de confusion du tableau précédent est bâtie en choisissant un seuil de décision $s=0.561$, qui permet le meilleur compromis dans la maximisation de la sensibilité et la spécificité.

13. Area Under Curve (aire sous la courbe ROC)

Enfin, un test de Hosmer-Lemeshow (détaillé en annexe 8) montre que notre modèle est bien calibré.

Ce premier modèle a permis d'avancer dans la compréhension des facteurs influençant l'absentéisme des conducteurs de bus. Toutefois, cette dichotomie entre « absence » et « présence » paraît trop tranchée au regard de la diversité des types d'absence. En effet, un agent sera classé « absent » qu'il aît été absent 1 jour , 50 jours ou 400 jours. Il apparaît donc opportun d'envisager un modèle plus fin.

Le temps de travail et l'activité le dimanche discriminent la population

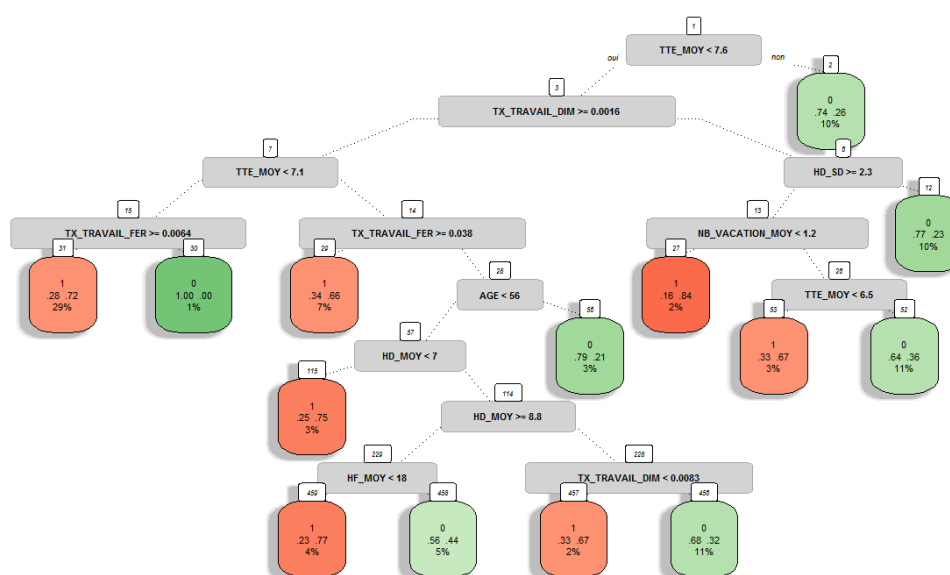


Figure 9 – Classification CART selon la modalité d'absence dans la base binaire

Source : Base Absentéisme Transdev, Base modèle binaire.

Comment lire cet arbre ? : La lecture se fait de haut en bas et les agents sont départagés selon le clivage binaire proposé à chaque nœud, à gauche pour la réponse positive, à droite pour la négative. Les agents classés 0 sont en vert et les 1 en rouge, couleurs dont l'intensité dépend de la probabilité de classement.

Note de lecture : En haut à droite, la feuille numéro 2 représente 10 % de la population, ce sont les agents dont le temps de travail journalier effectif moyen est supérieur ou égal à 7,6 heures (réponse non au nœud 1). Ils sont classés dans les non-absents dont la classe est 0 avec une probabilité de 0,74.

Les variables les plus discriminantes pour séparer la population apparaissent en premier dans l'arbre.

Ainsi la variable « temps de travail effectif journalier moyen » est la plus importante et sépare en une étape 10 % de la population qui travaille plus de 7,6 heures par jour en les classant dans les non-absents. Le 2ème clivage le plus important est le nœud d'indice 3 qui va séparer les agents travaillant le dimanche des autres. Les agents qui travaillent le dimanche et les jours fériés sont contenus dans les feuilles 29 et 31, ils représentent 36 % de la population et sont classés dans les absents.

Les agents sont classés dans la bonne modalité d'absence dans 71 % des cas.

Les plannings, un levier pour atténuer l'absentéisme

Afin de s'extraire de la vision binaire de l'absentéisme vue dans le modèle précédent, nous avons créé une variable taux d'absence. Elle prend trois modalités :

- 0 si l'agent n'a jamais été absent correspondant à 46 % des agents dans la base agrégée polytomique ;
- 1 si l'agent a un taux d'absence non nul et inférieur à 5 % soit 28 % de la population
- 2 si l'agent a un taux d'absence supérieur à 5 % soit 26 % des agents.

Le seuil de 5 % permet de diviser en deux groupes approximativement de même taille les conducteurs absents au moins une fois.

Un modèle qui nuance les résultats précédents

Nous allons utiliser un modèle multinomial : la variable d'intérêt prend les 3 modalités précédemment présentées. La modalité de référence sera la modalité $Y=0$.

L'utilisation d'un modèle à odds proportionnel, le modèle polytomique ordinal le plus usité, n'est pas adapté ici, le test d'égalité des pentes est rejeté sur notre jeu de données.

Sélection des variables

Nous avons sélectionné parmi les variables présentées dans le modèle binaire, celles qui sont significatives (sous SAS, via la méthode backward). Le modèle obtenu en utilisant l'échantillon d'apprentissage est le suivant :

$\text{logit}(\text{absence}) = \text{sexe} + \text{agence} + \text{temps de travail effectif moyen} + \text{amplitude horaire moyenne} + \text{écart type de l'amplitude horaire} + \text{ancienneté} + \text{âge} + \text{taux travail dimanche} + \text{taux travail férié}$

Le nombre moyen de vacances, l'écart-type du temps de travail effectif et le taux de travail le samedi ne sont plus significatives alors que le taux de travail les jours fériés le devient.

interprétations des odds-ratio :

L'interprétation des odds ratio dans un modèle polytomique n'est pas aisée. La lecture des résultats doit prendre en compte deux références : la catégorie de référence de la variable à expliquer et la modalité de référence de la variable explicative. Nous constatons, comme dans le modèle binaire et les statistiques descriptives, l'impact du sexe dans l'absentéisme.

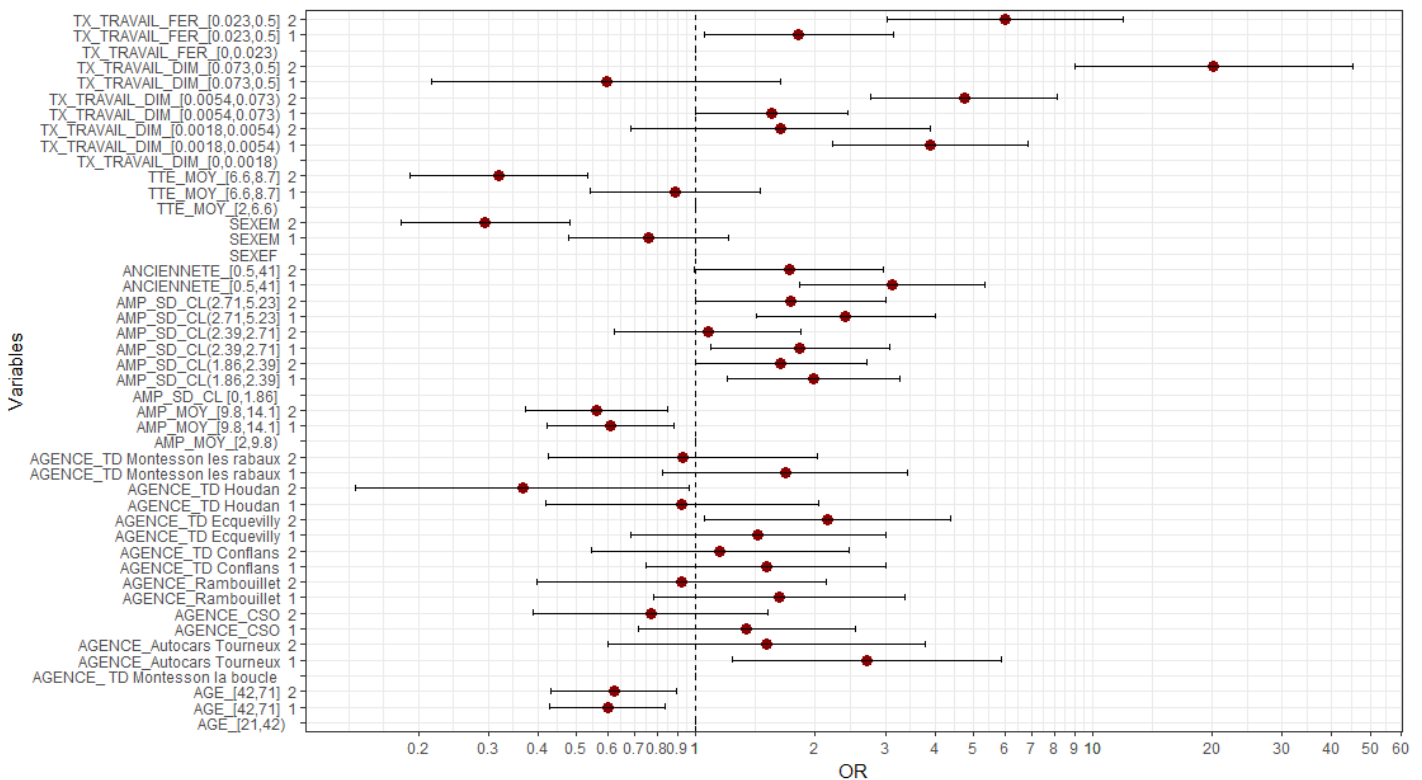


Figure 10 – Odds Ratios, modèle polytomique

Source : Base Absentéisme Transdev, Base polytomique.

Analysons les odds-ratio pour la variable sexe. Les statistiques descriptives et le modèle binaire indiquent un absentéisme plus important pour les femmes.

Le modèle nous donne : $OR_2(M/F) = 0,29$ et $OR_1(M/F) = 0,76$

Toutes choses égales par ailleurs, un homme a environ 71 % ($1-0,29$) de chances de moins qu'une femme d'être souvent absent plutôt que jamais absent. Un homme a 24 % de chances de moins qu'une femme d'être parfois absent plutôt que jamais absent.

La classe d'âge utilisée pour le modèle polytomique est différente de celle du modèle binaire, suite au travail de discrétisation réalisé en amont. Le message reste globalement le même : les conducteurs plus jeunes (ici, ceux ayant moins de 42 ans) ont davantage de chances d'être absent. Les odds ratio de la modalité 2 et de la modalité 1 sont proches.

Trois variables liées au planning sont significatives dans le modèle : le taux de travail le dimanche, le taux de travail les jours fériés et les changements d'amplitude de la durée d'une journée de travail.

Le taux de travail le dimanche (TTD) est découpé en quatre classes. Lorsque ce taux augmente, les OR d'être classé souvent absent augmentent également.

Toutes choses égales par ailleurs, un conducteur parmi ceux travaillant le plus le dimanche a environ 20 fois plus de chances d'être beaucoup absent plutôt que jamais comparé à un conducteur travaillant le moins le dimanche ($OR_2(TTD_{max}/TTD_{min}) = 20$).

Les changements d'amplitude de la durée d'une journée de travail influencent l'absentéisme. Un écart type inférieur à deux heures réduit les risques d'absence.

Contrairement aux autres modèles de ce rapport, le taux de travail lors des jours fériés est ici significatif : son augmentation accentue les risques d'absentéisme. Le mode de calcul de cette variable rend l'analyse difficile (nombre de jours fériés où le conducteur est au planning sur nombre de jours

où le conducteur est au planning), la proportion de jours fériés où Transdev inscrit le conducteur au planning aurait été plus simple à interpréter.

Concernant les agences, deux se démarquent relativement à l'agence de référence (Montesson la boucle). Les conducteurs de l'agence Houdan ont significativement moins de chances d'être souvent absent, avec un odd ratio de 0,36. Ceux de l'agence Ecquivilly au contraire ont davantage de chances d'être souvent absents avec un odd ratio de 2,14.

Ces résultats confirment ceux obtenus dans le modèle binaire tout en apportant un niveau de détail plus fin sur le niveau d'absence.

Adéquation du modèle

Contrairement au modèle binaire, il n'est pas possible de calculer un AUC global pour un modèle polytomique. On peut toutefois utiliser l'AUC moyen obtenu pour les modèles binaires combinant les modalités de Y (soit 3 modèles binaires ici : le modèle étudié au chapitre précédent, le modèle binaire moins de 5 % d'absentéisme contre plus de 5 % et le modèle aucune absence ou plus de 5 % contre moins de 5 %) ¹⁴. On obtient un AUC moyen de 61,5 % pour notre modèle polytomique.

En testant le modèle sur l'échantillon de validation, on obtient la matrice de confusion ci-dessous. Pour chaque agent de cet échantillon, une probabilité de classement dans les différentes modalités de Y est calculée en fonction des coefficients du modèle. La modalité de Y la plus probable pour le conducteur est ainsi prédite.

		Y prédits		
		0	1	2
Y observés	0	153	21	38
	1	53	44	20
	2	48	17	48

Tableau 8 – **Matrice de confusion, modèle polytomique**

Source : Base Absentéisme Transdev, Base polytomique.

Le taux de bien classés est de 55,4 %. Ce taux peut sembler faible mais est à comparer à un modèle classant au hasard les individus dans les différentes valeurs de Y, qui donne un taux de bien classés de 33 %.

14. Adams, N. M. & Hand, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain. Pattern Recognition, 32, 1139–1147.

L'ancienneté, première variable discriminante

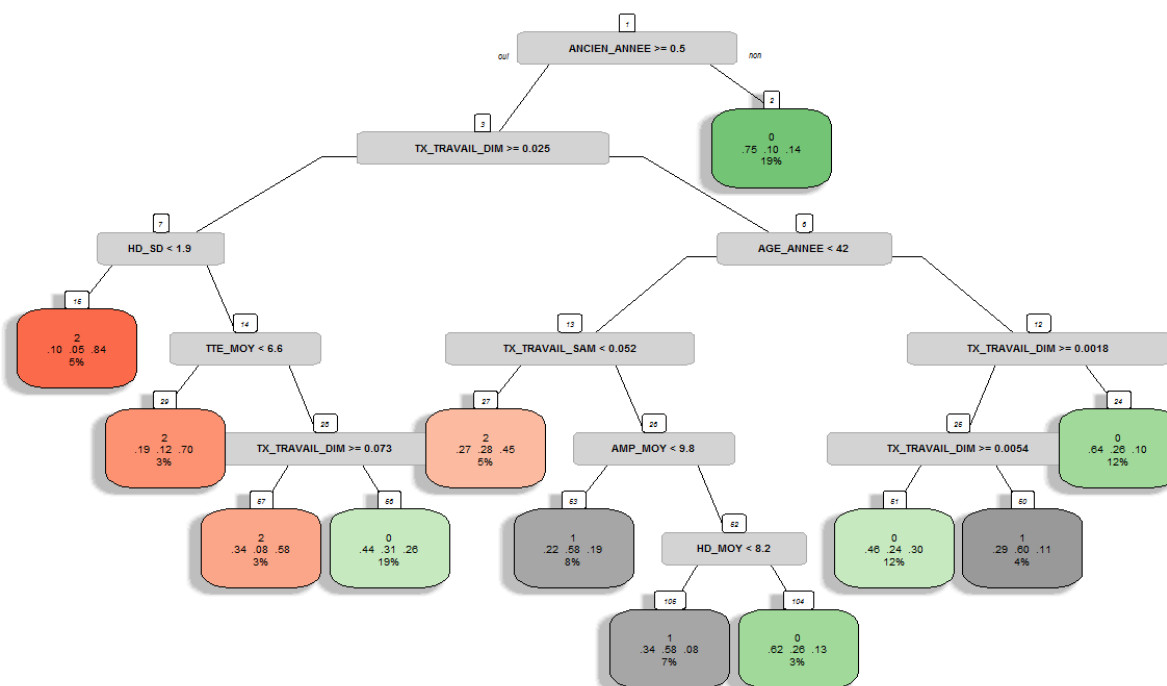


Figure 11 – **Classification CART** selon la modalité d'absence dans la base polytomique

Source : Base Absentéisme Transdev, Base polytomique.

Note de lecture : Le nœud 2 représente 19 % de la population, ce sont les agents dont l'ancienneté est inférieure à 6 mois. Ils sont classés avec une probabilité de 0,76 dans la classe 0. Les probabilités de classement dans les classes 1 et 2 sont respectivement de 0.1 et 0.14.

Les agents présents depuis moins de 6 mois chez Transdev sont directement classés dans les non-absents. Le clivage naturel attendu aurait été de un an correspondant à la durée légale pour bénéficier d'indemnités maladies selon la convention collective des transports urbains de voyageurs.

Comme dans la base binaire, le taux de travail le dimanche est une variable très discriminante. Les agents fréquemment absents présents dans les feuilles de gauche de l'arbre sont des agents qui travaillent le dimanche avec un faible temps de travail effectif quotidien et/ou un faible écart-type sur l'heure de début du poste.

59 % des agents et même 81 % des non-absents sont bien classés dans cet arbre. La classification est moins efficace pour les absents « occasionnels » et « réguliers » où le taux de bien classés est d'environ 40 %.

Les fluctuations de planning accroissent l'absentéisme

Dans les modèles précédents, pour chaque agent, nous avons regroupé un grand nombre d'observations journalières afin d'en tirer des indicateurs synthétiques (moyenne, écart-type, ...). Cette agrégation forte nous empêche d'observer des variations au cours du temps et nous fait donc perdre de l'information temporelle. Ainsi, nous n'avons pour le moment pas pu correctement mesurer l'influence des horaires ou des variations d'emploi du temps sur différentes périodes de temps. Il paraît donc légitime de regrouper les données journalières de chaque agent selon une période de référence, par exemple une semaine. Ce faisant, nous obtenons plusieurs périodes pour chaque individu. Cependant, lorsque l'on est en présence de mesures répétées pour chaque agent, l'hypothèse d'indépendance des données n'est pas réaliste (l'absence une période donnée peut être corrélée à l'absence sur une autre période). Pourtant, cette hypothèse est indispensable pour utiliser une régression logistique. Il faut alors utiliser d'autres méthodes tenant compte de la corrélation des données. Nous avons donc choisi d'utiliser la méthode GEE : Generalized Estimating Equations (l'annexe 9 détaille la méthodologie) dans une modélisation logistique à modalités binaires.

Choix de la période de référence optimale

Tout d'abord, il est nécessaire de choisir la durée de la période de référence. En effet, regrouper les données par semaine, par mois ou par semestre ne donnera pas les mêmes résultats. Nous avons donc bâti plusieurs modèles selon des périodes de durées différentes puis nous avons étudié l'efficacité prédictive de chacun. Pour chaque modèle, nous avons discrétisé les variables continues dans la même optique que pour les modèles précédents, à la différence que nous n'avons pas utilisé les arbres de classement, mais un algorithme de discretisation appelé MDLP (voir annexe 4). Cet algorithme permet également d'obtenir une discrétisation optimale selon les valeurs de notre variable à expliquer, l'absence. Le tableau suivant regroupe les résultats des essais sur différentes période de temps.

Période (en jours)	Auc (Aire sous la courbe ROC)	Taux de mal classés
07	0.75	27.0 %
15	0.71	34.0 %
30	0.71	35.6 %
60	0.71	32.5 %
90	0.70	35.6 %
180	0.70	35.4 %

Tableau 9 – Comparaison de l'efficacité prédictive des modèles GEE selon la période de référence

Source : Base Absentéisme Transdev, Base GEE7.

Excepté pour le modèle à 7 jours, les deux indicateurs de performance (Auc et taux de mal classés) sont très proches et les résultats sont également très voisins. Lorsque les données sont agrégées par semaine, les performances sont sensiblement augmentées. Il est possible qu'un choix de période très court permette de mieux extraire l'information temporelle contenue dans les plannings des conducteurs de bus. Il aurait été souhaitable de continuer les essais pour chercher une durée optimale, mais le manque de temps nous a contraint à limiter les expérimentations ¹⁵.

15. Plus la durée choisie est courte, plus les tables, les modèles et la discrétisation des variables sont longs à créer.

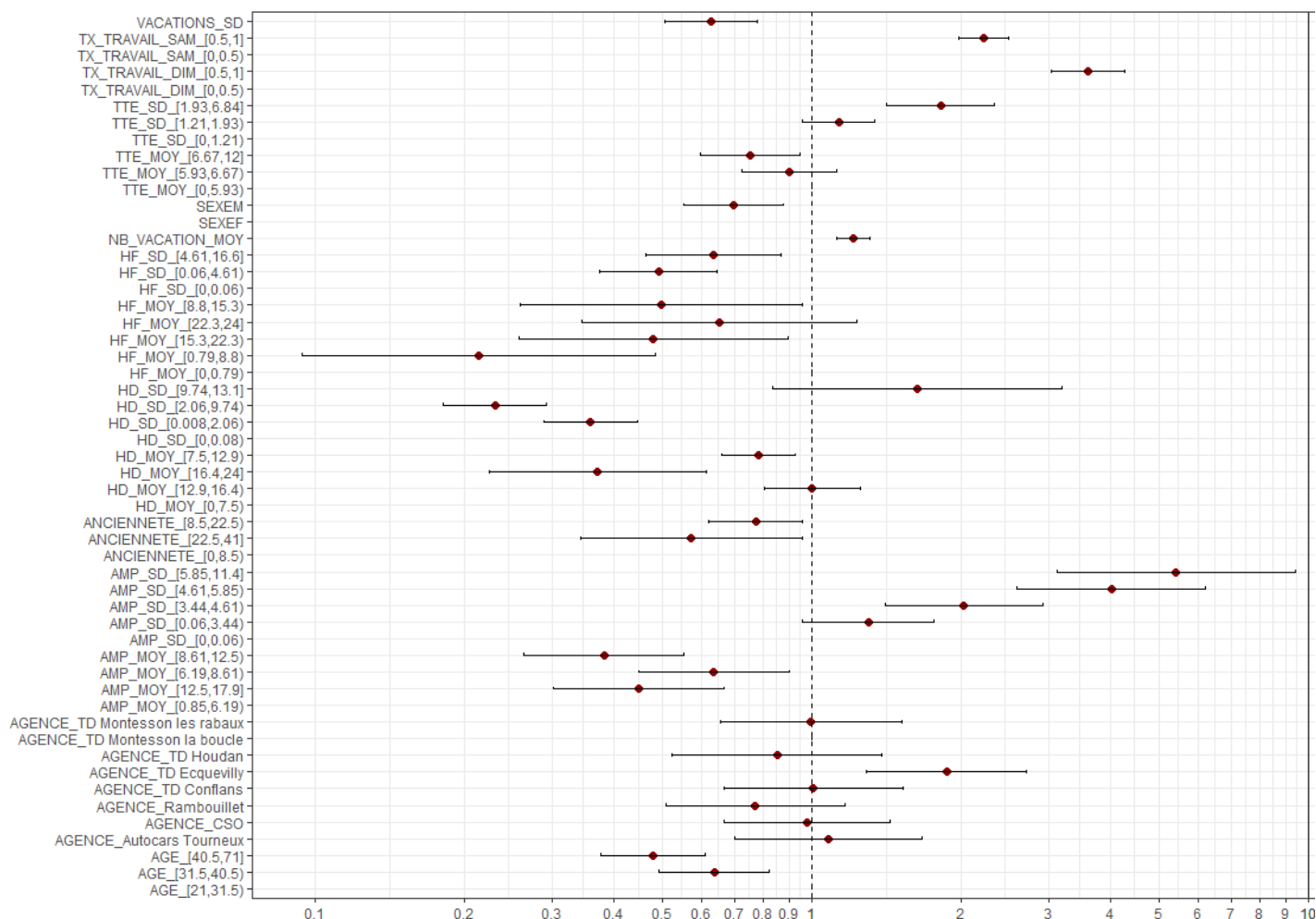


Figure 12 – Odds Ratios, modèle GEE7

Source : Base Absentéisme Transdev, Base GEE7.

Variable	Libelle	Variable	Libelle
VACATIONS SD	écart-type du nb de vacances	HF MOY	moyenne de l'heure de fin de journée
TX TRAVAIL SAM	taux de travail le samedi	HD SD	écart-type de l'heure de début de journée
TX TRAVAIL DIM	taux de travail le dimanche	HD MOY	moyenne de l'heure de début de journée
TTE SD	écart-type du temps de travail effectif	ANCIENNETE	ancienneté
TTE MOY	moyenne du temps de travail effectif	AMP SD	écart-type de l'amplitude horaire
SEXE	sexe	AMP MOY	moyenne de l'amplitude horaire
NB VACATION MOY	moyenne du nombre de vacances	AGENCE	agence
HF SD	écart-type de l'heure de fin de journée	AGE	âge

Tableau 10 – Légende, modèle GEE7

Variable	Odds Ratios	Intervalle de confiance
Constante	0.55	[0.23 , 1.30]
AGE [21,31.5)	ref	ref
AGE [31.5,40.5)	0.64 ***	[0.49 , 0.82]
AGE [40.5,71]	0.48 ***	[0.38 , 0.61]
ANCIENNETE [0,8,5)	ref	ref
ANCIENNETE [8.5,22.5)	0.77 **	[0.62 , 0.96]
ANCIENNETE [22.5,41]	0.57 **	[0.34 , 0.96]
Agence TD Montesson la boucle	ref	ref
AGENCE Autocars Tourneux	1.08	[0.70 , 1.67]
AGENCE CSO	0.98	[0.67 , 1.44]
AGENCE TD Conflans	1.01	[0.66 , 1.53]
AGENCE TD Houdan	0.85	[0.52 , 1.39]
AGENCE TD Montesson les rabaux	1.00	[0.66 , 1.52]
AGENCE TD Ecquevilly	1.87 ***	[1.29 , 2.71]
AGENCE Rambouillet	0.77	[0.51 , 1.16]
SEXE F	ref	ref
SEXE M	0.70 ***	[0.55 , 0.88]
TTE MOY [0,5.93)	ref	ref
TTE MOY [5.93,6.67)	0.90	[0.72 , 1.12]
TTE MOY [6.67,12]	0.75 **	[0.60 , 0.95]
TTE SD [0,1.21)	ref	ref
TTE SD [1.21,1.93)	1.13	[0.96 , 1.34]
TTE SD [1.93,6.84]	1.82 ***	[1.42 , 2.33]
AMP MOY [0,6.19)	ref	ref ,
AMP MOY [6.19,8.61)	0.63 **	[0.45 , 0.90]
AMP MOY [8.61,12.5)	0.38 ***	[0.26 , 0.55]
AMP MOY [12.5,17.9]	0.45 ***	[0.30 , 0.67]
AMP SD [0,0.06)	ref	ref
AMP SD [0.06,3.44)	1.30 *	[0.96 , 1.76]
AMP SD [3.44,4.61)	2.03 ***	[1.41 , 2.92]
AMP SD [4.61,5.85)	4.02 ***	[2.59 , 6.22]
AMP SD [5.85,11.4]	5.42 ***	[3.11 , 9.42]
VAC MOY	1.21 ***	[1.12 , 1.31]
VAC SD	0.63 ***	[0.51 , 0.78]
HD MOY [0,7.5)	ref	ref
HD MOY [7.5,12.9)	0.78 ***	[0.66 , 0.92]
HD MOY [12.9,16.4)	1.00	[0.80 , 1.25]
HD MOY [16.4,24]	0.37 ***	[0.22 , 0.61]
HD SD [0,0.01)	ref	ref
HD SD [0.01,2.06)	0.36 ***	[0.29 , 0.44]
HD SD [2.06,9.74)	0.23 ***	[0.18 , 0.29]
HD SD [9.74,13.1]	1.63	[0.83 , 3.19]
HF MOY [0,0.79)	ref	ref
HF MOY [0.79,8.8)	0.21 ***	[0.09 , 0.48]
HF MOY [8.8,15.3)	0.50 **	[0.26 , 0.96]
HF MOY [15.3,22.3)	0.48 **	[0.26 , 0.90]
HF MOY [22.3,24]	0.65	[0.34 , 1.23]
HF SD [0,0.06)	ref	ref
HF SD [0.06,4.61)	0.49 ***	[0.37 , 0.64]
HF SD [4.61,16.6]	0.63 ***	[0.46 , 0.87]
SAM [0,0.5)	ref	ref
SAM [0.5,1]	2.22 ***	[1.98 , 2.49]
DIM [0,0.5)	ref	ref
DIM [0.5,1]	3.60 ***	[3.04 , 4.27]

Tableau 11 – Odds ratios, modèle GEE7

* p-value <10%, ** p-value <5%, *** p-value < 1%

Note de lecture : l'odds ratio correspond au rapport de l'odds de la modalité divisé par l'odds de la modalité de référence

Source : Base Absentéisme Transdev, Base GEE7.

Les variations de planning aggravent l'absentéisme

Le travail le week-end : un effet marqué

L'influence du travail le week-end est confirmée et nuancée. Plus les agents sont contraints de travailler le week end, plus ils sont absents : les conducteurs qui travaillent **souvent** le samedi (resp. le dimanche) ont un risque d'absence multiplié par 2.2 (resp. 3.6) par rapport à ceux qui travaillent **rarement** le samedi (resp. le dimanche)¹⁶. Toutefois, il existe une minorité de salariés (113 agents soit 7.6 % de l'ensemble) qui travaillent tous les samedis pendant plusieurs mois et qui souffrent très peu d'absentéisme. C'est particulièrement visible avec les modèles sur longue période (GEE 180j, GEE 90j). On peut supposer que ces agents se sont habitués à travailler le samedi qui est alors considéré comme un jour traditionnel.

Taux de travail le samedi (%)	Présence (nb de périodes)	Absence (nb de périodes)	Taux d'absence (%)
0 à 8	1230	140	10.2
8 à 56	5421	1464	21.2
56 à 72	357	171	32.4
72 à 96	75	104	58.1
96 à 100	106	8	7.0

Tableau 12 – **Taux d'absence selon le taux de travail le samedi, modèle GEE à 90 jours**
Source : Base Absentéisme Transdev, Base GEE90.

L'instabilité des horaires, cause majeure d'absentéisme

Les effets des horaires de début et de fin de journée sont assez faibles ; les agents qui terminent entre minuit et 1h du matin sont les plus absents. À l'inverse, les conducteurs qui terminent entre 1h et 9h sont les moins absents. Les salariés qui commencent leur journée l'après midi ou en soirée (entre 16h30 et minuit) sont très peu absents. Les figures 4 et 5 détaillent la distribution de ces horaires.

L'influence des fluctuations d'horaires¹⁷ de début et de fin de journée est très net. Jusqu'à un certain point, l'augmentation des variations d'horaires provoque une diminution du risque d'absence, mais au-delà de ce seuil, le risque d'absence est très élevé.

L'amplitude horaire et le temps de travail effectif ont des effets similaires : l'augmentation du temps de travail quotidien fait diminuer le risque d'absence. En revanche, tous les modèles montrent un fort effet des variations d'amplitude horaire sur le risque d'absentéisme. Par exemple, en prenant pour référence les agents avec les horaires les plus stables, les agents avec les horaires les plus variables ont un odd ratio proche de 6. Le tableau 13 montre les taux d'absence selon l'écart-type de l'amplitude horaire, indicateur de l'instabilité des horaires ; la dernière classe concerne 132 agents sur les 1 459 de l'échantillon.

L'agence TD Ecquevilly se démarque

Quelque soit le modèle testé, l'agence TD Ecquevilly est celle dont l'odd ratio est le plus élevé, et l'effet est encore plus marqué à mesure que la période de référence se raccourcit. Cette agence se démarque donc véritablement des autres, ses agents ayant les plus forts risques d'absence.

16. Dans ce modèle, on peut approcher les odds ratios aux rapports des probabilités estimées puisque $OR(A \text{ vs } B) = \frac{p_A}{1-p_A} \div \frac{p_B}{1-p_B} \simeq \frac{p_A}{p_B}$ car p_A et p_B sont proches de zéro.

17. L'effet est net dans le modèle GEE à 7 jours avec un OR (heure début très variable / heure début stable) = 1.63

Ecart type de l'amplitude horaire	Présence (semaines)	Absence (semaines)	Taux d'absence (%)
0 à 3.4	58 583	1 131	1.9
3.4 à 4.6	4 707	138	2.8
4.6 à 5.9	934	75	7
5.9 à 11.4	185	41	18

Tableau 13 – Taux d'absence selon la variation d'amplitude horaire, modèle GEE à 7 jours

Source : Base Absentéisme Transdev, Base GEE7.

Adéquation modèle GEE7

La courbe ROC du modèle GEE à 7 jours est légèrement au-dessus de celle du modèle binaire ce qui est mesuré par une aire sous la courbe supérieure au modèle binaire ($Auc = 0.75$), dénotant un meilleur pouvoir prédictif (voir tableau 14). Le taux de bien classés est également légèrement meilleur (71.7 %).

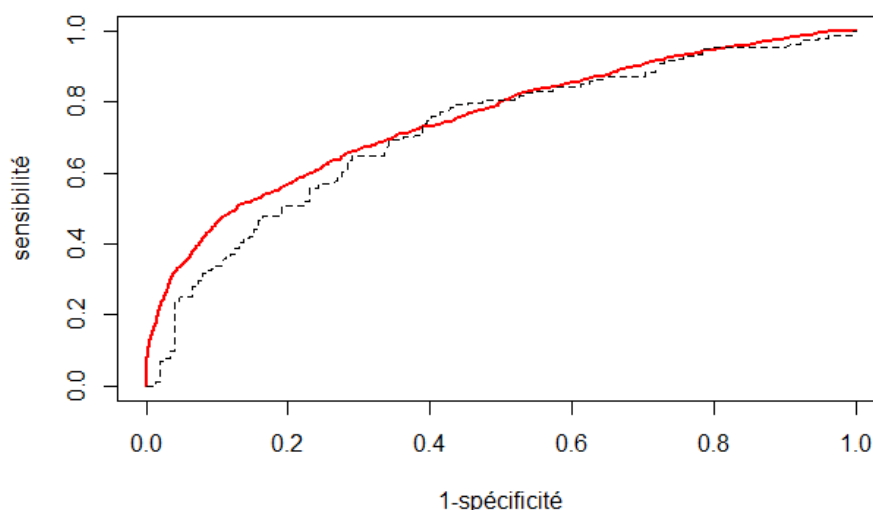


Figure 13 – Comparaison des courbes ROC

Note : Modèle binaire en pointillé, Modèle GEE7 en plein

	Absence prédite	Présence prédite
Absence réelle	483	254
Présence réelle	7 775	19 849

Tableau 14 – Matrice de confusion, modèle GEE à 7 jours

Source : Base Absentéisme Transdev, Base GEE7.

- Taux d'erreur= 28.3 %
- Taux de bien classés= 71.7 %
- Spécificité= 71.9 %
- Sensibilité= 65.6 %
- seuil de décision optimal $s=0.0262$
- $Auc=0.75$

Pistes d'amélioration et à explorer

À l'issue d'un travail de plusieurs mois, il est agréable de se retourner sur les épreuves traversées. On peut notamment se réjouir des avancées ou regretter les chemins que l'on n'a pas pu suivre. Cette partie se penchera plutôt sur les voies d'amélioration et sur des pistes non empruntées faute de temps.

Voies d'amélioration

Issue de registres internes destinés à assurer la gestion des ressources humaines, la base de données a nécessité des opérations de mise en forme et d'apurement mais sa qualité est globalement satisfaisante (entre 5 et 10 % de valeurs incohérentes ou manquantes initialement). Plusieurs actions simples permettraient toutefois d'en améliorer l'exploitation statistique.

- Certaines variables comportent des valeurs aberrantes nécessitant une correction très limitée. Il s'agit notamment de la date de naissance du conducteur (postérieure au 29 décembre 1998 pour une poignée d'agents) ou des variables de temps de travail effectif (TTE) ou d'amplitude horaire (AMP) qui ont dû être écrêtées à un seuil plausible (respectivement 12 et 18 heures).
- Dans le même ordre d'actions probablement peu contraignantes, la saisie d'une amplitude horaire (AMP) « planifiée » pour les agents absents permettraient de quantifier l'absence non plus seulement en nombre de jours mais plutôt en équivalent temps-plein. Cet indicateur est couramment utilisé et facilite les comparaisons. Une alternative consisterait à proposer un modèle d'imputation performant pour cette variable que nous avons hésité à affecter statistiquement du fait de la variabilité des horaires observée.
- Plus ambitieux mais probablement plus fructueux également, l'intégration de nouvelles variables pourrait améliorer significativement la capacité descriptive des modèles. Certains paramètres individuels (nombre d'enfants notamment en bas âge, distance domicile-travail) ou professionnels (tranches de salaires, appréciation de la pénibilité des trajets ou des services effectués) contribueraient sans doute à enrichir l'étude du phénomène.

Même modestes, la mise en forme et l'apurement ont été rondement menés pour rendre possible l'analyse de l'absentéisme. Des pistes d'améliorations substantielles subsistent également sur ces opérations :

- Les données du mois de décembre 2016 sont partielles (les deux derniers jours du mois ne sont pas intégrés dans la base) mais elles paraissent également étonnantes en termes d'absence notamment. Le taux d'absentéisme pour ce mois s'élève à 4,8 % alors qu'il est compris entre 8,1 % et 11,2 % pour les autres mois. Cette valeur « extrême » aurait mérité une analyse et un traitement spécifique que nous n'avons pas réalisés car nous avons découvert cette anomalie tardivement.
- L'imputation réalisée sur la variable âge sur un mode déterministe est une première étape intéressante qui nous a notamment permis de mesurer l'importance des données manquantes et de s'initier à leur traitement. La méthode retenue reste toutefois grossière. À défaut de mettre en oeuvre des techniques complexes, un outil plus robuste comme les plus proches voisins (knn ou k-nearest neighbors) est réputé donner des résultats plus satisfaisants.

Pistes non développées

Trois démarches distinctes auraient pu être adoptées qui enrichiraient les analyses réalisées. Il s'agit notamment d'une **analyse par profil d'absentéisme**, une **prise en compte des interactions entre variables** et, enfin, de **l'attention à certaines caractéristiques temporelles**.

L'encadré sur la page suivante initie une **analyse par profils** sans l'approfondir. On distingue ainsi quatre types d'agents selon la fréquence et la durée des épisodes d'absentéisme pendant la période. Le profil 2 rassemble par exemple des conducteurs peu souvent absents mais sur des durées longues. Une caractérisation rapide les place parmi les agents qui travaillent plus souvent le dimanche et dont l'ancienneté dans l'entreprise est plus élevée. Cette ébauche de profils n'a été réalisée que sur quelques variables à titre d'exemple et aurait pu faire l'objet d'une modélisation dédiée.

Les **interactions entre variables** n'ont également pas été prises en compte. L'effet du travail le dimanche est ainsi peut-être différent selon l'âge des agents. L'analyse descriptive et les échanges avec le tuteur, disposant d'une expertise fonctionnelle, auraient pu être développés dans ce sens afin d'enrichir les modèles.

Enfin, **l'étude de certains phénomènes saisonniers** (maladies, épisodes climatiques) ou plus exceptionnels (attentats) auraient pu être intégrés à l'analyse. La base de données couvre deux années quasiment complètes (2015 et 2016), elle autorisait donc ce type d'examen. Cette lecture chronologique paraît avec le recul intéressante d'autant que l'absentéisme est plus élevé à certaines périodes (novembre, février).

Quatre profils d'absentéisme

On peut distinguer 4 profils d'agents si l'on catégorise selon la durée et le nombre d'épisodes d'absence :

- Profil 0 : jamais absent (46,2 % des agents) ;
- Profil 1 : peu absent et peu longtemps, cela représente la majorité des absents avec 42,2 % des agents.
- Profil 2 : peu souvent absent mais longtemps absent (5,7 % des agents) ;
- Profil 3 : souvent absent (5,9 % des agents) ;

Les profils 2 et 3 représentent les cibles d'intérêt afin de réduire l'absentéisme.

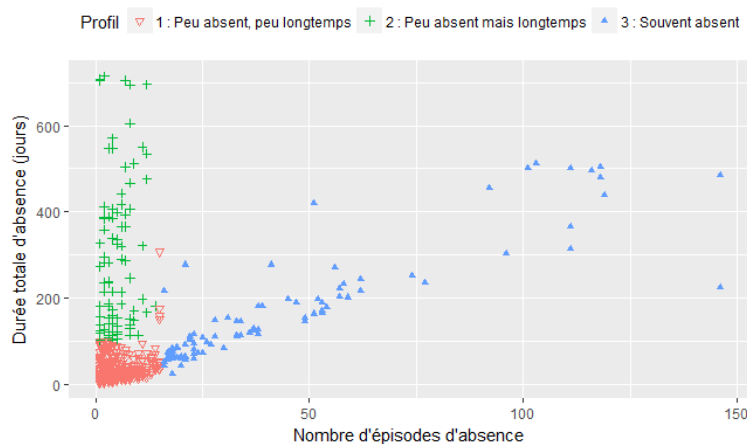


Figure 14 – Durée d'absence par épisode

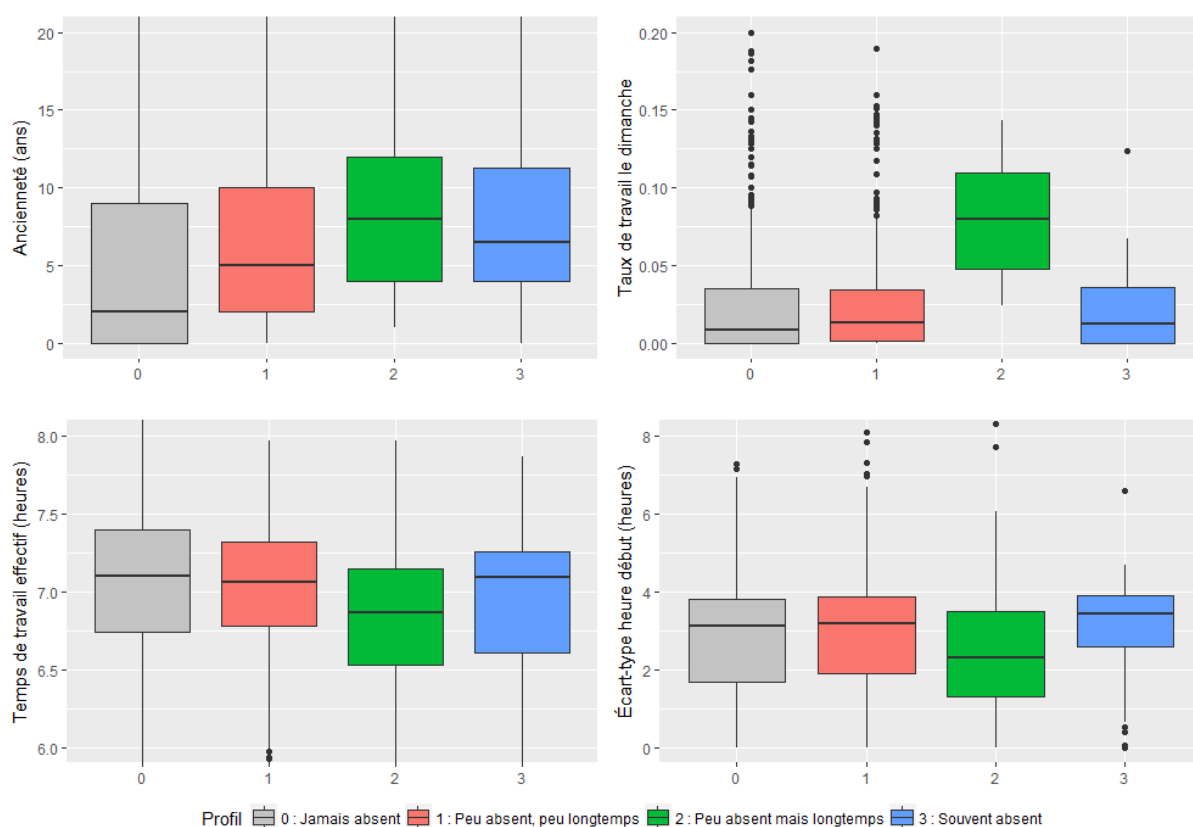


Figure 15 – 4 profils selon 4 variables d'intérêt

Source : figures 14 et 15 : Base Absentéisme Transdev, Base agrégée et apurée par matricule, 1 567 agents.

Note de lecture : La moitié des agents jamais absents (profil 0) ont un temps de travail effectif moyen entre 6.75 et 7.4 heures, avec une médiane à 7.1 heures.

Le profil 2 (peu souvent absent mais longtemps) semble le plus atypique. Il est présent depuis plus longtemps dans l'entreprise et travaille fréquemment le dimanche. Son temps de travail effectif et sa variabilité d'heure de prise de poste sont plus faibles que les autres profils. L'âge n'a pas été représenté, les différences étant minimales excepté pour le profil 2 légèrement plus âgé.

Conclusion

L'analyse exploratoire de l'absentéisme dans une population de conducteurs de bus urbains de la société Transdev révèle des relations plus ou moins intuitives avec les « *conditions de travail* ». À grands traits, on peut ainsi retenir trois grands éléments de vigilance pour limiter ce phénomène, représentés par le graphique 16.

Tout d'abord, la **variabilité des indicateurs d'absence entre les entités locales** invite à poursuivre les actions de prévention auprès du management de proximité. Ensuite, le **poids de certaines contraintes d'organisation de l'activité** comme le travail en fin de semaine ou les fluctuations importantes des grilles de planning plaide pour une gestion équilibrée et partagée de ces exigences de service, sans doute plus facile à recommander qu'à mettre en place. Enfin, **l'attention aux caractéristiques personnelles** parmi lesquelles l'ancienneté, l'âge ou le sexe de l'agent est probablement décisive tant ce phénomène mêle des aspects professionnels et individuels.

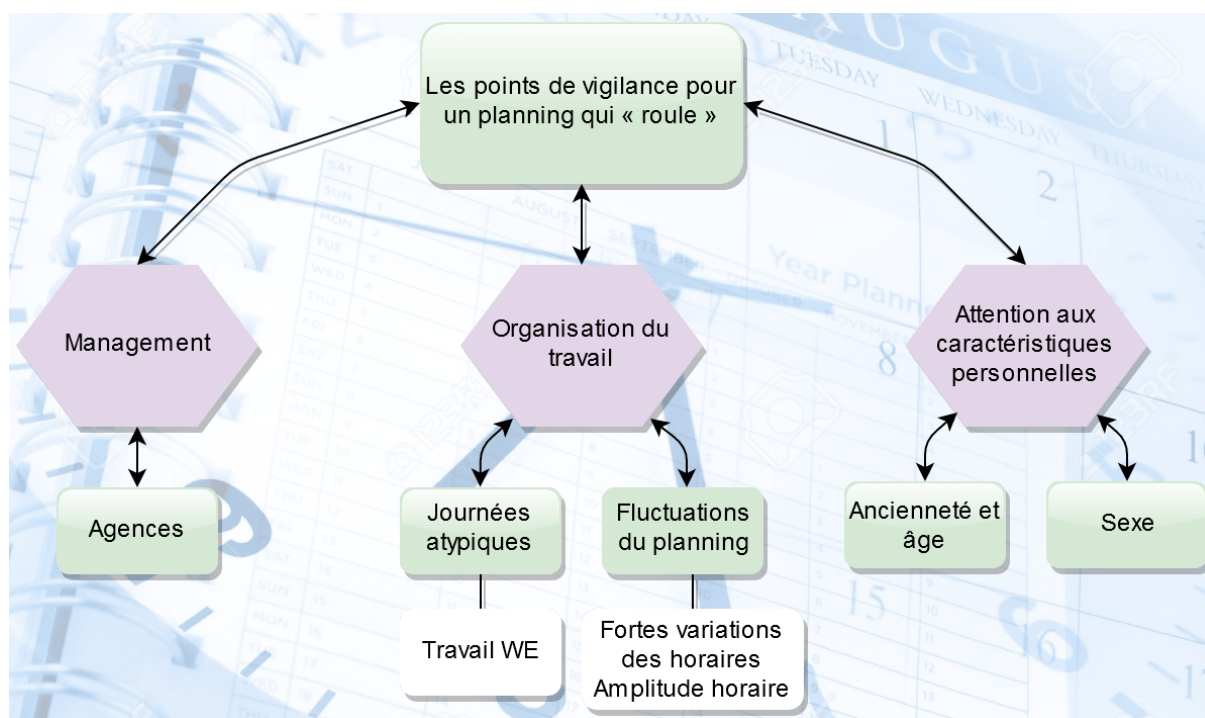


Figure 16 – Points de vigilance pour un planning optimisé

Ces premiers résultats semblent d'ailleurs concorder avec certaines actions déjà engagées par Transdev dans son plan initié en 2015. Nos indicateurs qui établissent une légère hausse de l'absentéisme entre 2015 et 2016 soulignent que ce travail s'inscrit dans la durée. Ils montrent en tout cas que la méthode statistique peut sans doute compléter une vision pragmatique déjà mise en oeuvre par l'entreprise.

L'exploration des liens entre absentéisme et conditions de travail pourrait encore être approfondie. Nous avons ainsi souligné l'intérêt d'étoffer les informations disponibles sur les individus (situation familiale, nombre d'enfants, distance domicile-travail, ...) ou de développer les variables décrivant les spécificités des trajets (difficulté, dangerosité, ...). Cette attention à la perception qualitative des agents pourrait conduire à des actions plus ciblées, tout en stimulant l'échange et l'écoute des réalités du terrain. De la même façon, les notions de fluctuations des plannings peuvent sans aucun doute être abordées de manière plus fine, révélant d'autres interactions.

Bibliographie

- Adams, N. M. & Hand, D. J. *Comparing classifiers when the misallocation costs are uncertain*. *Pattern Recognition*, 1999.
- Anact, *L'absentéisme : Outils et Méthodes pour agir*, 2016.
- Laurence Dauphin, Marie-Anne Le Garrec et Frédéric Tardieu, *Le Tourisme en France. Les vacances des Français depuis 40 ans*, Insee, 2008.
- Direccte Nord-Pas-de-Calais, Mission Synthèse, *Le travail en horaires décalés. Estimation des effectifs salariés concernés en région Nord-Pas-de-Calais*, 2012.
- Robin Genuer et Jean-Michel Poggi, *Arbres CART et Forêts aléatoires, Importance et sélection de variables*, ISPED, LMO, 2016.
- David W. Hosmer and Stanley Lemeshow, *Applied Logistic Regression*. A Wiley-Interscience Publication from John Wiley & Sons, 2nd Edition, 2000.
- David G. Kleinbaum, Mitchel Klein, *Logistic Regression, Statistics for Biology and Health*, Springer, 2010.
- Frédéric Santos, *Arbres de décision*, CNRS, 2015.
- UTP, *Bilan social des transports urbains 2015*, 2016.
- Site internet : <https://eric.univ-lyon2.fr/>, cours et tutoriels de Ricco Rakotomalala, Université Lumière Lyon 2.

Annexes

1 Taux d'absentéisme par variable d'intérêt

	Taux d'absentéisme
<i>Classes d'âge</i>	
21 à 35 ans	10,8 %
35 à 43 ans	9,8 %
43 à 52 ans	9,2 %
52 à 71 ans	8,4 %

Tableau 15 – Taux d'absentéisme selon la classe d'âge de l'agent

Note de lecture : Selon la définition retenue, le taux d'absentéisme pour les conducteurs âgés de 21 à 35 ans s'élève à 10,8 %

Source : Base Absentéisme Transdev, Base agrégée et apurée par matricule, 1 567 agents.

	Taux d'absentéisme
<i>Agences Transdev</i>	
TD Ecquevilly	16,8 %
Proxyway Versailles	16,7 %
CSO	13,4 %
TD Conflans	8,0 %
Autocars Tourneux	7,5 %
Rambouillet	6,9 %
TD Montesson les rabaux	6,8 %
TD Montesson la boucle	5,3 %
TD Houdan	4,0 %

Tableau 16 – Taux d'absentéisme selon l'agence territoriale Transdev de l'agent

Note de lecture : Selon la définition retenue, le taux d'absentéisme des conducteurs de l'agence TD Ecquevilly s'élève à 16,8 %.

Source : Base Absentéisme Transdev, Base agrégée et apurée par matricule, 1 567 agents.

	Taux d'absentéisme
<i>Taux de travail matinal</i>	
Entre 0 et 2,1 %	13,9 %
Entre 2,1 et 10,2 %	6,7 %
Entre 10,2 et 100 %	4,9 %

Tableau 17 – Taux d'absentéisme selon la proportion de travail matinal de l'agent

Note de lecture : Selon la définition retenue, le taux d'absentéisme pour les conducteurs dont la proportion de travail matinal dans le planning est inférieure à 2,1 % s'élève à 13,9 %

Source : Base Absentéisme Transdev, Base agrégée et apurée par matricule, 1 567 agents.

	Taux d'absentéisme
<i>Taux de travail dominical</i>	
Entre 0 et 0,2 %	5,1 %
Entre 0,2 et 1,4 %	4,4 %
Entre 1,4 et 3,7 %	5,9 %
Entre 3,7 et 40 %	22,9 %

Tableau 18 – **Taux d'absentéisme selon le taux de travail dominical de l'agent**

Note de lecture : Selon la définition retenue, le taux d'absentéisme pour les conducteurs dont la proportion de travail dominical dans le planning est inférieure à 0,2 % s'élève à 5,1 %

Source : Base Absentéisme Transdev, Base agrégée et apurée par matricule, 1 567 agents.

	Taux d'absentéisme
<i>Journée</i>	
Lundi	8,2 %
Mardi	8,3 %
Mercredi	8,4 %
Jeudi	8,4 %
Vendredi	8,6 %
Samedi	14,8 %
Dimanche	35,8 %

Tableau 19 – **Taux d'absentéisme selon la journée de la semaine**

Note de lecture : Selon la définition retenue, le taux d'absentéisme s'élève à 8,2 % pour la journée du lundi en tenant compte de l'ensemble des agents

Source : Base Absentéisme Transdev, Base mise en forme et apurée, 812 286 observations

	Taux d'absentéisme
<i>Classes d'ancienneté</i>	
0 à 3 ans	6,6 %
3 à 7 ans	13,0 %
7 à 12 ans	9,9 %
12 à 41 ans	9,5 %

Tableau 20 – **Taux d'absentéisme selon l'ancienneté en classes de l'agent**

Note de lecture : Selon la définition retenue, le taux d'absentéisme pour les conducteurs dont l'ancienneté dans l'entreprises est inférieure à 3 ans s'élève à 6,6 % en tenant compte de l'ensemble des agents

Source : Base Absentéisme Transdev, Base agrégée et apurée par matricule, 1 567 agents

	Taux d'absentéisme
<i>Type de régime horaire</i>	
CPS	4,7 %
Temps complet	9,4 %
Temps partiel	17,1 %
Multiple	16,7 %

Tableau 21 – **Taux d'absentéisme selon le régime horaire de l'agent**

Note de lecture : Selon la définition retenue, le taux d'absentéisme pour les conducteurs en période scolaire (CPS) s'élève à 4,7 % en tenant compte de l'ensemble des agents

Source : Base Absentéisme Transdev, Base agrégée et apurée par matricule, 1 567 agents

	Taux d'absentéisme
<i>Temps de travail effectif moyen</i>	
2 à 6,74 heures	11,7 %
6,74 à 7,07 heures	8,5 %
7,07 à 7,36 heures	6,7 %
7,36 à 8,83 heures	5,5 %

Tableau 22 – **Taux d'absentéisme selon la durée moyenne de travail effectif par jour de l'agent**

Note de lecture : Selon la définition retenue, le taux d'absentéisme pour les conducteurs dont la durée de la journée de travail effectif est comprise en moyenne entre 2 et 6 heures 44 s'élève à 11,7 % en tenant compte de l'ensemble des agents

Source : Base Absentéisme Transdev, Base agrégée et apurée par matricule, 1 567 agents

	Taux d'absentéisme
<i>Période de vacances scolaires</i>	
Hors Vacances scolaires	9,6 %
Vacances scolaires	9,5 %
Vacances d'été	8,9 %
Vacances d'été (Juillet)	8,9 %
Vacances d'été (Août)	8,9 %
Vacances de Noël	9,5 %
Vacances de Paques	9,5 %
Vacances de Toussaint	10,1 %
Vacances de Février	11,3 %

Tableau 23 – **Taux d'absentéisme selon la période de vacances scolaires**

Note de lecture : Selon la définition retenue, le taux d'absentéisme hors vacances scolaires s'élève à 9,6 %

Source : Base Absentéisme Transdev, Base mise en forme et apurée, 812 286 observations.

2 Création de la base binaire

À partir de la base initiale, nous avons tout d'abord agrégé les informations de chaque agent afin de résumer les caractéristiques dans le but d'obtenir une observation par conducteur de bus. Certaines variables stables dans le temps n'ont posé aucune difficulté (âge, sexe, agence, etc ...) en revanche, pour les variables qui évoluent selon les jours (amplitude horaire, horaire de fin de journée, nombre de vacations, etc ...), nous avons calculé la moyenne et l'écart-type afin de disposer d'indicateurs synthétiques.

Dans un second temps, pour pouvoir assurer la comparabilité des résultats, nous avons sélectionné les données des agents sur un même nombre de jours. En effet, notre période d'étude dure deux années, mais tous les agents n'étaient pas présents dans l'entreprise tout ce temps. Plus précisément, seulement 15 % des agents ont été présents 729 jours (la totalité des deux années). Le graphique ci-dessous montre la proportion d'agents présents en fonction de leur durée dans l'entreprise. Ainsi, il n'était guère pertinent de comparer l'absence ou la présence d'agents travaillant un nombre de jours différents. Nous avons donc choisi dans ce premier modèle de ne conserver que les agents ayant été présents au moins 400 jours. Ils représentent plus des 2/3 des agents pour lesquels des informations sont disponibles.

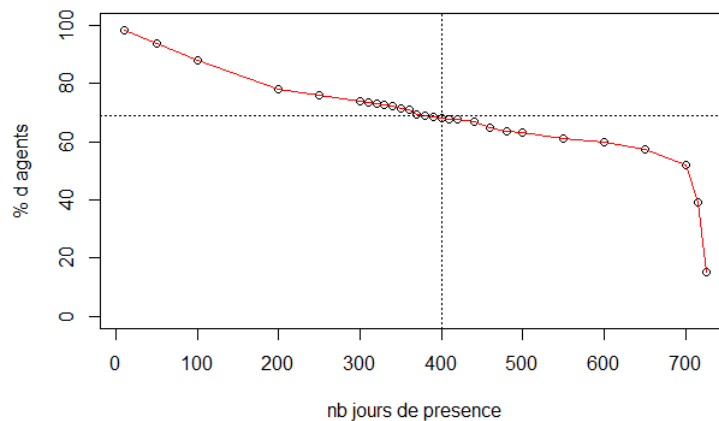


Figure 17 – Nombre de jours de présence des agents durant la période étudiée

Source : Base Absentéisme Transdev, Base binaire, 992 observations

3 Classification CART

La méthode de classification utilisée pour visualiser les variables clivantes en introduction des modèles binaires et polytomiques est l'algorithme CART qui signifie « Classification And Regression Trees ».

Un arbre est composé d'une racine qui contient l'ensemble de la population qui va se répartir selon des divisions binaires successives à travers des nœuds jusqu'à ce que plusieurs feuilles représentent une partition exhaustive.

Plus la connaissance de la variable d'intérêt Y est grande selon les variables prédictrices X_i , moins l'entropie (i.e. le désordre) est grande. Le critère de division utilisé est l'indice de Gini qui évalue la pureté de chaque feuille et permet la création de feuilles le plus homogène possible. L'algorithme s'arrête lorsqu'une division supplémentaire ne réduit plus le désordre. Il faut ensuite élaguer l'arbre pour éviter le sur-apprentissage associée à une complexité trop importante de l'arbre qui décrirait trop parfaitement la base de données empêchant une généralisation à un nouveau jeu de données. L'élagage s'effectue par validation croisée où des sous-ensembles de la base permettent tour à tour d'apprendre et de tester les prédictions de Y par rapport à sa vraie valeur. L'arbre optimal correspond à l'erreur minimale de prédiction et à l'arbre le plus simple possible correspondant. Les arbres CART sont très sensibles au tirage de l'échantillon. Et comme les bases binaire et polytomique comportent peu d'agents, les arbres CART diffèrent selon le tirage de l'échantillon d'apprentissage-validation, il a donc été choisi d'utiliser CART uniquement pour classer la population et non prédire l'absence ce qui aurait pu permettre de comparer la prédiction de CART avec les modèles binaires, polytomiques et GEE développés par la suite.

4 Méthodes de discrétisation supervisées utilisées

Tout d'abord, la discrétisation non supervisée consiste à découper une variable quantitative X en classes. Le regroupement se fait selon des intervalles d'une certaine longueur, des tranches de population (quantiles...), selon la distribution de X ou encore selon une classification ascendante hiérarchique, etc.

Alors que la discrétisation supervisée consiste à utiliser la variable d'intérêt ou variable « cible » Y pour découper X en classes. Deux approches sont à distinguer : les approches ascendantes et descendantes.

L'approche ascendante du Chi merge de Kerber a été testée, elle regroupe des intervalles adjacents

en utilisant un test du Khi2 d'équivalence distributionnelle. C'est la même logique qu'en analyse factorielle sur le regroupement des profils lignes à l'aide de la métrique du Khi2 mais avec des intervalles adjacents qui fusionnent ensemble. L'algorithme s'arrête lorsque tous les intervalles adjacents sont de profils significativement différents. La difficulté de cette approche réside dans le paramétrage de la significativité des tests et le traitement des bases contenant plus d'une centaine d'observations. Cette approche s'est avérée peu adaptée à nos bases.

Les deux autres approches utilisées : CART et MDLP sont des méthodes descendantes i.e. on part de l'ensemble de la population pour effectuer des découpages successifs aboutissant à la partition.

L'algorithme CART, dont le principe a été évoqué dans l'annexe précédente, est à réaliser sur chaque variable X selon Y. Dans l'exemple ci-dessous, l'ancienneté est ainsi découpée à l'aide de deux clivages binaires et sera donc discrétisée en 3 intervalles.

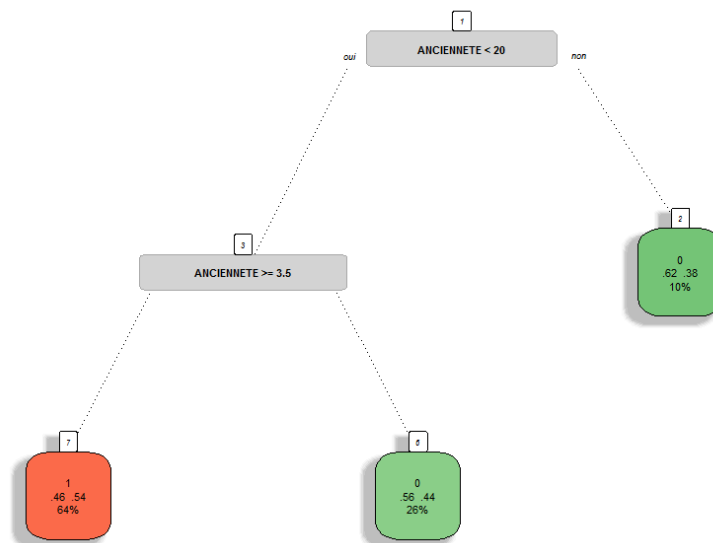


Figure 18 – **Exemple de discrétisation CART sur l'ancienneté**

Source : Base du modèle binaire 992 agents présents sur 400 jours

Note de lecture : Les agents dont l'ancienneté est supérieure ou égale à 20 ans représentent 10 % de la population et sont classés avec une probabilité de 0,62 dans la classe 0 des non-absents.

L'avantage de cette méthode est sa rapidité sur les grandes bases. Mais elle ne fonctionne pas lorsque la variable d'intérêt Y contient une modalité avec peu d'éléments. C'est le cas pour les modèles GEE sur 7 ou 15 jours où le taux d'absence est trop faible, entre 0 et 10 % de modalités « Y = 1 », i.e. d'absents, selon la période de référence.

L'algorithme MDLP (Minimum Description Length Principle) utilisé pour discrétiser les variables des modèles GEE est également basé sur la notion d'entropie et la quantité d'information contenue dans la base. Le principe de la longueur de description minimale est le critère d'arrêt, c'est un principe de compression de données qui minimise la quantité d'information utilisée en recherchant des similarités. À chaque étape, MDLP teste 2 hypothèses : découper ou non un intervalle, l'hypothèse retenue est celle qui va être représentée par le minimum d'information tout en tenant compte des éléments exclus ou inclus dans le mauvais intervalle.

5 Corrélations entre les variables explicatives

Il est nécessaire de vérifier l'absence de colinéarité entre les variables explicatives. Il faut donc calculer la matrice de corrélation pour tous les prédicteurs. En effet, si deux variables sont hautement corrélées ($\text{corr} > 0.8$), l'indépendance des deux prédicteurs ne sera plus assurée et si l'un est significatif

l'autre le sera aussi.

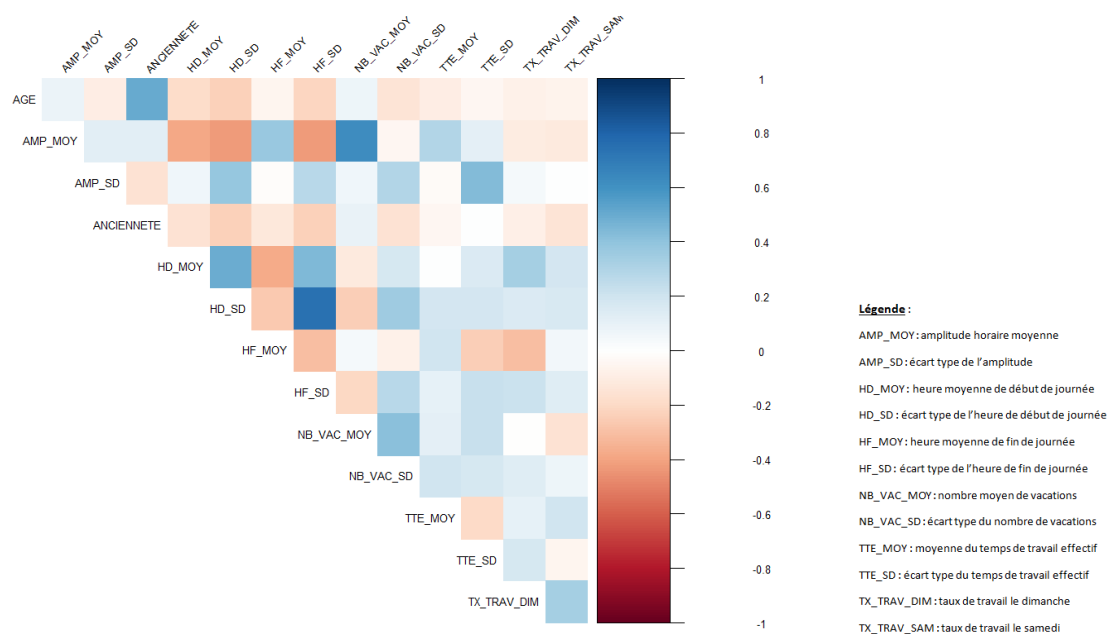


Figure 19 – Matrice de corrélation des variables explicatives

Note de lecture : La corrélation entre l'âge et l'ancienneté est d'environ 0.5
Source : Base Absentéisme Transdev

L'analyse de la corrélation entre les variables ne montre aucune liaison forte. On remarque juste une corrélation assez forte entre « l'écart type de l'heure de début de journée » et « l'écart type de l'heure de fin de journée » (corr=0.74), ainsi qu'entre « l'amplitude moyenne » et « le nombre moyen de vacances » (corr=0.62). Cette dernière relation s'expliquant facilement : les agents qui ont une plus grande amplitude horaire ont logiquement davantage de pauses durant leur journée de travail. Cette absence de corrélation dans nos variables explicatives nous permet de réaliser une régression logistique.

6 Note méthodologique : test du χ^2
d'indépendance entre l'ancienneté et l'amplitude horaire

	Ancienneté [1 - 3.5]	Ancienneté [3.5 - 19.5]	Ancienneté [19.5 - 41]	Total
AMP [3.3 - 8.06]	16	75	9	100
AMP [8.06 - 14]	239	560	93	892
Total	255	635	102	992

Source : Transdev
Tableau 24 – Tableau des effectifs observées

	Ancienneté [1 - 3.5]	Ancienneté [3.5 - 19.5]	Ancienneté [19.5 - 41]	Total
AMP [3.3 - 8.06]	25.7	64.0	10.3	100
AMP [8.06 - 14]	229.3	571.0	88.6	892
Total	255	635	102	992

Source : Transdev
Tableau 25 – Tableau des effectifs théoriques

Soit X : « la valeur de ancienneté » à p = 3 modalités ; Y : « l'amplitude horaire » à q = 2 modalités.

$$\begin{cases} H_0 & : \text{la valeur de l'ancienneté et l'amplitude sont indépendantes} \\ H_1 & : \text{la valeur de l'ancienneté et l'amplitude ne sont pas indépendantes} \end{cases}$$

Notre statistique de test est la suivante :

$$D = \sum_{i=1}^p \sum_{j=1}^q \frac{(n_{ij}^* - n_{ij})^2}{n_{ij}}$$

où n_{ij} est l'effectif théorique sous H_0 et n_{ij}^* les effectifs observés.

On sait qu'ici $p = 3$ et $q = 2$ donc $D \underset{H_0}{\sim} \chi_{(3-1)(2-1)}^2$.

Nous rejeterons donc si $[D \geq q_{(1-\alpha)}]$ avec :

$$\mathbb{P}(D \geq q_{(1-\alpha)}) = \alpha$$

Nous prenons ici un risque $\alpha = 5\%$. Or, sous H_0 , la statistique de test suit une χ^2 à $2 \times 1 = 2$ degrés de liberté. Ainsi nous rejeterons le test si $D > 5.99$. Nous avons :

$$D = (3.66 + 1.9 + 0.16) + (0.4 + 0.2 + 0.2) = 6.50$$

Ici $6.5 > 5.99$, on rejette H_0 , les variables sont dépendantes.

7 Validité du modèle binaire

Pour chacun des trois tests suivants, les deux hypothèses alternatives sont :

$$\begin{cases} H_0 : \text{tous les paramètres sont nuls;} \\ H_1 : \text{au moins un des paramètres est non nul.} \end{cases}$$

- Test de Wald : $\hat{\beta}_q' \widehat{\sum_q \hat{\beta}_q} \underset{H_0}{\sim} \chi_q^2$ avec $\hat{\beta}_q'$: le vecteur des coefficients estimés et $\widehat{\sum_q}$ le premier bloc d'ordre q de $\widehat{In}(\hat{\beta})$, un estimateur de la matrice d'information de Fisher. La statistique de test donne 118.7 et avec 27 ddl, H_0 est rejetée avec une p-value < 0.0001 ;
- Test de la déviance (ou du rapport de vraisemblance) : on compare le modèle complet au modèle H_0 , dont on a enlevé les régresseurs :
 $2[\ln L(\hat{\beta}) - \ln L(\hat{\beta}_{H_0})] \underset{H_0}{\sim} \chi_q^2$
 La statistique de test donne 165.2, et avec 27 ddl, on rejette H_0 avec une p-value < 0.0001 ;
- Test du score :
 $\nabla \ln L(\hat{\beta}_{H_0})' In(\hat{\beta}_{H_0}) \nabla \ln L(\hat{\beta}_{H_0}) \underset{H_0}{\sim} \chi_q^2$
 La statistique de test donne 145.8, et avec 27 ddl, H_0 est rejetée avec une p-value < 0.0001 .

Ces trois tests valident donc globalement le modèle. Ensuite, nous avons également testé la significativité de chacune des variables du modèle ajusté ; elles sont toutes significatives en présence des autres variables, au seuil $\alpha = 5\%$, à l'exception de l'agence et du sexe que nous avons conservé à titre d'information.

Effet	DDL	Khi-2 de Wald	p-value
Agence	7	8.46	0.29
Sexe	1	1.72	0.19
TTE moyen	4	33.92	<0.001
AMP moyen	1	8.23	0.004
Nb moyen vacations	3	8.18	0.042
TTE écart-type	2	11.30	0.003
AMP écart-type	3	17.41	0.0006
Ancienneté	2	7.84	0.02
Age	2	13.95	0.0009
Taux travail samedi	1	23.05	<0.0001
Taux travail dimanche	1	24.88	<0.0001

Tableau 26 – **Test de significativité de chaque variable, modèle binaire**

Source : Absentéisme Transdev, base modèle binaire, modèle binaire

8 Test de Hosmer-Lemeshow

Le test de Hosmer-Lemeshow permet de mesurer la qualité de l'ajustement de notre modèle : On forme des groupes (habituellement 10) en rangeant les observations par décile selon les probabilités prédites. On compare ensuite dans chaque groupe les effectifs observés et théoriques.

H_0 : les probabilités théoriques sont proches de celles observées ;

H_1 : les probabilités théoriques sont différentes des observées.

Dans chaque groupe g , d'effectif n_g , on définit :

n_{g1} , le nombre d'absents observés ;

n_{g0} , le nombre de présents observés ;

$\hat{n}_{g1}$: la somme des proba. prédites des observations du groupe g ;

$$p_{g1} = \frac{\hat{n}_{g1}}{n_g}.$$

La statistique de test : $c = \sum_g \frac{n_{g1} - \hat{n}_{g1}}{\hat{n}_{g1}(1 - p_{g1})}$

Sous H_0 , la statistique de test suit une loi du χ^2 à (nb de groupes-2) degrés de liberté. Pour nos données, on trouve $c=3.37$ et pour 8 degrés de libertés, au seuil 5% on ne rejette pas l'hypothèse H_0

Soit n observations et Y une variable à expliquer, $X_1, \dots, X_p$ des variables explicatives ainsi qu'un vecteur β de paramètres.

9 Modèle GEE

Le modèle logistique classique est défini par la relation :

$$\text{Logit}(p(X)) = X \cdot \beta \text{ avec } X = (1, X_1, \dots, X_p)$$

$$\text{On a alors : } p(X) = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

Pour calculer le vecteur de paramètres β on utilise la vraisemblance des observations :

$$L(\beta) = \prod_i p(X_i)^{Y_i} \cdot (1 - p(X_i))^{1-Y_i}$$

$$\ln L(\beta) = \sum_i Y_i X_i \beta - \ln(1 + e^{X_i \beta})$$

$$\frac{\partial \ln(L)}{\partial \beta} = \sum_i Y_i X_i - \frac{X_i e^{X_i \beta}}{1 + e^{X_i \beta}}$$

$$\frac{\partial \ln(L)}{\partial \beta} = \sum_i X_i (Y_i - p(X_i)) = 0$$

Ensuite, il s'agit de résoudre cette équation afin d'obtenir les paramètres β

Mais dans notre cas, nos observations sont corrélées, c'est à dire $\text{Corr}(Y_i, Y'_i) \neq 0$, nous devons donc utiliser une extension du modèle GLM, le modèle GEE (Generalized Estimating Equations)

développé par Liang et Zeger (1986). Ces équations d'estimation généralisées évitent le calcul de la vraisemblance jointe (qui s'avère compliquée à calculer dans le cas de données non indépendantes). Le modèle part du principe que les données sont corrélées pour un même individu, mais indépendantes entre deux individus. Les corrélations intra-individu sont modélisées par une matrice de corrélation dite de travail.

L'estimateur proposé est solution du système d'équations :

$$\sum D'W^{-1}(Y - p(X)) = 0$$

Où D représente la matrice des dérivées de $p(X)$ par rapport à β et W est la matrice de variance-covariance de travail définie par : $V = A^{\frac{1}{2}}.R(\alpha).A^{\frac{1}{2}}$

A est la matrice diagonale composée des variances $\text{Var}(Y)$. La matrice $R(\alpha)$ représente une matrice de corrélation de travail (working correlation matrix) qui dépend d'un paramètre α .

Pour minimiser le nombre de paramètres à estimer, la structure de la matrice de corrélation est commune à tous les individus. L'intérêt de la méthode GEE repose sur la robustesse des estimations obtenues.

L'utilisateur doit déterminer au préalable le type de matrice de corrélation, en fonction des données :

> Matrice Indépendante

$$R(\alpha) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Ce choix d'une matrice Identité permet de se ramener au cas où les données sont indépendantes.

> Matrice échangeable

$$R(\alpha) = \begin{pmatrix} 1 & \alpha & \dots & \alpha \\ \alpha & 1 & \dots & \alpha \\ \dots & \dots & \dots & \dots \\ \alpha & \alpha & \dots & 1 \end{pmatrix}$$

On considère que toutes les mesures répétées chez un individu ont une corrélation commune.

> Matrice autorégressive

$$R(\alpha) = \begin{pmatrix} 1 & \alpha & \dots & \alpha^{n-1} \\ \alpha & 1 & \dots & \alpha \\ \dots & \dots & \dots & \dots \\ \alpha^{n-1} & \alpha & \dots & 1 \end{pmatrix}$$

Cette structure de matrice est utilisée quand on considère que les corrélations entre les mesures diminuent avec le temps..

> Matrice non-structurée

$$R(\alpha) = \begin{pmatrix} 1 & \alpha_{12} & \dots & \alpha_{1J} \\ \alpha_{12} & 1 & \dots & \alpha_{2J} \\ \dots & \dots & \dots & \dots \\ \alpha_{1J} & \alpha_{2J} & \dots & 1 \end{pmatrix}$$

Aucune structure particulière n'est imposée.