JMS Journées de méthodologie statistique de l'Insee 2018

Méthodes de Machine Learning dans le cadre de la détection d'anomalies dans les Déclarations Sociales Nominatives (DSN)

Thomas Deroyon(*), Julie Djiriguian(*), Marie Villoing-Cordier (**)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale
(**) Insee, Direction des Statistiques Démographiques et Sociales

Mots-clés. Machine learning, big data, données administratives

Résumé

Cet article s'inscrit dans le projet global de refonte des Déclarations Annuelles de Données Sociales (DADS) qui vise à générer automatiquement des fichiers mensuels et un ensemble de fichiers annuels de Déclarations Sociales Nominatives (DSN), analogues pour ces derniers aux Déclarations Annuelles de Données Sociales qui servent de source de référence pour l'analyse des salaires en France. Cette refonte, qui se traduit par une reconstruction complète des chaînes de production des données sur les emplois et les salaires, est l'occasion de repenser la méthodologie des différents processus et traitements statistiques qui leur sont appliqués. Cet article porte sur un de ces traitements particuliers, la détection des erreurs, qui constitue la première étape du contrôle et redressement (ou *data editing*) des données de la DSN ou des DADS. Ce sujet a en effet fait l'objet de nombreux développements méthodologiques, notamment issus du champ de l'apprentissage statistique. La disponibilité de nouvelles données mensuelles massives incite à tester des méthodes de *machine learning* pour détecter automatiquement les anomalies.

Cette présentation a pour but de rendre compte des travaux qui ont été réalisés dans ce cadre, notamment des difficultés que nous avons rencontrées. Ces difficultés tiennent aux données elles-mêmes, et à la nature du problème que nous essayons de traiter.

Nous ne disposons en effet pas d'un échantillon de données de la DSN ou des DADS dans lesquelles les erreurs et les observations valides sont identifiées. À partir des données actuellement disponibles, il est seulement possible de rechercher des anomalies, *i.e.* des observations atypiques, *via* un apprentissage non supervisé. Cet exercice est cependant beaucoup plus difficile et très dépendant de la manière dont est défini ce qui constitue ou pas une anomalie. Chaque algorithme disponible repose en effet sur une représentation spécifique de ce qu'est une donnée normale et une donnée atypique et cherche, sur la base de cette définition, à identifier les secondes. Il est de plus impossible, sans échantillon d'observations dont la qualité est connue, de juger de la pertinence de la détection opérée par les différentes méthodes : quelle part des anomalies qu'elles identifient correspond effectivement à des erreurs ? Quelle part des erreurs dans les données arrivent-elles à détecter ?

Les travaux que nous avons réalisés nous ont permis de comparer les caractéristiques des anomalies identifiées par trois méthodes issues de l'apprentissage statistique avec les méthodes actuellement utilisées dans la chaîne de traitement des DADS, reprises dans celle de la DSN. Ils montrent que les algorithmes détectent en grande partie des anomalies différentes, en fonction de la manière dont ils définissent et identifient les erreurs présumées. L'utilisation combinée de plusieurs algorithmes de détection d'erreurs permettrait ainsi de couvrir un spectre plus large d'erreurs potentielles.

Abstract

This article joins in the global project of revision of the Annual Declarations of Social Data (DADS), wage bill information firms have had to fill in annually for each of their employees for payroll and fiscal tax purposes and which are now progressively replaced by new monthly Social Nominative Declarations files. These individual Annual Declarations of Social Data and Social Nominative Declarations are the standard national source for the analysis of wages in France. The reconstruction of the usual production line gives the opportunity to rethink the methodology of this process and more particularly the process for anomaly detection. Indeed, the management of big data each month urges us to test methods of machine learning to detect automatically anomalies.

This presentation aims at reporting works which were realized in this context. One part will focus on the specifities of of the data themselves, and the nature of the problem ; and another one on the methods tested and the results.

In particular, we will see that he statistical processing of 8 these data, as DADS, and their control, turn out complex, because two of the three main variables, which are the gross and net salaries, are not directly observed but calculated from data reported by companies to respect their social obligations. Therefore, the anomalies may arise either at the level of the elementary variables, or at the level of aggregated variables. In this paper we choose to detect anomalies directly on the synthetic variables (gross salary, net salary) since the detection stage occurs after a consistency analysis if the elementary variables. Indeed, the study performed at the level of reported variables does not exempt us from an analysis of aggregated variables of salaries. Furthermore, the available reported variables in the individual declarations depend on the standard according to which companies have to fill their declarations. Yet, if this standard changed, the reported variables could change and thus, the detection of outliers based on their values could be unusable.

For now, only an unsupervised learning, which consists of the detection of anomalies without labeled data, is feasible with the available data. This work is however much more complex because it is extremely hard to distinguish the outliers from normal data without ever assessing the real success of the process. We plan to use and test several methods like association rules, isolation forest, Local Outlier Factor (LOF). The presentation will present the algorithms tested for anomaly detection and compare them on observed data in a attempt of performance evalution.

Introduction

La déclaration sociale nominative (DSN) est une déclaration obligatoire dont les employeurs doivent s'acquitter auprès des organismes de sécurité sociale. Elle a remplacé à partir du 1er janvier 2017 la quasi-totalité des déclarations sociales obligatoires des employeurs : déclarations obligatoires des mouvements de main d'œuvre par exemple, mais aussi déclarations annuelles de données sociales. Cette déclaration comporte des informations sur les événements de la vie professionnelle du salarié (embauche, arrêt de travail, fin de contrat) et des informations sur les périodes d'emploi qu'il a réalisées auprès de ses employeurs avec les caractéristiques de celles-ci (rémunérations versées, jours de début et de fin, nombre d'heures de travail...). Ces déclarations sociales répondent à un double objectif : elles permettent le recouvrement des cotisations sociales auprès des employeurs par l'agence centrale des organismes de sécurité sociale (Acoss) qui les réceptionne ; elles permettent également l'ouverture et l'alimentation des droits des salariés auprès des organismes de sécurité sociale et de l'assurance chômage.

Les déclarations annuelles de données sociales (DADS), que la DSN remplace à compter de 2017, constituaient pour l'Insee la source de référence sur l'emploi et les salaires. Aussi, le remplacement des DADS par la DSN a un impact important sur la statistique publique ; elle nécessite des adaptations importantes des systèmes d'information existants, qui s'inscrivent dans un programme plus global de refonte du système d'information sur l'emploi et les revenus d'activité. Cette refonte, qui permet la construction d'une chaîne de production adaptée au nouveau format des déclarations, est aussi l'occasion de repenser et d'adapter la méthodologie des différents processus et traitements statistiques appliqués aux données, notamment leurs contrôles et redressements.

Ces traitements ont par ailleurs fait l'objet d'avancées méthodologiques importantes au cours des années récentes. La détection des erreurs dans les données, sur laquelle nous allons nous concentrer dans cet article, a notamment fait l'objet d'une attention particulière de la part de spécialistes de nombreux domaines, tels que le secteur bancaire, le secteur informatique et le secteur médical, dans lesquels la sécurité, la détection de fraudes, le décèlement d'une intrusion ou encore le repérage d'une cellule anormale, constituent un enjeu majeur. Ces développements ont favorisé l'émergence d'un vaste ensemble de méthodes innovantes, relevant du champ de l'apprentissage statistique ou *machine learning*, qui, en retour, peuvent être également appliquées à la détection plus classique des erreurs dans les données utilisées pour des exploitations statistiques. L'objectif de cet article est d'étudier l'intérêt et l'apport de quelques unes de ces méthodes (forêt d'isolation, *local outlier factor*, règles d'association floues) pour la détection d'erreurs dans les données.

La détection d'erreurs dans les données est un exercice complexe, qui peut être fortement contraint par la quantité et la nature des informations disponibles. Il peut s'exercer dans trois approches assez différentes :

— APPROCHE SUPERVISÉE : nous savons, pour un échantillon d'observations, si les informations qu'elles contiennent sont correctes ou non, et les erreurs qui les affectent. Dans ce cas, il est possible d'essayer de décrire et d'identifier ces erreurs dans l'échantillon à partir des informations disponibles, l'objectif étant de construire des algorithmes permettant de prédire, pour chaque nouvelle observation, si elle est une erreur ou pas en fonction des valeurs des variables disponibles. Le large éventail des méthodes de classification, permettant la prédiction d'une variable qualitative (ici l'indicatrice de présence d'une erreur), qui s'étendent de la simple régression logistique aux réseaux de neurones en passant par des méthodes ensemblistes comme les forêts aléatoires, peut être alors mobilisé. De plus, il est possible de tester la qualité de la détection opérée par les différents algorithmes en comparant leurs performances sur une fraction de l'échantillon d'observations labellisées non utilisée pour construire les modèles. Cette approche dite supervisée n'est cependant que rarement utilisable car elle suppose un travail préalable d'analyse fine d'un échantillon de données par des spécialistes du sujet, ce qui est en général extrêmement coûteux ; cela impliquerait par exemple qu'un échantillon de DSN soit examiné à la main par des gestionnaires pour vérifier si elles contiennent ou pas des erreurs. De plus, cette approche ne peut détecter que les erreurs déjà connues et observées dans les données. Elle est, par construction, impuissante à identifier des erreurs qui ne se seraient jamais produites ;

— APPROCHE SEMI-SUPERVISÉE : dans cette approche, nous disposons d'un échantillon d'observations dans lequel toutes les observations sont considérées comme normales. L'idée est alors, à l'aide de techniques algorithmiques adaptées (comme la méthode des *one-class SVM*), d'identifier la frontière de l'ensemble, ou *cluster*, que forment les données disponibles, englobant toutes les observations normales. Par la suite, les nouvelles observations situées en dehors de cette frontière sont automatiquement considérées comme des erreurs ou au moins des anomalies. Toutefois, la constitution d'un échantillon sans anomalie reste coûteuse en termes de labellisation pour les gestionnaires. Elle nécessite un travail minutieux car la présence d'erreurs ou de points atypiques au sein de cet échantillon pourrait fausser la construction du *cluster* et de sa frontière. Ces erreurs conduiraient à détecter des points comme atypiques à tort (faux positifs) et à omettre des anomalies (faux négatifs). Par rapport aux méthodes supervisées, les approches semi-supervisées peuvent être malgré tout moins coûteuses, surtout si les erreurs sont rares. Dans le cas des méthodes supervisées, il est en effet nécessaire de disposer d'un échantillon suffisamment large pour qu'il soit susceptible de contenir tous les types d'erreurs possibles, alors que l'échantillon labellisé nécessaire aux approches semi-supervisées n'a à décrire que les observations normales, formant le cœur de la distribution des données. Les approches semi-supervisées ont également une capacité plus importante que les méthodes supervisées à détecter des erreurs qui ne se se-

raient jamais produites dans les données, dès lors que celles-ci se traduisent par un écart important aux observations correctes.

— APPROCHE NON SUPERVISÉE : dans cette approche, aucune information sur le statut des données n'est disponible. Dans ce cas, la détection des erreurs ne peut se faire qu'à l'aveugle, de manière dite non supervisée. Il faut alors, pour identifier les déclarations erronées, partir d'une définition *a priori* de ce à quoi peut correspondre une erreur : les algorithmes non supervisés visent alors à identifier des anomalies, *i.e.* des points ne ressemblant pas au reste de la population. Dans ce cadre, la comparaison et l'évaluation des différents algorithmes est plus complexe. En l'absence d'échantillon labellisé, il est impossible de savoir avec certitude quelles méthodes permettent d'identifier des données réellement erronées et si elles identifient bien toutes les erreurs. En revanche, ces approches sont plus aptes que les autres à détecter de nouveaux types d'erreurs dans les données, dès lors que ces erreurs se traduisent par des écarts avec le reste des observations.

La détection des erreurs dans la DSN et les DADS s'effectue dans la troisième approche, non supervisée. Il n'existe en effet pas d'échantillon labellisé de données de la DSN ou des DADS qui puisse être considéré comme représentatif de l'ensemble des données, notamment de l'ensemble des erreurs susceptibles de se produire.

Ainsi, avant d'exposer les résultats des premiers travaux menés, il paraît indispensable de chercher à définir plus précisément la notion d'« anomalie ». Traditionnellement, dans des données, les anomalies, ou points anormaux, ou atypiques, appelés également *outliers*, correspondent à deux situations : à des erreurs ou à des données correctes mais très atypiques voire totalement singulières. Certains de ces points anormaux, appelés unités influentes, sont susceptibles d'avoir une incidence forte sur les résultats des exploitations des données en jouant sur les niveaux des indicateurs diffusés ou en altérant les estimateurs de paramètres de modèles. Le contrôle et le redressement des données consiste à identifier, caractériser et corriger autant que possible les points anormaux, notamment les unités influentes, de façon à produire en sortie un fichier de données individuelles ne contenant que des informations correctes et dans lequel les seuls points anormaux résiduels sont des unités atypiques mais valides.

Selon [Chandola2009], une anomalie correspond à une observation qui n'est pas conforme au schéma standard attendu. Cette conception des points atypiques repose sur la notion de comportement « normal » duquel s'éloignent certains points perçus alors comme des anomalies. Ces dernières sont alors atypiques relativement à une représentation de ce à quoi doivent ressembler des observations normales. Cette représentation peut résulter de la modélisation explicite d'une variable cible par un modèle de régression réalisé avec un ensemble des variables explicatives. Une unité atypique est alors une observation pour laquelle la valeur prédite par le modèle s'écarte sensiblement de la valeur effectivement observée pour cette variable. En l'absence de variable d'intérêt centrale, le « modèle » peut correspondre à une représentation multidimensionnelle des variables sur lesquelles est opérée la détection des erreurs. Une anomalie est alors définie comme une observation isolée des autres en termes de distance ou de densité dans l'espace défini par ces variables.

Selon une définition plus stricte posée par Grubbs en 1969 dans son article [Grubbs1969] et reprise par des articles récents dédiés aux forêts d'isolation (cf. partie 2.2), les anomalies sont des observations rares au sein du jeu de données et qui diffèrent fortement des autres en termes de leurs caractéristiques. Cette définition restrictive écarte toutefois un type d'anomalies, celles qui sont plus nombreuses et qui peuvent être regroupées au sein de *clusters* par des méthodes de partitionnement.

Il existe de fait de multiples manières de définir ce qui constitue une anomalie et donc une erreur potentielle dans un jeu de données, chacune d'entre elles pouvant être détectée avec un algorithme adapté. Il convient donc de se demander quelles sont les méthodes les plus appropriées pour la détection des anomalies au sein des DADS [1]. Pour ce faire, nous comparons sur des données extraites des DADS plusieurs méthodes de détection d'anomalies issues de l'apprentissage statistique avec des méthodes plus classiques, reprises des traitements appliqués dans la chaîne actuelle de production des données. Dans un premier temps, nous présentons plus en détail les données sur lesquelles nous réalisons nos comparaisons et les méthodes de détection d'anomalies actuellement incorporées dans la chaîne de traitement des DADS. Puis nous détaillons les trois méthodes issues du *machine learning* que nous avons testées sur ces données, avant de décrire les résultats de ces comparaisons.

1. En raison de l'indisponibilité des données de la DSN lors de notre expérimentation, notre étude est réalisée sur les DADS. Les méthodes et les spécificités relevées sont toutefois généralisables et applicables au cadre des données de la DSN.

1 Les données sur lesquelles porte la détection d'anomalies et les techniques actuellement utilisées

Les données sur lesquelles nous comparons les méthodes de détection d'anomalies sont issues des déclarations annuelles de donnés sociales (DADS). Dans cette partie, nous présentons cette source, les observations et les variables sur lesquelles nous faisons porter la détection et les méthodes actuellement utilisées dans la chaîne de traitement.

1.1 Les déclarations annuelles de données sociales

La DADS est une formalité administrative que doit accomplir chaque année tout établissement employant des salariés et qui permet d'alimenter les droits sociaux des salariés (droits aux allocations du régime d'assurance chômage, droit à la retraite, à l'assurance maladie...). Les déclarations des établissements employeurs alimentent une base dans laquelle chaque ligne correspond à une période d'emploi. Toutes les modifications importantes du statut ou des conditions d'emploi d'un salarié au cours d'une année font l'objet d'une nouvelle déclaration de période d'emploi (modification d'un contrat, arrêt, etc.). La déclaration d'un établissement désigne alors l'ensemble des périodes d'emploi déclarées par un établissement sur une année pour l'ensemble de ses salariés. Les périodes d'emploi déclarées dans les DADS sont recueillies par l'Agence centrale des organismes de sécurité sociale (Acoss) et transmises annuellement à l'Insee pour la production de statistiques sur les emplois et les salaires.

La DSN vient en remplacement de ce dispositif, et fait l'objet de déclarations mensuelles dont les périodes sont agrégées sur l'année afin de constituer une base semblable à celle des DADS. Le remplacement des DADS par la DSN conduit donc à un changement majeur du système d'information de la statistique publique sur l'emploi et les salaires. La construction de ce nouveau système d'information est l'occasion de revoir les traitements statistiques appliqués aux données collectées, notamment les traitements de contrôle et redressement des données (ou *data editing*). Au moment où les travaux présentés dans cet article ont été réalisés, les données issues de la DSN n'étaient cependant pas encore disponibles, aussi les comparaisons de méthodes de détection d'anomalies ont-elles été conduites sur les DADS les plus récentes disponibles, *i.e.* les DADS 2015.

Plus précisément, nous travaillons sur les périodes d'emploi transmises par les entreprises qui correspondent aux données les plus élémentaires disponibles dans les DADS. Dans le processus actuel de traitement des données, dont nous détaillerons plus loin la partie réalisant la détection d'anomalies, les périodes d'emploi, une fois validées, sont agrégées au niveau des postes de travail pour permettre des analyses sur les emplois et les salaires. Ces données sont très volumineuses. Les DADS 2015 contiennent par exemple 45 101 240 périodes d'emploi [2]. L'objet de nos travaux est de comparer des algorithmes de détection d'anomalies afin de mieux comprendre leurs caractéristiques et leur intérêt potentiel, en préalable à leur éventuelle intégration à la chaîne de production. Aussi, afin de ne pas être contraint dans nos tests par la limitation des ressources informatiques à notre disposition, nous avons choisi de nous limiter à un échantillon de périodes beaucoup plus restreint : nous travaillons sur 50 000 périodes d'emploi réalisées par des salariés d'établissements du secteur du commerce de détail [3] et sélectionnées aléatoirement [4]. Pour limiter nos travaux aux cas de périodes pour lesquelles la détection d'erreurs pose un réel problème, nous nous limitons également à des périodes ne présentant pas d'erreurs manifestes ; les périodes pour lesquelles au moins une des variables d'intérêt, détaillées *infra*, est manquante ou nulle, ne sont pas prises en compte dans notre échantillon.

1.2 Les variables sur lesquelles porte la détection d'anomalies

Les DADS sont principalement utilisées pour étudier l'évolution et la structure des salaires, bruts et nets, ramenés au nombre d'heures travaillées. Ces concepts sont définis dans la note [Cordier-Villoing2018] :

— le salaire brut correspond à l'intégralité des sommes perçues par le salarié au titre de son contrat de travail, avant toute déduction de cotisations obligatoires. Il correspond au concept de rémunération des salariés défini par la comptabilité nationale (opération D1, voir par exemple [Eurostat2014]) ;

— le salaire net (de prélèvements sociaux) est le salaire que perçoit effectivement le salarié. Il est net de toutes cotisations sociales, y compris CSG (contribution sociale généralisée) et CRDS (contribution au remboursement de la dette sociale).

Aussi les variables centrales sur lesquelles porte la détection des anomalies sont les éléments de rémunération et le volume horaire de travail sur les périodes d'emploi.

2. Il s'agit des DADS émises par les employeurs du secteur privé.
3. division 47 de la nomenclature d'activité française
4. L'échantillon de périodes a été sélectionné suivant un plan de sondage systématique à probabilités égales sur fichier trié par les variables suivantes : sexe et tranche d'âge quinquennal du salarié, indicatrice de présence dans les DADS n-1 du même salarié dans le même établissement, catégorie sociale sur deux positions, secteur d'activité de l'unité légale employeuse au niveau A88, statut d'association de l'unité légale employeuse

Les DADS, comme toutes les sources administratives, ne sont pas conçues pour répondre aux besoins de la production statistique, mais aux besoins des administrations qui les collectent. Si le nombre d'heures travaillées sur une période est directement disponible dans les déclarations, il n'en va ainsi pas de même pour les rémunérations. Les variables disponibles dans les DADS ne correspondent de fait pas strictement aux concepts statistiques de salaires bruts et nets, mais aux assiettes utilisées par l'Acoss pour la détermination des montants de cotisation sociale et des droits des salariés. Le salaire brut et le salaire net doivent donc être approchés à partir des concepts renseignés dans les déclarations, selon les formules suivantes, expliquées dans les notes [Chaput2015] et [Cordier-Villoing2018] :

— SALAIRE BRUT : Base CSG $+ \ min\{ \frac{\text{Base CSG} - \text{Sortie d'abattement}}{0,9825}, \text{plafond abattement}\} \times 0,0175$

— SALAIRE NET : Net fiscal $-$ (Base CSG $\times$ (Taux CRDS $+$ Taux CSG non déductible))

Dans ces équations, le salaire net fiscal désigne le salaire imposable au titre de l'impôt sur le revenu des personnes physiques, renseigné sur la fiche de paie du salarié. Pour obtenir un montant correspondant au salaire net que vise l'analyse statistique, il faut retirer du net fiscal les cotisations sur lesquelles le salarié paie un impôt sur le revenu, la CRDS et la part non déductible de la CSG.

Le calcul du salaire brut à partir des informations disponibles dans les DADS est plus complexe. Il est approché par l'assiette de la CSG, *i.e.* la somme des revenus soumis à la contribution sociale généralisée. Cette assiette n'est pas directement disponible dans les DADS, mais peut être estimée.
En effet, les employeurs renseignent dans leurs déclarations la base CSG, qui est la somme à laquelle est appliqué le taux de la contribution sociale généralisée pour déterminer le montant de celle-ci. La base CSG est obtenue à partir de l'assiette de la CSG en appliquant un abattement sur certains revenus. Plus précisément, l'assiette de la CSG est découpée en deux ensembles de revenus :

— des revenus soumis à un abattement, *i.e.* qui ne sont pas pris en compte dans leur totalité pour le calcul du montant de CSG ;

— des revenus non soumis à abattement, appelés sorties d'abattement [5].

Les revenus soumis à abattement sont plafonnés, et le taux d'abattement en 2015 est de 1,75 %. Ainsi, l'assiette de la CSG s'exprime comme la somme de la base CSG et des abattements, calculés en appliquant un taux de 1,75 % à la partie de l'assiette soumise à abattement.
Celle-ci peut de même être évaluée : la base CSG diminuée des sorties d'abattement est en effet égale à 98,25 % du montant de l'assiette soumise aux abattements, dans la limite du plafond des abattements. La partie de l'assiette de la CSG soumise par abattement peut donc être estimée par $min(\frac{\text{Base CSG} - \text{Sorties abattement}}{0,9825}, \text{ plafond abattement})$.

Le graphique 1 représente les distributions, sur l'échantillon de 50 000 périodes d'emploi du commerce de détail en 2015, des trois variables d'intérêt sur lesquelles nous comparons les méthodes de détection des anomalies : le nombre d'heures de travail déclaré, les salaires brut et net horaires, obtenus en divisant les salaires bruts et nets définis plus haut par le nombre d'heures travaillées. Les deux variables de salaire brut ont des distributions unimodales, avec des queues de distribution très longues (le salaire brut horaire médian, par exemple, est de 11,85 € / h, tandis que le salaire brut horaire maximal est de 4 872,73 € / h). Le nombre d'heures par contre a une distribution bimodale, avec une forte concentration sur les nombres d'heures très faibles et un deuxième mode autour de 1 600 heures, soit la durée de travail annuelle légale à temps plein. Le premier mode correspond aux nombres d'heures travaillées sur les périodes d'emploi courtes (en CDD ou en interim principalement), tandis que le second mode correspond au nombre d'heures travaillées sur les périodes d'emploi des salariés à temps complet sur l'année entière.

5. Il s'agit des sommes attribuées au titre de la participation, de l'intéressement, les montants versés au titre de l'abondement de l'entreprise aux versements sur plan d'épargne, les montants de dividende du travail, des indemnités de rupture, des actions gratuites distribuées et des chèques vacances.

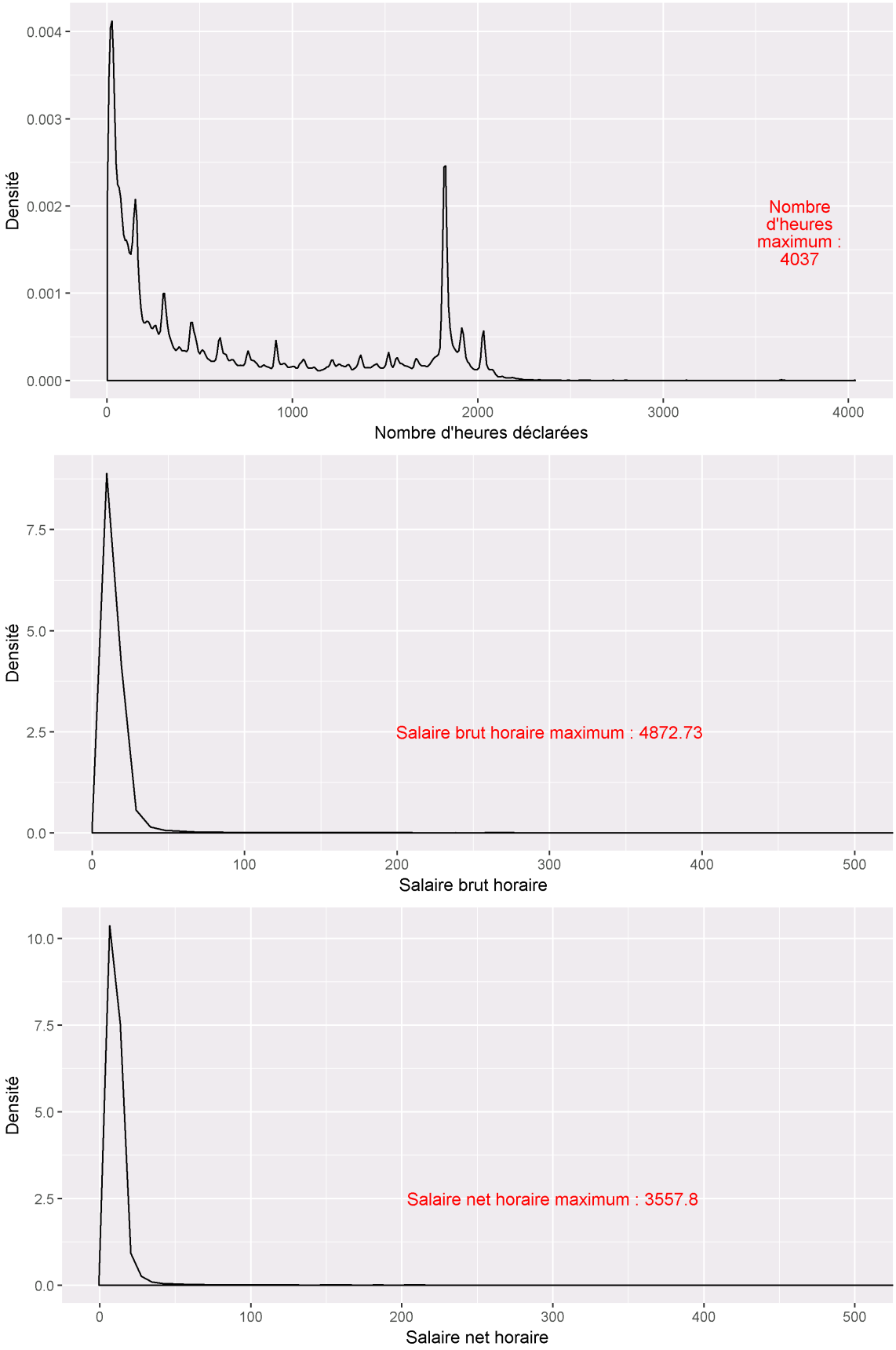


FIGURE 1 – Distributions du nombre d'heures travaillées, des salaires bruts et nets horaires sur l'échantillon de 50 000 périodes d'emploi du commerce de détail

L'échantillon comprend une majorité de périodes d'emploi réalisées par des femmes et des employés, comme le montre le tableau 1. Les périodes d'emploi de l'échantillon sont également majoritairement effectuées sous le régime du CDI et à temps complet.

Variable	Modalité	Statistique / Proportion
Durée de la période d'emploi (en jours)	P5	3.00
Durée de la période d'emploi (en jours)	Q1	26.00
Durée de la période d'emploi (en jours)	Médiane	88.00
Durée de la période d'emploi (en jours)	Moyenne	149.63
Durée de la période d'emploi (en jours)	Q3	356.00
Durée de la période d'emploi (en jours)	P95	360.00
Genre	Femme	65.68
Genre	Homme	34.32
Âge	Moins de 30 ans	51.70
Âge	30-50 ans	35.18
Âge	Plus de 50 ans	13.11
PCS	Cadres	5.11
PCS	Employés	78.13
PCS	Indépendants	0.25
PCS	Ouvriers	7.58
PCS	Prof. Intermédiaires	8.92
Contrat de travail	Autres	2.59
Contrat de travail	CDD	37.90
Contrat de travail	CDI	59.51
Tps travail	Autres	42.33
Tps travail	Temps complet	57.67

TABLE 1 – Caractéristiques de l'échantillon de 50 000 périodes d'emploi du commerce de détail en termes de variables auxiliaires

1.3 Les méthodes actuelles de détection

La chaîne de production actuelle des DADS détecte différents types d'anomalies :

1. **Nombre d'heures nul :**
   Un nombre d'heures nul n'est accepté que pour certaines professions particulières comme les journalistes à la pige ou les travailleurs à domicile par exemple. Dans les cas contraires, un nombre d'heures manquant ou nul signale une anomalie devant être corrigée.
2. **Anomalies financières (incohérences entre salaires brut et salaires nets) :**
   Plusieurs types d'anomalies financières sont distinguées. D'une part, il est vérifié que les variables servant au calcul des salaires bruts et nets comme les assiettes brutes de sécurité sociale, le montant de la base de la contribution sociale généralisée (CSG), *i.e.* l'assiette utilisée pour calculer le montant de CSG applicable à la période, les salaires net et brut fiscaux ne sont pas nuls, et d'autre part qu'il y a cohérence (deux à deux) entre elles (le rapport de l'une sur l'autre ne doit pas dépasser des bornes minimales et maximales définies *a priori* comme paramètres dans l'application gérant la chaîne de traitements). Les populations particulières comme les stagiaires et apprentis sont prises en compte et font l'objet de traitement spécifiques.
3. **Salaires horaires nets trop élevés ou trop faibles ou une incohérence entre les heures :**
   Ce type d'anomalies concerne à la fois les salaires horaires nets (calculés comme le ratio des rémunérations nettes versées sur la période et du nombre d'heures de travail sur la période) anormalement élevés ou faibles par rapport à un seuil fixe (en dessous du SMIC ou au-dessus d'une limite prédéfinie) mais également par rapport à un salaire horaire prédit par un modèle de régression (sur lequel nous reviendrons plus en détail dans la section 1.4). Certaines professions, du fait de revenus horaires particuliers comme les professions de l'information, des arts et des spectacles, ont des règles qui leur sont propres.
4. **Incohérences entre nombre d'heures, durée et condition d'emploi** :
   La chaîne actuelle identifie aussi des incohérences entre le nombre d'heures, la durée et la condition d'emploi déclarée (temps complet/temps partiel). Ces contrôles visent par exemple à détecter des situations dans lesquelles le salarié est déclaré à temps complet mais où le nombre d'heures de travail par jour sur la période d'emploi est inférieure à un certain seuil.

Ces différents types d'anomalies n'ont pas le même niveau de gravité : les anomalies sur le nombre d'heures sont plus graves que les anomalies financières, elles-mêmes plus graves que les anomalies sur les salaires horaires et enfin que les incohérences entre durée

de la période d'emploi, nombre d'heures et condition d'emploi.

Cela se traduit par le calcul d'un score d'atypie résumant la détection des anomalies. Ce score est d'autant plus élevé que les anomalies détectées sont graves. Dans le chaîne de traitement des DADS, ce score est calculé au niveau d'un établissement : elle tient donc compte des anomalies sur l'ensemble des périodes d'emploi reportées par l'établissement. En pratique, pour cette étude, nous avons recalculé un score d'atypie inspiré des traitements de la chaîne DADS, mais spécifique à chaque période. Ce score est défini à partir des anomalies listées précédemment et qui concernent nos données. Par construction de notre échantillon de travail, aucune période d'emploi n'est concernée par les anomalies sur le nombre d'heures nul. De même, nous ne prenons pas en compte les derniers contrôles intégrés à la chaîne, portant sur la cohérence entre durée des périodes d'emploi, nombre d'heures travaillées et conditions d'emploi. Ces contrôles visent en effet à vérifier et éventuellement redresser les informations disponibles sur les conditions d'emploi, alors que nous nous concentrons dans cette étude sur la validation du nombre d'heures et des salaires brut et net horaires. Ainsi, afin de réaliser une comparaison cohérente de cette méthode avec les autres méthodes de détection d'anomalies testées, les anomalies issues de ces contrôles ne sont pas prises en compte dans le calcul du score d'atypie associé à la chaîne de production des DADS. Ce score est égal à :

— 1 pour les périodes d'emploi présentant une anomalie financière ;
— 0,83 si le salaire net horaire est inférieur à 0,5 SMIC ;
— 0,66 si le salaire net horaire est compris entre 0,5 et 0,8 SMIC ;
— 0,5 si le salaire net horaire est compris entre 0,8 et 1 SMIC ;
— 0,33 si le salaire net horaire est faible, mais supérieur au SMIC (ces cas sont identifiés grâce à la régression détaillée dans la section 1.4) ;
— 0,16 si le salaire net horaire est très élevé (ces cas sont identifiés grâce à la régression détaillée dans la section 1.4).

1.4 Zoom sur la régression réalisée pour la détection au sein de la chaîne de production

Comme évoqué dans la partie 1.3, la détection des anomalies sur les salaires nets horaires fait intervenir un modèle de régression, utilisé pour caractériser les salaires nets horaires trop faibles mais supérieurs au SMIC et les salaires nets horaires trop élevés. Ce modèle est donné par :

$$\text{Log(Snh)} = \text{OLS(CS, Âge, Region, Sexe, treffect, CJ)} \tag{1}$$

où

$$\text{Snh} = \frac{NetFiscal}{NbHeur}$$

Dans cette équation, « CS » désigne la catégorie socio-professionnelle du salarié, « Region » la région d'implantation, « Treffect » les tranches d'effectifs et « CJ » la catégorie juridique de l'établissement employeur.

Les coefficients de cette régression sont obtenus à partir des données n-1. Les coefficients permettent de calculer une prédiction pour le logarithme du salaire. Les anomalies détectées dans la chaîne DADS correspondent alors aux observations pour lesquelles l'écart entre le logarithme du salaire net horaire effectivement observé et sa prédiction à l'aide du modèle est supérieur à 4 fois l'écart-type estimé du modèle. Ces anomalies correspondent aux situations où le salaire net horaire est très élevé [6] ou très faible sans être inférieur au SMIC [7].

La comparaison entre les valeurs observées et prédites par un modèle est un outil fréquent pour la détection d'anomalies aussi nous avons souhaité étudier ses propriétés séparément de celles de la chaîne de contrôle des DADS afin de les comparer avec les autres algorithmes. Pour ce faire, nous avons donc réestimé le modèle de régression (1) sur les DADS 2014 et l'avons appliqué aux périodes d'emploi de notre échantillon. Nous avons ainsi pu associer à chaque observation un score d'atypie obtenu par régression et égal à la valeur absolue de l'écart entre les logarithmes du salaire net observé et de sa prédiction, normalisé entre 0 et 1 par la formule suivante :

$$Score_{\text{regression}}(i) = \frac{|log(Snh) - \hat{log}(Snh)| - min_i(|log(Snh) - \hat{log}(Snh)|)}{max_i(|log(Snh) - \hat{log}(Snh)|) - min_i(|log(Snh) - \hat{log}(Snh)|)}$$

où $\hat{log}(Snh)$ désigne la valeur prédite par le modèle de régression du logarithme.

Nous allons décrire dans la partie suivante les trois méthodes de détection d'erreurs issues de l'apprentissage statistique pour lesquelles nous avons comparé leurs performances avec les contrôles de la chaîne DADS et le modèle de régression qu'elle incorpore. Ces trois méthodes, que nous allons à présent détailler, reposent sur des algorithmes et des approches très différentes de ce qu'est une anomalie.

6. Dans ce cas, le logarithme de la valeur observée est supérieur à la valeur prédite plus 4 fois l'écart-type du modèle.
7. Dans ce cas, le logarithme de la valeur observée est inférieur à la valeur prédite moins 4 fois l'écart-type du modèle.

2 Les nouvelles méthodes de détection d'anomalies testées au sein de cet article

Face à la chaîne de production actuelle des DADS et à la régression sur le salaire net horaire explicitées dans la partie 1, trois méthodes non supervisées de détection d'anomalies, présentées successivement dans cette partie, ont été étudiées.

2.1 Le *Local Outlier Factor* (LOF)

En parallèle des méthodes de *clustering*, d'autres méthodes inspirées de celles-ci ont émergé telles que LDBSCAN (*Local Density Based Spatial Clustering Algorithm with Noise*) [8]. Cette dernière repose sur l'évaluation d'un *Local Outlier Factor* (LOF). Ce facteur, qui s'appuie sur des différences de densités entre les *clusters* et évalue la distance des observations par rapport aux *clusters* relativement à leurs densités respectives, mesure ainsi le degré d'atypie d'une observation.

Le facteur LOF repose donc sur le concept d'*outliers* locaux, par opposition aux *outliers* globaux généralement détectés avec une définition courante du terme d'*ouliers*. En effet, dans une approche basée sur la densité globale, une observation est généralement perçue comme un *outlier* dès qu'elle se situe trop loin d'un pourcentage élevé des observations du jeu de données relativement à la densité globale du nuage de points complet. Or, cette définition, adaptée à la localisation d'*outliers* globaux dans le cadre d'un jeu de données de densité uniforme, ne capte pas les points atypiques dans des situations plus complexes. Par exemple, une observation peut se révéler atypique au sein de son voisinage avec des voisins caractérisés par une densité spécifique. Sur la Figure 2, les deux points o_1 et o_2 sont tous les deux atypiques. Si l'approche globale permet d'identifier facilement o_1 comme un *outlier*, aucun paramètre global ne permet d'isoler o_2 sans considérer aussi de multiples observations normales du premier *cluster* C_1 comme des *outliers*. En revanche, le LOF, qui considère les densités locales, détecte dans son cas les deux points atypiques.

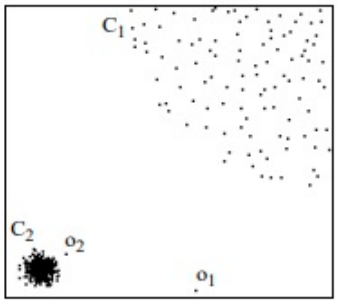


FIGURE 2 – Exemples d'*outliers*

Notons que cette méthode ne pose donc pas d'hypothèse sur la distribution multidimensionnelle des variables. Les différentes distributions sont, en effet, estimées localement.

Afin d'évaluer le caractère atypique d'une observation, un facteur d'atypie locale, reflétant la proximité de celle-ci avec les observations voisines relativement à la densité des voisins, est estimé à partir de l'indicateur LOF défini ci-dessous.

Dans l'optique d'obtenir ce facteur, nous commençons par définir la **k-distance**, notée $Dist_k(x)$ qui correspond à la distance maximale entre l'observation x et ses k plus proches voisins. Pour le voisin y associé à cette distance maximale, elle désigne la distance entre x et y pour qu'au moins k observations y' du jeu de données soient telles que $d(x, y') \leq d(x, y)$ et qu'au plus $k - 1$ observations y' du jeu de données soient telles que $d(x, y') < d(x, y)$. En raison de l'existence potentielle d'observations équidistantes par rapport à x, le nombre de voisins considérés, qui forment l'ensemble noté $N_{k-distance}(x)$, peut être supérieur à k.

8. Les méthodes de *clustering*, très souvent utilisées en analyse de données notamment pour la segmentation des observations, peuvent permettre d'isoler des observations atypiques. La méthode des k-moyennes initialement utilisée dans l'article [MacQueen1967] ou une version plus robuste vis-à-vis des valeurs aberrantes, la méthode des k-médoïdes, constituent des méthodes classiques de *clustering* applicables pour la détection d'anomalies qui s'appuient sur la distance entre les observations relativement aux variables d'intérêt. Toutefois, ces méthodes nécessitent de fixer initialement le nombre de *clusters* et ne tiennent compte intrinsèquement que de l'éloignement entre les observations. Il est alors possible de s'appuyer sur des méthodes basées sur la densité des observations telles que DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) proposée dans l'article [Ester1996] ou OPTICS (*Ordering Points to Identify the Clustering Structure*) développée dans l'article [Ankerst1999]. Contrairement à DBSCAN, la méthode OPTICS permet de détecter des *clusters* de densités différentes. Cependant, face à une construction parfois difficile des *clusters* avec la méthode OPTICS, d'autres méthodes ont émergé telles que LDBSCAN (*Local Density Based Spatial Clustering Algorithm with Noise*) qui s'appuie sur l'évaluation d'un *Local Outlier Factor* (LOF).

De ce premier indicateur se déduit la distance d'atteignabilité, ou **R-distance**, notée $RDist_k(x, y)$. Elle est égale, pour une observation y, au maximum entre la k-distance et la distance entre x et y. L'utilisation du terme *distance* constitue toutefois un abus de langage car la R-distance ne respecte pas la condition de symétrie nécessaire pour caractériser une distance. La distance d'atteignabilité est toutefois privilégiée car elle constitue un indicateur plus lisse de la proximité avec ses voisins. En effet, une variation du nombre k de voisins considérés n'implique ainsi pas de fortes disparités de cet indicateur entre deux observations appartenant au même voisinage.

La **densité locale d'atteignabilité** d'une observation x résulte de ces deux premiers indicateurs et se définit comme :

$$\text{localDensity(x)} = \frac{1}{\frac{\sum_{y \in N_{k-distance}(x)} RDist_k(x,y)}{|N_{k-distance}(x)|}} \tag{2}$$

Enfin, à partir de ces indicateurs intermédiaires, nous obtenons, selon la formule 3, l'indicateur **LOF** d'une observation x qui mesure le niveau relatif de normalité et d'atypie d'une observation.

$$\text{LOF(x)} = \frac{\sum_{y \in N_{k-distance}(x)} \frac{localDensity(y)}{localDensity(x)}}{|N_{k-distance}(x)|} \tag{3}$$

Plus x est éloigné de ses voisins relativement à leur densité, plus la distance d'atteignabilité entre x et ses k voisins est élevée. Dans cette situation, l'éloignement relatif de x par rapport au *cluster* constitué par ses k voisins se traduit par une densité locale d'atteignabilité de x faible notamment en comparaison de la distance d'atteignabilité de ses k voisins. L'indicateur LOF, qui est inversement proportionnel à ce rapport de densités locales d'atteignabilité, augmente. Ainsi, plus x est éloigné de ses voisins relativement à leur densité, plus le facteur LOF est élevé.

2.2 Les forêts d'isolation

À l'image des forêts aléatoires, les forêts d'isolation, *isolation forests (IF)* en anglais, (voir [Liu2008] et [Liu2012]) consistent en l'agrégation de nombreux arbres, dits *isolation trees*. Chaque arbre utilise un échantillon aléatoire d'observations et, pour chaque scission en deux branches, une variable ainsi qu'un seuil de découpage sont aussi tirés aléatoirement. Toutefois, contrairement aux forêts aléatoires, cette méthode est non supervisée.

Les forêts d'isolation reposent toutefois sur une hypothèse forte selon laquelle les anomalies sont peu nombreuses et différentes. Selon cette hypothèse, les anomalies sont donc des observations très vite isolées lors de la construction des arbres alors que les observations normales sont isolées bien plus profondément (scissions de l'arbre très loin de la racine). En effet, une anomalie sera rapidement séparée des autres observations en raison de différences importantes des valeurs associées aux prédicteurs. En revanche, l'arbre commencera par affecter, dans des branches différentes, des observations normales associées à des valeurs similaires qu'une fois que les données seront épurées des valeurs atypiques.

Ainsi, comme les anomalies se caractérisent par une proximité de leur nœud final avec la racine des arbres, la longueur du chemin, noté $h(x)$, correspondant au nombre de nœuds traversés entre x et la racine de l'arbre, constitue un premier indicateur pertinent pour détecter les anomalies à l'échelle de chaque arbre. Au niveau d'une forêt d'isolation, on calcule la moyenne de ces longueurs, notée $\mathbb{E}(h(x))$. En utilisant cette moyenne, nous déduisons un score d'anomalie, noté $s(x, n)$, d'une observation x dans un jeu de données de cardinal n, égal à :

$$s(x, n) = 2^{-\frac{\mathbb{E}(h(x))}{c(n)}} \tag{4}$$

Où $\begin{cases} c(n) = 2H(n-1) - 2\frac{n-1}{n} \\ \text{H(n) nombre harmonique, c'est-à-dire } H(n) = \sum_{k=1}^{n} \frac{1}{k} \\ \text{n est le nombre d'observations utilisées pour la construction de chaque arbre d'isolation} \end{cases}$

Au sein de l'équation 4, le terme $c(n)$, qui correspond à la moyenne de $h(x)$ sachant n, permet de normaliser $\mathbb{E}(h(x))$.

Les arbres d'isolation (*isolation trees*) peuvent souffrir de deux défauts majeurs. D'une part, ils peuvent identifier à tort des observations comme des anomalies. Ce risque, appelé le *swamping*, apparaît dès que les anomalies sont trop proches des observations normales. D'autre part, les anomalies peuvent être difficilement identifiables en présence de *clusters* larges et denses d'anomalies. On parle alors de *masking*. Ces deux effets pervers résultent notamment du nombre excessif d'observations. Toutefois, afin d'atténuer le *swamping* et le *masking*, les *isolation forests* doivent impérativement résulter de la collection d'arbres construits avec des sous-échantillons de données. Par conséquent, à l'inverse de la majorité des méthodes, les arbres d'isolation ont des plus grandes performances s'ils utilisent des échantillons de données de petite taille.

Finalement, ces deux défauts soulignent un enjeu crucial de la détection d'anomalies. En fait, le praticien doit s'interroger sur la nature des anomalies qu'il souhaite détecter et surtout celles qu'il veut absolument mettre en exergue en raison de leur incidence sur les analyses qui découleront de la correction des variables sur lesquelles la détection est réalisée.

Sur le plan pratique, les forêts d'isolation, qui ne reposent pas sur le calcul de distance ou de densité, ont donc de faibles coûts d'implémentation et d'exécution. Alors que les méthodes les plus performantes atteignent un temps d'exécution linéaire seulement si l'utilisation de la mémoire est optimisée, les forêts d'isolation parviennent à cette performance sans exigence requise d'optimisation de la mémoire. Par ailleurs, cet algorithme est facilement parallélisable. Cet dernier atout autorise son application à de jeux de données de grande dimension.

2.3 Les règles d'association

Si les méthodes de *clustering* isolent les observations atypiques lors de la construction des *clusters*, la détection d'anomalies peut aussi résulter de la construction de règles mettant en exergue des relations fréquentes entre les variables. Les observations atypiques sont alors celles qui ne respectent pas ces règles.

L'apprentissage de règles d'association a été développé originellement par [Agrawal1993] pour l'analyse de données de transaction. Une transaction est une liste d'objets ou de produits qui ont été achetés simultanément par un client ; le but de l'apprentissage de règles d'association est d'identifier les objets qui sont fréquemment achetés ensemble dans les mêmes transactions. Pour ce faire, l'apprentissage de règles d'association repose sur deux notions principales : les ensembles fréquents et les règles d'association.

— ENSEMBLES FRÉQUENTS : un ensemble d'objets est fréquent si ces objets sont contenus dans plus de $s\,\%$ des transactions, s étant un paramètre fixé a priori appelé support minimal. Le nombre de transactions auquel appartient un ensemble d'objets A est appelé son support, et noté $s(A)$;

— RÈGLES D'ASSOCIATION : deux ensembles d'objets A et B [9] forment une règle d'association $A \Rightarrow B$ si $A \cup B$ est un ensemble fréquent et si le ratio $\frac{s(A \cup B)}{s(A)}$ du nombre de transactions contenant à la fois A et B et du nombre de transactions contenant A est supérieur à un paramètre fixé a priori c, appelé confiance minimale. Ce ratio, appelé niveau de confiance de la règle, peut s'interpréter comme la probabilité conditionnelle que la transaction contienne l'ensemble d'objets B, sachant qu'elle contient les objets de l'ensemble A.

Les règles d'association sont en général identifiées en deux étapes : d'abord, par une recherche exhaustive des ensembles d'objets fréquents présents dans les données puis par une recherche dans ces ensembles des règles d'association de la forme $X \Rightarrow A - X$, où A est un ensemble fréquent et X un sous-ensemble d'objets de A.

Les nombreux algorithmes d'apprentissage de règles d'association diffèrent selon la manière dont ils mettent en oeuvre la recherche exhaustive des ensembles fréquents dans les données. L'algorithme le plus connu, apriori (voir Agrawal1996, [Tan2006], [Borgelt2002]) et [Borgelt2012]) s'appuie sur la propriété suivante, appelée propriété apriori : si un ensemble d'objets est fréquent, alors tous les sous-ensembles d'objets qu'il contient le sont également [10]. De ce fait, si un ensemble d'objets n'est pas fréquent, aucun des ensembles d'objets dans lesquels il est inclus ne peut être fréquent non plus.
Apriori identifie tous les ensembles fréquents contenus dans des données en itérant sur leur taille. D'abord, il identifie tous les ensembles fréquents ne contenant qu'un objet en comptant le nombre de transactions qui les contiennent. Puis les ensembles fréquents de taille 2 sont identifiés en deux étapes :

1. d'abord, l'algorithme crée une liste d'ensembles candidats en appariant deux à deux les ensembles fréquents de taille 1 ;

2. puis il calcule les supports de ces candidats en comptant le nombre de transactions qui les contiennent.

La première étape permet de réduire le nombre d'ensembles de taille 2 dont le support doit être calculé. Cette opération, qui nécessite de parcourir l'intégralité du fichier des transactions, est coûteuse en temps de calcul. L'algorithme est appliqué en augmentant progressivement la taille des ensembles fréquents considérés, jusqu'à ce qu'aucun ensemble fréquent d'une taille donnée ne puisse être identifié dans les données.
D'autres algorithmes comme eclat ([Zaki1997], [Zaki2003] et [Borgelt2012]) ou FP-growth ([Han2000] et [Han2004]), proposés ultérieurement, sont plus efficaces qu'apriori sur la quasi-totalité des bases de données. Le package arules ([Hahsler2005]), sous R, contient des implémentations efficaces des algorithmes apriori et eclat.

9. Ces ensembles doivent être disjoints
10. En effet, si une transaction contient tous les objets d'un ensemble A, elle contient forcément tous les objets de n'importe lequel de ses sous-ensembles B. Le support de B est donc nécessairement supérieur à celui de A.

Le nombre de règles d'association dans un fichier peut être très élevé selon le nombre d'observations et d'objets qu'il contient. Toutes les règles que peuvent identifier les algorithmes précédemment décrits ne sont cependant pas pertinentes. Par exemple, une règle $A \Rightarrow B$ peut avoir un niveau de confiance supérieur à 90 %. Mais si la part des transactions qui contiennent B est supérieure à ce niveau de confiance, cela signifie que savoir qu'une transaction contient A n'apporte pas réellement d'information sur le fait qu'elle contienne également B. De nombreux auteurs ont proposé des mesures de pertinence des règles d'association (voir [Tan2006] pour une présentation). Une de ces mesures, appelée *lift*, est définie, pour une règle d'association $A \Rightarrow B$ par :

$$L(A \Rightarrow B) = \frac{s(A \cup B)}{s(A)\, s(B)}$$

Les règles d'association pertinentes sont caractérisées par des valeurs du *lift* supérieures à 1. Dans ce cas, en effet, le niveau de confiance de $A \Rightarrow B$ est supérieur au support de B : les transactions qui contiennent A ont une probabilité supérieure à la moyenne de contenir également les objets de B.

a) Utilisation des règles d'association pour le data editing

Les données pour lesquelles l'apprentissage de règles d'association a été développé sont des données de transaction. Ce fichier contient autant de lignes que de transactions, et de colonnes que d'objets pouvant figurer dans ces transactions. Pour une transaction, la colonne correspondant à un objet est égale à 1 si la transaction contient cet objet, 0 sinon.

L'apprentissage de règles d'association, même s'il a été développé dans ce cadre, n'est cependant pas limité aux données de transaction. Les transactions peuvent ainsi être remplacées par des observations, et les objets composant les transactions par les valeurs des variables décrivant ces observations.
Par exemple, si le fichier sur lequel nous travaillons est constitué d'établissements décrits par leur localisation géographique et leur secteur d'activité, les objets associés à ces observations sont les différentes valeurs possibles de la localisation et du secteur d'activité dans le fichier. Les données peuvent alors être représentées par un fichier contenant autant de lignes que d'établissements, et autant de colonnes que de valeurs possibles pour la localisation et le secteur.

Dans ce cas, les ensembles fréquents représentent le fait que les valeurs de certaines variables sont fréquemment observées simultanément ; les règles d'association, des relations entre variables, i.e. le fait que si certaines valeurs sont observées pour des variables, alors il est plus fréquent d'observer certaines autres valeurs pour d'autres variables. Les unités atypiques sont alors les observations ne respectant pas ces relations.

Cependant, seules les variables prenant un nombre limité de valeurs peuvent être prises en compte aisément dans ce cadre. Pour prendre en compte les variables quantitatives dans l'apprentissage de règles d'association, il est nécessaire d'apporter des modifications aux algorithmes de bases évoqués plus haut.

b) Le traitement des variables quantitatives dans l'apprentissage de règles d'association

La première solution pour intégrer les variables quantitatives est de transformer celles-ci en variables qualitatives en partitionnant la population suivant leurs modalités (voir [Srikant1996]). Les règles d'association obtenues sur ces variables discrétisées dépendent cependant fortement de la partition, notamment de la manière dont les unités proches des bornes des classes (ou *clusters*) de la partition sont traitées.

Considérons en effet que nous souhaitions discrétiser une variable de salaire mensuel, et que 2 000 € / mois constitue le frontière entre deux classes de la partition que nous avons constituée. Dans ce cas, des individus dont les salaires mensuels sont de 1 999 € et de 2 001 € ne seront pas placés dans la même classe de la partition. Il est cependant très probable que leurs situations soient proches et que, pour l'identification de relations fortes et fréquentes entre caractéristiques des individus, il soit plus pertinent de considérer que ces individus appartiennent à la même classe.
Ce problème se pose quelle que soit la frontière considérée et la manière de la définir, dès lors que le population n'est pas aisément divisible en classes naturelles, i.e. en groupes de population dont les frontières sont clairement délimitées et séparées les unes des autres.

Pour contourner cette difficulté, une possibilité serait d'autoriser les classes entre lesquelles sont divisées les variables quantitatives à se recouper. Pour revenir à notre exemple, les individus dont les salaires sont égaux à 1 999 € et 2 001 € pourraient ainsi participer à deux classes, autour de l'ancienne frontière de 2 000 €. Le problème tient alors au fait que les observations des zones où les *clusters* se recoupent pèsent plus que les autres observations, puisqu'ils participent au calcul du support de chacune des deux classes. Elles contribuent alors plus fortement aux résultats et à l'identification des règles d'association que les autres observations.

La solution que nous avons retenue repose sur l'utilisation des ensembles flous (voir [Zimmermann1991]), qui permettent de découper des variables quantitatives en *clusters* non disjoints, sans surpondérer les unités situées au niveau des frontières entre classes.
Un ensemble usuel peut être défini par la donnée d'une fonction d'appartenance. Cette fonction est une variable indicatrice, égale à 1 pour les unités qui appartiennent à l'ensemble, et 0 pour les autres. Un ensemble flou est une extension de ce concept d'ensemble : un ensemble flou est défini par la donnée d'une fonction, appelée degré d'appartenance, à valeurs dans l'intervalle $[0; 1]$. Les observations dont le degré d'appartenance à l'ensemble est proche de 1 appartiennent fortement à l'ensemble, tandis que les observations dont le degré d'appartenance est proche de 0 n'ont qu'un lien faible avec celui-ci.

Les ensembles flous permettent de découper une population en classes non disjointes sur la base des valeurs d'une variable quantitative, sans déséquilibrer les poids des différentes observations : il suffit pour cela de garantir que la somme des degrés d'appartenance de chaque observation aux différents ensembles flous entre lesquels est découpée une variable est toujours égale à 1. La figure 3 montre le résultat obtenu en découpant les salaires nets horaires observés dans 50 000 périodes d'emploi effectuées auprès d'employeurs du commerce de détail en 2015 en huit sous-ensembles flous. Le graphique représente les degrés d'appartenance aux huit *clusters* flous, ainsi que le dernier graphique, la densité de la variable. Pour chaque observation, la somme des huit degrés d'appartenance est égale à 1 : chaque observation a donc le même poids et contribue de la même manière à l'identification des associations dans la population.

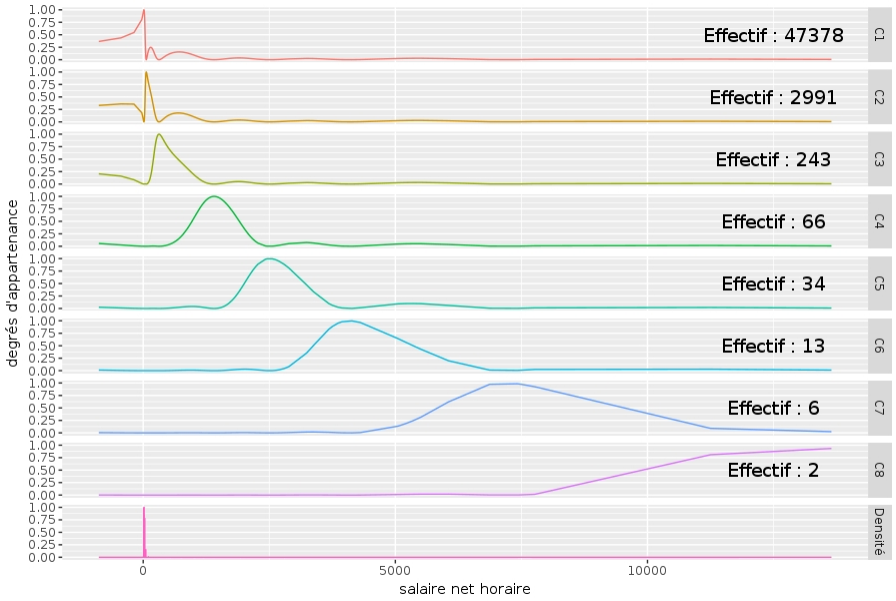


FIGURE 3 – Densité et degrés d'appartenance aux clusters de la partition du salaire net horaire en huit groupes

Les *clusters* flous de la figure 3 ont été obtenus avec l'algorithme des fuzzy c-means, proposé par [Dunn1973] et amélioré par [Bezdek1981], qui généralise l'algorithme classique des centres mobiles (ou k-means) à la théorie des ensembles flous.
Supposons que nous disposions d'une population de N observations $X_i, i = 1..N$ que nous cherchons à découper en K *clusters* flous. Nous voulons donc générer K points, $C_j, j = 1..K$ correspondant aux centres des *clusters* et K variables $u_j, j = 1..K$ représentant les degrés d'appartenance à ces classes : $u_{ij}, i = 1..N, j = 1..K$ représente ainsi le degré d'appartenance de l'individu i au *cluster* j. Ces différents éléments (centres des clusters et degrés d'appartenance) sont alors déterminés, dans l'algorithme des fuzzy c-means, de façon à minimiser la fonction suivante :

$$O(m) = \sum_{i=1}^{N} \sum_{j=1}^{K} u_{ij}^{m} \, ||X_i - C_j||^2$$

avec m un paramètre compris entre 1 et N et permettant de contrôler à quel point les classes constituées seront floues.

Pour trouver la solution de ce programme, l'algorithme, comme celui des k-means, est itératif. Les itérations sont initialisées avec des valeurs aléatoires des degrés d'appartenance u_{ij}. A chaque itération :

— Les points C_j minimisant $O(m)$ conditionnellement aux valeurs disponibles des degrés d'appartenance u_{ij} sont déterminés,

par application de la formule :

$$C_j = \frac{\sum_{i=1}^{N} u_{ij}^m X_i}{\sum_{i=1}^{N} u_{ij}^m}$$

— Puis, les degrés d'appartenance u_{ij} minimisant $O(m)$ conditionnellement aux centres des classes C_j déterminés à l'étape précédente sont calculés à l'aide de la formule suivante :

$$u_{ij} = \frac{1}{\sum_{k=1}^{K} \left(\frac{||X_i - C_j||}{||X_i - C_k||} \right)^{\frac{2}{m-1}}}$$

Les centres des classes et les degrés d'appartenance convergent au cours des itérations. L'algorithme s'arrête quand les modifications des centres et des degrés d'appartenance d'une itération à l'autre sont inférieures à des seuils fixés a priori.
A chaque itération, les degrés d'appartenance sont des fonctions croissantes du paramètre m, ce qui explique pourquoi celui-ci permet de contrôler de degré de flou de la partition : plus m est faible, plus les degrés d'appartenance de chaque individu aux différents *clusters* sont faibles.

D'autres algorithmes de construction de *clusters* flous ont été développés, notamment l'algorithme F-CLARANS (*Fuzzy Clustering Large Applications based on RANdomized Search*, voir [Mahfouz2009]). L'algorithme CLARANS (voir [Ng2002]) s'inspire de l'algorithme CLARA (*Clustering LARge Applications, voir [Kaufman1990]*) qui correspond à un clustering par k-medoids sur des sous-échantillons des données.
Les résultats sur lesquels nous nous appuyons ont tous été obtenus en utilisant l'algorithme des fuzzy c-means.

c) Apprentissage de règles d'association floues

Nous discutons à présent la manière dont les algorithmes d'apprentissage de règles d'association présentés plus haut peuvent être adaptés à la prise en compte des variables floues. L'objectif est d'aboutir à des algorithmes pouvant s'appliquer à des données mixtes, composées de variables qualitatives classiques et de variables quantitatives représentées par les degrés d'appartenance aux *clusters* flous d'une partition.

Pour ce faire, nous nous appuyons sur les modifications de l'algorithme apriori proposées par [Kuok1998] et [Hong2001].
D'abord, toutes les variables quantitatives sont découpées en *clusters* flous à l'aide d'un algorithme adapté (les fuzzy c-means en ce qui nous concerne). Les données en entrée de la procédure contiennent alors :

— un ensemble de variables qualitatives $X_j, j \in J$, décrites dans le fichier par les m_j variables indicatrices $U_{jk}, k = 1..m_j$, avec m_j le nombre de modalités de la variable X_j ;

— un ensemble de variables quantitatives $Y_i, i \in I$, représentées dans le fichier par les n_i degrés d'appartenance $Y_{it}, t = 1..n_i$, avec n_i le nombre de *clusters* entre lesquels est découpée la variable Y_i.

Dans le cas où toutes les variables du fichier sont qualitatives, les algorithmes usuels d'apprentissage de règles d'association peuvent être appliqués aux données, en considérant que les transactions sont les différentes observations du fichier et les objets qui les composent les différentes modalités des variables du fichier. Les ensembles d'objets sont alors des ensembles de modalités des variables d'intérêt, et les règles d'association décrivent des relations entre ces modalités observées dans la population.

Dans le cas de l'apprentissage de règles d'association floues sur données mixtes, les objets sont les différentes modalités des variables qualitatives et les *clusters* entre lesquels sont découpées les variables quantitatives. Les ensembles d'objets (colonnes de la base de données) sont ainsi constitués de modalités de variables qualitatives, représentées dans les données par des variables indicatrices, et des classes floues des variables quantitatives, représentées dans les données par leur degré d'appartenance. Les règles d'association décrivent des relations entre ces modalités et ces classes.

Les différentes étapes de l'algorithme apriori flou (fuzzy apriori) sont les mêmes que celles de l'algorithme originel de [Agrawal1996].
D'abord, les ensembles d'objets fréquents de taille 1 sont identifiés. Puis les ensembles d'objets de taille 2 pouvant être fréquents sont constitués par appariement des ensembles fréquents de taille 1. Leur support est ensuite calculé en passant en revue l'ensemble des transactions, ce qui permet l'identification des ensembles fréquents de taille 2. Le processus est itéré jusqu'à ce qu'il ne soit pas possible d'identifier des ensembles fréquents d'une taille donnée dans le fichier.

La définition et le calcul du support d'un ensemble d'objet doivent cependant être adaptés pour tenir compte des classes floues et de leurs degrés d'appartenance. Dans le cas classique où toutes les variables sont qualitatives et sont donc représentées par des indicatrices d'appartenance à leurs modalités, le support d'un ensemble d'objets A est égal au nombre d'observations contenant tous les objets de l'ensemble A : il est en pratique calculé comme la somme sur l'ensemble du fichier de la variable $A(i)$ égale à 1 si l'observation i contient tous les objets de l'ensemble A et 0 sinon. Si les variables $U_k, k = 1..K$ désignent les différentes indicatrices des modalités correspondant aux objets de l'ensemble A, $A(i)$ est une variable égale à 1 si toutes les indicatrices U_k sont égale à 1, et 0 si au moins une des variables U_k est nulle. De ce fait, $A(i)$ peut être définie comme : $A_i = min_{k=1..K} U_k$, ou comme $A_i = \prod_{k=1}^{K} U_k$. Ces deux formules vérifient la propriété essentielle de $A(i)$, qui doit être nulle dès lors qu'une des variables U_k est nulle.

Ces formules sont simples à étendre au cas où les variables U_k composant l'ensemble d'objets A ne sont pas uniquement des indicatrices, mais à la fois des indicatrices et des degrés d'appartenance. Pour chaque observation, nous pouvons définir le degré auquel l'ensemble d'objets A appartient à l'observation i (qui peut être appelé degré d'appartenance de l'ensemble A à l'observation i) par :

$$A(i) = \begin{cases} min_{k=1..K}(U_k) \\ \text{ou} \\ \prod_{k=1}^{K} U_k \end{cases}$$

La première définition de $A(i)$ est celle proposée par [Kuok1998], tandis que [Hong2001] utilise la seconde.
Le support de l'ensemble d'objets A dans un ensemble de données est alors défini comme :

$$s(A) = \frac{\sum_{i=1}^{N} A(i)}{N}$$

La définition du niveau de confiance d'une règle d'association floue est également simple à adapter :

$$c(A \Rightarrow B) = \frac{s(A \cup B)}{s(B)} = \frac{\sum_{i=1}^{N} A \cup B(i)}{\sum_{i=1}^{N} A(i)}$$

Avec ces définitions, l'algorithme apriori peut être totalement adapté au traitement des variables quantitatives représentées par des *clusters* flous. [Hong2001] suggère cependant une autre modification : les ensembles d'objets contenant plus d'un degré d'appartenance relatif à la même variable ne sont pas pris en compte dans l'identification des ensembles fréquents.
Ce problème ne se pose pas dans le cas non flou, dans lequel le fichier ne contient que des indicatrices d'appartenance aux modalités de variables qualitatives. Il est en effet impossible qu'une observation contienne simultanément deux modalités d'une même variable qualitative, le support des ensembles d'objets contenant au moins deux modalités d'une même variable est donc forcément nul.
Avec les ensembles flous, la situation est plus complexe : du fait des recouvrements entre *clusters* flous, il existe des observations appartenant en même temps à plusieurs *clusters*, au sens où les degrés d'appartenance de cette observation à chacune des deux classes sont non nuls. De ce fait, des ensembles d'objets contenant plusieurs classes floues d'une variable quantitative peuvent être fréquents. Ils ne représentent cependant aucun intérêt dans la recherche d'associations entre variables, et sont de ce fait éliminés de la recherche d'ensembles fréquents et de règles d'association.

Une limite importante des travaux de [Kuok1998] et [Hong2001] est que la « fuzzyfication » des variables quantitatives est effectuée en amont de la recherche d'ensembles fréquents et de règles d'association, sans prise en considération de cet objectif. Il n'y a de ce fait aucune garantie que les classes floues obtenues au terme de cette étape soient les plus adaptées pour l'identification de règles d'association. [Hong2006] et [Hong2008] ont proposé d'autre approches où la « fuzzyfication » est intégrée plus complètement à la recherche des règles, de façon à identifier les liens les plus pertinents entre variables. Nous n'avons cependant pas eu le temps de les tester sur nos données.

d) Identification des unités atypiques avec les règles d'association floues

Les traitements présentés dans les sections précédentes détaillent comment identifier les ensembles fréquents et les règles d'association dans les fichiers de données mixtes, regroupant variables qualitatives et quantitatives résumées via les degrés d'appartenance à des classes floues. Il reste à préciser comment utiliser ces règles pour attribuer un score d'atypie aux observations de l'échantillon.

Dans le cas de règles d'association classiques, l'identification des unités atypiques est simple : pour une règle $A \Rightarrow B$, les unités atypiques sont celles qui ne respectent pas la règle, i.e. les observations contenant les modalités de l'ensemble A mais pas celles de l'ensemble B. Il est ainsi possible d'associer à chaque observation et pour chaque règle une variable égale à 0 si l'observation respecte la règle et 1 sinon. Ces indicatrices peuvent être ensuite résumées par un score unique, égal au nombre ou à la part de règles non respectées.

Cette procédure peut être adaptée aux règles d'association floues. Il n'est plus possible de déterminer de manière certaine si une observation contient un ensemble A ou un ensemble B, et donc si elle respecte ou pas une règle floue. En effet, un ensemble d'objet A n'est présent dans une observation i qu'à un certain degré mesuré par la quantité $A(i)$ définie plus haut. La solution va consister à s'appuyer sur les degrés d'appartenance des ensembles A et B à l'observation i pour définir un degré d'observance par l'observation de la règle $A \Rightarrow B$.

Par construction, le ratio $\frac{A \cup B(i)}{A(i)}$ est compris entre 0 et 1. S'il est très proche de 1, cela signifie que l'ensemble $A \cup B$ appartient autant à l'observation i que le seul ensemble A : dès que A est présent dans l'observation, alors celle-ci contient également B. À l'inverse, le ratio est proche de 0 si le degré d'appartenance de $A \cup B(i)$ est beaucoup plus faible que celui de $A(i)$, donc si l'observation contient beaucoup A mais peu A et B. Le ratio $\frac{A \cup B(i)}{A(i)}$ est donc une mesure du degré d'observance par l'observation i de la règle $A \Rightarrow B$ [11].

Pour obtenir un score mesurant le non-respect par une observation d'une règle, il donc est possible de considérer le complémentaire à 1 du ratio précédent : $S_A(i, A \Rightarrow B) = \begin{cases} 1 - \frac{A \cup B(i)}{A(i)} & \text{si } A(i) > 0 \\ 0 & \text{sinon} \end{cases}$

Ce score mesure de manière absolue le respect par l'observation de la règle. Il n'est en effet égal à 1 que si $A \cup B(i) = A(i)$, donc que si les indicatrices ou degrés d'appartenance aux variables composant B sont égales à 1 dans l'observation. Nous proposons un autre score, qui tienne également compte du niveau global de confiance de la règle dans la population. L'idée est ici qu'un ratio $\frac{A \cup B(i)}{A(i)}$ de 0,9 n'a pas la même signification si le niveau de confiance de la règle $A \Rightarrow B$ dans la population est de 0,6 ou de 0,99. Dans le premier cas, l'observation respecte beaucoup plus la règle que la moyenne de la population tandis qu'elle la respecte sensiblement moins dans le second. Aussi, nous proposons de considérer plutôt une mesure relative de l'atypie d'une observation pour une règle, définie par :

$$s_R(i, A \Rightarrow B) = \begin{cases} c(A \Rightarrow B) \,/\, \frac{A \cup B(i)}{A(i)} & \text{si } \frac{A \cup B(i)}{A(i)} > 0 \\ +\infty & \text{sinon} \end{cases}$$

normalisé en :

$$S_R(i, A \Rightarrow B) = \begin{cases} \frac{s_R(i, A \Rightarrow B) - min_{j / s_R(j, A \Rightarrow B) < +\infty}[s_R(j, A \Rightarrow B)]}{max_{j / s_R(j, A \Rightarrow B) < +\infty}[s_R(j, A \Rightarrow B)] - min_{j / s_R(j, A \Rightarrow B) < +\infty}[s_R(j, A \Rightarrow B)]} & \text{si } \frac{A \cup B(i)}{A(i)} > 0 \\ 1 & \text{sinon} \end{cases}$$

Le score $s_R(i, A \Rightarrow B)$ est supérieur à 1 si le degré d'observance par l'observation i de la règle $A \Rightarrow B$, mesuré par le ratio $\frac{A \cup B(i)}{A(i)}$, est inférieur au niveau de confiance de la règle dans la population, i.e. si l'observation respecte moins la règle que la population. Plus il dépasse 1, moins l'observation respecte la règle au regard de la population. Le score S_R est obtenu en normalisant s_R de manière à garantir que ses valeurs soient comprises entre 0 et 1, les valeurs proches de 1 identifiant les observations ne respectant pas les règles. Au cas où l'observation ne respecte pas du tout la règle (i.e. son degré d'observance $\frac{A \cup B(i)}{A(i)}$ est nul), le score est directement forcé à sa valeur maximale de 1.

Au terme des étapes précédentes, nous disposons pour chaque observation et chaque règle de scores absolus $S_A(i, A \Rightarrow B)$ et relatifs $S_R(i, A \Rightarrow B)$ compris entre 0 et 1 et mesurant à quel point l'observation respecte la règle. Ces scores sont résumés en une valeur unique associée à chaque observation en calculant leur moyenne sur l'ensemble des règles d'association identifiées dans les données :

$$S_A(i) = \frac{\sum_{\mathcal{R} \in \mathcal{A}} S_A(i, R)}{|\mathcal{A}|}$$

$$S_R(i) = \frac{\sum_{\mathcal{R} \in \mathcal{A}} S_R(i, R)}{|\mathcal{A}|}$$

avec $\mathcal{A}$ l'ensemble des règles d'association floues identifiées dans le fichier pour un choix de support et de niveau de confiance minimaux.

11. Ce ratio est par convention considéré comme nul si $A(i)$ est nul : dans ce cas, l'observation ne contient pas du tout l'ensemble A et n'est de ce fait pas concernée par la règle

D'autres stratégies d'agrégation des scores relatifs à chaque règle en un score unique d'atypie pour chaque observation sont possibles. En effet, le score d'atypie ainsi construit est élevé si une observation respecte peu beaucoup de règles. Il risque cependant d'avoir plus de difficultés à identifier une observation qui respecte toutes les règles, sauf une pour laquelle son score est proche de 1. Dans ce cas, agréger les scores associés à chaque règle en prenant leur maximum aurait été plus efficace. A l'inverse, le score maximal aurait plus de mal à identifier comme atypique une observation respectant mal toutes les règles tout en les respectant toutes un peu. Nous nous en sommes tenus, dans l'analyse des résultats, au calcul du score moyen.

3 Comparaison des méthodes sur les données de la DADS

Si l'analyse méthodologique de ces trois méthodes semble prometteuse pour effectuer la détection des anomalies, il convient maintenant de mettre en œuvre ces algorithmes et d'étudier leurs résultats. Pour chaque algorithme, nous avons défini comme anomalies les n observations ayant les valeurs les plus élevées du score produit par la méthode, n étant le nombre d'anomalies détectées dans notre échantillon par la chaîne de production des DADS [12]. Nous avons utilisé, pour mettre en œuvre les différents algorithmes, les outils suivants :

— le package R *isofor* pour la forêt d'isolation ;

— le package R *rlof* pour le *Local Outlier Factor* ;

— un package R interne développé spécifiquement pour la mise en œuvre des règles d'association floues, les classes floues étant quant à elles constituées par application de l'algorithme des centres mobiles flous avec le package R *ppclust*.

Nous commencerons par nous interroger sur la nature des anomalies détectées. Pour cela, nous analyserons d'une part les anomalies au regard du triplet de variables d'intérêt. Nous décrirons également les anomalies à partir des principales variables disponibles dans les DADS. Puis, nous comparerons les résultats des différentes méthodes, en étudiant les anomalies qu'elles détectent en commun ou séparément.

Nous mettrons notamment en regard les résultats des trois algorithmes avec la méthode de détection implémentée dans la chaîne de production actuelle. Celle-ci identifie les anomalies sur la base d'un ensemble de règles définies a priori mais aussi en s'appuyant sur l'écart entre le salaire net horaire [13] observé et la prédiction d'un modèle de régression log-linéaire défini dans la partie 1.4. Nous comparons également les résultats des trois algorithmes avec ceux de cette régression.

3.1 Que détectent nos algorithmes ?

Pour prolonger l'étude théorique de la partie 2, nous allons observer comment les différences constatées entre les algorithmes se traduisent concrètement au regard des variables sur lesquelles les anomalies sont détectées, ainsi que sur les principales variables auxiliaires.

Comme l'illustre la figure 4, le facteur LOF détecte comme anomalies potentielles deux types d'observations :

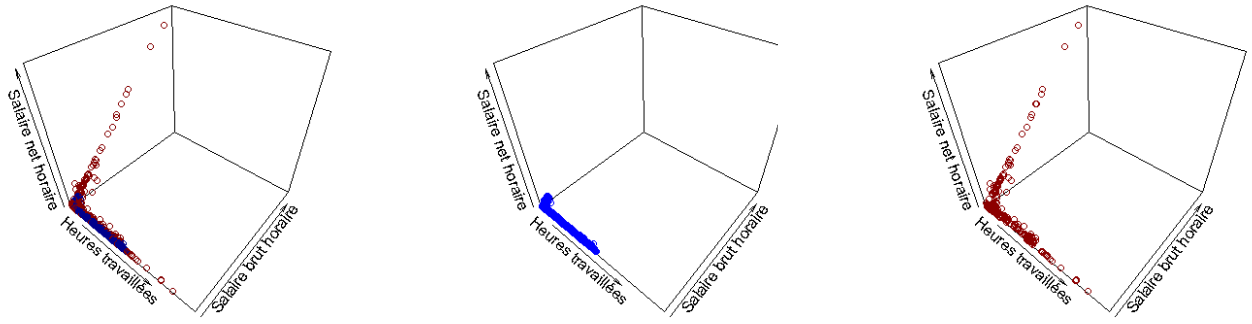
— des observations extrêmes, correspondant à des valeurs très élevées des salaires brut et net associées à un nombre d'heures faible ; mais aussi à des valeurs très faibles des salaires associées à un nombre d'heures élevé ;

— des observations formant une enveloppe autour des observations « normales » au regard du facteur LOF [14].

12. Les observations détectées par la chaîne DADS sont toutes celles dont le score, présenté dans la partie 1.3 est non nul, soit 645 observations sur notre échantillon.

13. Plus précisément, le modèle de régression utilisé dans la chaîne des DADS et que nous avons reproduit tente de prédire le salaire net fiscal horaire, tandis que nous travaillons sur des variables de salaire brut et net constituées à partir d'un ensemble plus large d'informations. Pour ne pas alourdir la rédaction, nous ne mentionnerons plus cette spécificité de la régression.

14. Les observations « normales » sont représentées en bleu sur la figure 4, les anomalies potentielles en rouge.



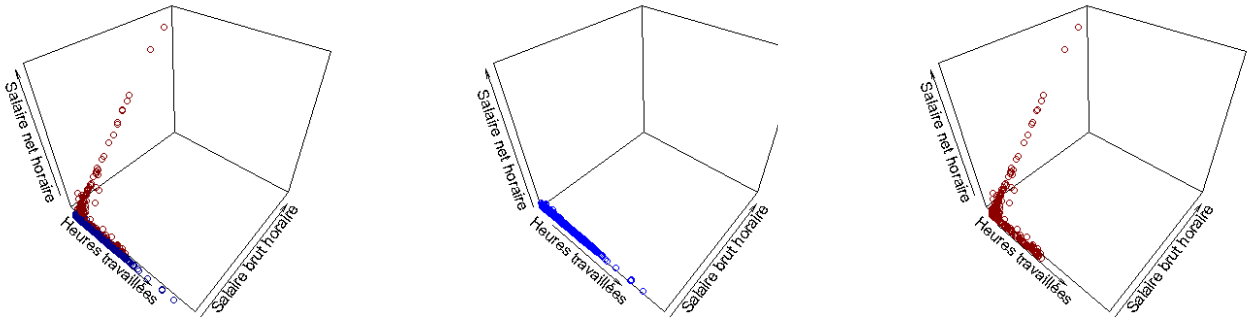
(a) Représentation des anomalies (en rouge) et des observations normales

(b) Représentation des observations normales (en bleu)

(c) Représentation des anomalies (en rouge)

FIGURE 4 – Détection des anomalies à partir du facteur LOF

Pour l'algorithme des forêts d'isolation, les anomalies détectées, signalées en rouge sur la Figure 5, correspondent à des observations pour lesquelles les salaires bruts et nets horaires sont élevés pour un nombre d'heures moyen ou faible. Contrairement aux résultats de la méthode LOF, la séparation graphique entre les anomalies et les observations normales est nette. En effet, la Figure 5a montre que les anomalies semblent séparées des observations normales avec des valeurs de salaires horaires supérieurs aux valeurs moyennes associées aux observations normales. Ce constat semble cohérent avec le principe des forêts d'isolation qui tendent à détecter des anomalies parmi des observations éloignées des autres pour les variables sur lesquelles est opérée la détection : les observations présentant des valeurs extrêmes pour au moins une variable d'intérêt sont par construction de l'algorithme aisément séparables des autres et donc identifiées prioritairement comme anomalies par la forêt d'isolation.



(a) Représentation des anomalies (en rouge) et des observations normales

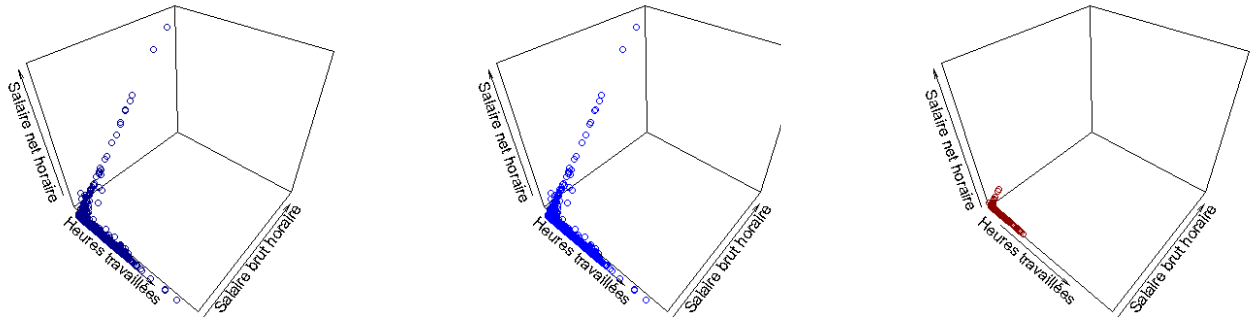
(b) Représentation des observations normales (en bleu)

(c) Représentation des anomalies (en rouge)

FIGURE 5 – Détection des anomalies à partir de l'indicateur d'anomalies issu d'une forêt d'isolation

Contrairement aux anomalies mises en évidence par le facteur LOF et la forêt d'isolation, celles issues des règles d'association ne se situent pas au niveau des valeurs extrêmes. Par exemple, la Figure 6c ne signale pas comme anomalies les observations associées à des salaires élevés et un nombre d'heures travaillées faible. Les anomalies détectées par les règles d'association floues concernent des observations avec des salaires horaires brut et net moyens ou faibles. Cela tient notamment au fait que les anomalies détectées avec les règles d'association correspondent à des observations satisfaisant l'antécédent d'une ou plusieurs règles mais pas le conséquent. Or, comme l'antécédent est nécessairement un ensemble fréquent, les observations associées à des valeurs extrêmes ne sont pas

concernées par celui-ci et ne peuvent donc être détectées comme des anomalies.



(a) Représentation des anomalies (en rouge) et des observations normales

(b) Représentation des observations normales (en bleu)

(c) Représentation des anomalies (en rouge)

FIGURE 6 – Détection des anomalies à partir de l'indicateur d'anomalies issu des règles d'association floues

Ces constats se retrouvent dans le tableau 2, qui met également en évidence les spécificités des anomalies détectées par la chaîne des DADS concernant les variables d'intérêt. Les anomalies identifiées par la chaîne actuelle se caractérisent par des valeurs très faibles des salaires, au contraire des autres méthodes, notamment du modèle de régression qui participe pourtant à la construction du score d'atypie dans la chaîne.

La chaîne de production des DADS identifie essentiellement comme anomalies des employés, conformément à la structure de la population étudiée, alors que les autres algorithmes détectent comme anomalies plutôt d'autres catégories socio-professionnelles (PCS) : cadres pour la forêt d'isolation et le LOF, ouvriers et professions intermédiaires pour les règles d'association floues. Les différences sur les autres variables auxiliaires sont moins marquées que sur la PCS ; forêt d'isolation et LOF détectent comme anomalies moins souvent des femmes (qui forment la majorité de la population de notre échantillon) que les autres algorithmes.

Variable	modalite	Dads	Régression	IF	LOF	FARM	Ensemble
Durée	Mediane	20.00	25.00	70.00	66.00	55.00	88.00
	Moyenne	90.53	64.72	145.09	152.72	92.77	149.63
Nb d'heures	Mediane	239.00	48.00	42.00	175.00	169.00	312.00
	Moyenne	577.45	231.75	452.80	697.96	349.98	665.98
Brut horaire	Mediane	2.87	62.43	75.16	9.79	12.62	11.85
	Moyenne	8.46	154.53	179.65	124.36	17.84	15.49
Net horaire	Mediane	2.77	50.78	56.64	9.83	9.70	9.16
	Moyenne	9.15	121.83	138.14	96.02	14.31	11.87
Genre	Femme	66.00	71.00	57.00	60.00	72.00	66.00
	Homme	34.00	29.00	43.00	40.00	28.00	34.00
PCS	Cadres	10.00	7.00	31.00	23.00	21.00	5.00
	Employes	69.00	77.00	52.00	56.00	0.00	78.00
	Independants	3.00	1.00	3.00	4.00	2.00	0.00
	Ouvriers	7.00	5.00	4.00	5.00	40.00	8.00
	Prof. inter.	12.00	11.00	10.00	12.00	37.00	9.00
secteur_ul	Comd	95.00	94.00	92.00	93.00	92.00	96.00
	Comg	1.00	2.00	3.00	2.00	2.00	1.00
	Industrie	2.00	2.00	3.00	2.00	1.00	1.00
	Services	3.00	2.00	2.00	3.00	4.00	2.00
Contrat de travail	Autres	7.00	5.00	4.00	4.00	7.00	3.00
	Cdd	29.00	36.00	29.00	25.00	57.00	38.00
	Cdi	64.00	59.00	68.00	70.00	36.00	60.00
Tps travail	Autres	58.00	47.00	34.00	40.00	53.00	42.00
	Tps complet	42.00	53.00	66.00	60.00	47.00	58.00

TABLE 2 – Caractéristiques des anomalies pour les trois méthodes testées et pour la régression actuellement utilisée dans la chaîne de production

Ces premiers résultats mettent en évidence les grandes caractéristiques des différents algorithmes et des anomalies qu'ils détectent. Ils montrent également qu'il peut exister une certaine hétérogénéité à l'intérieur des unités atypiques identifiées par une méthode.

3.2 Dans quelle mesure les différentes méthodes se recoupent-elles ?

Pour analyser plus finement ces premiers résultats, nous comparons les trois algorithmes à la méthode actuelle de détection des anomalies et entre eux. Pour cela, nous distinguons les anomalies simultanément détectées par deux méthodes de celles spécifiques à chacune d'elles.

Le tableau 3 résume les comptages de nombre d'anomalies détectées en commun par les différentes méthodes. Chaque ligne précise la méthode prise en référence et indique pour chaque autre méthode la part d'anomalies qu'elles détectent en commun parmi les anomalies que la méthode de référence détecte. Ainsi, la première ligne du tableau indique que 2 % des anomalies détectées par la chaîne de contrôle des DADS le sont aussi par la forêt d'isolation. Le LOF détecte également 38 % des anomalies identifiées par la chaîne DADS, les règles d'association floues, 1 % et la régression, 39 %. Comme le nombre d'anomalies identifiées, fixé à partir du nombre d'anomalies détectées par la chaîne des DADS, est le même pour toutes les méthodes, le tableau est une matrice symétrique : par exemple, la part des anomalies détectées par le LOF parmi les anomalies identifiées par la forêt d'isolation est égale à la part des anomalies détectées par la forêt d'isolation parmi les anomalies identifiées par le LOF.

	DADS	Forêt d'isolation	LOF	Règles floues	Régression
DADS	100.00	2.00	38.00	4.00	39.00
Forêt d'isolation	2.00	100.00	35.00	2.00	61.00
LOF	38.00	35.00	100.00	4.00	31.00
Règles floues	4.00	2.00	4.00	100.00	3.00
Régression	39.00	61.00	31.00	3.00	100.00

TABLE 3 – Part d'anomalies détectées en commun par les différentes méthodes

La régression et la chaîne de production des DADS détectent une part importante d'observations en commun, ce qui s'explique par le fait que la régression contribue à la détection des anomalies dans la chaîne des DADS. Celle-ci détecte également une part équivalente des anomalies identifiées par le LOF. En revanche, elle semble orthogonale aux algorithmes de forêt d'isolation et de règles d'association floues.

Seules onze anomalies parmi les 645 détectées par chaque méthode sont de fait identifiées simultanément par la chaîne des DADS et la forêt d'isolation. Ceci tient vraisemblablement au fait que la procédure intégrée à la chaîne de production des DADS se concentre sur les observations ayant des valeurs très faibles de salaires horaires. Ces valeurs, tout en étant particulières, sont proches du coeur de la distribution des salaires, en tout cas beaucoup moins éloignées de celles-ci que les salaires horaires extrêmes. De ce fait, la forêt d'isolation détecte essentiellement des périodes caractérisées par des valeurs très élevées de salaires comme le montrent les exemples d'anomalies identifiées par la forêt d'isolation du tableau 4. Étant donné la durée de la période et le fait que ces observations soient associées à des temps complets, le nombre d'heures renseigné pour les périodes présentées dans le tableau 4 est très probablement erroné.

Brut horaire	Nb heures	Net horaire	Durée	Genre	PCS	Type contrat	Tps travail
394.50	2	309.06	12	HOMME	Employes	CDD	temps_complet
184.20	5	146.06	17	HOMME	Employes	CDD	temps_complet
116.89	39	92.75	84	HOMME	Employes	CDD	temps_complet
71.50	16	57.55	19	HOMME	Employes	CDD	temps_complet
52.79	24	44.76	21	HOMME	Employes	autres	temps_complet

TABLE 4 – Anomalies détectées par la forêt d'isolation et pas par la chaîne des DADS présentant le score d'atypie le plus élevé

Les règles d'association floues ne détectent pas du tout les mêmes anomalies que les autres méthodes. La logique de construction des règles d'association floues est en effet totalement différente des autres algorithmes. Ces derniers visent plutôt à détecter des valeurs extrêmes et rares, ou éloignées des autres, alors que les règles identifient des observations qui ne satisfont pas aux relations entre variables identifiées comme régulières dans le jeu de données, tout en présentant des caractéristiques fréquemment observées dans celui-ci. Ces différences de résultats proviennent également du fait que, contrairement à la forêt d'isolation et au LOF, les règles d'association utilisent aussi des variables auxiliaires pour le calcul du score d'atypie. D'après le tableau 5, les anomalies détectées par les règles d'association floues et non par la chaîne actuelle des DADS sont difficiles à caractériser : les niveaux de salaires horaires ne sont a priori pas aberrants et le faible nombre d'heures travaillées peut aussi s'expliquer par la faible durée de la période et le fait que ces observations ne soient pas associées à des temps complets. Seules les sixième et septième lignes associées à des temps complet présentent vraisemblablement un nombre d'heures erroné.

Brut horaire	Nb heures	Net horaire	Durée	Genre	PCS	Type contrat	Tps travail
9.93	202	7.67	44	FEMME	Ouvriers	CDD	autres
9.63	991	7.48	285	FEMME	Prof.Inter	CDD	autres
12.14	29	9.20	7	FEMME	Prof.Inter	CDD	autres
23.54	28	18.42	21	FEMME	Cadres	CDD	autres
20.01	151	15.49	87	FEMME	Prof.Inter	CDD	autres
12.81	43	10.16	9	FEMME	Prof.Inter	CDD	temps_complet
11.49	35	9.01	6	FEMME	Ouvriers	CDD	temps_complet
11.50	1320	8.88	324	FEMME	Prof.Inter	autres	autres
13.09	112	10.02	31	FEMME	Ouvriers	CDD	autres
11.15	238	7.71	360	FEMME	Ouvriers	CDD	autres

TABLE 5 – Anomalies détectées uniquement par les règles d'association floues et pas par la chaîne des DADS présentant le score d'atypie le plus élevé

D'après le tableau 6, les anomalies que la chaîne des DADS et le LOF détectent tous deux correspondent à des périodes d'emploi pour lesquelles les variables d'intérêt (salaires brut et net horaires, nombre d'heures) ont des valeurs extrêmes, en majorité très faibles. Les moyennes sont très supérieures aux médianes, traduisant que certaines de ces anomalies ont des valeurs très élevées des variables d'intérêt. Ces anomalies présentent aussi des incohérences manifestes entre salaires brut et net, la médiane et la moyenne de ce dernier étant supérieures respectivement à la médiane et à la moyenne du salaire net.
Ces résultats sont vraisemblablement liés aux spécificités de chacun de ces deux algorithmes et aux anomalies détectées. D'une part, la chaîne des DADS, en raison de son implémentation, tend à détecter surtout des observations ayant des valeurs extrêmes des variables d'intérêt, notamment des valeurs très faibles de celles-ci. D'autre part, le LOF est également susceptible de détecter des observations présentant une incohérence entre les trois variables d'intérêt. Ainsi, les valeurs extrêmes généralement faibles associées à des incohérences se retrouvent parmi les anomalies détectées en commun par les deux méthodes. En revanche, les anomalies détectées uniquement par la chaîne des DADS se caractérisent par des valeurs très faibles de salaires horaires sans incohérence apparente entre les salaires bruts et nets. Les anomalies identifiées uniquement par le LOF sont essentiellement associées à des valeurs très élevées de salaires horaires et à des nombres d'heures légèrement plus élevés que la moyenne.

Variable	modalité	groupe 1	groupe 2	groupe 3	groupe 4
Durée	Mediane	7.00	32.50	220.50	89.00
	Moyenne	63.51	106.87	206.64	149.94
Nb d'heures	Mediane	151.00	441.00	382.50	312.00
	Moyenne	446.76	656.46	849.80	665.63
Brut horaire	Mediane	1.69	3.41	51.23	11.87
	Moyenne	11.83	6.42	192.39	14.13
Net horaire	Mediane	2.43	2.97	35.35	9.16
	Moyenne	16.01	5.00	144.39	10.81
Genre	Femme	69.50	63.70	54.50	65.80
	Homme	30.50	36.30	45.50	34.20
PCS	Cadres	8.20	10.70	31.80	4.80
	Employes	70.00	68.20	46.80	78.50
	Independants	1.20	3.50	5.50	0.20
	Ouvriers	7.00	6.20	4.20	7.60
	Prof. inter.	13.60	11.40	11.70	8.90
Contrat de travail	Autres	7.40	6.50	2.70	2.50
	Cdd	34.60	26.10	19.70	38.20
	Cdi	58.00	67.40	77.60	59.30
Tps travail	Autres	55.60	58.70	31.10	42.20
	Tps complet	44.40	41.30	68.90	57.80
Effectifs		243.00	402.00	402.00	48953.00

TABLE 6 – Caractéristiques des anomalies en croisant LOF et DADS
Note de lecture : Le groupe 1 correspond aux observations détectées en anomalies par le LOF et la chaîne des DADS, le groupe 2 aux observations identifiées en anomalies seulement par la chaîne DADS, le groupe 3 aux observations identifiées en anomalies seulement par le LOF et le groupe 4 aux observations considérées comme « normales » par les deux méthodes.

Le tableau 7 illustre la diversité des anomalies que détecte le LOF et pas la chaîne des DADS :

— Les quatrième, sixième, septième et dixième lignes correspondent à des valeurs très élevées en termes de salaires horaires, et très basses en termes de nombre d'heures ;

— Les huitième et neuvième lignes correspondent à des périodes présentant une incohérence entre les niveaux des salaires, le salaire brut étant inférieur au salaire net ;

— Les autres observations sont plus difficiles à caractériser, probablement liées à une anomalie dans le nuage de points.

Brut horaire	Nb heures	Net horaire	Durée	Genre	PCS	Type contrat	Tps travail
9.59	150	7.54	27	FEMME	Employes	CDD	temps_complet
9.61	152	7.49	30	HOMME	Employes	CDD	temps_complet
9.61	152	7.49	30	FEMME	Employes	CDD	temps_complet
451.00	2	350.42	14	HOMME	Employes	CDD	temps_complet
8.84	151	6.88	30	FEMME	Employes	CDD	temps_complet
394.50	2	309.06	12	HOMME	Employes	CDD	temps_complet
53.97	243	40.93	292	FEMME	Employes	CDD	temps_complet
9.46	91	18.24	11	HOMME	Employes	CDD	temps_complet
4.20	25	7.84	7	FEMME	Employes	CDD	temps_complet
44.89	225	35.32	149	HOMME	Employes	CDD	temps_complet

TABLE 7 – Anomalies détectées uniquement par le LOF et pas par la chaîne des DADS présentant le score d'atypie le plus élevé

Selon le tableau 3, la régression, la forêt d'isolation et le LOF détectent deux à deux une part non négligeable d'anomalies en commun, d'où l'intérêt d'étudier plus spécifiquement leurs interactions.

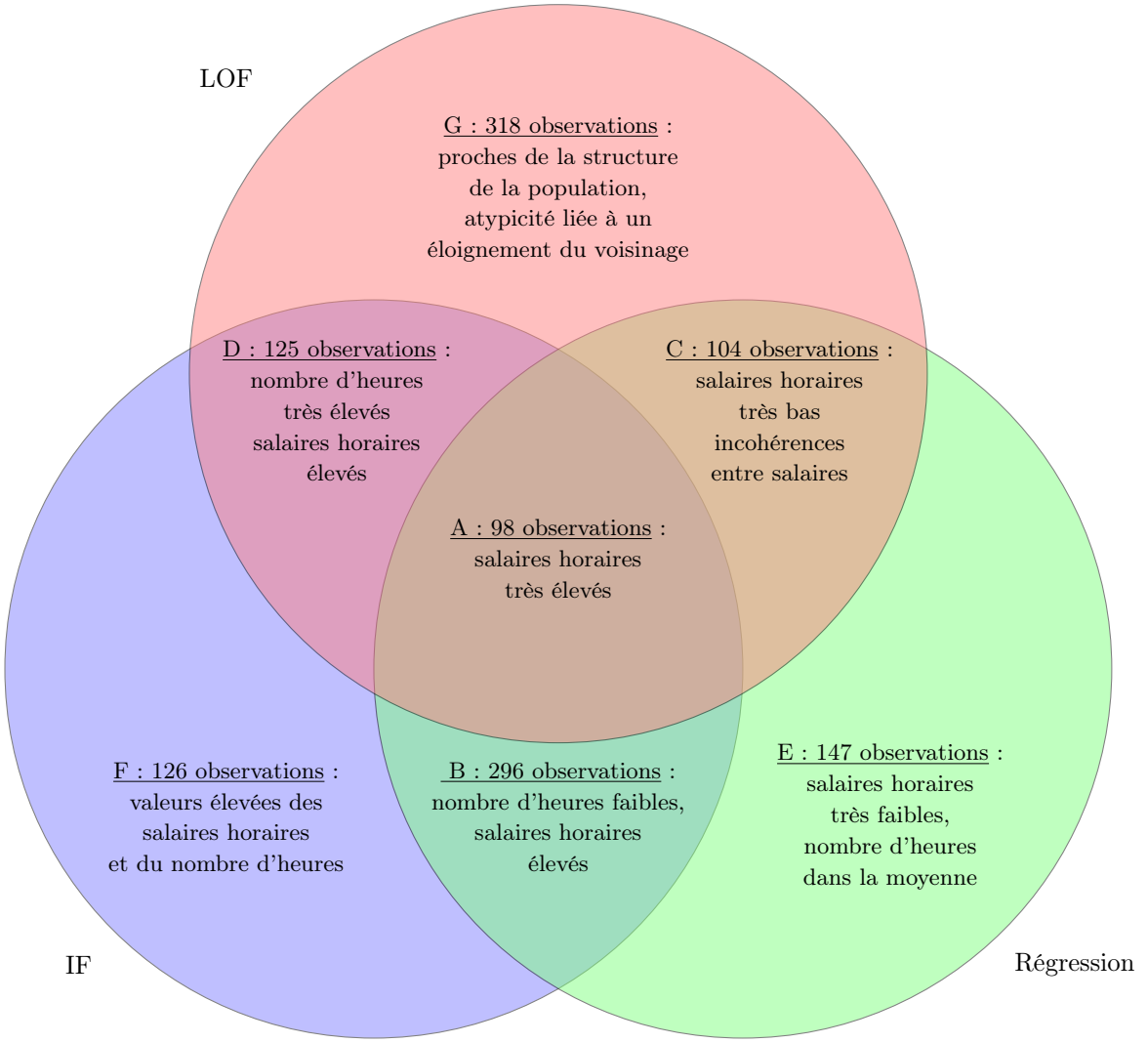


FIGURE 7 – Représentation des intersections entre ensembles d'anomalies détectées par la forêt d'isolation, le LOF et le modèle de régression log-linéaire

Les similitudes et les différences mises en exergue par le graphique 7 et le tableau 8 s'expliquent par les spécificités de chaque algorithme.

- Rappelons que la forêt d'isolation détecte en priorité des observations aisément séparables des autres ; ces observations sont celles qui ont des valeurs éloignées des modes des variables considérées. Comme elles sont appliquées directement sur les salaires nets et bruts horaires et le nombre d'heures, elles détectent d'abord des écarts élevés en valeur absolue sur l'une de ces variables. Or, les variables de salaires, et dans une moindre mesure de nombre d'heures, ont des distributions très dissymétriques avec des queues très étalées à droite. Ainsi, des valeurs très élevées de salaires se retrouvent très rapidement éloignées de la moyenne et sont donc détectées en anomalies, tandis que de faibles valeurs de salaires ont un écart relatif à la moyenne automatiquement de moindre ampleur.
- Le LOF évalue l'éloignement relatif d'une observation par rapport à ses voisins étant donné la densité de son voisinage. Ainsi, un même écart entre une observation et ses voisins se traduit par un score élevé si la densité de voisinage est très forte et que ses voisins sont très rapprochés et par un score faible si les voisins sont très dispersés. Comme pour la forêt d'isolation, le LOF met en évidence des observations avec des valeurs élevées sur l'une des nos trois variables d'intérêt. Mais il peut également détecter des observations éloignées d'une masse d'observations très dense, comme l'illustrait déjà le graphique 4c. Ces propriétés expliquent pourquoi le LOF est susceptible de détecter les incohérences entre les variables d'intérêt. En effet, les observations incohérentes ont tendance à être isolées et éloignées de leurs voisins.
- La régression met en exergue des observations avec des valeurs du logarithme de salaire net horaire éloignées de la moyenne conditionnellement aux valeurs des variables auxiliaires. La prise en compte explicite de ces variables conduit à détecter des observations atypiques en termes de salaires horaires au sein d'un groupe défini par les modalités de ces variables. De plus, l'application d'une régression log-linéaire, conformément à celle mise en œuvre dans la chaîne des DADS, accorde mécaniquement, en raison de la forme de la fonction logarithme, une importance plus élevée à un écart entre deux faibles valeurs qu'à un même écart entre deux valeurs plus élevées. La régression attribue des résidus équivalents aux salaires horaires de 1 et de 100 lorsque le salaire moyen du groupe est de 10 car il y a le même écart entre log(1) et log(10) qu'entre log(10) et log(100). Ainsi, les salaires doivent être très élevés pour être détectés en anomalies par la régression.

Ces différents éléments justifient les caractéristiques observées dans les colonnes du tableau 8.

Les anomalies détectées simultanément par les trois méthodes sont associées à des valeurs très élevées de salaires horaires.

Les anomalies détectées par la régression et le LOF, associées à la colonne C, ont de faibles valeurs de salaires horaires et des incohérences entre le salaire brut et le salaire net. Le premier point explique qu'elles soient identifiées par la régression et le second qu'elles le soient par le LOF. En revanche, les anomalies détectées uniquement par la régression ont de très faibles valeurs de salaires nets horaires. Elles ne semblent pas présenter d'incohérences entre les salaires bruts et nets et ne s'éloignent pas suffisamment de leur voisinage pour être perçues comme des anomalies par le LOF. De plus, aucune de ces anomalies n'est détectée par la forêt d'isolation car les salaires associés sont situés dans un intervalle de valeurs denses et proches de leur mode.

Les anomalies détectées uniquement par la forêt d'isolation (colonne F) le sont de nouveau du fait de salaires horaires élevés. Elles échappent cependant à la détection par la régression car elles concernent une majorité de cadres dont le salaire horaire moyen conditionnel est également élevé, situé autour de 25 euros. En effet, au sein de cette sous-population, l'écart en logarithme entre le salaire horaire net prédit et cette moyenne est trop faible. Les anomalies détectées par ces deux méthodes ont également des valeurs élevées de salaires horaires mais concernent en très grande majorité des employés dont le salaire net horaire moyen est de 10 euros. Ceci justifie leur identification par la régression contrairement aux observations de la colonne F.

Les anomalies détectées uniquement par le LOF ne se signalent pas par des valeurs très différentes de la moyenne de la population sur les variables d'intérêt. En termes de variables auxiliaires, ces anomalies ont une structure également proche de la moyenne de la population. Cela tient probablement au fait que le LOF détecte des observations atypiques relativement à leurs voisins.

Variable	modalite	A reg/IF/LOF	B reg/IF	C reg/LOF	D IF/LOF	E reg	F IF	G LOF	aucune
Durée	Mediane	199.00	30.00	3.00	360.00	11.00	178.00	30.00	90.00
	Moyenne	191.17	53.28	17.93	276.48	36.56	194.60	136.30	150.39
Nb d'heures	Mediane	11.00	15.00	151.00	1392.00	370.00	155.50	152.00	317.00
	Moyenne	189.86	25.52	356.14	1235.64	586.95	884.43	754.98	669.11
Brut horaire	Mediane	370.50	86.76	0.77	62.05	1.48	52.26	7.98	11.84
	Moyenne	678.93	109.84	1.49	75.28	3.20	58.90	12.94	13.41
Net horaire	Mediane	296.85	68.33	0.81	44.32	1.16	39.77	7.15	9.15
	Moyenne	529.39	86.87	5.45	49.62	2.86	42.10	10.34	10.25
Genre	Femme	54.10	70.60	74.00	39.20	81.00	45.20	65.70	65.70
	Homme	45.90	29.40	26.00	60.80	19.00	54.80	34.30	34.30
PCS	Cadres	26.50	3.70	2.90	71.20	2.70	56.30	9.40	4.80
	Employes	46.90	87.80	74.00	3.20	75.50	20.60	72.60	78.50
	Independants	3.10		1.90	12.80	2.00	2.40	1.30	0.20
	Ouvriers	5.10	4.40	4.80	1.60	5.40	2.40	6.90	7.70
	Prof. inter.	18.40	4.10	16.30	11.20	14.30	18.30	9.70	8.90
Contrat de travail	Autres	3.10	2.70	10.60	4.80	7.50	4.80	2.80	2.50
	Cdd	19.40	47.60	30.80		27.20	19.00	35.20	38.10
	Cdi	77.60	49.70	58.70	95.20	65.30	76.20	61.90	59.40
Tps travail	Autres	36.70	45.90	56.70	13.60	46.90	23.00	46.50	42.40
	Tps complet	63.30	54.10	43.30	86.40	53.10	77.00	53.50	57.60
Effectifs		98	296	104	125	147	126	318	48 786

TABLE 8 – Caractéristiques des anomalies en croisant la régression, la forêt d'isolation et le LOF

4 Conclusion

Ainsi, à l'issue de nos premiers travaux, nous constatons que les différentes méthodes permettent de détecter divers types d'anomalies. Si la forêt d'isolation met quasiment exclusivement l'accent sur les anomalies se distinguant nettement de la masse moyenne, le facteur LOF, qui évalue l'éloignement relatif d'une observation par rapport à ses voisins étant donné la densité de son voisinage, identifie, en plus de ses points très éloignés de la masse moyenne, des anomalies au sein du nuage de points mais localement éloignées d'une masse d'observations très dense. En revanche, les règles d'association qui appuient leur détection sur les règles suffisamment fréquentes détectent quasiment exclusivement des anomalies au sein de la masse des observations.

Face à ces différences, un échantillon de validation, qui fournit un label pour chaque période précisant si les salaires horaires net et brut et le nombre d'heures sont corrects ou en anomalie, serait indispensable afin d'évaluer les performances de chaque algorithme puis de sélectionner le plus pertinent ou de confirmer leur complémentarité. Toutefois, le coût élevé et la complexité de la constitution d'un tel échantillon qui, pour être totalement pertinent, devrait contenir un panel exhaustif des erreurs susceptibles de se produire, contraint à trouver d'autres alternatives d'évaluation. Ainsi, l'absence totale de labellisation nous contraint à juger de leur pertinence en croisant les résultats issus des différentes méthodes et en examinant manuellement les exemples d'anomalies les plus graves détectées par une méthode et omises par une autre. Il serait également possible d'analyser les anomalies au regard des variables auxiliaires. De plus, les analyses présentées dans ce document n'ont pas du tout étudié l'influence des observations sur les indicateurs majeurs issus des DADS. L'enjeu était en effet d'identifier des anomalies dans les données ; l'analyse de l'influence des anomalies est un complément utile, qui permet de juger de l'importance des éventuelles erreurs détectées et de prioriser leur traitement.

Les travaux réalisés et présentés dans ce document s'appuient sur des méthodes non paramétriques car elles ne supposent pas que les données suivent une distribution spécifique et connue. Toutefois, nous pourrions aussi envisager des méthodes paramétriques, qui sont par définition basées sur des distributions statistiques supposées dans les données, généralement des distributions gaussiennes. Sous cette hypothèse de distribution, une technique usuelle de détection d'anomalies consisterait à isoler les queues de distribution pour les variables d'intérêt étudiées séparément ou conjointement. Cette analyse est souvent appliquée aux résidus issus des régressions effectuées sur chacune des variables d'intérêt en supposant que les résidus suivent une loi gaussienne. Par ailleurs, sous une hypothèse paramétrique, le degré d'atypie d'une observation pourrait aussi être évalué par la distance de Mahalanobis suggérée par Hotelling dans l'article [Hotelling1931]. En s'appuyant également sur une hypothèse de normalité, la méthode ICS (*Invariant Coordinate Selection*) proposée dans l'article [Tyler2009] et appliquée dans l'article [Gazen2016] généralise la méthode en analyse en composantes principales pour calculer un score d'atypie des observations.

Pour prolonger cette étude, nous pourrions aussi envisager d'améliorer le modèle de régression utilisé dans la chaîne de production actuelle explicitée dans la partie 1.3, voire d'explorer d'autres méthodes de prédiction potentiellement plus performantes comme des méthodes ensemblistes, voire mêmes des réseaux de neurones.

Les limites de ces travaux seront cependant toujours les mêmes : en l'absence d'un échantillon de validation, il n'est pas possible de savoir laquelle des différentes méthodes de détection d'anomalies testées est la plus efficace, laquelle rate le moins d'erreurs et identifie, parmi les anomalies qu'elle signale, la part la plus élevée d'erreurs réelles. Les résultats de nos travaux montrent cependant que les algorithmes détectent des anomalies de nature différente suivant leurs spécifications. En l'absence de données permettant de valider et choisir les méthodes les plus efficaces, une solution peut être de combiner plusieurs algorithmes de détection d'anomalies, permettant de couvrir une gamme assez large d'erreurs possibles. Le contrôle de ces anomalies pourrait alors être priorisé suivant leur influence sur les principaux indicateurs issus des DADS.

Références

[Agrawal1993] R. Agrawal, T. Imielinski, and Swami A. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM Sigmod Conference*. Special Interest Group on the Management Of Data, 1994.

[Agrawal1996] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo. *Advances in knowledge discovery and data mining*, chapter Fast discovery of association rules, pages 307–328. Menlo Park : AAAI Press, 1996.

[Ankerst1999] Kriegel H.-P. Sander J. Ankerst M., Breunig M. M. Optics : Ordering points to identify the clustering structure. *ACM SIGMOD international conference on Management of data*, page 49–60, 1999.

[Bezdek1981] J. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Springer, 1981.

[Borgelt2002] C. Borgelt and R. Kruse. Induction of association rules : apriori implementation. In *Proceedings in computational statistics of the COMPSTAT 2002 Conference*, pages 487–489. COMPSTAT, Physica, 2002.

[Borgelt2012] C. Borgelt. Frequent item set mining. *WIREs Data Mining and Knowledge Discovery*, 2(6) :437–456, 2012.

[Chandola2009] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection : a survey. *ACM Computing Surveys*, 41(3), 2009.

[Chaput2015] H. Chaput. Description du mode de calcul des salaires nets et bruts dans les applications siasp et dads et diffusés dans le fichier dads grand format. Note Insee n°1237/DG75-F220/KAT-HC, 2015.

[Cordier-Villoing2018] M. Cordier-Villoing. Les concepts de salaire et mesure des concepts de rémunération. Note Insee n°18-F220-0296, 2018.

[Dunn1973] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well separated clusters. *Journal of Cybernetics*, 3(3) :32–57, 1973.

[Ester1996] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second International Conference on Knowledge Discovery and Data Mining*, pages 226–231. Association for the Advancement of Artificial Intelligence, 1996.

[Eurostat2014] Eurostat. *Essential System of National Accounts : Buiding the basics*. Eurostat Manual and Guidelines, 2014.

[Gazen2016] Ruiz-Gazena A. Archimbauda A., Nordhausenb K. Ics for multivariate outlier detection with application to quality control. *Computational Statistics & Data Analysis*, 128 :184–199, 2016.

[Grubbs1969] Grubbs F. E. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1) :1–21, 1969.

[Hahsler2005] M. Hahsler, B. Grün, and K. Hornik. arules - a computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, 14(15), 2005.

[Han2000] J. Han, H. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM Sigmod Conference*. Special Interest Group on the Management Of Data, 2000.

[Han2004] J. Han, H. Pei, and Y. Yin. Mining frequent patterns without candidate generation : a frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8(1) :53–87, 2004.

[Hong2001] T.-P. Hong, C.-S. Kuo, and S.-C. Chi. Tradeoff between computation time and number of rules for fuzzy mining from quantitative data. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 9(5) :587–604, 2001.

[Hong2006] T.-P. Hong, C.-H. Chen, Y.-L. Wu, and Y.-C. Lee. A ga-based fuzzy mining approach to achieve a trade-off between number of rules and suitability of membership functions. *Soft Computing*, 10(11) :1091–1101, 2006.

[Hong2008] T.-P. Hong, C.-H. Chen, Y.-L. Wu, and Y.-C. Lee. Genetic fuzzy data mining with divide-and-conquer strategy. *IEEE Transactions on evolutionary computation*, 12(2) :252–265, 2008.

[Hotelling1931] Hotelling H. The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(360–378), 1931.

[Kaufman1990] L. Kaufman and P. Rousseeuw. *Finding Groups in Data : An Introduction To Cluster Analysis*. John Wiley, New York, 1990.

[Kuok1998] C. M. Kuok, A. Fu, and M. H. Wong. Mining fuzzy association rules in databases. *ACM SIGMOD Record*, 27(1) :41–46, 1998.

[Liu2008] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *Proceedings of the 2008 Eight IEEE International Conference on Data Mining*, pages 1–6. Institute of Electrical and Electronics Engineers, 2008.

[Liu2012] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 276(1), 2012.

[MacQueen1967] MacQueen J. Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, 1 :281–297, 1967.

[Mahfouz2009] M. Mahfouz and M. Ismael. Fuzzy relatives of the clarans algorithm with application to text clustering. *International Journal of Computer and Information Engineering*, 3(1) :32–39, 2009.

[Ng2002] R.T. Ng and J. Han. Clarans : a method for clustering objects for spatial data mining. *IEEE Transaction on Knowledge and Data Engineering*, 14(5) :1003–1013, 2002.

[Srikant1996] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational databases. pages 229–234. Special Interest Group on the Management Of Data, 1996.

[Tan2006] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson Addison Wesley, 2006.

[Tyler2009] Dümbgen L.-Oja H. Tyler D. E., Critchley F. Invariant co-ordinate selection. 2009.

[Zaki1997] M. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. In *KDD-97 Proceedings*, pages 283–286. Association of the Advancement of Artificial Intelligence, 1997.

[Zaki2003] M. Zaki and K. Gouda. Fast vertical mining using diffsets. In *Proceedings of the 9th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*. Special Interest group on Knowledge Discovery and Data Mining, 1997.

[Zimmermann1991] H.J. Zimmermann. *Fuzzy Set Theory and Its Application*. Kluwer Academic Publisher, 1991.