5324 euros de l'heure : outlier ou footballeur ?

Méthodes d'apprentissage non supervisé pour la détection d'anomalies : application au cas de la Déclaration Sociale Nominative

Odran Bonnet, Marie Cordier-Villoing,
Thomas Deroyon, Julie Djiriguian

14 juin 2018

Introduction

- Détection des anomalies dans une base constitué à partir des déclarations de données sociales des entreprises : anciennement DADS et aujourd'hui DSN (travail sur les DADS).

- Base contenant l'ensemble des périodes d'emploi d'une année avec des variables de salaires (nets et bruts), individuelles (âge...) et d'entreprise (secteur...).

- Objectif : déterminer les observations pour lesquelles le triplet salaire brut, salaire net, nombre d'heures est aberrant.
  - Enjeu : distinguer les anomalies des valeurs atypiques (footballeurs vs outliers).

Introduction

- Difficulté principale : contexte d'apprentissage non supervisé.
  - Absence d'échantillon d'entraînement où les anomalies auraient été labellisées.
  - Nous ne pouvons que repérer des observations dont on soupçonne qu'elles sont des anomalies, mais pas de certitude !
  - Ainsi on va comparer quelles observations ressortent en tant qu'anomalies ou non selon différentes méthodes.
  - Dans les DADS, une fois la détection faite, des gestionnaires regardent spécifiquement les lignes identifiées comme anomalies.

- Autre difficulté : la taille des données, 63 millions de périodes.
  - Focus sur un sous-échantillon de salariés de la banque et des assurances contenant 50732 périodes.
  - À terme, adaptation possible des algorithmes.

Plan de la présentation

- La détection d'anomalies dans les DADS/DSN.
- Les méthodes de machine learning testées, présentation et comparaison des résultats :
  - Le *Local Outlier Factor* (LOF)
  - Les règles d'association floues
  - Les *isolation forests*

La détection d'anomalies dans les DADS/DSN

Les fichiers administratifs des salariés (DADS/DSN)

- Refonte des Déclarations Annuelles de Données Sociales (DADS) remplacées par les Déclarations Sociales Nominatives (DSN) qui sont **mensuelles**.
  - Dans le futur, travail sur des bases de taille comparable après agrégation des données mensuelles sur l'année.

- Spécificité des données : les salaires bruts et nets sont des données reconstruites à partir d'autres variables financières (base CSG, net fiscal...).

- A l'Insee, les DADS/DSN sont utilisées
  - pour fournir des statistiques économiques sur l'évolution des salaires et de l'emploi.
  - mettre à disposition des chercheurs (via le CASD) et des chargés d'études des données annuelles exhaustives sur les postes occupés + suivi d'un panel depuis 1967.

La détection d'anomalies actuellement réalisée

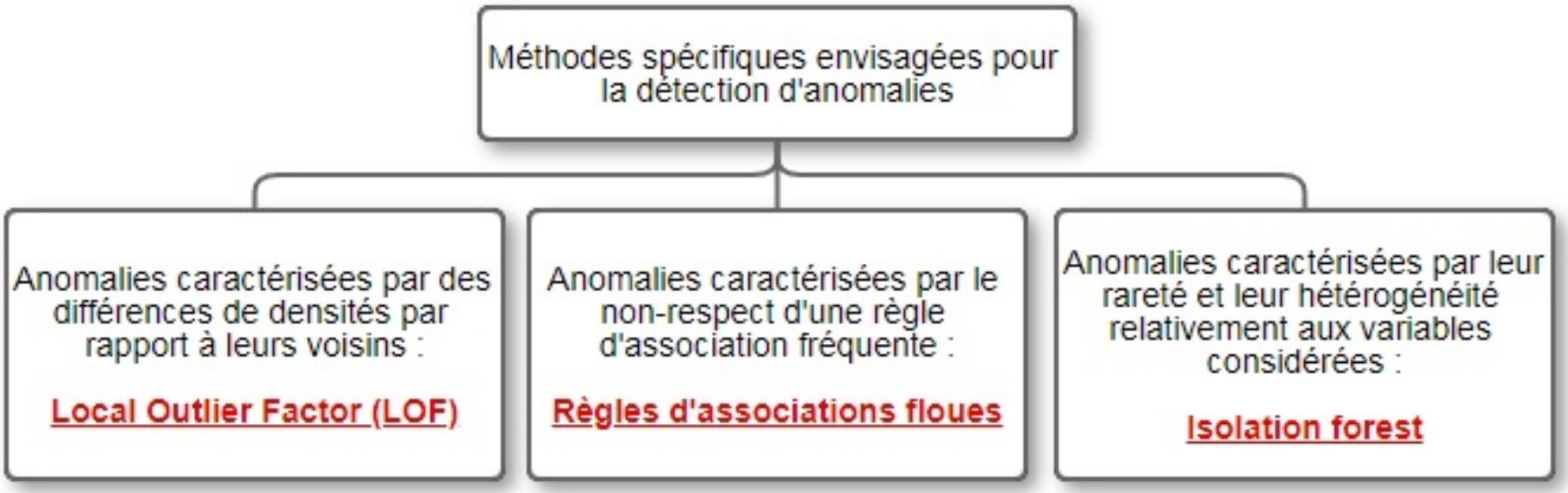
Les principales méthodes de détection utilisées :

- Utilisation de ratios de variables financières.
- Régression du salaire horaire sur les caractéristiques des entreprises (secteur...) et des individus (PCS, âge...).
  - Une anomalie est une observation dont le salaire observé s'éloigne trop du salaire prédit par la régression.

Des méthodes adaptées à la taille des données mais qui repose sur des hypothèses paramétriques (quel modèle de régression choisir ?).

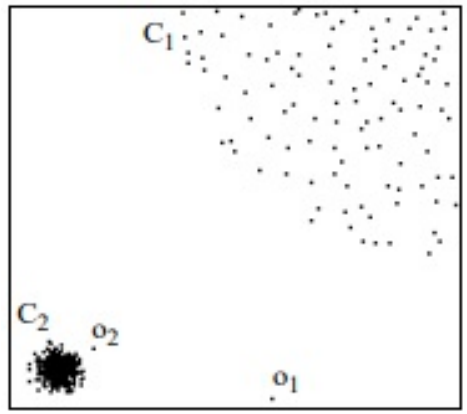
Les méthodes utilisées pour détecter les anomalies

Trois algorithmes : trois conceptions différentes des anomalies



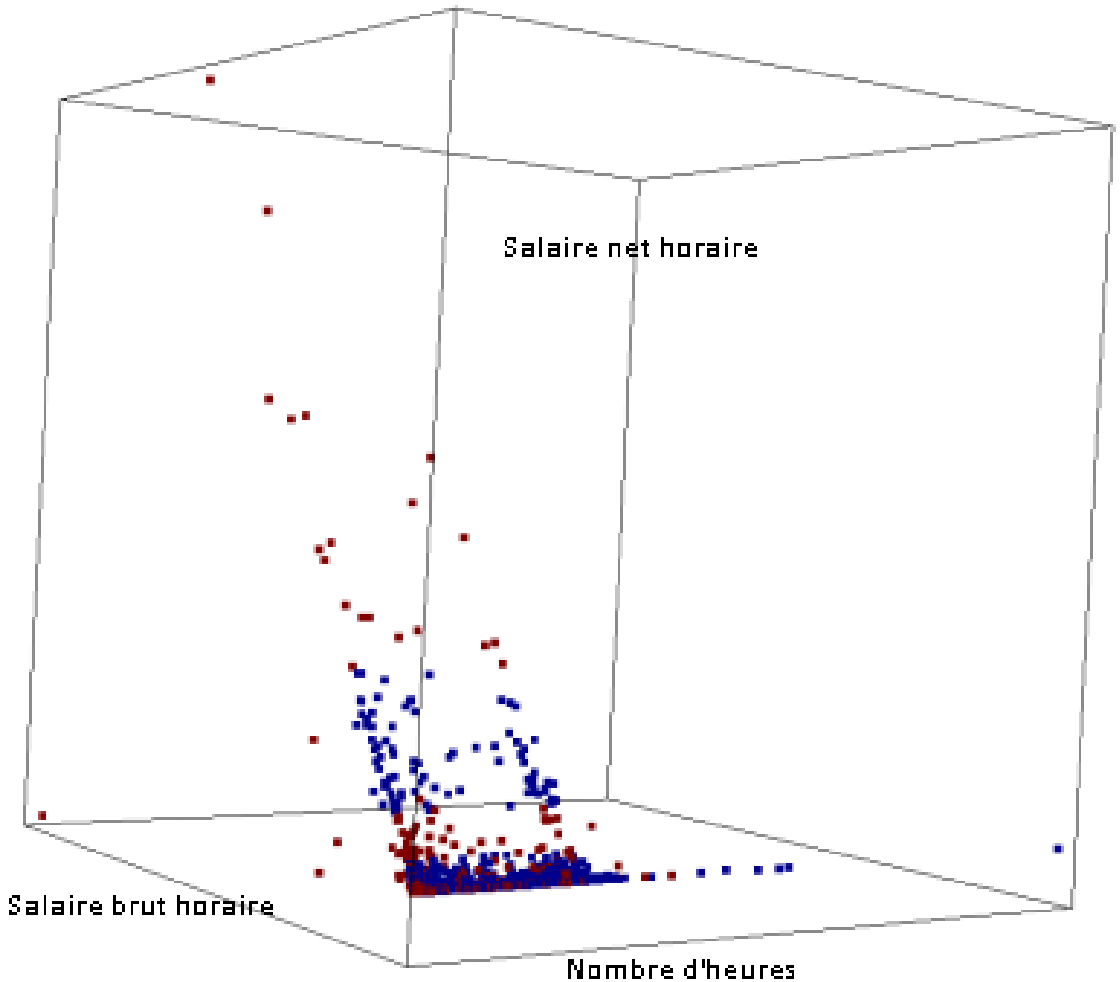
1) Le *Local Outlier Factor*

- Algorithme reposant sur le concept d'*outliers* locaux, par opposition aux *outliers* globaux généralement détectés avec une définition courante du terme d'*ouliers*



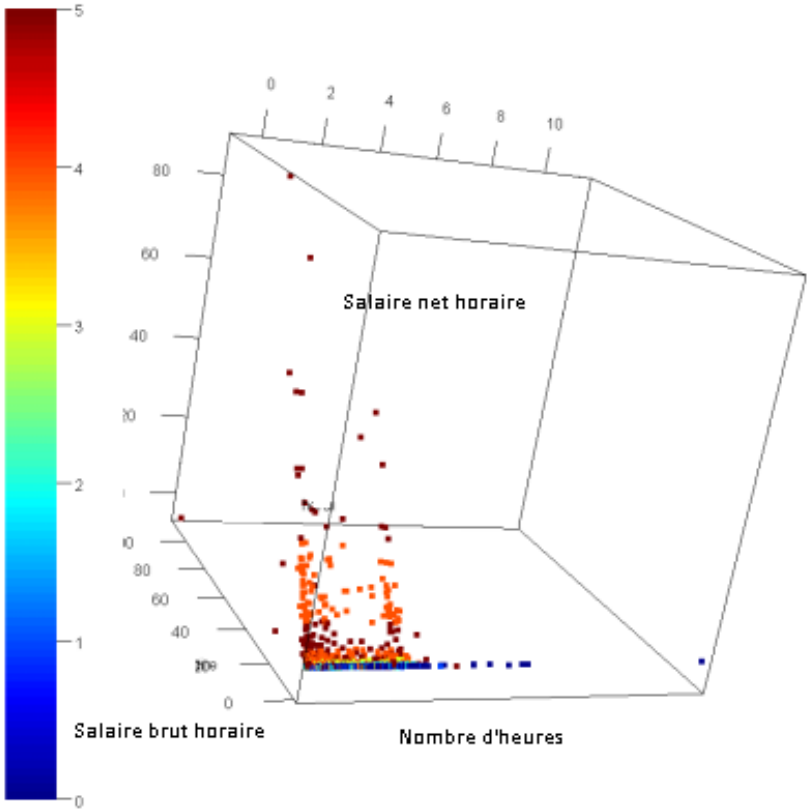
- Notions proches de celles caractéristiques à l'algorithme DBSCAN mais pas de constitution de *clusters* et pas une approche de densité globale
- Pas d'hypothèse sur la distribution multidimensionnelle des variables. Les différentes distributions sont, en effet, estimées localement

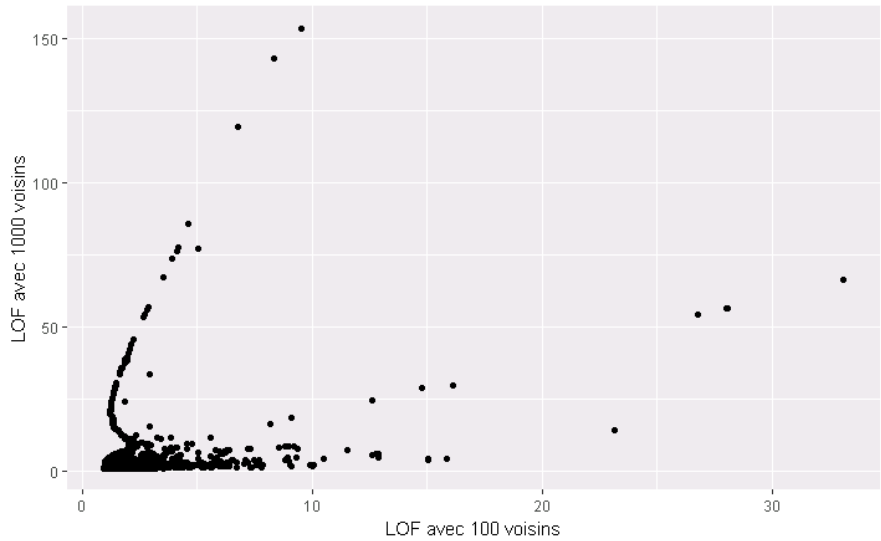
Que captons-nous avec une LOF relativement à nos variables d'intérêt ?



- Des couples de salaires (brut, net) élevés pour un faible nombre d'heures
- Mais pas que ...

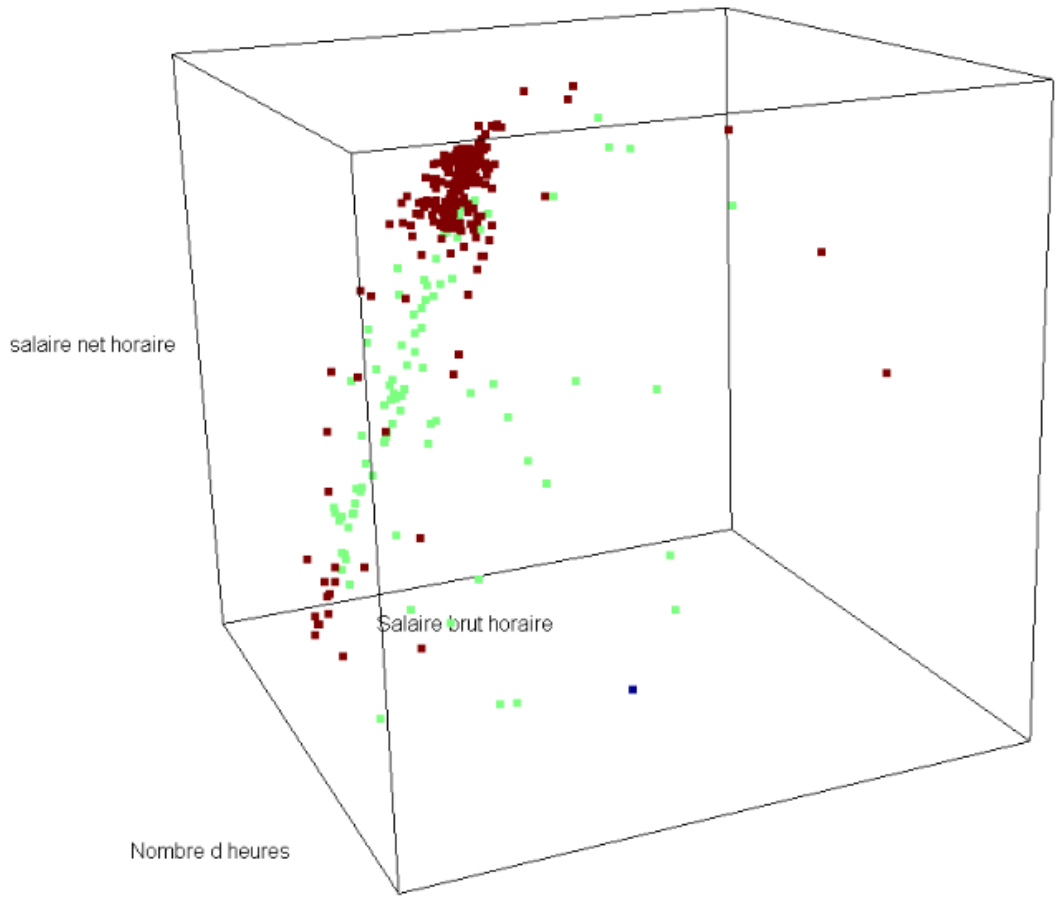
Comment choisir le nombre de voisins considérés ?





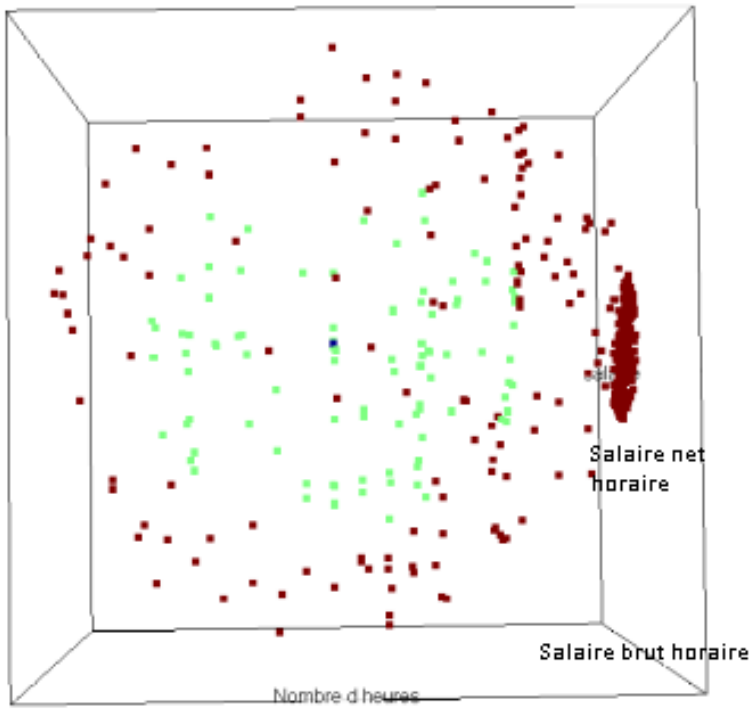
- Une cohérence globale :
  - Une masse d'observations autour de la moyenne jugées normales quel que soit le nombre de voisins
  - Des couples (brut, net) élevés jugés en anomalies quel que soit le nombre de voisins
- Des différences notables sur les anomalies selon le nombre de voisins

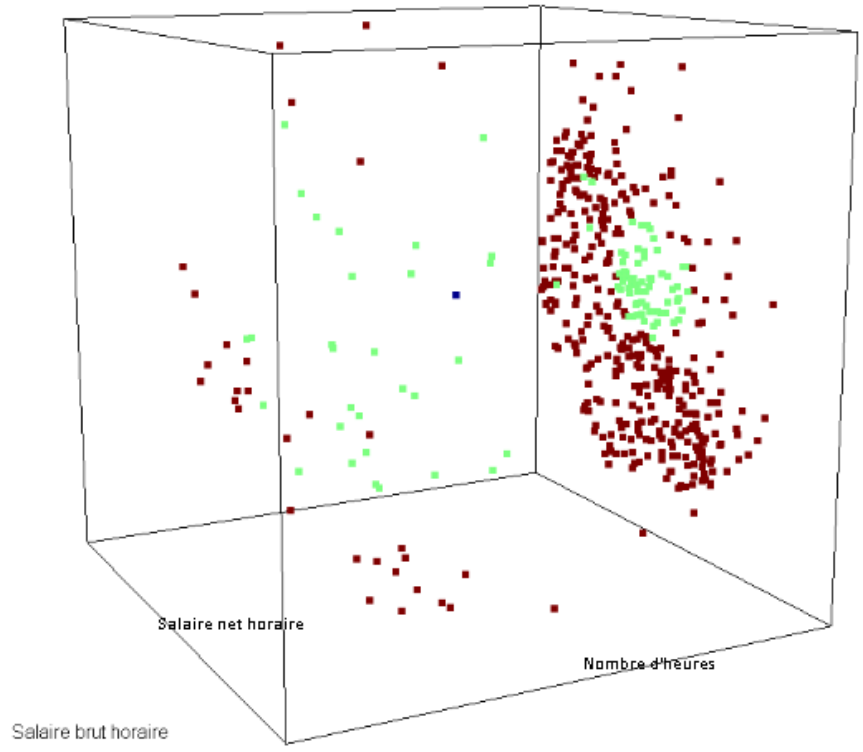
Des cohérences entre les LOF avec 100 et 500 voisins



- Une observation éloignée relativement aux densités des 100 et 500 observations voisines

Mais aussi des différences selon le nombre de voisins considérés

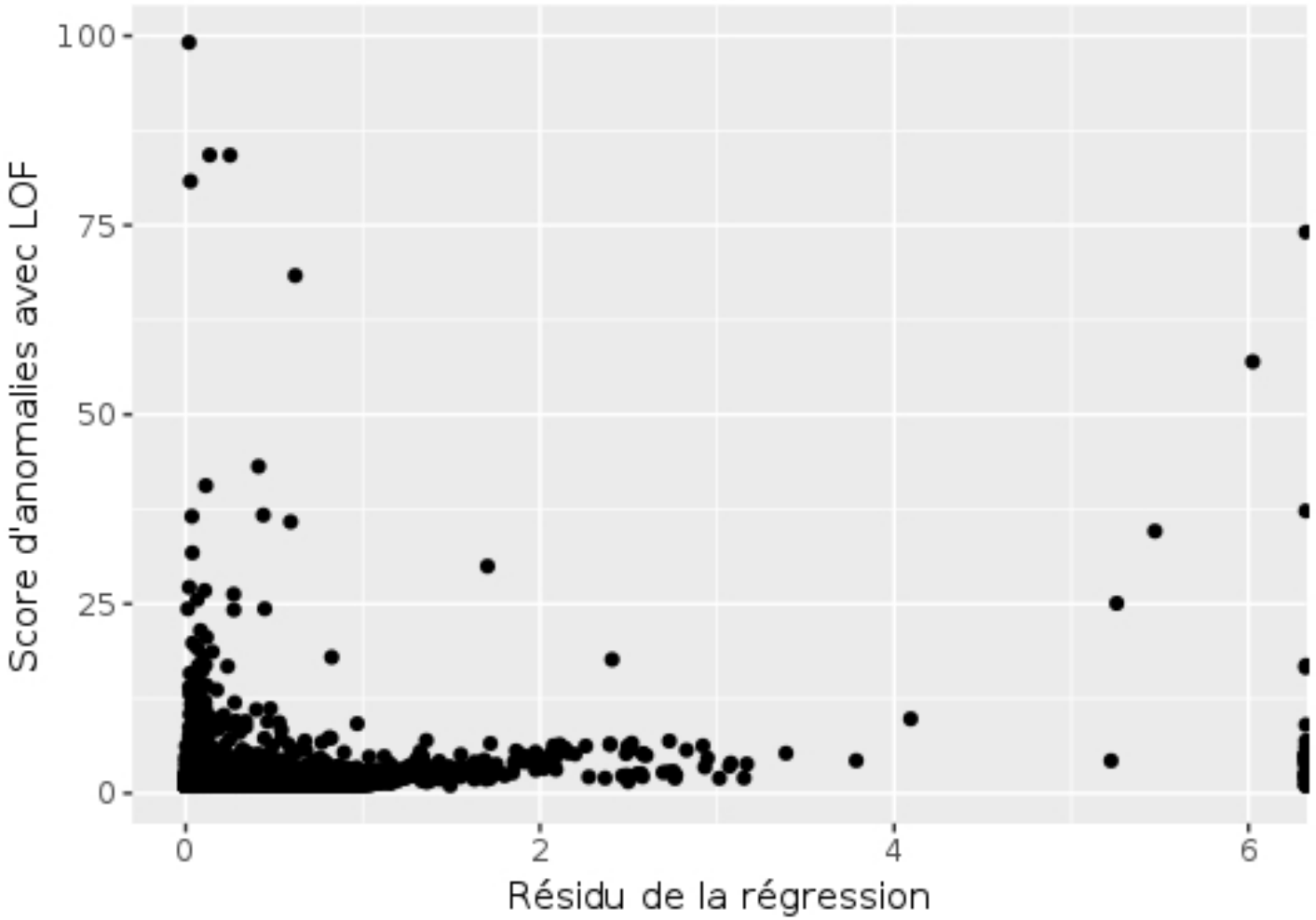




⇒ Ajout de voisins peut introduire des observations formant un *cluster* éloigné dense

⇒ Ajout de voisins peut introduire des observations moins denses

Des détections différentes avec la régression linéaire



$\Rightarrow$ Pas le même type d'anomalies détectées entre la régression et l'algorithme LOF

2) Les règles d'association floues

Format d'une règle d'association :
Catégorie sociale, sexe, tranche d'âge $\Rightarrow$ (salaire brut, salaire net, nombre d'heures)
Ou encore : (salaire brut, salaire net) $\Rightarrow$ nombre d'heures

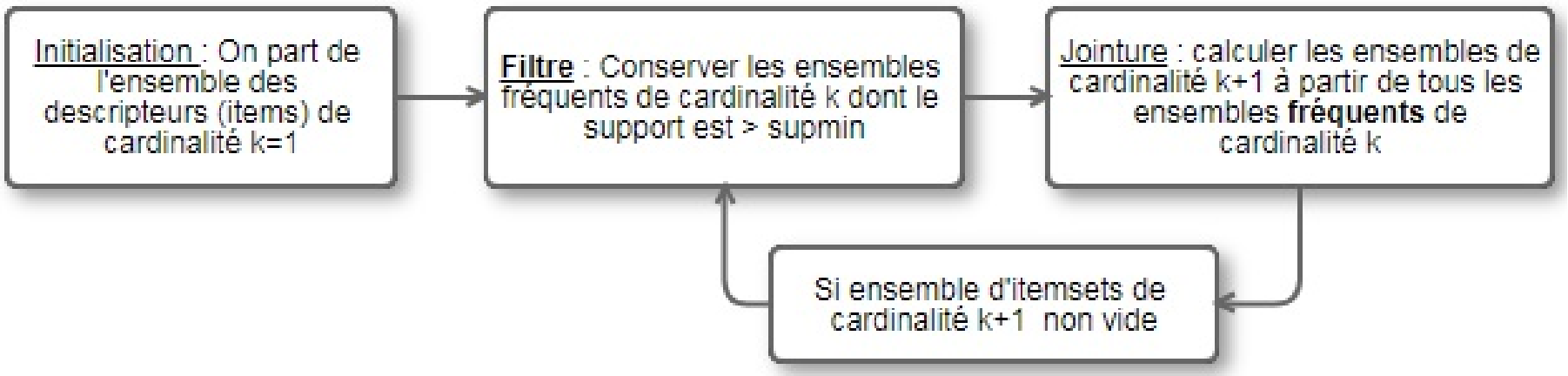
Deux critères de sélection des règles :

$$\begin{cases} \text{Support : } \sigma(A \Rightarrow B) = \text{nombre de fois où l'}\textit{itemset} \\ \text{considéré apparaît présent dans le jeu de données.} \\ \text{Niveau de confiance : } \frac{\sigma(A \Rightarrow B)}{\sigma(A)} \end{cases}$$

Principe des règles d'association classiques :

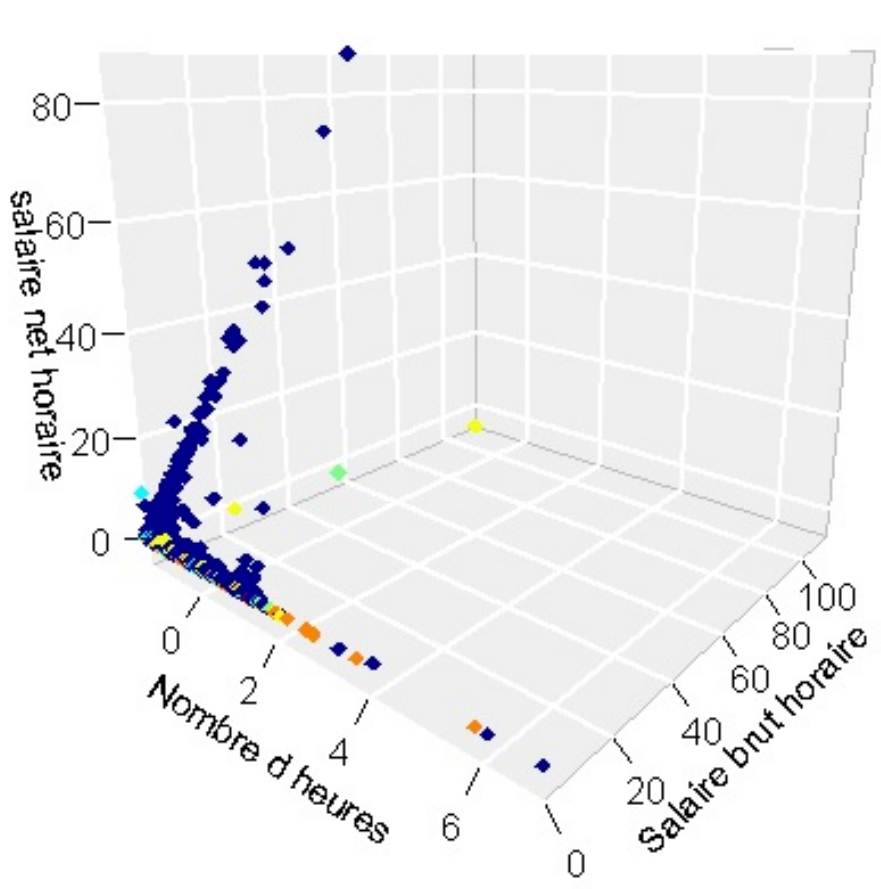
- Recherche des *itemsets* (ensemble de produits, par exemple le couple (salaire brut, salaire net)) fréquents avec le critère d'un support supérieur au support minimum
- Construction de règles à partir des *itemsets* fréquents

Algorithme *A priori*



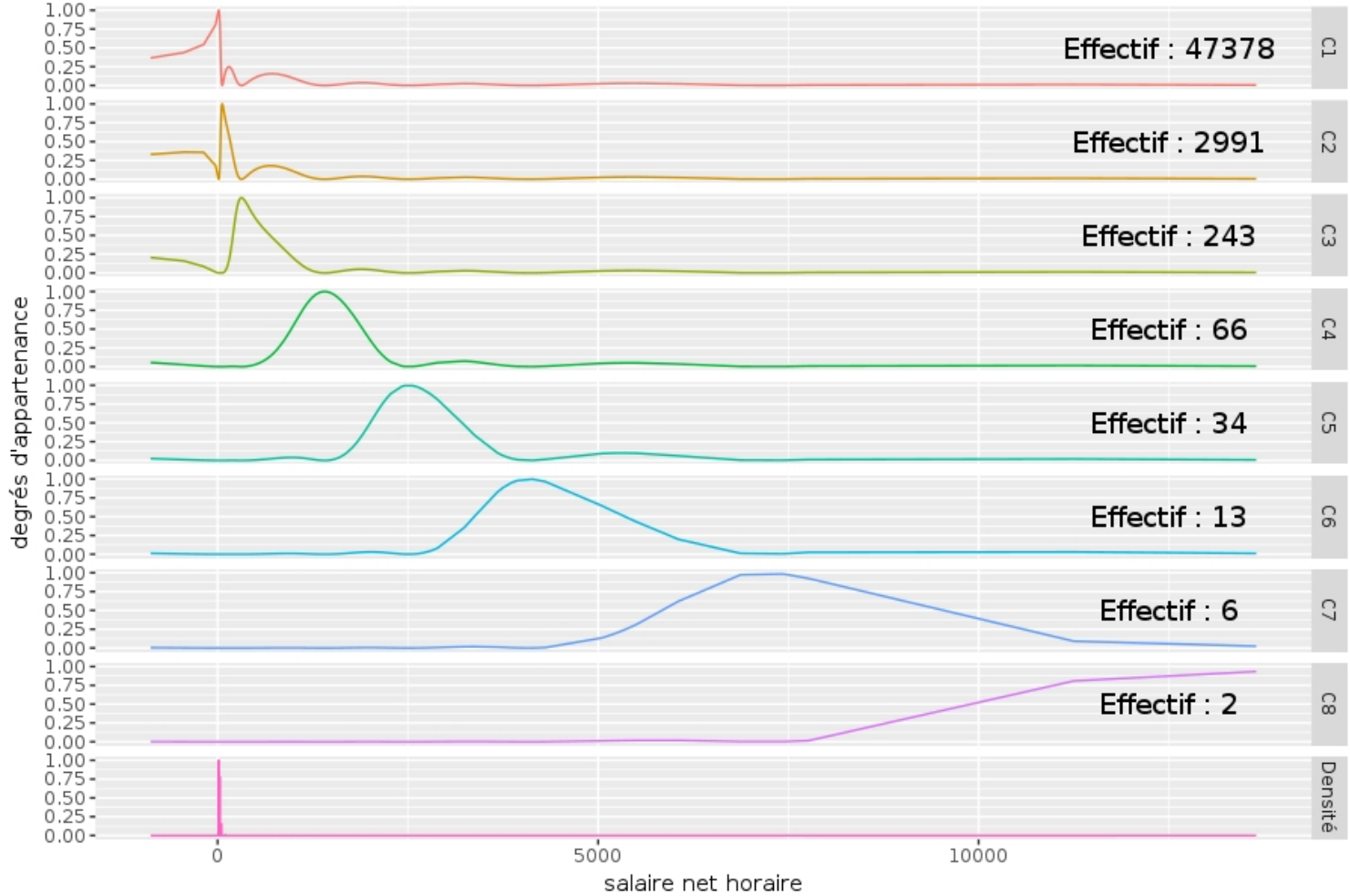
- **Propriété d'antimonotonie de la condition de support** :
  - tous les sur-itemsets d'un itemset non fréquent sont non fréquents
  - tous les sous-itemsets d'un itemset fréquent sont fréquents

Des anomalies détectées différentes au regard du triplet d'intérêt



- Des observations au nombre d'heures élevé pour un couple de salaires (brut, net) faibles
- Des anomalies détectées parmi la masse des observations

L'intérêt d'un découpage plus souple des variables quantitatives



Individu similaire avec des salaires (par exemple) de chaque côté du seuil $\Rightarrow$ Règles d'association potentiellement différentes

Des règles d'association mi-classiques, mi-floues

- Affectation d'une observation à plusieurs intervalles
- Affectation à l'intervalle I dès que le degré d'appartenance de l'observation dépasse le quantile de la distribution des degrés d'appartenance associés à I

- **Avantage** : Possibilité d'utiliser la *package arules* implémenté sous R
- **Inconvénient majeur** : Génération de règles absurdes :

```
lhs                    rhs
{f10_net_10=1} => {f10_brut_4=1}
{f10_net_10=1} => {f10_brut_8=1}
{f10_net_10=1} => {f10_brut_2=1}
{f10_net_10=1} => {f10_brut_9=1}
{f10_net_10=1} => {f10_brut_10=1}
{f10_net_10=1} => {f10_brut_7=1}
{f10_net_10=1} => {f10_brut_6=0}
{f10_net_10=1} => {f10_brut_5=1}
{f10_net_10=1} => {f10_brut_3=1}
{f10_net_10=1} => {f10_brut_1=1}
```

L'application de règles d'association floues

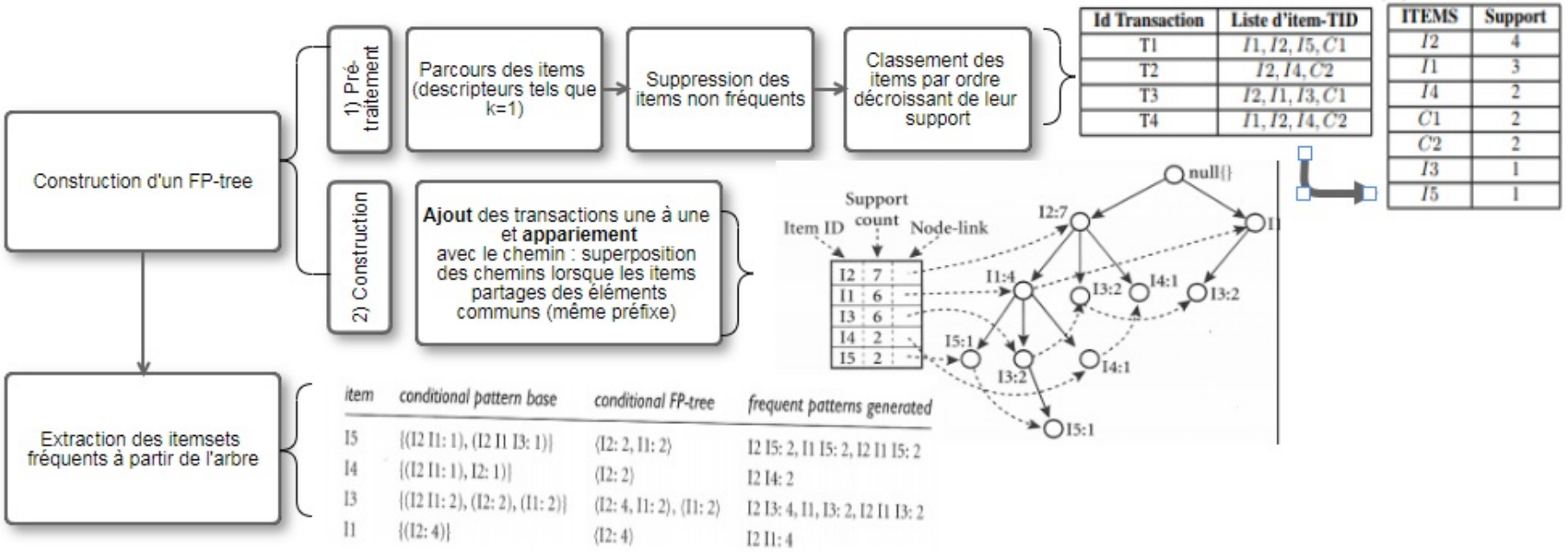
Parallèle entre les règles d'association floues et classiques :

- *significance* : équivalent du support
  = ratio entre le nombre d'observations satisfaisant l'*itemset* sur le nombre total d'observations

- *certainty factor* : homologue de la confiance
  $= \frac{\textit{significance de } \langle Z,C \rangle}{\textit{significance de } \langle A,X \rangle}$

- Implémentation de Helm sous R utilisant l'algorithme *FP-Growth* plutôt qu'*A priori*
- Un arbitrage difficile entre le critère de support minimal et le coût computationnel

Algorithme FP-Growth

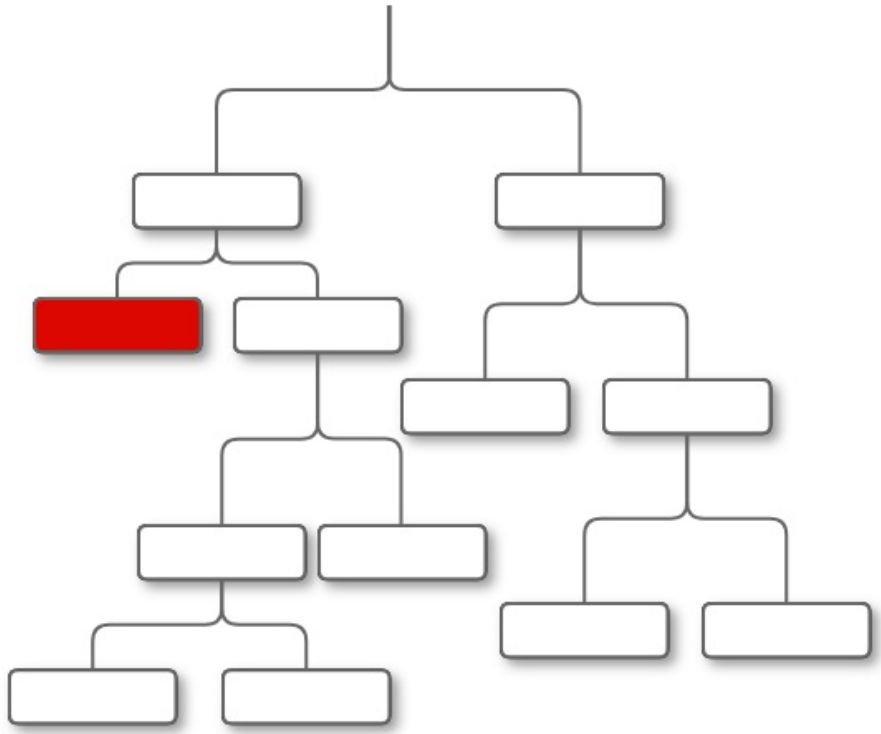


Différences majeures avec l'algorithme *A-Priori* dont :

- Pas de génération des *itemsets* candidats
- Seulement deux parcours des données

3) Les *isolation forests*

- Agrégation de nombreux arbres, dits *isolation trees*. Chaque arbre utilise un échantillon aléatoire d'observations et, à chaque scission, tire aléatoirement une variable puis la valeur de découpage



- Hypothèse relative aux *outliers* : Les anomalies sont **peu nombreuses** et **différentes**

- Anomalies : 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)

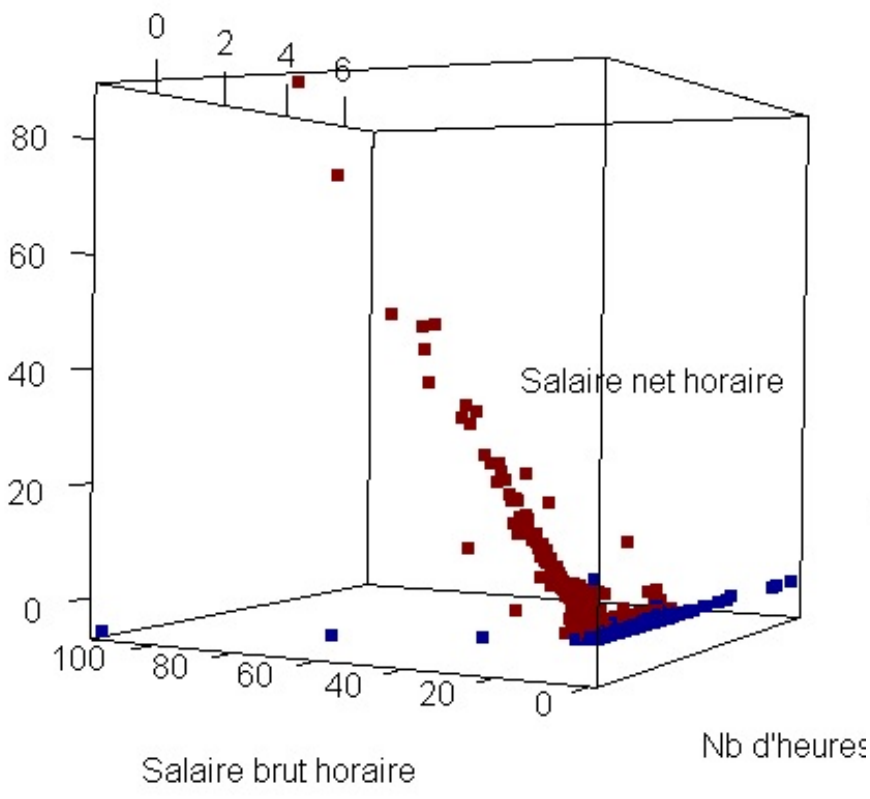
Avantages des *isolation forests* :

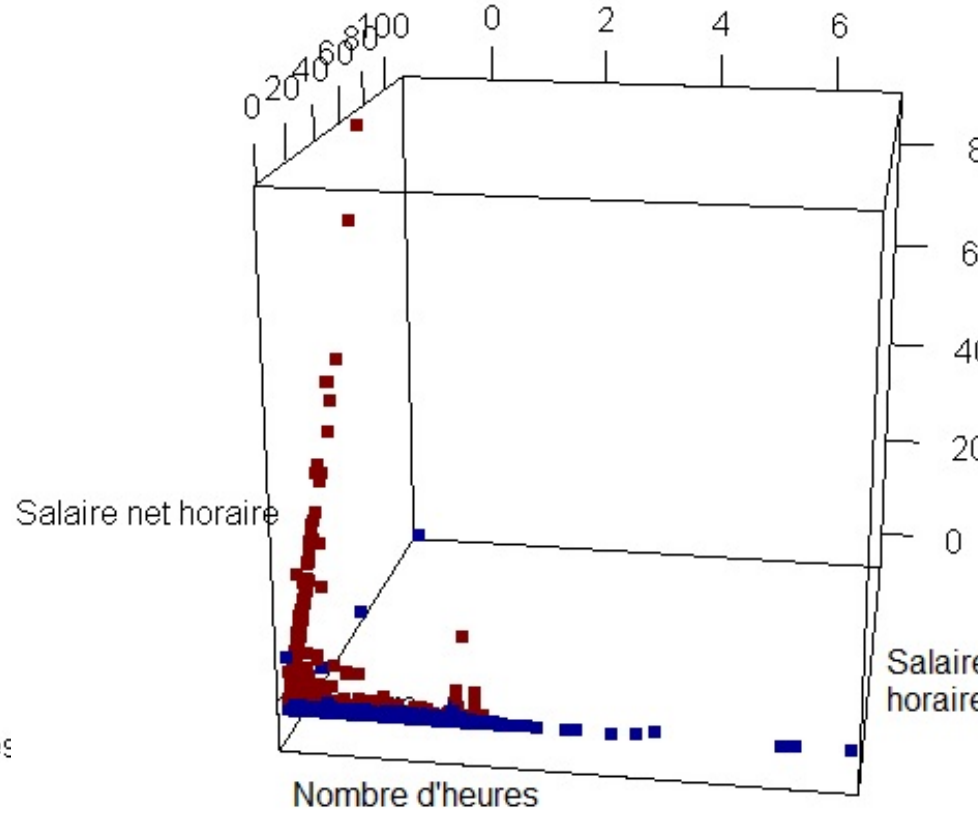
- Pas d'utilisation de distance, ni de mesure de densité pour détecter les anomalies $\Rightarrow$ Suppression d'un coût computationnel majeur comparativement aux autres méthodes

- Faible complexité algorithmique

- Capacité à gérer le passage à l'échelle avec des données de très grandes tailles et un grand nombre de variables peu/pas pertinentes

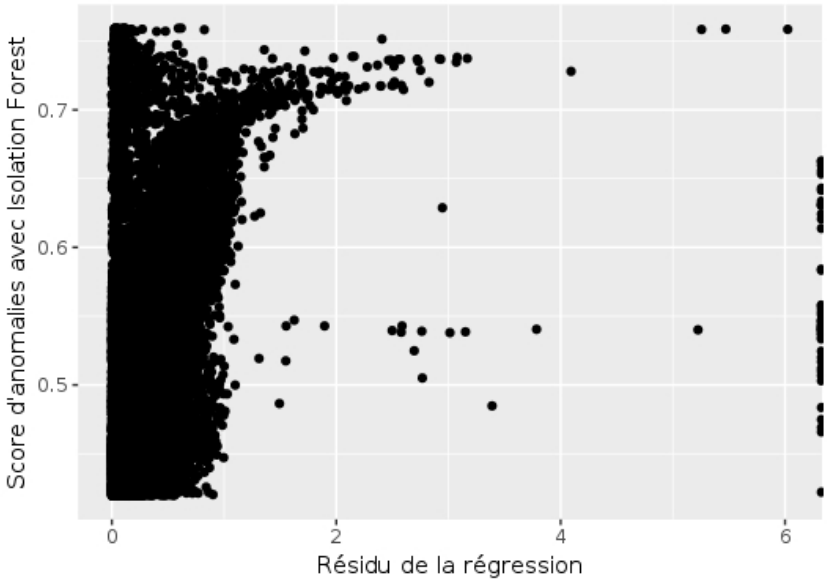
Les anomalies détectées par les *isolation forests*

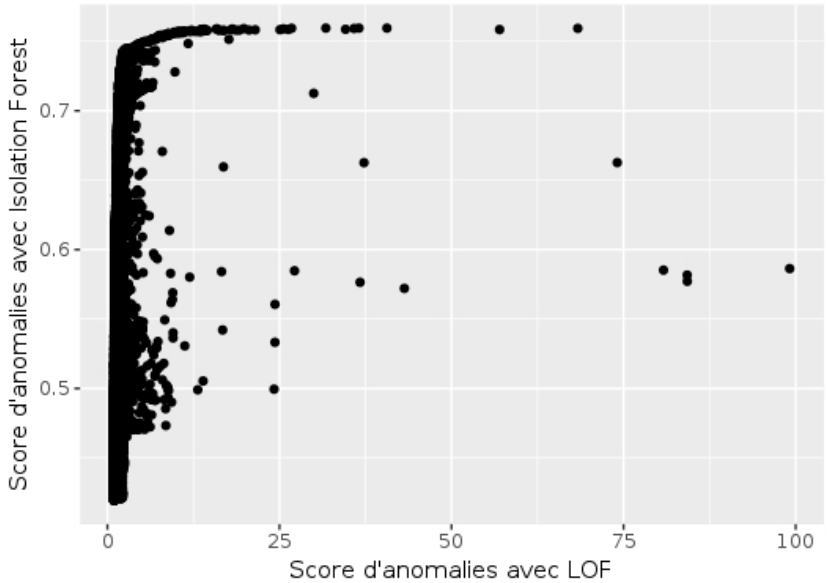




- Une masse d'observations autour de la moyenne jugées normales
- Des couples (brut, net) élevés jugés en anomalies

Quelles différences avec la régression et la LOF ?





$\Rightarrow$ Anomalies détectées différentes selon les méthodes : complémentarité entre les algorithmes ? Ou manque de pertinence d'un des algorithmes ? Ou mauvais paramétrage des algorithmes ?

En bref ...

Retour sur les résultats obtenus :

- Différences parmi les anomalies détectées avec les différents algorithmes notamment liées à des différences intrinsèques de ces derniers
- Difficultés notables à évaluer les performances absolues des algorithmes en raison de l'absence d'échantillon d'évaluation

Enjeux pour la suite :

- Nécessité d'un échantillon d'évaluation, voire d'un échantillon d'apprentissage pour explorer des méthodes supervisées ou semi-supervisées telles que l'*one class SVM*
- Exécution de règles d'association floues car constat que ARM permettent de détecter des anomalies mêlées à la masse au regard du triplet d'intérêt (notamment en raison de l'introduction facile de variables de contrôle)
- Améliorer le modèle de référence actuel (régression)
- Application à des échantillons de plus grande taille avec la parallélisation des tâches