Détection d'anomalies

Résumé

Présentation schématique du cadre général des méthodes de détection d'anomalies multidimensionnelles ou défauts, fraudes, défaillances... Description plus avancée de celles non supervisées et plus précisément non paramétriques, sans hypothèses sur la distribution ds variables ou de celles des résidu à un modèle : LOF, OCC SVM et Random Forest pour la détection d'observations atypiques. Illustration sur les données de prévision du dépassement du seuil d'ozone.

Retour au plan du cours

1 Introduction

1.1 Historique

La détection d'anomalies (*outliers*) est en enjeu ancien et majeur des applications industrielles de la Statistique notamment pour la détection d'une défaillance ou défaut de fabrication. Historiquement très présente dans les services de suivi de la qualité par contrôle statistique des procédés (*Statistical Process Control*), la détection d'anomalies utilise des techniques largement répandues et imposées par la normalisation : diagramme boîte (cf. figure 1), dépassement d'un seuil fixé par un expert, tests séquentiels et *cartes de contrôles*, tests paramétriques de *discordance* (Rosner, 1983)[11].

1.2 Objectif

Le but initial du contrôle de qualité prend de l'ampleur avec la recherche d'une explication voire même idéalement d'une *prévision de la défaillance* en vue d'une *maintenance prédictive*. De plus, l'afflux de données issues d'une multitude de capteurs ou objets connectés impose de remplacer l'approche unidimensionnelle par une conception *multidimensionnelle* de l'anomalie (cf. figure 1) pour prendre en compte volume, variété, vélocité des données. Par ailleurs, le même but de détection d'anomalies se décline en beaucoup objec-

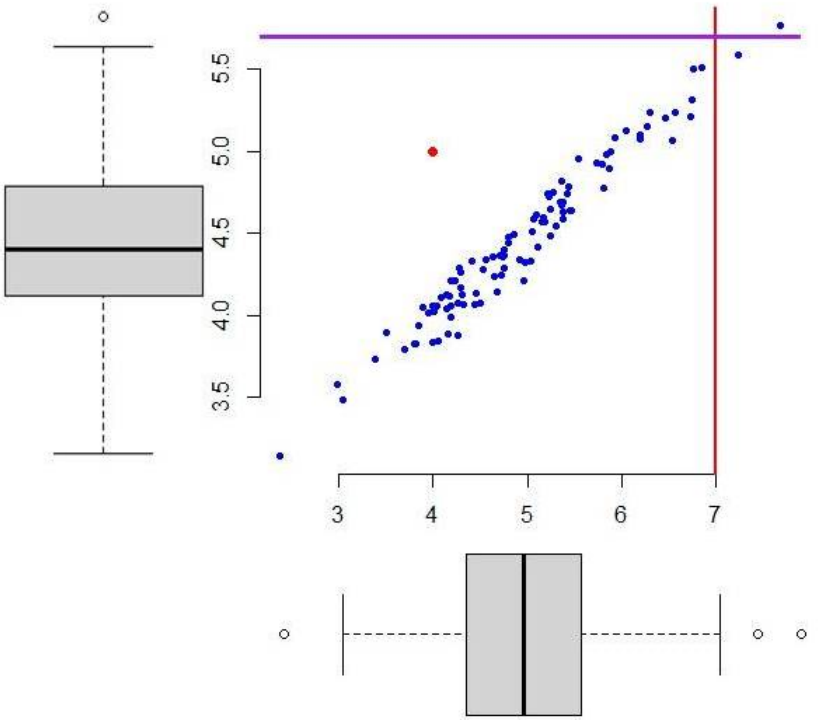

FIGURE 1 – *Exemple trivial : une anomalie détectée simultanément par les diagrammes en boîte unidimensionels semble moins atypique que celle en rouge au regard de la distribution bidimensionnelle.*

tifs analogues de détection dans d'autres domaines que celui strict de la production industrielle : alarme, défaut, fraude, intrusion... Cela concerne donc tous les secteurs de production de données et d'informations.

L'objectif de cette vignette est de tenter de donner un aperçu synthétique des méthodes les plus utilisées tout en les situant dans un schéma global des approches et stratégies de détection d'anomalies.

2 Aperçu des méthodes de détection

2.1 Choix opérés

Le rapide tour d'horizon des méthodes et technologies concernées se focalise sur celles récentes, multidimensionnelles, non paramétriques et implémentées dans les librairies facilement accessibles. Se reporter à la bibliographie et aux normes usuelles pour aborder les approches unidimensionnelles par cartes de contrôles et multidimensionnelles paramétriques.

Attention, la théorie des *valeurs extrêmes* qui estime des probabilités d'occurrences d'évènements rares en lien avec des lois de probabilités spécifiques est un tout autre problème pas du tout abordé.

2.2 Taxonomie de la détection d'anomalies

Quelque soit la méthode utilisée, une *anomalie des données* est toujours définie, implicitement ou explicitement relativement à un *modèle sous-jacent*. Le choix de la méthode et donc de ce modèle dépend complètement du contexte, de l'objectif visé, des données disponibles, de leurs propriétés. Voici un aperçu schématique de quelques possibilités. Se référer à Aggarwal (2017)[1] pour approfondir la réflexion.

La question de la détection d'anomalie peut être abordée de deux points de vue :

Supervisé : à condition de disposer d'une bases de données historiques contenant d'une part des observations jugées correctes à opposer à d'autres observations identifiées par un expert comme étant des anomalies. Apprendre à identifier, expliquer, prévoir ces anomalies se ramène alors à un objectif classique de classification binaire supervisée et renvoie aux méthodes exposées dans les autres vignettes. Attention, la classe des anomalies est très généralement (heureusement) sous-représentées, engendrant les problèmes classiques de déséquilibres des classes de la variable à prédire. Problème dont la résolution doit prendre en compte le coût relatif entre celui d'une détections à tort et celui de la non détection ; rapport dépendant du contexte métier de l'étude.

Non-supervisé : Dans le cas où aucune donnée ou caractéristique définit l'anomalie, la vaste littérature parle aussi de classification à une classe (*One Class Classification, OCC*) ou détection de nouveauté (*novety detection*). Ces dernières appellations introduisent une nuance dans l'objectif. Il s'agit soit de la détection d'anomalies (*outliers*) dans un ensemble de données, soit de déterminer si une nouvelle observation est cohérente ou non (*novety detection*) avec les données déjà disponibles regroupées en une seule classe. Néanmoins les mêmes techniques, objets de cette vignette, peuvent être utilisées dans les deux cas.

Attention, même dans ce cas non-supervisé, il et très utile voire indispensable de disposer d'un historique, ou de simulations réalistes, relatant des cas authentifiés d'anomalies. Ils permettent de valider un choix de méthode et même optimiser un seuil de détection.

Comme déjà écrit, une observation est considérée anormale ou **atypique par rapport à un modèle**. De façon schématique et pour structurer la présentation, ce modèle peut-être :

- *paramétrique*, très généralement gaussien (*e.g.* modèle de mélanges),
- *non paramétrique* défini à partir des données (*e.g.* voisinage au sens des k plus proches voisins).

D'autre part, ce *modèle* peut être

- relatif à la présence d'une *variable* cible à expliquer, *modéliser*, prévoir par régression ou discrimination.
- Dans le cas contraire, il est celui de la *densité de probabilité* ou distribution multidimensionnelle des variables.

Par ailleurs, certaines méthodes prennent en compte des mélanges de variables mixtes : quantitatives ou qualitatives, d'autres ne sont adaptées qu'à des variables quantitatives, notamment les méthodes paramétriques. Il est nécessaire dans ce dernier cas de rendre quantitatives les variables qualitatives susceptibles d'être présentes : remplacement par des variables indicatrices, composantes de l'AFCM.

2.3 Cas non supervisé

Voici un résumé succinct et quelques exemples spécifiques au cas *non-supervisé* ; il n'existe pas d'ensemble d'observations identifiées *a priori* comme des anomalies.

Cas paramétrique

La loi des variables explicatives, ou celle des résidus à un modèle, est supposée explicitement ou implicitement gaussienne multidimensionnelle.

Relatif au modèle d'une variable cible Dans le cas d'un modèle linéaire gaussien, les observations atypiques ou mal ajustées sont, par exemple, celles présentant de grands résidus studentisés ou de grandes valeurs de coefficients h_{ii} sur la diagonale de $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ où $\mathbf{X}$ est la matrice de design du modèle. Néanmoins, la détection d'observations atypiques rejoint plutôt celle des *observations influentes*, par exemple à l'aide de la *distance de Cook*, bien connue déjà en régression linéaire simple. Les autres observations mal ajustées peuvent en effet passer inaperçues sans risque pour la prévision.

Sans variable cible à modéliser une généralisation du T-test de student a été proposée par Hotteling (1931)[7]. L'anormalité à une densité multidimensionnelle peut être caractérisée par la distance (D_M) de Mahalanobis estimée à partir de la matrice inverse de celle de covariance empirique : ($\mathbf{S}^{-1}$) de la distribution. La librairie `mvoutlier` de R (voir la bibliographie associée) calcule des quantiles au sens de D_M.

Une autre approche (Ruiz-Gazen et Caussinus ; 2007)[5], basée sur l'*analyse en composantes principales*, utilise une estimation robuste de ($\mathbf{S}^{-1}$) :

$$\begin{aligned}
\overline{\mathbf{x}} &= \sum_{i=1}^{n} \mathbf{x}_i \\
\mathbf{S} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' \\
w_i &= \exp(-\beta/2||\mathbf{x}_i - \overline{\mathbf{x}}||^2_{\mathbf{S}^{-1}}) \\
\mathbf{R} &= \frac{\sum_{i=1}^{n} w_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'}{\sum_{i=1}^{n} w_i}
\end{aligned}$$

L'ACP calculée en utilisant pour métrique de l'espace des individus celle de matrice $\mathbf{M} = \mathbf{R}^{-1}$ à pour effet de faire ressortir les observations atypiques au sens de cette distance sur le premier plan factoriel. L'hypothèse de normalité n'est pas explicite pour exécuter l'ACP mais seule une distribution sous-jacente approximativement gaussienne peut conduire à des résultats raisonnables. Des distributions trop exotiques ou des structures de liaisons non linéaires complexes entre les variables peuvent sévèrement perturber les représentations.

Cas non Paramétrique

Il n'y pas d'hypothèse sur la distribution multidimensionnelle des variables, celle-ci est estimée localement de différentes façon.

Relatif au modèle d'une variable cible *Random forest* (Breiman, 2001)[2] inclut une solution originale adaptée à la prise en compte de variables mixtes.

Sans variable cible à modéliser C'est dans ce dernier cas que le plus de méthodes ont été proposées et la littérature la plus vaste. Quelques mots clés : LOF, GLOSH, OCC SVM... pour des variables quantitatives ou rendues quantitatives, *random forest* et *isolation forest* pour variables mixtes.

Il ne s'agit évidemment pas d'une *liste exhaustive* des méthodes et algorithmes disponibles. Ainsi les méthodes de classification non supervisée (CAH, k-means,...) sont aussi candidates à la détection d'anomalie lorsqu'une

ou des classes se réduisent à une seule observation ou que des observations restent en marge (DBSCAN). Consulter Aggarwal (2017)[1] pour avoir un aperçu plus vaste et plus détaillé d'un ensemble des méthodes et de leur justification.

Le *choix* opéré par la suite met l'accent sur les méthodes les plus généralement utilisables donc acceptant des variables mixtes, et facilement accessibles dans les librairies classiques (R et pour certaines en Python). Les méthodes choisies, non paramétriques, sont illustrées sur les données de prévision de dépassement du seuil d'ozone par adaptation statistique.

3 Méthodes non-paramétriques

Cette section développe donc les situations dans lesquelles

- il n'existe pas de base de données suffisamment renseignée des anomalies pour les apprendre,
- et qui ne nécessitent pas d'hypothèse de nature probabiliste (normalité) sur la distribution des variables ou de celle des résidus à un modèle.

Deux cas sont donc considérés selon qu'une anomalie est définie par rapport à un modèle explicatif ou prédictif d'une variable cible ou celui d'estimation non paramétrique de la distribution des observations.

3.1 Anomalie par rapport à un modèle prédictif

Il s'agit donc d'identifier des observations atypiques par rapport à un modèle expliquant une variable Y (régression ou discrimination) par des variables mixtes quantitatives ou qualitatives et sans hypothèse sur leur distribution. Breiman (2001)[2] propose de la faire en définissant une notion de *proximité* ou similarité puis de distance entre les observations participant à l'apprentissage d'une forêt aléatoire.

Une matrice de similarité entre les observations prises deux à deux qui ont participé à l'apprentissage d'une forêt aléatoire est simplement obtenue en comptant le nombre de fois où deux observations appartiennent à la même feuille d'un arbre. Ces effectifs sont ensuite normalisés par le nombre d'arbres de la forêt pour obtenir une matrice symétrique, positive dont les termes sont bornés par 1, les valeurs de la diagonale.

Breiman (2001)[2] définit ensuite un *score d'anomalie* d'une observation relativement à sa classe. Soit $\bar{P}$ la somme des carrés des proximités de l'observation en question à toutes les observations de sa classe. Le score est le rapport du nombre d'observations divisé par $\bar{P}$ puis normalisé en soustrayant la médiane et divisant par *MAD (mean absolute deviation)* : la médiane des écarts absolus à la médiane.

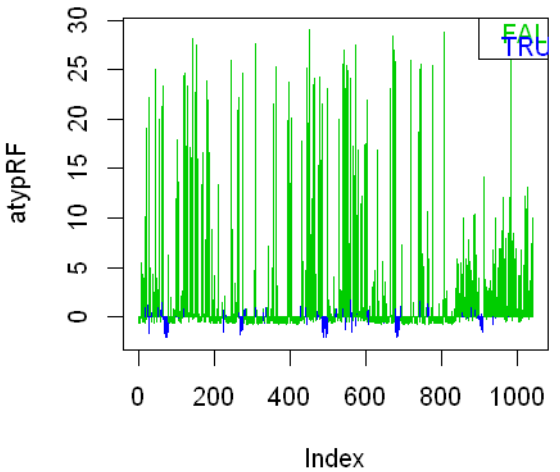

FIGURE 2 – *Ozone : Score d'anomalie des observations au sens des forêts aléatoires et par rapport à la prévision du dépassement du seuil d'ozone. En bleu dépassement, en vert pas de dépassement.*

La figure 2 représente les scores d'anomalies pour chaque observation au sens de cette définition lors de la prévision de dépassement du pic d'ozone. Seules des observations sans dépassement de seuil conduisent à des scores élevées ; observations dont les conditions météorologiques s'apparentent à celles observées lors d'un dépassement.

Le graphique suivant figure 3 représentent les observations jugées atypiques dans le premier plan factoriel de l'analyse en composantes principales

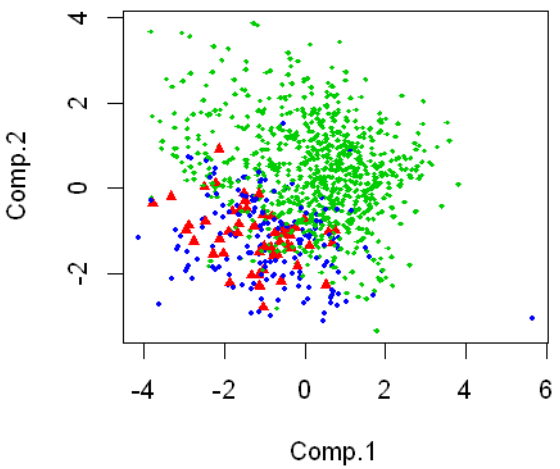

FIGURE 3 – *Ozone : observations atypiques (en rouge) au sens du critère issue d'une forêt aléatoire expliquant la variable de dépassement (bleu vs. vert) du seuil d'ozone.*

(ACP) ; ce sont celles avec un score supérieur à un seuil choisi arbitrairement à 20. Sans surprise, ces observations sans dépassement se projettent sur le premier plan de l'ACP dans le proche voisinage des observations avec dépassement.

Attention, supprimer ou modifier les observations atypiques à un modèle sans justification serait totalement contraire à l'éthique. L'objectif est avant tout de les identifier car ce sont celles, les plus susceptibles d'être la conséquence d'une erreur (à confirmer) de mesure, de libellé, ou encore une anomalie, défaillance ou tentative de fraude, d'intrusion, selon le contexte.

3.2 Anomalie par rapport à une distribution

Il s'agit dans ce cas d'évaluer l'incohérence ou l'*isolement* d'une ou de plusieurs observations par rapport à l'ensemble des autres observations. Plusieurs approches sont décrites sous les appellations de *One Class Classification* ou *novelty detection*.

Densité locale

Notations Beaucoup de travaux sont basés sur une estimation locale de la densité des observations en considérant leurs distances mutuelles. Dans cette section, les données sont un ensemble D d'individus x issus des observations de n vecteurs $\mathbf{x}$ de $\mathbb{R}^p$ muni d'une métrique (L_1, L_2, L_p...) définissant une distance $d(\mathbf{x}, \mathbf{y})$. Les données peuvent également être, directement, la connaissance d'une matrice $\mathbf{D}_{n\times n}$ de distances des individus, observations ou instances pris 2 à 2 dans D.

L'anomalie ou l'isolement d'une observation est apprécié par la proximité des points de son voisinage. Ramaswamy et al. (2000)[10] ordonnent les observations $\mathbf{x}$ de D selon la distance : $\text{Ddist}_k(x)$ de leur k-ième voisin. Les plus grandes valeurs désignent les observations les plus atypiques. Knorr et al. 2000)[8] proposent une autre approche en considérant atypique une observation x si un grand pourcentage, à fixer, des observations y de D est à une distance $d(x, y)$ plus grande qu'une borne minimale également à fixer. Les auteurs se focalisent également sur la complexité des algorithmes proposés.

LOF De très nombreux aménagements ont été proposés à ces versions de base notamment pour stabiliser les résultats ou pour réduire la sensibilité à des situations singulières comme des mélanges de distributions présentant des niveaux de densité très différents. La version la plus populaire est le LOF (*local outlier factor* ; Breunig et al. 2000)[3] basé sur des notions proches de l'algorithme DBSCAN (*density-based spatial clustering of applications with noise* ; Ester et al. 1996)[6] de classification non supervisée sans pour autant chercher des classes.

Breuning et al. (2000)[3] commencent par définir deux quantités. La k-distance, notée Dist_k, plus complexe que celle ci-dessus afin de prendre en compte les possibles équidistances entre les observations.

$\text{Dist}_k(x)$ est égale à $d(x, y)$ pour une observation y de sorte que pour au moins k observations y' de $D : d(x, y') \leqslant d(x, y)$ et que pour au plus $k - 1$ observations y' de $D : d(x, y') < d(x, y)$.

$V_k(x)$, le voisinage de k-distance de x est l'ensemble des observations vérifiant :

$$V_k(x) = \{x \in D \quad | \quad d(x, y) \leqslant \text{Dist}_k(x)\}.$$

Du fait de possibles équidistances entre les observations, le cardinale de $V_k(x)$ peut être plus grand que k.

La *distance d'atteignabilité* (*reachability distance*) entre deux observations est définie ci-dessous. *Attention*, malgré une appellation communément admise, cette quantité n'est pas une *distance* car elle n'est pas symétrique.

$$\text{RDist}_k(x, y) = \max\{\text{Dist}_k(y), d(x, y)\}.$$

$\text{RDist}_k(x)$ est la distance $d(x, y))$ si y est assez éloigné de x mais, si x est dans le k-voisinage de y, $\text{RDist}_k(x)$ est minoré par $\text{Dist}_k(y)$. Les observations du k-voisinage de y sont considérées équidistantes afin d'apporter une forme de lissage, contrôlé par le paramètre k, à la conception des critères.

La densité locale d'atteignabilité (*local reachability density* en x est ensuite définie par l'inverse de la moyenne de la distance d'atteignabilité dans le k-voisinage de x :

$$\text{LRDens}(x) = 1/\left(\frac{\sum_{y \in V_k(x)} \text{RDist}_k(x, y)}{\text{card}(V_k(x))}\right).$$

Finalement :

$$\text{LOF}_k(x) = \frac{\sum_{y \in V_k(x)} \frac{\text{LRDens}(y)}{\text{LRDens}(x)}}{\text{card}(V_k(x))}$$

La valeur du LOF est difficile voire impossible à interpréter dans l'absolu. Une valeur de 1 correspond à une observation dans la norme de la distribution, mais une borne au delà de laquelle une observation est atypique n'est pas explicite, cela dépend du contexte et des dispersions relatives.

Motivées par les critiques, des variantes sont proposées afin d'y remédier. *LoOP (Local Outlier Probability)* tente d'être moins sensible au choix de k et est normalisé dans l'intervalle $[0, 1]$. L'*Interpreting and Unifying Outlier Scores* est présentée comme une amélioration du précédent. D'autres approches tentent une démarche similaire à l'agrégation de modèles : échantillons bootstrap et *bagging* des critères, combinaison de critères différents. Enfin, le *Global-Local Outlier Score from Hierarchies (GLOSH)* (Campello et al. 2015)[4] est basé sur une version hiérarchique de l'algorithme DBSCAN plutôt que sur DBSCAN comme le LOF.

Comme souvent, il faudra un peut de temps et d'expérimentations pour qu'une sélection naturelle opère entre toutes les approches publiées et retienne le critère garantissant un meilleur compromis entre pertinence des résultats et complexité algorithmique.

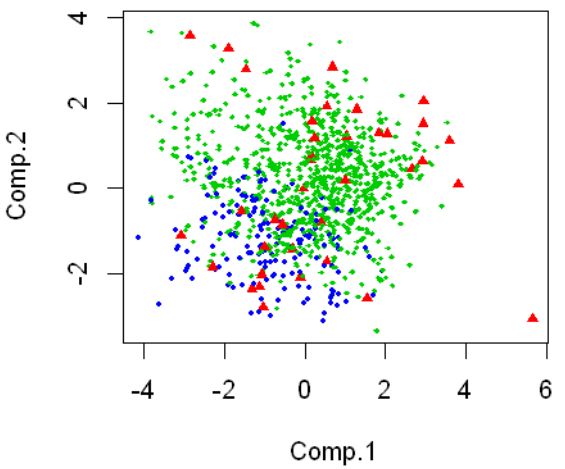

FIGURE 4 – *Ozone : Anomalies au sens du LOF (Local Outlier Factor).*

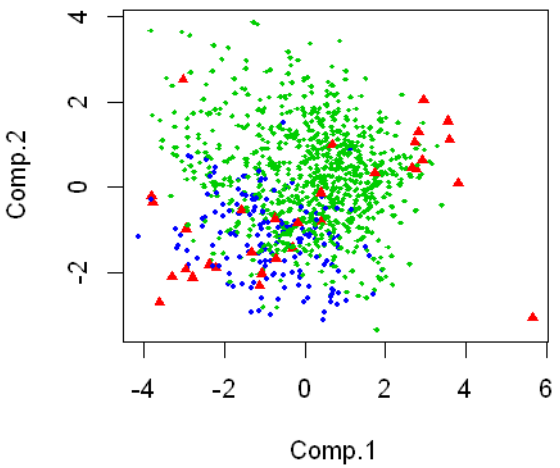

FIGURE 5 – *Ozone : Anomalies au sens du GLOSH (Global-Local Outlier Score from Hierarchies).*

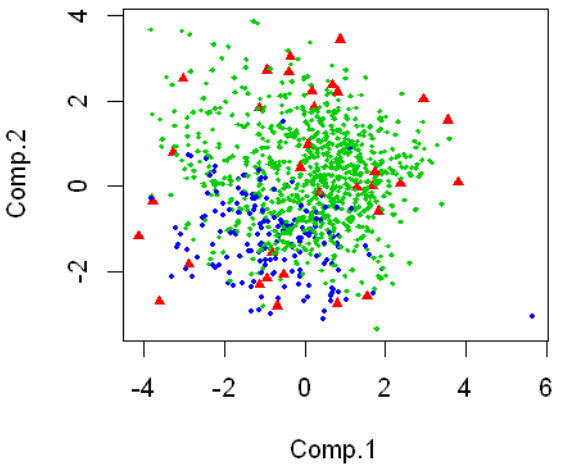

FIGURE 6 – *Ozone : Anomalies au sens de OCC SVM (One Class Classification SVM).*

OCC SVM

Une autre approche cherche une enveloppe, ou un support des observations jugées normales, définie par des séparateurs à vaste marge (*One Class Classification SVM*) (Schölkopf et al. 1999)[12]. Le principe consiste à poser le problème d'optimisation des SVM avec pour objectif de séparer les données, toutes les observations, de l'origine, dans l'espace de représentation (*feature space*) en maximisant la marge, à savoir la distance entre l'hyperplan et l'origine. La solution produit une fonction binaire qui vaut $+1$ dans la plus petite région captant les données et -1 ailleurs. Le paramètre de pénalisation à optimiser établit un équilibre entre la régularité de la frontière et la proportion d'observations considérées comme atypiques.

OCC RF

Enfin, la version de *random forest* (Breiman, 2001) pour la classification non supervisée est adaptée de façon très spécifique à l'objectif visé de détection d'anomalies.

Forêt aléatoire non supervisée La version non-supervisée des forêts aléatoires est une application de la notion de proximité entre les observations. Par défaut, lorsqu'aucune variable explicative ou cible n'est fournie à l'algorithme, celui-ci génère deux classes d'observations. La première désigne les observations initiales, la deuxième est obtenue par *permutation aléatoire des valeurs de chaque colonne*. Chaque colonne ou variable possède les mêmes propriétés de centrage et dispersion mais, dans ce deuxième jeu de données, la structure de corrélation ou de liaison entre les variables est évacuée.

La première classe, données initiales sont les observations normales, la deuxième classe constitue un ensemble d'observations synthétiques d'atypiques ou anomalies par rapport à la distribution des données initiales.

À l'issue de ces simulations, un forêt est apprise sur ces données en cherchant à ajuster au mieux la variable classe ainsi construite. Il en découle comme précédemment la construction d'une mesure de proximité, donc de distance, entre les observations deux à deux. La matrice de distance obtenue est enfin utilisée dans un algorithme de classification ascendante hiérarchique pour l'approche non supervisée issue d'une forêt aléatoire et de positionnement multidimensionnel pour une représentation plane de ces distances.

Anomalies selon les forêts aléatoires Cette même matrice de proximités entre les observations deux à deux est utilisée pour construire le score d'anomalie de chaque observation comme dans la section 3.1. Ce score indique donc pour chaque observation sa plus ou moins grande proximité avec toutes les autres observations de la classe normale.

Comme précédemment, un seuil arbitraire est choisi et les observations atypiques sont projetées dans le premier plan factoriel de l'ACP. Avec des résultats très différents de ceux de la section 3.1.

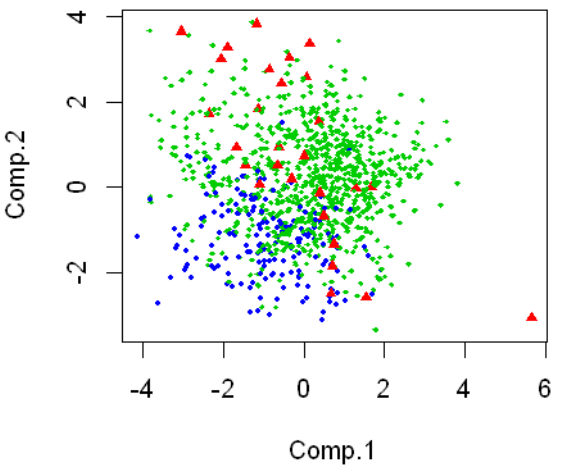

FIGURE 7 – *Ozone : observations atypiques (en rouge) au sens du critère issue d'une forêt aléatoire non supervisée c'est-à-dire sans chercher à modéliser le dépassement de seuil.*

Isolation Forest

La version d'*OCC random forest* précédente n'est disponible que dans l'implémentation historique : librairie R `randomForest` de cet algorithme. Une version suivant un principe complètement différent est proposée par Liu et al. (2008)[9] et disponible dans `Scikit-learn`. Le principe repose sur la constructions d'un ensemble d'arbres complètement aléatoires : *isolation tree*. La division opérée dans chaque nœud est issue du tirage aléatoire d'une variable et d'un seuil aléatoire d'une variable quantitative ou répartition toujours aléatoires des modalités en deux groupes. La construction de l'arbre est poursuivie jusqu'à l'obtention d'une observation par feuille. La quantification de l'isolement ou de l'anomalie d'une observation est obtenue par la longueur du chemin atteignant cette observation. Plus celui-ci est court, plus l'observation est considérée isolée ou atypique.

Liu et al. (2008) proposent de construire B (par défaut 100) arbres d'isolation sur un sous -échantillon aléatoire (par défaut de taille 256) des données puis de calculer, pour chaque observation, la moyenne des longueurs des chemins comme score d'anomalie.

4 Conclusion

La détection d'anomalies est un problème complexe sans solution uniformément meilleure car le choix de la méthode à utiliser dépend largement du contexte, des propriétés des variables et données observées, ainsi que de l'objectif poursuivi. La recherche d'anomalies sur les données de prévision de dépassement de seuil d'ozone montre, sur cet exemple, que chaque méthode projette sa conception de ce qu'est une anomalie. Néanmoins, il est probable et rassurant que les différentes méthodes vont s'accorder sur la détection d'observations très atypiques pas ou peu présentes dans les données de concentration d'ozone. D'autre part, cf. Aggarwal (2017)[1] pour une revue, de nouvelles approches cherchent à conjuguer ou agréger plusieurs méthodes de détection d'anomalies pour conduire à des résultats plus robustes comme en apprentissage avec le *bagging* ou le *boosting*.

Naturellement, la différence est encore plus marquée entre les deux types d'anomalies détectées par les forêts aléatoires ; anomalies par rapport à un modèle expliquant le dépassement ou par rapport à la distribution globales des observations.

En conséquence, même sans mettre en œuvre une approche supervisée, il est important de disposer d'un historique, sinon de simulations, identifiant des anomalies afin de pouvoir rétrospectivement évaluer l'efficacité de la ou des méthodes tout en optimisant la valeur du paramètre de sensibilité. Que l'approche soit paramétrique ou non, ce paramètre est toujours présent.

Cette très courte présentation n'aborde pas les problèmes plus complexes pouvant émerger. La prise en compte de signaux, courbes, images (*autoencoder et deep learning*), chemins sur un graphe, nécessite d'adapter ou sélectionner la bonne base de représentation (Fourier, splines, ondelettes) ou encore la bonne distance entre les observations concernées, avant de mettre en œuvre l'une des méthodes ci-dessus. Un traitement en ligne des données nécessitent la mise en place de décision séquentielle ou adaptative ; autant de sujets de recherche en cours.

Références

[1] Charu C. Aggarwal, *Outlier Analysis*, Springer Publishing Company, Incorporated, 2013, ISBN 1461463955, 9781461463955.

[2] L. Breiman, *Random forests*, Machine Learning **45** (2001), 5–32.

[3] Markus Breunig, Hans Peter Kriegel, Raymond T. Ng et Jörg Sander, *LOF : Identifying Density-Based Local Outliers*, PROCEEDINGS OF THE 2000 ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, ACM, 2000, p. 93–104.

[4] Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek et Jörg Sander, *Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection*, ACM Trans. Knowl. Discov. Data **10** (2015), n° 1, 5 :1–5 :51.

[5] Henri Caussinus et Anne Ruiz-Gazen, *Classification and generalized principal component analysis*, Selected contributions in data analysis and classification (Brito, Bertrand, Cucumel et Carvalho, réds.), Stud. Classification Data Anal. Knowledge Organ., Springer, 2007, p. 539–548, https://hal.archives-ouvertes.fr/hal-00635541.

[6] Martin Ester, Hans Peter Kriegel, Jörg S Sander et Xiaowei Xu, *A density-based algorithm for discovering clusters in large spatial databases with noise*, AAAI Press, 1996, p. 226–231.

[7] Harold Hotelling, *The Generalization of Student's Ratio*, Ann. Math. Statist. **2** (1931), n° 3, 360–378, https://doi.org/10.1214/aoms/1177732979.

[8] Edwin M. Knorr, Raymond T. Ng et Vladimir Tucakov, *Distance-based Outliers : Algorithms and Applications*, The VLDB Journal **8** (2000), n° 3-4, 237–253, ISSN 1066-8888, http://dx.doi.org/10.1007/s007780050006.

[9] F.T. Liu, K. M. Ting et Z. H. Zhou, *Isolation Forest*, Proceedings of the Eighth IEEE International Conference on Data Mining, 2008, p. 413–422.

[10] Sridhar Ramaswamy, Rajeev Rastogi et Kyuseok Shim, *Efficient Algorithms for Mining Outliers from Large Data Sets*, Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (New York, NY, USA), SIGMOD '00, ACM, 2000, p. 427–438, ISBN 1-58113-217-4, http://doi.acm.org/10.1145/342009.335437.

[11] Bernard Rosner, *Percentage Points for a Generalized ESD Many-Outlier Procedure*, Technometrics **25** (1983), n° 2, 165–172.

[12] Bernhard Schölkopf, Robert C. Williamson, Alex J. Smola, John Shawe-Taylor et John C. Platt, *Support Vector Method for Novelty Detection*, Advances in Neural Information Processing Systems 12 (S. A. Solla, T. K. Leen et K. Müller, réds.), MIT Press, 2000, p. 582–588, http://papers.nips.cc/paper/1723-support-vector-method-for-novelty-detection.pdf.