



**HAL**  
open science

## **PANDA : Personnaliser les ANomalies Détectées par Apprentissage**

Grégory Smits, Marie-Jeanne Lesot, Véronne Yepmo, Olivier Pivert

► **To cite this version:**

Grégory Smits, Marie-Jeanne Lesot, Véronne Yepmo, Olivier Pivert. PANDA : Personnaliser les ANomalies Détectées par Apprentissage. LFA 2022 - Rencontres francophones sur la logique floue et ses applications, Oct 2022, Toulouse, France. hal-03777547

**HAL Id: hal-03777547**

**<https://inria.hal.science/hal-03777547>**

Submitted on 14 Sep 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# PANDA : Personnaliser les ANomalies Détectées par Apprentissage

Grégory Smits<sup>1</sup>

Marie-Jeanne Lesot<sup>2</sup>

Véronne Yepmo Tchaghe<sup>1</sup>

Olivier Pivert<sup>1</sup>

<sup>1</sup> Université de Rennes - IRISA {gregory.smits, veronne.yepmo-tchaghe, olivier.pivert}@irisa.fr

<sup>2</sup> Sorbonne Université, CNRS, LIP6 marie-jeanne.lesot@lip6.fr

## Résumé :

La personnalisation par prise en compte de l'expertise de l'utilisateur des processus de détection d'anomalies est le sujet principal de cet article. Les anomalies sont identifiées dans deux espaces de représentation des données, l'espace initial composé des valeurs brutes et l'espace induit par un vocabulaire flou défini par l'utilisateur sur ces attributs initiaux. L'application de toute méthode de détection d'anomalies sur ces deux espaces conduit à différencier des anomalies issues des données brutes de celles issues du vocabulaire. Afin à la fois d'améliorer la détection des exceptions et de disposer d'outils pour les expliquer, ces deux ensembles d'anomalies sont confrontés. Trois situations émanant de cette comparaison sont étudiées afin de montrer comment l'utilisation combinée des deux espaces améliore l'efficacité de la procédure de détection d'anomalies ainsi que l'interprétabilité des résultats qu'elle génère.

## Mots-clés :

détection et explication d'anomalies, approche orientée utilisateur, vocabulaire flou, description linguistique

## Abstract:

The paper addresses the tasks of anomaly detection and explanation simultaneously, in the human-in-the-loop paradigm integrating the end-user expertise : it first proposes to exploit two complementary data representations to identify anomalies, namely the description induced by the raw features and the description induced by a user-defined vocabulary. These representations respectively lead to identify so-called data-driven and knowledge-driven anomalies. The paper then proposes to confront these two sets of instances so as to improve the detection step and to dispose of tools towards anomaly explanations. It distinguishes and discusses three cases, underlining how the two description spaces can benefit from one another, in terms of accuracy and interpretability.

## Keywords:

outlier detection, outlier explanation, human-in-the-loop, fuzzy vocabulary, linguistic description

## 1 Introduction

Une manière d'améliorer l'interprétabilité des méthodes d'Intelligence Artificielle (IA) est d'intégrer l'utilisateur final et son expertise au cœur des procédures de traduction des données en connaissances. Ce travail s'intéresse à la tâche d'identification d'anomalies dans un jeu de données pour lequel l'utilisateur dis-

pose d'un vocabulaire flou décrivant linguistiquement des propriétés observables dans ces données. Un point est considéré comme une anomalie si ses valeurs « dévient tellement des autres observations que l'on puisse suspecter qu'il a été généré par un mécanisme différent » [8]. De nombreuses méthodes existent pour traiter spécifiquement cette tâche (voir par exemple [12]). Cependant, ces approches ne prennent pas en compte l'utilisateur en charge de l'analyse des anomalies extraites. Plus récemment, de nombreux travaux complémentaires proposent des stratégies *a posteriori* pour présenter les anomalies dans un cadre d'IA explicable [15].

Cet article propose, à l'inverse, de considérer l'utilisateur final comme point de départ d'une procédure de détection d'anomalies, afin d'aboutir à une méthode de détection d'anomalies guidée par des connaissances expertes. La prise en compte du vocabulaire de l'utilisateur permet également de décrire linguistiquement les anomalies identifiées.

Pour atteindre cet objectif, l'approche nommée PANDA, acronyme de Personnaliser les ANomalies Détectées par Apprentissage, considère en entrée un jeu de données ainsi qu'un vocabulaire flou subjectif et contextuel. Les termes de ce vocabulaire, dont la définition traduit une forme de connaissance experte sur la façon dont les valeurs brutes doivent être interprétées, sont utilisés pour formuler des descriptions linguistiques des données. Outre le fait d'être associé à une étiquette linguistique, un terme du vocabulaire regroupe différentes valeurs brutes qui ne doivent pas être traitées différemment mais bien au contraire interprétées de la même façon. La prise en compte de relations d'indistinguabilité

dans les données sur la base d'un vocabulaire expert conduit à des comparaisons plus pertinentes des données, du point de vue de l'utilisateur [6].

La méthode PANDA exploite un tel vocabulaire pour construire un second espace de représentation des données, complémentaire à l'espace initial de description des données brutes. Dans ce second espace, les données sont décrites par des vecteurs formés par la concaténation des degrés d'appartenance aux différentes modalités définies sur les attributs. Ces vecteurs constituent une représentation des données guidée par des connaissances expertes, mais apportent également une stratégie de normalisation de valeurs initialement non-commensurables, permettant ainsi la prise en compte de données numériques et catégorielles.

PANDA applique une méthode de détection d'anomalies quelconque dans ces deux espaces de représentation des données, à la manière de la stratégie employée par [9] pour une tâche de classification non supervisée des données. Cette dernière vise à quantifier l'adéquation entre les structures extraites des deux espaces. En revanche, l'analyse conduite par l'approche PANDA confronte les deux ensembles d'anomalies, identifiés dans chacun des deux espaces, afin d'améliorer l'identification globale des anomalies et d'offrir des outils ouvrant la voie à une explication de ces exceptions. PANDA est une stratégie générique, agnostique de la méthode de détection d'anomalie employée, améliorant, par personnalisation et analyse croisée, la pertinence des anomalies détectées dans les deux espaces.

Après avoir présenté quelques travaux connexes dans la section 2 portant sur la détection et l'explication d'anomalies, la section 3 décrit la méthode PANDA et illustre son fonctionnement sur des données synthétiques. Avant de conclure (section 5), la section 4 présente un cas concret d'utilisation sur des données réelles.

## 2 Travaux connexes

Cette section présente quelques approches de détection et d'explication d'anomalies.

### 2.1 Détection d'anomalies

De nombreuses méthodes sont dédiées à l'identification d'anomalies, ces points qui diffèrent des phénomènes réguliers; de nombreuses synthèses ont également été proposées, voir par exemple [12]. Bien qu'il n'existe pas de consensus sur une structuration de ces méthodes, elles sont souvent caractérisées selon qu'elles mettent en œuvre une approche supervisée ou non. Au delà de cet aspect, on peut distinguer les méthodes exploitant le voisinage des anomalies, la structure intrinsèque des données, la densité locale ou des modèles spécifiques de séparation des points [10]. Parmi les méthodes les plus connues, on peut citer LOF (Local Outlying Factor) et ses variantes [3], les forêts d'isolation [10], les one-class SVM [1] et les auto-encodeurs [4].

L'approche PANDA a été expérimentée, et est illustrée dans ce document, sur les Forêts d'Isolation (FI) [10]. Ce choix est motivé par deux raisons : la robustesse des valeurs par défaut des deux hyper-paramètres nécessaires, ainsi que la possibilité de visualiser, au moins sur des données en deux dimensions, le fonctionnement de l'approche. Il s'agit d'une méthode d'apprentissage non supervisé exploitant les informations structurelles sur les données capturées par un ensemble d'arbres d'isolation. Son principe est d'effectuer des coupes récursives aléatoires afin de construire des sous-espaces dont les bornes sont représentées dans les nœuds des arbres. Par définition, les anomalies sont des points distants des points dits réguliers, donc plus facilement séparables du reste des données. Les anomalies correspondent aux points isolés le plus rapidement, se trouvant ainsi dans des feuilles proches des racines. Un score d'anomalie est alors calculé, pour chaque point, en fonction de la profondeur moyenne,

au sein de la forêt, des feuilles auxquelles il est affecté.

## 2.2 Explication d’anomalies

À partir d’un ensemble de points identifiés comme des anomalies, une attente légitime d’un utilisateur final est de connaître les raisons du caractère anormal de ces points. Fournir des explications en réponse à cette attente est le rôle des méthodes dites d’explication d’anomalies. Selon une récente analyse [15], quatre catégories d’approches peuvent être distinguées selon la nature des explications fournies. La première, la plus répandue, vise à quantifier l’importance des différents attributs, voir par exemple [11], afin d’expliquer sur quelle(s) dimension(s) les points possèdent des valeurs suspectes. Cette famille de méthodes peut également être séparée en sous-groupes selon que les anomalies sont expliquées localement, point par point, ou globalement, par groupes, ou selon qu’elles sont génériques ou spécifiques à certaines méthodes de détection d’anomalies [5]. La seconde famille génère des explications plus précises en décrivant les sous-espaces dans lesquels se trouvent les anomalies [2]. La troisième catégorie associe à chaque anomalie un ou plusieurs exemples de points contre-factuels dont la comparaison permet de comprendre les valeurs qui caractérisent l’anomalie [7]. Finalement, la quatrième famille fournit des explications plus riches en comparant les caractéristiques des anomalies avec celles des groupes de données régulières [13].

Cependant, il ne semble pas exister de méthode de détection et d’explication d’anomalies qui s’appuie sur des connaissances expertes pour améliorer la compréhension que l’utilisateur a des données.

## 3 L’approche PANDA

Après avoir introduit les notations utilisées, cette section décrit les deux étapes de l’approche PANDA : l’application d’une méthode

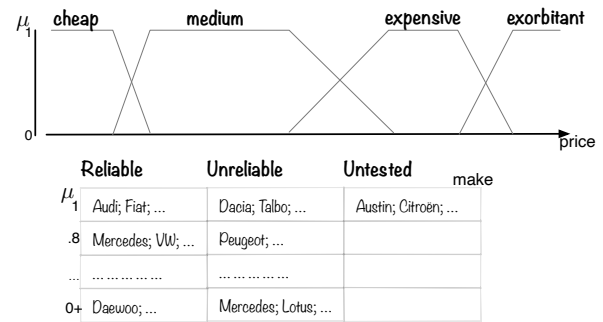


FIGURE 1 – Exemples de partitions floues, pour attribut numérique (en haut) et catégoriel (en bas).

de détection d’anomalies sur deux espaces de représentation des données puis l’analyse contrastive des anomalies trouvées.

### 3.1 Notations et données illustratives

$\mathcal{D} = \{x_1, x_2, \dots, x_n\}$  est un ensemble de  $n$  points décrits sur  $m$  attributs, notés  $A_1$  à  $A_m$ ,  $D_1$  à  $D_m$  étant leurs domaines de définition respectifs.  $\mathcal{V} = \{P_1, \dots, P_m\}$  représente un vocabulaire défini comme un ensemble de variables linguistiques : pour  $i = 1..m$ ,  $P_i$  est un triplet  $\langle A_i, \{\mu_i\}, \{l_i\} \rangle$  avec  $q_i$  modalités.

Les  $\mu_{ij}$ ,  $j = 1..q_i$ , sont les fonctions d’appartenance associées aux modalités définies sur l’univers  $D_i$  et les  $l_{ij}$  sont leurs étiquettes linguistiques. La figure 1 illustre deux exemples de partitions floues : en haut pour un attribut numérique décrivant le prix d’une voiture d’occasion avec  $q = 4$  modalités ayant pour étiquettes ‘cheap’, ‘medium’, ‘expensive’ et ‘exorbitant’. La partie basse de la figure décrit une interprétation possible des marques selon leur fiabilité décrite par trois modalités ‘reliable’, ‘unreliable’ et ‘untested’. La fonction d’appartenance est alors définie sous forme d’ $\alpha$ -coupes.

Chaque  $P_i$  forme une partition forte, i.e.  $\forall y \in D_i, \sum_{j=1}^{q_j} \mu_{ij}(y) = 1$ . Il est en effet imposé que chaque valeur ne puisse pas satisfaire plus de deux modalités (qui doivent être adjacentes si la valeur est numérique).

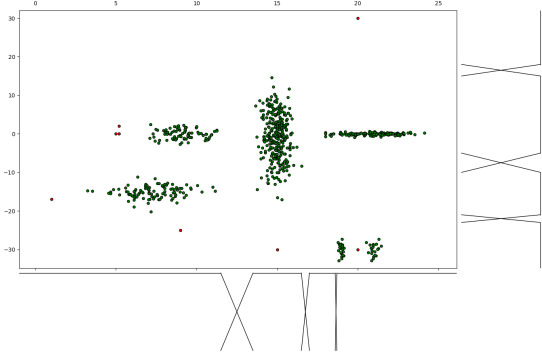


FIGURE 2 – Jeu de données illustratif

Dans cette section, nous considérons comme exemple illustratif les  $n = 651$  points de la figure 2 décrits par deux attributs numériques  $m = 2$  :  $A_1$  et  $A_2$  de domaines respectifs  $D_1 = [0, 25]$  et  $D_2 = [-40, 30]$ .  $P_1$  contient  $q_1 = 4$  modalités dont les fonctions d'appartenance sont indiquées en bas de la figure.  $P_2$  inclut également quatre modalités illustrées sur la droite de la figure. Le jeu de données contient plusieurs zones denses et des points isolés.

### 3.2 Données et espaces de représentation

**Réécriture des données initiales.** Chaque point est tout d'abord réécrit en un vecteur contenant les degrés d'appartenance aux différentes modalités des différents attributs. Un tel vecteur composé de  $Q = \sum_{j=1}^m q_j$  degrés a la forme suivante  $\langle \mu_{v_{11}}(x^1), \dots, \mu_{v_{1q_1}}(x^1), \dots, \mu_{v_{m1}}(x^m), \dots, \mu_{v_{mq_m}}(x^m) \rangle$ . En raison des contraintes imposées sur les partitions, ce vecteur contient au plus  $2m$  valeurs non nulles. Les données  $\mathcal{D}$  sont ainsi réécrites en  $\mathcal{D}^\mathcal{V}$ , dans un nouvel espace défini par le vocabulaire  $\mathcal{V}$ , en temps constant par rapport à sa taille  $|\mathcal{D}|$ , cette réécriture pouvant être distribuée [14].

**Détection d'anomalies dans les deux espaces.** Afin de tirer parti des connaissances dont dispose l'expert sur les données, connaissances formalisées sous forme du vocabulaire, l'approche PANDA repose sur l'application d'une même méthode de détection d'anomalies à la fois sur  $\mathcal{D}$  et sur  $\mathcal{D}^\mathcal{V}$ . Les ensembles  $\mathcal{A}$  et  $\mathcal{A}^\mathcal{V}$  désignent

respectivement les anomalies découvertes dans l'espace initial de définition des données,  $\mathcal{D}$ , et dans l'espace réécrit,  $\mathcal{D}^\mathcal{V}$ .

### 3.3 Analyse croisée des anomalies détectées

L'analyse des anomalies consiste en une comparaison des ensembles  $\mathcal{A}$  et  $\mathcal{A}^\mathcal{V}$ , sur la base de leur intersection et de leurs différences. L'objectif de cette comparaison est d'aider l'utilisateur à mieux comprendre les données et son vocabulaire. Nous montrons que cette comparaison permet d'identifier des anomalies issues des données initiales qui correspondent à des régularités contextuelles. Inversement des points qui paraissent réguliers dans l'espace initial peuvent être transformés, par exploitation du vocabulaire de l'utilisateur, en anomalies contextuelles. Le vocabulaire sert également d'outil pour décrire ces cas particuliers.

La figure 3 illustre deux exemples d'arbres d'isolation construits sur  $\mathcal{D}$  (à gauche) et sur  $\mathcal{D}^\mathcal{V}$  (à droite). Les niveaux de bleu indiquent les scores d'anomalies, plus élevés dans les zones claires. Les lignes noires sont les séparatrices qui forment les nœuds des arbres d'isolation. La première observation est que le profil des scores d'anomalie varie fortement d'un espace à l'autre. Guidées par les modalités du vocabulaire, les zones d'isolation construites dans l'espace  $\mathcal{D}^\mathcal{V}$  (à droite) forment des hyper-rectangles parallèles aux axes : la relation d'indistinguabilité introduite par les modalités floues conduit à des zones de même score au sein de chaque  $\alpha$ -coupe. Considérés comme indistinguables, les points satisfaisant une modalité à un même degré ne peuvent ainsi pas être séparés. Les scores d'anomalie calculés dans l'espace initial  $\mathcal{D}$  varient évidemment davantage en fonction de la densité locale des données.

**Description linguistique des anomalies :  $\mathcal{A} \cap \mathcal{A}^\mathcal{V}$ .** Le premier cas analysé correspond aux anomalies identifiées comme telles dans les deux espaces : ces points sont décrits comme des anomalies confirmées pour lesquelles l'approche PANDA

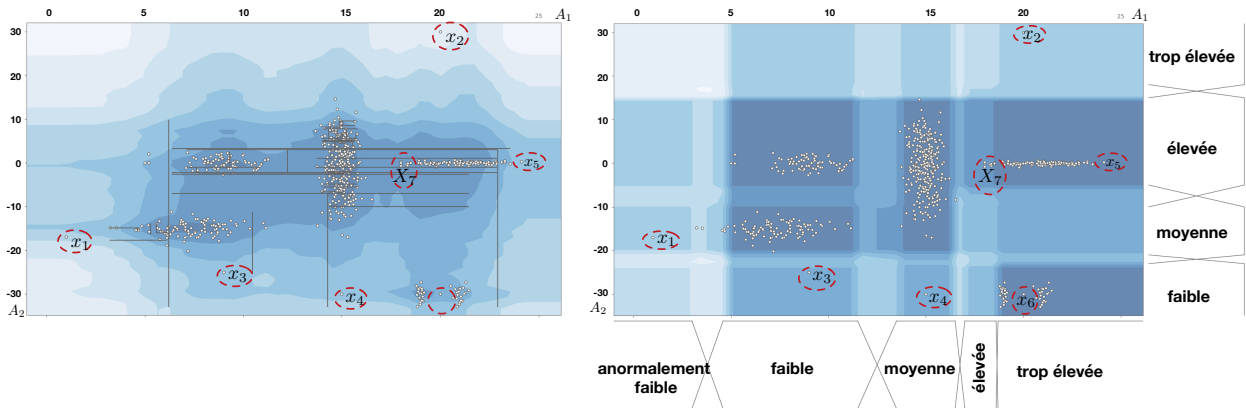


FIGURE 3 – Profil des scores d’anomalie dans les deux espaces :  $\mathcal{D}$  à gauche,  $\mathcal{D}^v$  à droite.

fournit en plus une description linguistique. En effet, un point  $x \in \mathcal{A} \cap \mathcal{A}^v$  correspond à une anomalie identifiée dans l’espace initial pour lequel la description linguistique selon le vocabulaire de l’utilisateur est également rare. La réécriture de ce point  $x$  selon les termes du vocabulaire apporte une description subjective et contextuelle de l’anomalie.

Ce cas est illustré par les points  $x_1$  de coordonnées (1, -17),  $x_2$  (20, 30),  $x_3$  (9, -25) et  $x_4$  (15, -30) dans la figure 3. Ils sont en effet isolés rapidement par la forêt d’isolation dans les deux espaces. Deux situations sont à différencier :  $x_1$  et  $x_2$  sont des anomalies car ils possèdent une valeur extrême et rare, alors que  $x_3$  et  $x_4$  possèdent une combinaison rare de valeurs par rapport aux autres points. Le vocabulaire peut être exploité pour décrire ces situations :  $x_1$  est décrit de la façon suivante “la valeur en  $A_1$  est anormalement faible et la valeur en  $A_2$  est moyenne’ alors que  $x_3$  est décrit de la façon suivante “la valeur en  $A_1$  est faible et la valeur en  $A_2$  est faible”.

Ces descriptions linguistiques peuvent être considérées comme un point de départ pour une démarche d’explication interprétable des anomalies. Cependant, il faut souligner que l’intégralité des dimensions est considérée pour générer la description linguistique, ce qui peut conduire à des expressions trop complexes en présence d’un grand nombre de dimensions. Une solution serait d’appliquer une étape

supplémentaire d’analyse des dimensions pertinentes pour caractériser chaque anomalie (voir section 2.2). L’approche PANDA génère une description linguistique des anomalies plutôt qu’une explication de leur provenance.

**Anomalies inattendues :  $\mathcal{A} \setminus \mathcal{A}^v$ .** Certains points peuvent être considérés comme des anomalies dans l’espace initial de définition des données sans pour autant correspondre à des anomalies une fois le vocabulaire pris en compte. Cette situation, qui fait apparaître des anomalies pouvant être considérées comme inattendues selon les connaissances de l’utilisateur, est illustrée par les points  $x_5$  (24, 0) et  $x_6$  (20, -30) sur la figure 3. Leur score d’anomalie est en effet très faible dans  $\mathcal{D}^v$  car leur réécriture sur les deux dimensions est partagée, de manière indistinguable, avec de nombreux autres points qui sont pourtant distants dans l’espace numérique.

Ces anomalies issues de l’espace initial sont dites inattendues pour l’utilisateur, car son vocabulaire ne comprend pas de modalités permettant de les décrire en les différenciant des autres points. Cette situation de désaccord peut s’avérer très informative dans la mesure où elle permet d’attirer l’attention de l’utilisateur sur des zones précises de l’espace initial de définition des données pouvant conduire à des révisions du vocabulaire.

En effet, ces anomalies inattendues peuvent amener l’utilisateur à revoir l’étendue d’une

zone d'indistinguabilité, notamment en ajoutant une nouvelle modalité dédiée à la description de la zone dans laquelle des anomalies ont été trouvées. Ainsi  $x_5$  constitue un exemple d'anomalie « numérique » pour laquelle il peut être suggéré à l'utilisateur d'analyser la définition de la modalité *très élevée* sur l'attribut  $A_1$ . Ce cas d'inadéquation entre les deux espaces, pour une tâche d'identification et de description des anomalies, complète l'approche [9] qui ne prend pas en compte la description de sous-espaces de faible densité.

Cependant, réciproquement, cette situation peut mettre en lumière la nécessité de disposer de connaissances expertes pour comprendre que certaines valeurs, pourtant rares, ne doivent pas être considérées comme des anomalies : outre le caractère contextuel de ces connaissances, le vocabulaire peut également pallier l'absence de données d'apprentissage dans un sous-espace. L'interprétation subjective des domaines, à travers l'introduction d'une relation d'indistinguabilité, permet de connecter ces points apparaissant comme séparés numériquement à un groupe dense de points. Ainsi, la modalité *très élevée* sur  $A_2$  permet de raccrocher  $x_5$ , malgré sa valeur élevée sur  $A_2$ , au groupe de points de valeurs inférieures. Cette même modalité permet d'indiquer que le point  $x_6$  n'est pas à dissocier des deux groupes qui l'encadrent.

**Vocabulaire non approprié :**  $\mathcal{A}^V \setminus \mathcal{A}$ . Symétriquement au cas  $\mathcal{A} \setminus \mathcal{A}^V$ , une troisième situation pouvant résulter de la comparaison des deux ensembles d'anomalies correspond à des anomalies identifiées dans l'espace réécrit selon le vocabulaire mais pas considérées comme telles dans l'espace initial. Ce cas est par exemple illustré par l'ensemble de points  $X_7$  sur la figure 3, au voisinage du point (17, 0), décrit par la modalité *élevée* de l'attribut  $A_1$ .

Cette situation peut être vue comme un autre type d'inadéquation entre le vocabulaire et la distribution des données dans leur espace initial de définition. Dans de tels cas, la coupure entre deux modalités adjacentes, ici *élevée* et

*très élevée* pour l'attribut  $A_2$ , peut sembler inappropriée dans la mesure où elle scinde un ensemble dense de points. Indiquer à l'utilisateur cette possible inadéquation peut conduire à une fusion des modalités concernées ou à la redéfinition de leurs bornes.

Cependant, le vocabulaire étant une formalisation des connaissances dont l'expert dispose [6], cette situation peut également être analysée sous un autre angle : ces points considérés comme réguliers dans  $\mathcal{D}$  mais anormaux dans  $\mathcal{D}^V$  peuvent correspondre à des points dont l'identification comme anomalies requiert des connaissances supplémentaires sur le contexte applicatif, allant au delà de la distribution du jeu de données disponible. A titre d'exemple, considérons un ensemble de mesures de la température d'un moteur à explosion dont la valeur idéale est aux alentours de 90°C. Capter quelques températures basses dans l'intervalle [60, 91] pendant la courte période de chauffe ne doit pas conduire à la levée d'alertes, alors que l'observation d'une température à 95°C traduit un comportement anormal. Une solution pour différencier ces plages de valeurs difficilement observables et peu fréquentes est d'avoir recours à un vocabulaire spécifique.

## 4 Exemple de cas d'utilisation

Cette section illustre, sur des données et un cas d'usage réel concernant des annonces de voitures d'occasion à vendre, comment l'approche PANDA permet de mieux comprendre les données et d'influencer une procédure de détection d'anomalies : PANDA permet, à l'aide de connaissances expertes, de distinguer des valeurs erronées, car aberrantes, d'annonces exceptionnelles comme des voitures rares.

### 4.1 Données et connaissances expertes

Le jeu de données utilisé contient 49188 annonces de voitures d'occasion à vendre décrites sur six attributs *price*, *mileage*, *year*, *priceNew*, *make* et *model*. L'attribut *priceNew* correspond

TABLEAU 1 – Termes du vocabulaire sur les voitures d’occasion

Attr.	Étiquettes linguistiques
Price	almostOffered, veryLow, low, medium, expensive, veryExpensive, exorbitant
Mileage	almostNull, veryLow, low, medium, high, veryHigh, huge
Year	vintage, old, acceptable, recent, almostNew
PriceNew	veryLow, low, medium, expensive, veryExpensive, exorbitant
Make	luxury, highClass, mediumClass, lowClass

au prix de vente du véhicule neuf. Le tableau 1 indique uniquement les étiquettes associées aux variables linguistiques définies sur les cinq premiers attributs. Les fonctions d’appartenance associées correspondent à des définitions consensuelles des propriétés modélisées.

L’approche PANDA a été appliquée aux annonces qui forment l’ensemble  $\mathcal{D}$  ainsi que sur leur version réécrite  $\mathcal{D}^v$ . La méthode de détection d’anomalies par forêt d’isolation est utilisée ici avec les valeurs par défaut proposées dans [10] pour les hyper-paramètres : 100 arbres, 256 points pour construire chaque arbre et 0,5 pour le seuil d’anomalie.

## 4.2 Analyse des résultats

Afin de comprendre l’impact de la prise en compte du vocabulaire sur les données, le tableau 2 donne les cinq annonces de score d’anomalie maximal dans  $\mathcal{D}$  et  $\mathcal{D}^v$  respectivement. Les valeurs aberrantes sont indiquées en gras. Il est à noter que ces deux ensembles d’anomalies à comparer ont une intersection vide, les anomalies en commun arrivant au-delà du top-5. Cette section en explique la raison et détaille les deux autres situations.

Les anomalies identifiées dans  $\mathcal{A}$  peuvent être divisées en deux groupes : les annonces erronées qui contiennent des valeurs aberrantes (annonces 1, 4, 5), et les annonces non-erronées décrivant des voitures rares possédant des valeurs atypiques. Ce second groupe est composé de voitures de sport luxueuses et de voi-

tures de collection dont le prix reste très élevé malgré le nombre d’années ou de kilomètres (annonces 2, 3).

L’application du même algorithme de détection d’anomalies aux données réécrites conduit à un résultat très différent. L’existence de modalités pour décrire des marques de luxe (*luxury*) et des prix très élevés (*very expensive*) conduit au regroupement des voitures de sport et de collection, les rendant ainsi moins isolées et associées à des scores d’anomalie moins élevés.

Les annonces considérées comme les plus anormales une fois réécrites selon le vocabulaire sont de deux types. Il s’agit soit d’erreurs de saisie conduisant à des valeurs aberrantes, comme un kilométrage de 1 pour une voiture *vintage* (annonce 3), soit de valeurs individuellement fréquentes mais qui, une fois combinées sont invraisemblables (annonce 9). Les détails de l’annonce révèlent un moteur cassé, d’où un prix anormalement bas.

## 5 Conclusion et perspectives

L’approche PANDA permet d’intégrer les connaissances dont dispose un utilisateur sur les données analysées dans un processus d’identification d’anomalies. Outre le fait d’éviter que des propriétés rares mais légitimes, pour un contexte applicatif donné, soient considérées comme des anomalies, le vocabulaire utilisateur peut être utilisé comme un outil de description des anomalies, fournissant des outils pour leur interprétation et compréhension.

Les travaux en cours visent à compléter PANDA, pour la génération de descriptions focalisées sur les attributs les plus pertinents et mettant en parallèle données régulières et anomalies. Les outils ainsi développés, et intégrés dans des systèmes de gestion de bases de données, feront l’objet d’expériences avec des données et des utilisateurs réels, pour valider leur contribution à la compréhension et la satisfaction des utilisateurs lors de l’exploration de données massives.



TABLEAU 2 – Top-5 des anomalies dans  $\mathcal{A}$  en haut et dans  $\mathcal{A}^y$  en bas

	Price	Mileage	Year	PriceNew	Make	Model	Score
1	<b>7,500,000</b>	112,000	1993	98,754	mercedes	500 SL A	0.705
2	<b>110,000</b>	15,000	<b>1984</b>	80,570	ferrari	BB 512 5	0.696
3	<b>62,000</b>	<b>50,000</b>	<b>1992</b>	168,174	ferrari	F 512 4.9i	0.69
4	<b>42,600,000</b>	22,000	2010	44,020	mercedes	Classe C 350 CDI	0.688
5	<b>17,490,000</b>	202,000	2005	54,440	mercedes	Classe CLS 320 CDI	0.682

	Price	Mileage	Year	PriceNew	Make	Model	Score
1	<b>450</b>	<b>100</b>	<b>1988</b>	8,232	renault	Super 5 Tiga	0.609
2	<b>850</b>	229,000	1983	<b>8,345</b>	bmw	315	0.604
3	<b>25</b>	<b>1</b>	2010	26,798	audi	A3 Sportback 2.0 TDI	0.602
4	<b>2,350</b>	<b>4,801</b>	2009	9,639	dacia	Sandero 1.5 dCi 70	0.599
5	<b>2,990</b>	290,000	1988	<b>76,441</b>	bmw	750 iL	0.596

## Références

- [1] Amer, M., Goldstein, M., Abdennadher, S. : Enhancing one-class support vector machines for unsupervised anomaly detection. In : Proc. of the ACM SIGKDD Workshop on Outlier Detection and Description. pp. 8–15 (2013)
- [2] Barbado, A., Corcho, O., Benjamins, R. : Rule extraction in unsupervised anomaly detection for model explainability : Application to OneClass SVM. Expert Systems with Applications 189 (2022)
- [3] Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J. : LOF : identifying density-based local outliers. ACM SIGMOD record 29(2), 94–104 (2000)
- [4] Chen, J., Sathe, S., Aggarwal, D., Turaga, D. : Outlier detection with autoencoder ensembles. In : Proc. of the SIAM Int. Conf. on Data Mining. pp. 90–98 (2017)
- [5] Duan, L., Tang, G., Pei, J., Bailey, J., Campbell, A., Tang, C. : Mining outlying aspects on numeric data. Data Mining and Knowledge Discovery 29, 116–1151 (2014)
- [6] Guillaume, S., Charnomordic, B., Loisel, P. : Fuzzy partitions : a way to integrate expert knowledge into distance calculations. Information sciences 245, 76–95 (2013)
- [7] Haldar, S., Johnand, P.G., Saha, D. : Reliable counterfactual explanations for autoencoder based anomalies. In : Proc. of the 8th ACM IKDD CODS and 26th COMAD Conf. pp. 83–91. ACM (2021)
- [8] Hawkins, D.M. : Identification of outliers, Springer (1980)
- [9] Lesot, M.-J., Smits, G., Pivert, O. : Adequacy of a user-defined vocabulary to the data structure. In : Proc. of the IEEE Int. Conf. on Fuzzy Systems. IEEE (2013)
- [10] Liu, F.T., Ting, K.M., Zhou, Z.H. : Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD) 6(1), 3 (2012)
- [11] Myrtakis, N., Tsamardinos, I., Christophides, V. : Proteus : Predictive explanation of anomalies. In : Proc. of the 37th IEEE Int. Conf. on Data Engineering (ICDE). pp. 1967–1972. IEEE (2021)
- [12] Ruff, L., Kauffmann, J., Vandermeulen, R., Montavon, G., Samek, W., Kloft, M., Dietterich, T., Müller, K.R. : A unifying review of deep and shallow anomaly detection. Proc. of the IEEE 109(5), 756–795 (2021)
- [13] Shukla, A.K., Smits, G., Pivert, O., Lesot, M.-J. : Explaining data regularities and anomalies. In : Proc. of the Int. Conf. on Fuzzy Systems. IEEE (2020)
- [14] Smits, G., Pivert, O., Yager, R.R., Nerzic, P. : A soft computing approach to big data summarization. Fuzzy Sets and Systems 348, 4–20 (2018)
- [15] Tchaghe, V.Y., Smits, G., Pivert, O. : Anomaly explanation : A review. Data & Knowledge Engineering 137 (2021)