

Humanités causales

Analyse et causalités sur des données de qualité de
vie au travail

Diviyan Kalainathan

Encadré par M. Sébag, P. Caillou, I. Guyon,
P. Tubaro

INRIA - TAO (05/2016 - 11/2016) - Stage PFE-Master

Copyright © 2016 Diviyan Kalainathan

STAGE DE FIN D'ÉTUDES, ISAE-ENSMA

GITHUB.COM/DIVIYAN-KALAINATHAN/CAUSAL-HUMANS

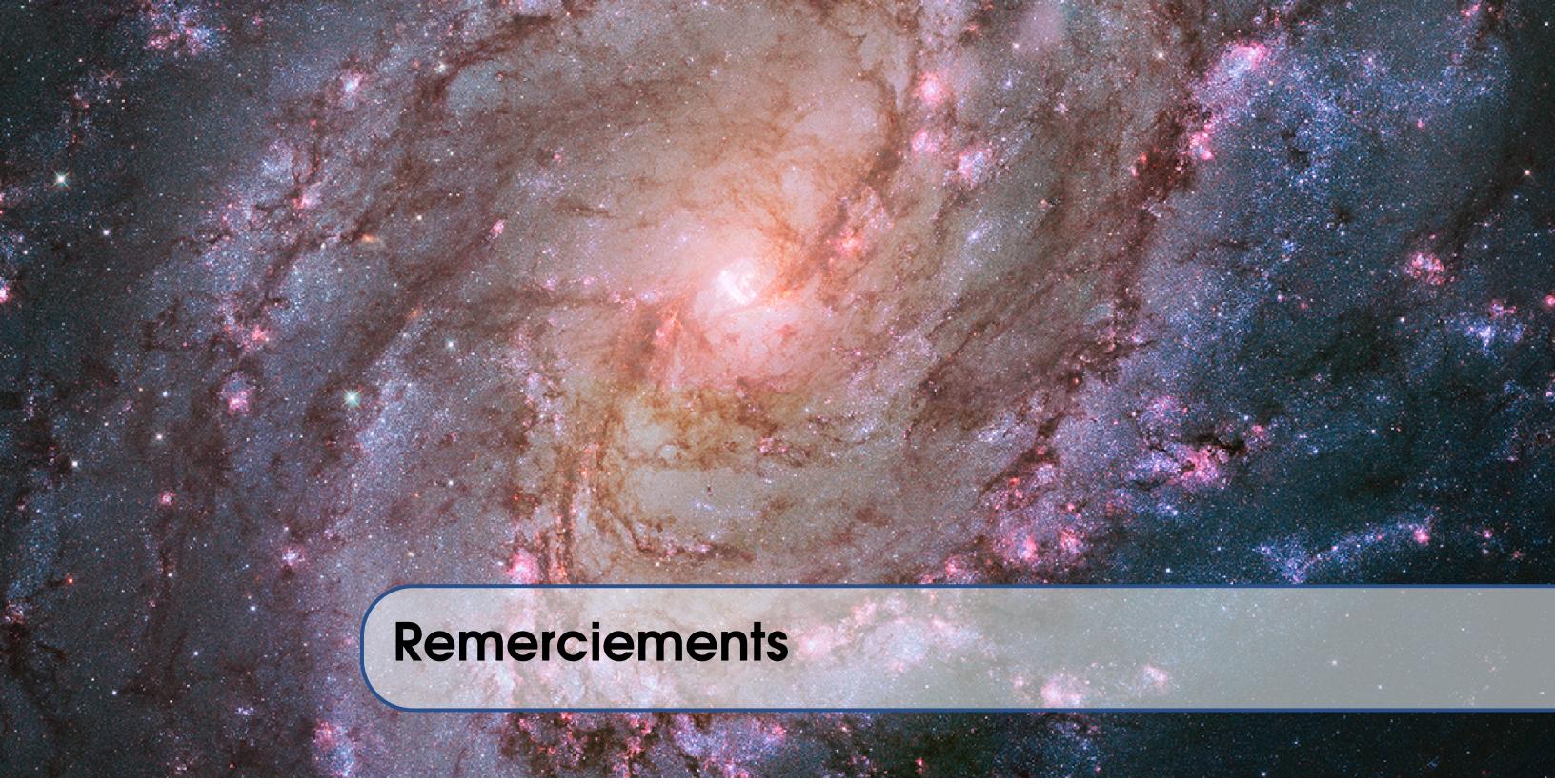
Ce travail de recherche a été effectué sous la supervision des chercheurs Michèle Sebag, Phillippe Caillou, Isabelle Guyon et Paola Tubaro, avec la collaboration d'Olivier Goudet au sein de l'équipe TAO, ainsi qu'avec l'aide de la DARES et le support de l'INRIA pour un stage de 27 semaines, du 17 Mai au 15 Novembre 2016.



Sommaire

Remerciements	5
1 Introduction	6
1.1 État de l'art	7
1.2 Présentation des données	7
1.3 Choix méthodologiques	7
1.4 Démarche	8
1.4.1 Clustering et analyses	8
1.4.2 Causalité	9
2 Analyse descriptive des données	10
2.1 Analyse en composantes principales	10
2.1.1 Principe	10
2.1.2 Obtention des nouveaux axes	11
2.1.3 Étude des axes	11
2.1.4 Interprétation des axes	12
2.2 Clusters et analyses	15
2.2.1 Méthode employée	15
2.2.2 Clusters objectifs	16
2.2.3 Clusters subjectifs	19
2.3 Correspondance entre clusters objectifs et subjectifs	22
2.3.1 Croisement des populations de clusters	22
2.3.2 Autonomie et clusters	23

3	Analyse causale	26
3.1	Méthodologie	26
3.1.1	Hétérogénéité des données	26
3.1.2	Nécessité d'une déconvolution	27
3.1.3	Coefficient de causalité	28
	Bibliographie	29
	Table des figures	30
	Liste des tableaux	31



Remerciements



1. Introduction

La qualité de vie au travail est un aspect de la vie en société qui est peu quantifiable, mais qui tout de même n'a pas cessé de croître en France au cours de ces dernières décennies, contrairement à d'autres pays qui favorisent la production aux dépends des conditions de travail. Toutefois, les sociologues se sont souvent posé la question du lien entre cette qualité de vie au travail (QVT) et de la satisfaction d'un employé : comment évolue la satisfaction au travail avec la qualité de vie ? Sont-ils liés ? Comment l'amélioration de la qualité de vie au travail va-t-elle impacter l'environnement de travail de ses employés ? C'est à ces questions que nous souhaitons répondre au cours de ce projet de fin d'études.

Pour y parvenir, nous disposons d'une quantité importante de données : les réponses de 33673 personnes sur un questionnaire effectué par la DARES¹ en collaboration avec l'INSEE², portant sur divers aspect de la vie des enquêtés dans le cadre de l'enquête conditions de travail 2013. Nous allons donc étudier ces données sous différents angles afin de pouvoir en tirer des interprétations sur les raisons derrière les différences entre la satisfaction au travail des enquêtés et leur situation.

Ce sujet de projet de fin d'études peut paraître être plus une étude de sociologie qu'une étude d'ingénieur en traitement et analyse de données. Toutefois, la méthode et l'angle d'approche du problème correspondent à celles utilisées en ingénierie des données . En effet, le but est de tirer des conclusions de l'analyse des données, c'est-à-dire les réponses au questionnaire. A l'aide de ces réponses et l'analyse de celles-ci par le biais de techniques et d'algorithmes issus de la recherche en informatique, il s'agit de non seulement interpréter les résultats pour obtenir des éléments de réponse à notre problématique et constituer une méthodologie opérationnelle pour analyser l'impact de la QVT sur la performance de l'entreprise, mais aussi permettre de tester la validité des techniques sur un ensemble de données avec les connaissances des sociologues.

1. Direction de l'animation de la recherche, des études et des statistiques du Ministère du travail, de l'emploi, de la formation professionnelle et du dialogue social.

2. Institut national de la statistique et des études économiques collecte, produit, analyse et diffuse des informations sur l'économie et la société françaises

1.1 État de l'art

Les études de données de la Dares faites jusqu'à nos jours, sont non seulement des analyses descriptives,

1.2 Présentation des données

Comme mentionné à la section précédente, nous avons accès aux réponses de 33673 personnes sur un questionnaire de 520 questions. Dans cette ensemble d'enquêtés, on ne considère que les actifs, c'est-à-dire ceux occupant un emploi à la date du questionnaire, ce qui nous laisse 31112 enquêtés. Les questions portant sur les aspects de la vie de l'enquêté, sont regroupées en 7 rubriques (Table 1.1). Notons de plus que les données issues du questionnaire sont hétérogènes : les enquêtés avaient le choix de refuser de répondre à une question, ou de dire qu'ils ne connaissaient pas la réponse ; mais la principale source d'hétérogénéité est la nature du questionnaire lui-même, qui est un questionnaire à multiples branchements. Par ailleurs, la nature des réponses aux questions peut varier. En effet, on peut demander au questionné soit une réponse numérique (p. ex. *Quel est le montant de votre revenu ?*), soit une réponse parmi un choix de réponses multiples (QCM). Les données comportent aussi une quantité importante de données manquantes (21%), dû au fait que le questionnaire est un questionnaire à branches.

- 1. Activité professionnelle/Statut
- 2. Organisation du temps de travail
- 3. Contraintes physiques, prévention et accidents
- 4. Organisation du travail
- 5. Santé
- 6. Parcours familial et professionnel
- 7. Auto-questionnaire sur les risques psychosociaux

TABLE 1.1: Catégories des questions du questionnaire de la Dares

1.3 Choix méthodologiques

Dans un premier temps, il s'agit de recoder les variables, pour éviter les biais quantitatifs, notamment sur les variables catégorielles (issues des questions à choix multiples)³. Une question comprenant N options est ainsi représentée par $N + 1$ variables booléennes (la dernière permettant de caractériser les cas où l'enquêté n'a pas pu ou pas voulu répondre à la question⁴). Dans le cas des questions à réponse continue (e.g. ancienneté ou salaire), celles-ci sont représentées par une variable continue et une variable booléenne, cette dernière codant la non-réponse de l'enquêté, afin de prendre en compte les valeurs manquantes dans les données.

Indépendamment de leur nature, catégorielle ou continue⁵, nous avons choisi de partitionner les questions en deux groupes, correspondant respectivement aux éléments factuels (questions objectives)

3. Une option codée '3' (= pays du Maghreb) ne vaut ni plus, ni moins, qu'une option codée '4' (= Extrême Orient). Les options sont ainsi codées par des variables booléennes X_{opt} , prenant la valeur vrai si la variable X prend la valeur opt et faux sinon.

4. Formellement, les options de fiabilité, définissant des variables drapeaux, prennent 4 valeurs : Réponse (1), Sans objet (0, par exemple si la question n'a pas été posée), Ne sait pas (-1) et Refuse de se prononcer (-2).

5. Cet aspect n'a été pris en compte que dans le pré-traitement, pour recoder les données.

et au ressenti des personnes (questions subjectives). La nature objective ou subjective d'une question dépend principalement de sa formulation. Par exemple "Pensez-vous que", ou "À votre avis" sont des marqueurs de questions subjectives. D'autres questions peuvent être plus ambiguës ; par exemple "Êtes-vous-obligés de vous dépêcher ?" a été classée dans les questions subjectives parce qu'elle fait intervenir le ressenti de l'enquêté (la question peut être reformulée en "L'enquêté se sent-il obligé de se dépêcher ?"). Environ 20% des questions sont considérées comme subjectives ; leur répartition en fonction des rubriques est indiquée Figure 1.1.

Cette distinction constitue à notre connaissance l'un des points originaux de la méthodologie proposée ; elle est motivée par le fait que la notion de QVT dépend clairement à la fois d'éléments factuels (les variables objectives) et de leur ressenti (les variables subjectives). Cette méthodologie nous permet d'analyser indépendamment les deux blocs de données (objectives et subjectives) avant d'examiner les liens entre les situations objectives et leur ressenti.

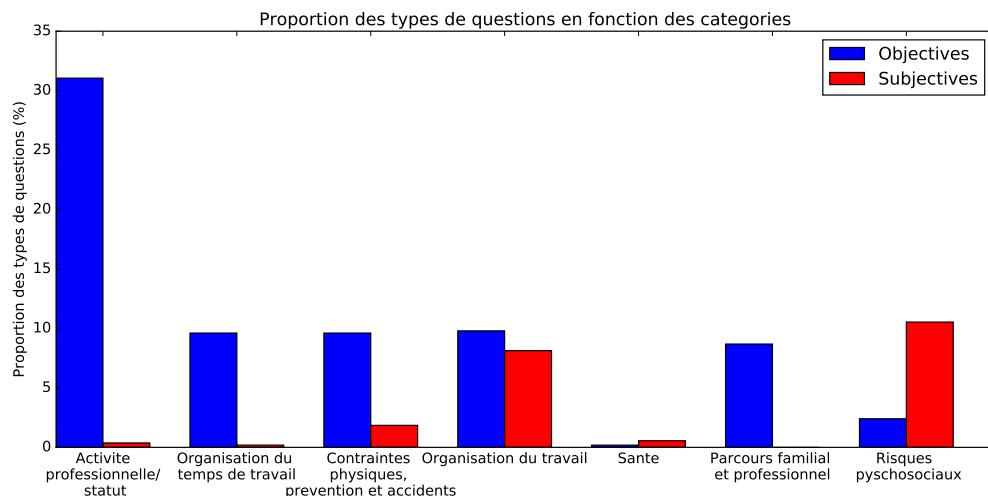


FIGURE 1.1: Répartition des types de questions en fonction des catégories

1.4 Démarche

Après cette phase de prétraitement, l'étude est divisée en deux : dans un premier temps on effectuera une analyse descriptive des données poussée, avant d'effectuer une analyse causale, où nous chercherons à déterminer les relations d'implication entre les variables du questionnaire.

1.4.1 Clustering et analyses

Une analyse en composantes principales permet de remédier à la redondance des variables, pour définir un petit nombre d'axes (variables agrégées, définie par une somme pondérée des variables initiales) capturant la variabilité des données. L'interprétation d'un axe se fait en considérant les variables initiales les plus importantes (valeurs absolues des poids les plus élevés).

Dans l'espace des axes, chaque individu est un vecteur de \mathbb{R}^d . On utilise la catégorisation (clustering) pour identifier les sous-groupes de données homogènes ; l'algorithme employé est un K-means++ [Arthur et Vassilvitskii, 2007]. Avant d'analyser les clusters, on s'assure de leur stabilité selon les critères définis par [Meilă, 2006].

Chaque cluster est interprété par ses variables significatives au sens du v-test [Lebart *et al.*, 2006] ; formellement, une variable est significative pour un cluster lorsque sa valeur moyenne sur ce cluster est significativement distincte de la valeur moyenne sur l'ensemble des données (compte tenu de la taille du cluster). Après avoir établi des clusters sur les variables de situation et de ressenti (respectivement sur les variables objectives et subjectives), il s'agit d'analyser comment évoluent les groupes à travers des variables choisies, telles que le revenu ou le score de bien-être défini par l'OMS⁶ ; mais aussi étudier le croisement des populations entre les clusters de situation et de ressenti est une analyse qui peut s'avérer intéressante.

1.4.2 Causalité

La deuxième partie de l'étude consiste à analyser la causalité au sens de [Granger, 1969] ; la causalité inclut plus d'informations qu'une simple corrélation, par la présence d'une hiérarchie entre les variables reliées causalement. En effet, la présence d'une corrélation traduit juste la "ressemblance entre deux courbes", et ne permet pas de conclure sur l'existence d'un réel lien entre les deux variables⁷. L'étude de la causalité, se basant sur des techniques complexes et variées, entre prédition par machine learning et inférence par l'étude des distributions de probabilités permettent de déterminer la présence ou non d'une relation de causalité, mais aussi du sens de cette relation. Ainsi, on aura pour but de construire le graphe le plus complet et le plus fiable des variables et de leurs liens causaux, afin de comprendre les phénomènes moteurs dans le questionnaire et dans l'étude du bien-être au travail.

6. cf www.euro.who.int.

7. Par exemple, la corrélation entre le nombre de pirates en activité et le réchauffement climatique est importante alors que ces deux variables ne sont pas directement liées causalement.

2. Analyse descriptive des données

2.1 Analyse en composantes principales

2.1.1 Principe

L'analyse en composantes principales (ACP) est une procédure statistique inventée en 1901 par Karl Pearson qui permet de déterminer un ensemble de variables décorrélées à partir d'un ensemble de variables possiblement corrélées. De plus, étudier l'inertie des valeurs propres associées à ces nouvelles variables permet de mettre en évidence la dimension intrinsèque des données, et donc de réduire la dimension de nos données initiales, afin de remédier à la redondance des variables et d'obtenir un nombre de variables pour représenter les individus dans un espace intelligible. L'algorithme est le suivant :

Algorithme 2.1.1 — ACP

Données : Données pré-traitées de taille $m_{variables} \times n_{exemples}$

Résultat : Données à p dimensions ($p << m_{variables}$)

// Normalisation & centrage des données

pour $i \leftarrow 1$ à $m_{variables}$ faire

$\mathbf{D} \leftarrow$ Données $[i,:]$ // Vecteur de données pour la variable i
 $\mathbf{M}[i,:] \leftarrow \frac{\mathbf{D}-\text{moyenne}(\mathbf{D})}{\text{variance}(\mathbf{D})}$

fin

$\mathbf{M} \leftarrow \text{matrice_covariance}(\mathbf{M})$

$\mathbf{W} \leftarrow \text{vecteurs_propres}(\mathbf{M})$

$p \leftarrow \text{compromis}(\text{inertie}, \text{bruit})$ $\mathbf{W} \leftarrow \mathbf{W}[:, p,:]$ // On tronque les p vecteurs propres qu'on a choisis

$\mathbf{R} \leftarrow \mathbf{M} \times \text{Données}^T$

retourner \mathbf{R}

Le choix de p s'effectue avec un compromis entre inertie expliquée et nombre de valeurs propres conservées (bruit). En effet, plus on rajoute des dimensions, moins on perd en information, mais plus on rajoute du bruit.

Toutefois, il y a la présence d'un biais assez important lié aux questions catégorielles : en effet, du fait que les variables catégorielles ont été éclatées en n variables booléennes (n étant le nombre de modalités de la question) et par conséquent, ces variables ont une variance très faible, voire nulle, si personne n'a répondu cette modalité à la question. Par conséquent, lors de la normalisation, certaines variables ont tendance à croître de manière très importante et donc fausser la PCA. Pour remédier à ce problème, on applique une normalisation différente aux variables catégorielles : l'*Inverse Document Frequency* (IDF) de [Jones, 1972], qui revient à effectuer :

$$\mathbf{M}[i,:] \leftarrow \mathbf{D} \times \ln \left(1 + \frac{n_{exemples}}{\text{occurrences}(\mathbf{D}_i = i)} \right)$$

avec les notations de l'algorithme 2.1.1 ; en ayant de plus $\text{occurrences}(\mathbf{D}_i = i) = \text{somme}(\mathbf{D})$ car en éclatant nos questions catégorielles en variables booléennes, une somme de \mathbf{D}_i revient à avoir le nombre d'occurrences de la modalité i .

2.1.2 Obtention des nouveaux axes

Le recodage des questions catégorielles et l'ajout des variables booléennes conduit à un total de 2463 variables (numériques et booléennes) ; et on réduit la dimension de ces données à l'aide de l'ACP. L'ensemble ordonné de ses valeurs propres, utilisé pour choisir la dimension de sortie des données, est représenté Fig.2.1a et Fig.2.2a. On se restreint à considérer les premiers vecteurs propres de cette matrice. Chacun de ces vecteurs propres (somme pondérée des variables initiales) définit une variable agrégée.

ACP sur les variables objectives

Le spectre des valeurs propres de la matrice de covariance des variables objectives est représenté Fig.2.1a. Nous avons choisi de sélectionner les 8 premiers vecteurs propres, comme un compromis entre la taille de la représentation réduite et l'inertie capturée (62%) ; notons qu'il faut pratiquement doubler le nombre de vp. pour arriver à 70% d'inertie.

ACP sur les variables subjectives

Dans le cas des variables subjectives (environ 20% des variables), le spectre est représenté Fig.2.2a. Le fait de retenir les 5 premiers vecteurs propres permet de capturer 80% de l'inertie des données.

2.1.3 Étude des axes

13 nouveaux axes d'étude¹ sont ainsi obtenus, dont nous allons étudier les caractéristiques. Ces 13 axes sont obtenus par une combinaison linéaire de multiples variables. Afin de connaître la nature de ces nouveaux axes, le poids des différentes variables sur les axes en fonction de leur catégories

1. 8 axes objectifs et 5 axes subjectifs

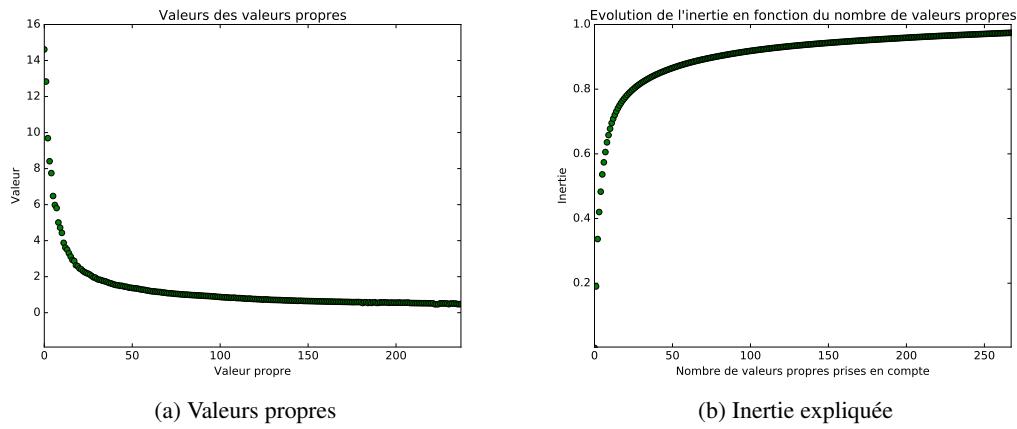


FIGURE 2.1: Données DARES, variables objectives : Spectre de la matrice de covariance.

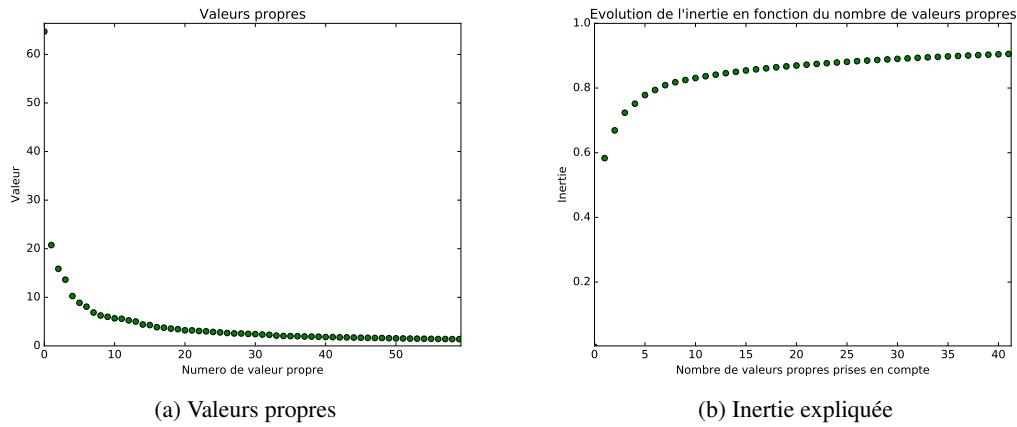


FIGURE 2.2: Données DARES, variables subjectives : Spectre de la matrice de covariance.

est un bon indicateur, qui est représenté sur la figure 4 pour les axes objectifs et sur la figure 5 pour les axes subjectifs.

Les risques psychosociaux, l'organisation du travail et la santé n'apparaissent pas de manière évidente dans les axes objectifs : ceci est du au faible nombre de questions objectives dans ces catégories qu'on peut voir sur la figure 1, ce qui réduit la variance expliquée par ces catégories. Ainsi, les poids des catégories de variables sur les nouveaux axes subjectifs ressemble à un graphe complémentaire sauf pour la "Santé", car elle ne comporte que 4 questions dans le questionnaire. Toutefois, pour déterminer précisément la nature des axes, une analyse des valeurs-test permet de relever les caractéristiques qu'ils mettent en valeur pour différencier les personnes enquêtées.

2.1.4 Interprétation des axes

Les valeurs-test donnent directement les variables les plus significatives sur les 10% des personnes les plus élevées et 10% des personnes les plus basses sur l'axe, ce qui permet d'identifier ces

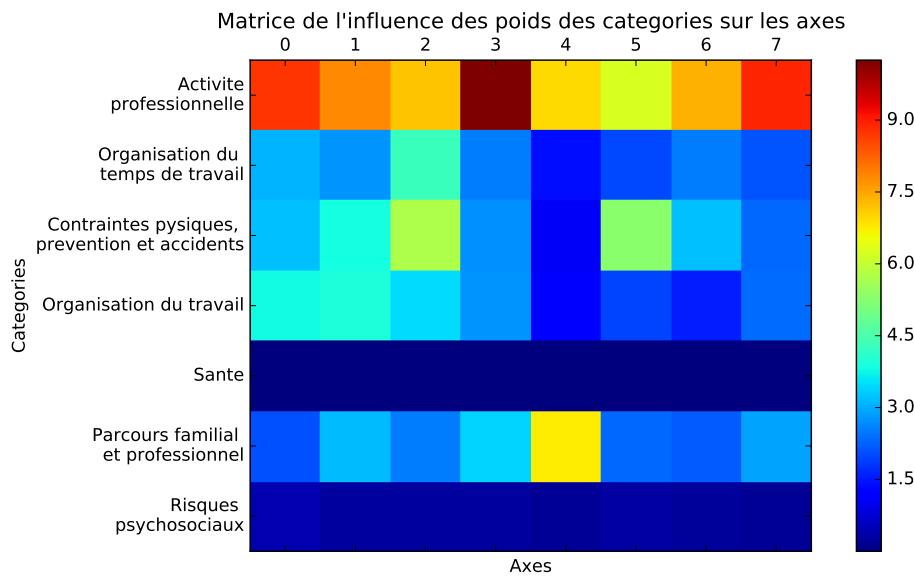


FIGURE 2.3: Poids total des catégories de variables sur les nouveaux axes objectifs

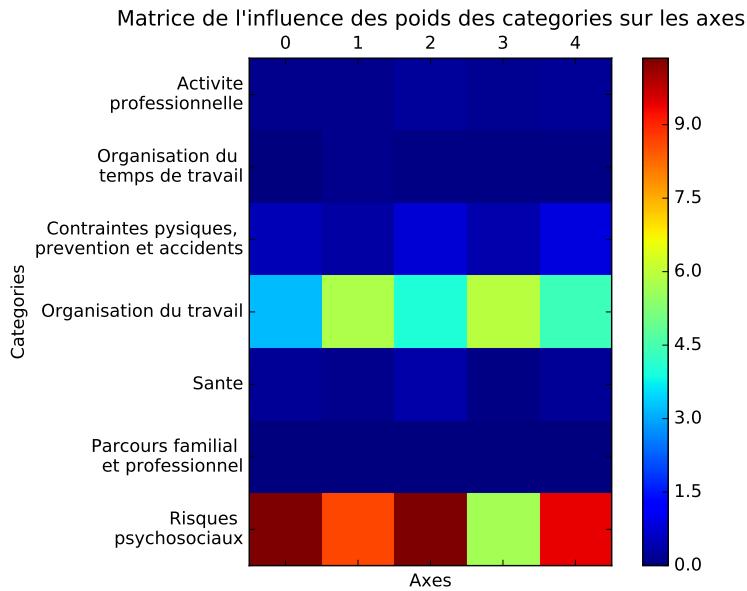


FIGURE 2.4: Somme des valeurs absolues des poids des variables dans la définition de chaque axe subjectif, par catégorie de variable

sous-populations, et donc d'interpréter l'axe. Ainsi, nous avons pour chaque axe un graphique en annexe des valeurs-test des variables les plus représentées, et nous pouvons résumer l'analyse des axes dans les tables 1 & 2.

Axes de l'ACP objectif	Variables corrélées positivement	Variables corrélées négativement
Axe 1 : Taille de l'entreprise employant l'enquêté	Ancienneté Possibilité de congés Entretiens d'évaluation Présence de ressources humaines	Statut indépendant Pas de collègues Non syndiqué
Axe 2 : Rémunération et niveau de qualification	Pas de mails, d'intranet Doit effectuer des mouvements fatigants Nécessité de rester longtemps debout	Revenus Temps passé devant l'informatique, mails Travail non pénible physiquement
Axe 3 : Temps de travail et sécurité	Nombre d'heures de travail par semaine Nombre de dimanches/samedis travaillés Nombre de nuits travaillées	Pas de port de protection Pas de risque de blessure/accident Pas de consignes de sécurité
Axe 4 : Nature de l'organisme employeur	Salarié du privé Entreprise de grande taille Cadres d'entreprise	Employé d'administration publique, enseignement, santé, social Salarié de l'État
Axe 5 : Immigration	Père/Mère nés en France Pas de lien à la migration	Mère/Père immigré(e) Immigré Naturalisé ou étranger
Axe 6 : Accidents du travail	Age Information sur les risques du travail Origine de ces informations	Date de l'accident de travail Accident signalé à l'employeur L'employeur n'a pas pris de mesures pour réduire les risques
Axe 7 : Ancienneté/ Taille de la famille	Année de naissance Nombre de personnes au foyer Année de début de contrat	Age Personne seule Date du dernier accident de travail
Axe 8 : Situation familiale	Nombre de personnes au foyer Nombre d'actifs au foyer Revenus En couple et marié	Seul(e) au foyer Pas en couple Pas marié

TABLE 2.1: Tableaux des principales contribution des variables pour les 8 premiers axes de l'ACP des variables objectives

Axes de l'ACP subjectif	Variables corrélées positivement	Variables corrélées négativement
Axe 1 : Risques psychosociaux	Personnes provenant de l'entreprise ont des comportements inappropriés Personne ignorée, critiquée, a son travail saboté	Score de bien-être de l'OMS
Axe 2 : Indépendance/ Présence de collègues/ supérieurs	Possibilité de discuter avec son supérieur Parfois en désaccord avec ses collègues A été consulté pour un changement de l'environnement de travail	Pas de collègues Indépendant
Axe 3 : Bon management	Score de bien-être de l'OMS Le supérieur prête attention aux propos de l'enquêté et lui apporte de l'aide	Pense que son travail est mauvais pour la santé Pas souvent de bonne humeur, calme et tranquille pas de possibilité de coopérer Doit se dépêcher
Axe 4 : Changement du milieu de travail	Informé des changements Consulté pour effectuer les changements Pense que ces changements sont positifs	Pas de changement de poste Le travail ne permet pas d'apprendre des choses nouvelles
Axe 5 : Satisfaction du travail en équipe	Bonne humeur Frais et disposé, calme et tranquille Pas de pression Fier du travail	Pas de collègues Pas de supérieurs

TABLE 2.2: Tableaux des principales contributions des variables pour les 5 premiers axes de l'ACP des variables subjectives

2.2 Clusters et analyses

2.2.1 Méthode employée

Dans la suite, la représentation considérée est celle définie par les axes ci-dessus (i.e. chaque personne est projetée dans l'espace \mathbb{R}^d , où $d = 8$ ou $d = 5$ selon que l'on considère les données objectives ou subjectives). Les personnes sont ensuite partitionnées en communautés (clusters) à l'aide de l'algorithme *k-means++*² se fondant sur la distance classique de \mathbb{R}^d [Arthur et Vassilvitskii, 2007]. On obtient ainsi 8 groupes objectives et 6 groupes subjectifs. Notons que le fait de distinguer les données objectives et subjectives conduit à une meilleure stabilité des clusters obtenus ; le fait de considérer toutes les données conduit à des interférences entre situation objective et ressenti. Parmi

2. L'implémentation utilisée est celle de la librairie *Scikit-Learn*.

nos outils d'analyse nous avons utilisé la valeur-test (v-test) de [Lebart *et al.*, 2006], défini par les formules suivantes pour les variables numériques (V_n) et catégorielles (V_c) :

$$V_n = \frac{\mu_g - \mu}{\sqrt{\frac{n-n_g}{n-1} \times \frac{\sigma^2}{n_g}}}$$

$$V_c = \frac{n_{jg} - \frac{n_g \times n_j}{n}}{\sqrt{\frac{n-n_g}{n-1} \times \left(1 - \frac{n_j}{n}\right) \times \frac{n_g \times n_j}{n}}}$$

avec :

μ : Moyenne globale de la variable

σ : Variance totale

n : Nombre d'individus total

index g : valeur sur un cluster

index j : valeur sur une catégorie

Finalement, la valeur du v-test peut-être interprétée comme une différence de moyenne entre les valeurs des variables entre le cluster et la population globale, dans le but de mettre en valeur les variables significatives.

2.2.2 Clusters objectifs

Chaque cluster est interprété en fonction de son centre (représenté en coordonnées parallèles en fonction des axes de l'ACP³ à la Fig.2.5), et considérant les variables significatives au sens du v-test pour ce cluster : dont la valeur sur le cluster est soit significativement plus élevée, soit moins élevée que pour l'ensemble des données. Une variable particulière, le code NAF17 (Fig.2.6) permet d'avoir une idée de la répartition des classes socioprofessionnelles dans les différentes communautés. Les résultats de l'analyse sont résumés dans la table 2.3.

Groupe 1 : Indépendants (INDEP.)

Ce cluster représente des personnes étant dans des entreprises très petites, un temps de travail assez élevé ainsi qu'un temps de travail élevé ; ce cluster est représenté par les classes NAF *Agriculture, sylviculture et pêche, Commerces, Construction, Hébergement et restauration*. Cette communauté représente donc les gens indépendants. Les caractéristiques de ce groupe sont, mis à part le fait que la taille de l'organisation qui emploie l'enquêté est très petite, le nombre de congés disponibles aux enquêtés (14,58 jours contre 36,58 dans la population globale), et le nombre de jours d'absence correspondant à des arrêts maladie sont peu importants (3,66 jours contre 8,34 jours).

Groupe 2 : Services aux particuliers (SERVPART)

Les caractéristiques de ce deuxième groupe sont un faible niveau de qualification, un temps de travail et une sécurité faibles, travailleurs du secteur public, plutôt dans le domaine des activités de services. En analysant plus en détail les valeurs des v-test sur les codes NAF, la catégorie socioprofessionnelle la plus représentée est celle des services aux particuliers. Les revenus moyens de ce cluster sont bien inférieurs aux revenus moyens de l'ensemble des enquêtés (1163€/mois

3. Ce qui correspond à la moyenne du cluster ou encore l'individu représentatif du cluster projeté sur les axes de l'ACP

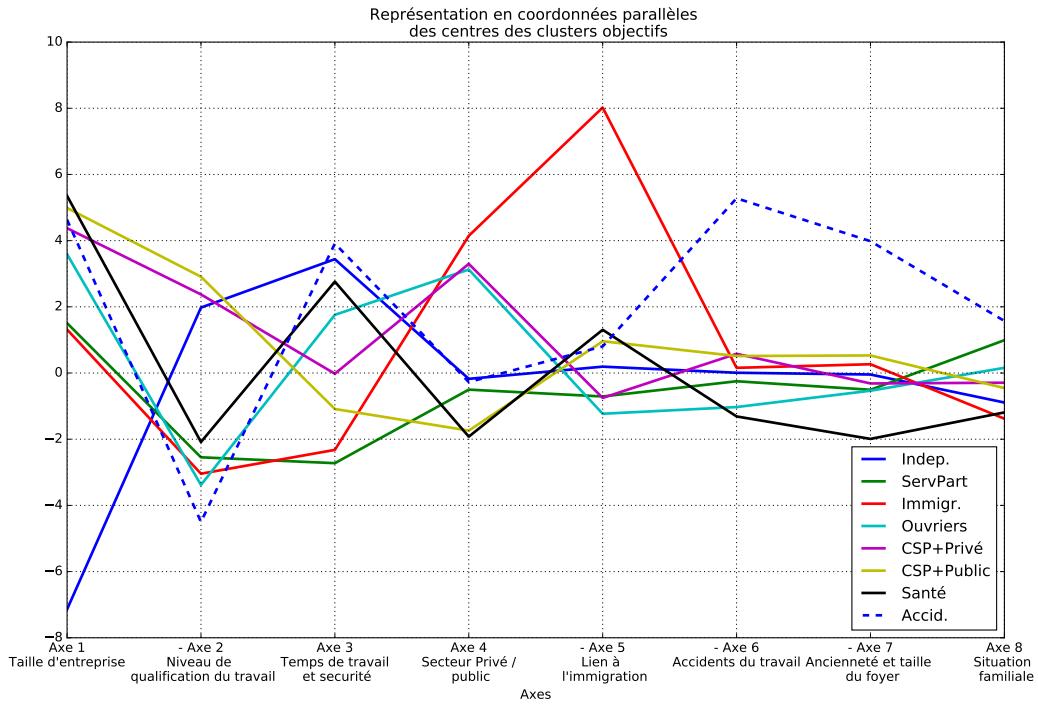


FIGURE 2.5: Représentation en coordonnées parallèles des centres des clusters objectifs sur les axes de l'ACP

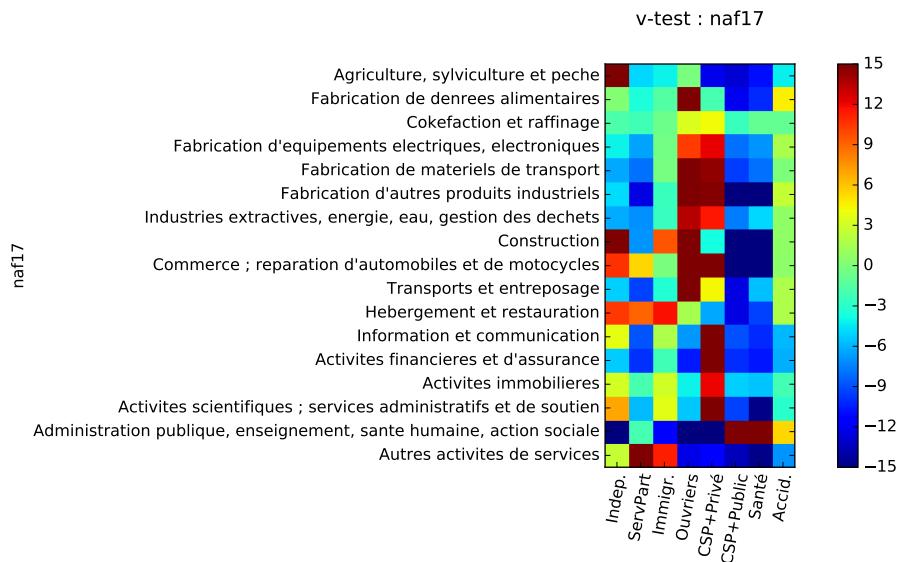


FIGURE 2.6: Valeurs V-test des clusters objectifs sur les codes NAF17

contre 1833€ en moyenne), avec une qualification assez faible (19% sans diplôme, 39% avec un CAP, BEP ou équivalent).

Groupe 3 : Lien à l'immigration (IMMIGR.)

Une caractéristique principale de ce groupe, qui apparaît à la Fig.2.5, est le lien à l'immigration. En effet, 53% des personnes de ce cluster sont étrangers, et 42% sont français par naturalisation, mariage, déclaration ou option à la majorité. Ils ont, d'après les moyennes calculées sur le cluster, travaillent principalement dans le secteur privé, comme l'indique la Fig.2.5.

Groupe 4 : Ouvriers (OUVRIERS)

Le troisième cluster est représenté par des personnes employées dans une grande entreprise, ayant un faible niveau de qualification, et plutôt du secteur privé. Les codes NAF sur-représentés dans ce cluster sont souvent des secteurs de fabrication de produits, d'industrie de l'énergie et des transports. Ce cluster met donc en valeur les ouvriers, techniciens, agents de maîtrise et contremaîtres du secteur privé. Les enquêtés sont à 93% salariés d'une entreprise, d'un artisan, ou d'une association. Ce cluster est composé majoritairement d'hommes (76%), qui mettent en avant des inconvénients du travail et des conditions de travail telles que la saleté (53%), des courants d'air (62%), des secousses ou vibrations (40%), de l'humidité (44%) et une température basse (56%).

Groupe 5 : Employés de bureau du secteur privé (CSP+PRIVÉ)

Cette communauté est similaire à celle du cluster 4, à l'exception du niveau de qualification qui est élevé, ainsi que les secteurs d'activité, qui incluent ici les secteurs des services et des activités scientifiques. Ce cluster est identifiable à une population d'employés de bureau du privé, comprenant les cadres. Les salaires de ce groupe sont par ailleurs bien supérieurs à la moyenne (2328€/mois contre 1833€/mois pour l'ensemble de la population étudiée). 90% des enquêtés n'ont pas à rester longtemps debout pour effectuer leur travail, 95% disposent d'une boîte aux lettres électronique professionnelle et plus de 80% des personnes du cluster sont satisfaits des conditions de travail.

Groupe 6 : Employés de bureau du secteur public (CSP+PUBLIC)

Ce cluster a des caractéristiques très proches du cluster 5 ; à la différence du secteur d'activité, qui est public. Ce groupe peut donc être interprété comme le cluster des employés de bureau du secteur public. Les enquêtés de ce groupe sont à 59% des salariés de l'état, et sont aussi mieux payés que la moyenne : 2357€/mois contre 1833€/mois en moyenne. Contrairement au cluster 5, les personnes constituant ce cluster bénéficient d'un grand nombre de congés (60 jours contre 37 jours en moyenne).

Groupe 7 : Santé (SANTE)

Dans ce groupe, les enquêtés sont dans des entreprises de grande taille, avec aussi des temps de travail et une sécurité assez élevée, principalement dans le secteur public. Le code NAF le plus présent dans ce cluster est *Action publique, enseignement, santé humaine, action sociale*, mais en affinant notre analyse la catégorie la plus représentée ici est celle de la santé humaine. Ce cluster peut être appelé "Santé". 62% des enquêtés de ce groupe travaillent dans le soin des personnes et la plupart ont un grand nombre d'heures de travail, et travaillent aussi le matin, le soir et les fins de semaine. De plus, ces personnes ont souvent de grandes responsabilités : les erreurs de 85% des personnes peuvent entraîner des conséquences dangereuses pour leur sécurité ou celle d'autre personnes.

Groupe 8 : Accident du travail (ACCID.)

Cette dernière communauté possède aussi une caractéristique distincte des autres : l'accident au travail. Les enquêtés formant cette communauté sont souvent des personnes ayant un faible niveau de qualification, travaillent beaucoup et insistent sur la sécurité, mais ont subi un accident du travail. La plupart de ces enquêtés critiquent par ailleurs les conditions de travail pénibles et le manque de sécurité dans leur travail, ainsi que des situations de tension avec les supérieurs.

Cluster	Abréviation	Distribution	Intitulé
1	INDEP.	9.2%	Indépendants
2	SERVPART	14.9%	Services aux particuliers
3	IMMIGR.	6.2%	Lien à l'immigration
4	OUVRIERS	13.6%	Ouvriers, techniciens, agents de maîtrise et contremaîtres du secteur privé
5	CSP+PRIVÉ	18.5%	Employés de bureau du secteur privé
6	CSP+PUBLIC	16.8%	Employés de bureau du secteur public
7	SANTE	12.9%	Santé
8	ACCID.	7.8%	Accident du travail

TABLE 2.3: Identification des clusters objectifs

2.2.3 Clusters subjectifs

Après avoir analysé les clusters objectifs, il faut maintenant analyser les clusters subjectifs, avant de pouvoir comparer les deux analyses. On dispose aussi de la représentation en coordonnées parallèles à la Fig.2.7 et de la répartition des v-test avec le code NAF17 à la Figure ???. Le résumé de ces analyses se retrouve à la Table 2.4.

Groupe 1 : Indépendants (INDEP.)

Ce premier groupe est caractérisé par les enquêtés qui sont indépendants à leur travail, et donc isolés. Ils sont caractérisés par les secteurs d'agriculture, de sylviculture, de pêche, ainsi que des activités de service. L'organisation du travail est plutôt stable, et ils sont assez satisfaits de leur travail, et ce malgré des revenus bien inférieurs à la moyenne (1512€/mois contre 1877€/mois en moyenne) et peu de congés (17,55 jours contre 38,48 en moyenne), et une moyenne d'âge supérieure à la moyenne globale (47 ans contre 43 en moyenne). Les axes 3 et 4 ont des valeurs pour ce cluster assez faibles car l'enquêté n'a pas de supérieur ni de collègues.

Groupe 2 : Heureux (HEUR.)

Ce cluster est représenté par les personnes qui n'ont pas de problèmes avec leur environnement de travail, qui sont satisfaites de leur travail et ont une bonne vie de groupe. Les enquêtés constituant ce cluster sont légèrement moins payées que la moyenne (1753€/mois contre 1877€/mois en moyenne) mais ont un score de bien-être défini par l'OMS bien supérieur à la moyenne (20,38 contre 15,65). Par ailleurs, la notion de tension et pression est très peu présente dans ce groupe : la plupart des enquêtés sont jamais sous pression et n'ont aucune tension avec leur équipe.

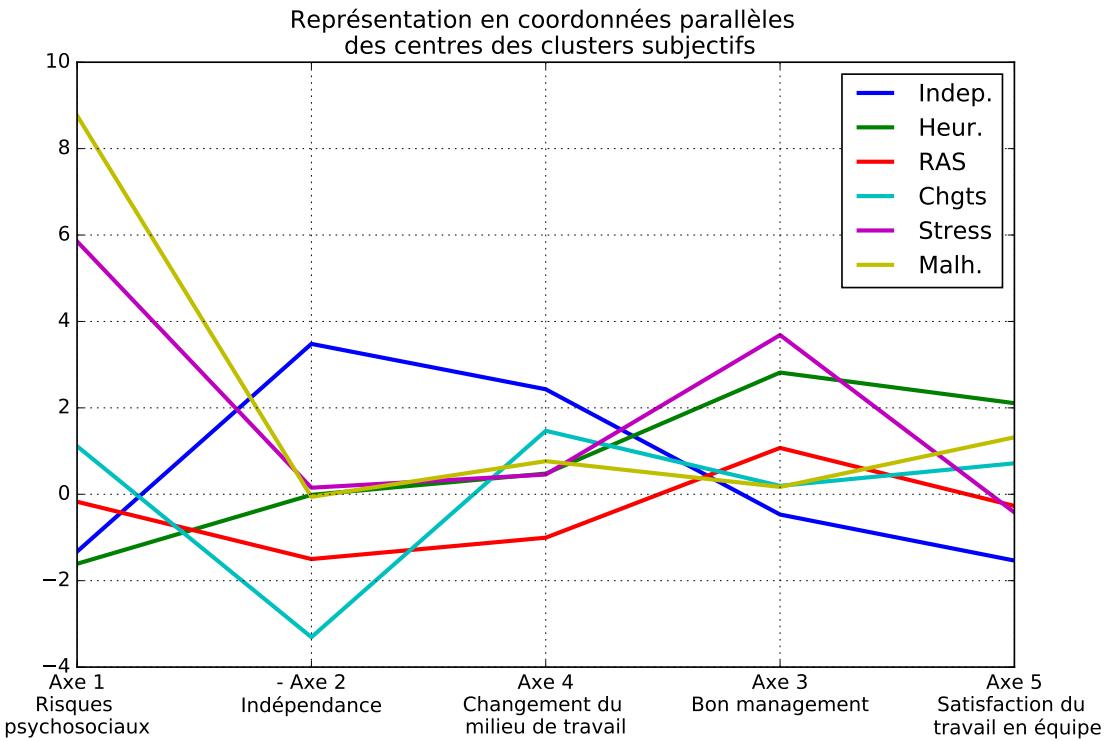


FIGURE 2.7: Représentation en coordonnées parallèles des centres des clusters subjectifs sur les axes de l'ACP

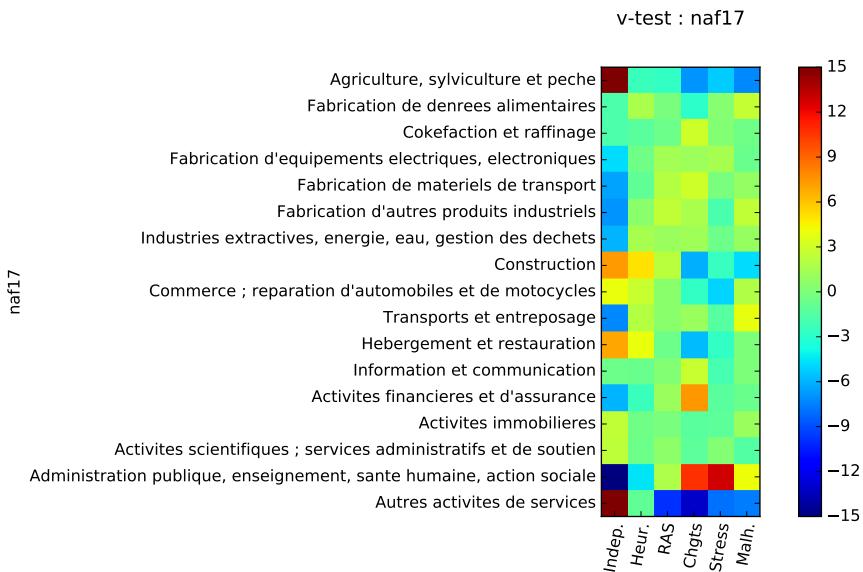


FIGURE 2.8: Valeurs V-test des clusters subjectifs sur les codes NAF17

Groupe 3 : Rien à signaler (RAS)

Ce cluster semble être assez vague, alors que c'est l'un des plus peuplés. Dans cette communauté, les personnes sont plutôt satisfaites, n'ont pas subi de changement d'environnement de travail au

cours des douze derniers mois, et proviennent à peu près que chaque catégorie socioprofessionnelle. Ce cluster représente donc les personnes qui n'ont rien à signaler de particulier et son plutôt satisfaites de leur vie au travail. Elles ont pourtant un salaire supérieur à la moyenne (2023€/mois) et plus de congés que la moyenne (41 jours).

Groupe 4 : Changements de l'environnement de travail (CHGTS)

Cette communauté se caractérise par le fait que ses constituants proviennent de grandes organisations, et qu'ils ont récemment subi des changements dans leur milieu de travail au cours des douze derniers mois. Ils sont assez satisfaits du travail en équipe, et pensent que les changements ont été plutôt positifs et bien effectués. Ils proviennent principalement des catégories socioprofessionnelles "*Administration publique, enseignement, santé humaine, action sociale*" et "*Activités financières et d'assurance*", et sont mieux payés que la moyenne : 2094€/mois.

Groupe 5 : Environnement stressant (STRESS)

Cette communauté représente ceux qui sont satisfaits du management, mais ont des problèmes liés à leur environnement de travail qui leur est, selon leur point de vue, néfaste et source de risques psychosociaux. Ces sentiments sont éprouvés à cause de situations de tension avec les collègues et de comportement nuisibles, par exemple l'enquêté est ignoré, ou critiqué injustement.

Groupe 6 : Malheureux (MALH.)

Enfin, ce groupe représente les personnes qui ont beaucoup de risques psychosociaux malgré une satisfaction du travail en équipe assez bonne. La pression ainsi que la tension avec les supérieurs sont très présentes dans ce cluster. Il y a un grand ressenti d'injustice, et un sentiment d'exploitation. Ceci se traduit sur la représentation en coordonnées parallèles par un mauvais score sur l'axe 3 ; ou encore un taux d'absentéisme assez élevé (17,25 jours contre 7,95 en moyenne). Ces éléments traduisent aussi un rejet du travail actuel de l'enquêté : 79% de la population de ce groupe ne seraient pas heureux si l'un de leurs enfants s'engagent dans la même activité professionnelle qu'eux et 62% ne se sentent pas capables de continuer leur travail jusqu'à leur retraite.

Cluster	Abréviation	Distribution	Intitulé
1	INDEP.	9.5%	Indépendants
2	HEUR.	15.7%	Heureux
3	RAS	21.8%	Rien à signaler
4	CHGTS	17.5%	Changement de l'environnement de travail
5	STRESS	22.5%	Environnement stressant
6	MALH.	13.0%	Malheureux

TABLE 2.4: Identification des clusters subjectifs

2.3 Correspondance entre clusters objectifs et subjectifs

2.3.1 Croisement des populations de clusters

Après avoir étudié les différents clusters individuellement, il est intéressant de voir comment se recoupent les clusters subjectifs et les clusters objectifs. Entre autres, cette étude nous permet de voir directement comment se projettent la situation professionnelle des enquêtés et le ressenti de leur situation. La Figure 2.9 est une représentation de ce croisement qui permettent d'analyser les caractéristiques communes à ces deux clustering, qui est pour chaque case,

$$M_{i,j} = \frac{\text{Card}(O_j \cap S_i)}{\text{Card}(O_j)}$$

avec O_j le j^{eme} cluster objectif et S_i le i^{eme} cluster subjectif.

La normalisation est effectuée sur les clusters objectifs, c'est-à-dire que la somme des éléments sur une colonne de la Fig.2.9 vaut 1.

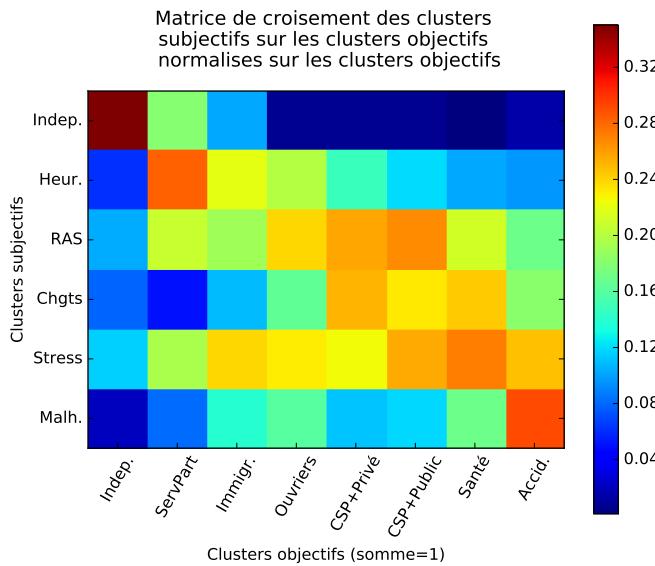


FIGURE 2.9: Matrice de correspondance entre les clusters subjectifs et objectifs

Le premier élément qui apparaît de manière claire est le recouvrement des clusters INDEP. objectifs et INDEP. subjectifs : il y a en effet plus de 70% de la population de ces groupes en commun.

Le cluster objectif SERVPART est ventilé en quatre clusters subjectifs : INDEP. (17%), HEUR. (28%), RAS(20%), et STRESS (18%).

Les clusters objectifs OUVRIERS et IMMIGR. se répartissent de manière similaire entre les clusters subjectifs HEUR. & RAS, et STRESS. Les clusters CSP+PRIVÉ et CSP+PUBLIC de même se répartissent sur les clusters subjectifs RAS, CHGTS et STRESS. Enfin, le cluster ACCID. recoupe essentiellement les clusters subjectifs MALH. (30%) et STRESS(26%).

La spécificité des indépendants ici peut être perçue comme un artefact du questionnaire : beaucoup de questions font référence au travail en équipe et au management d'équipe. Ainsi, ces questions ne concernent pas les indépendants ; et donc un grand nombre de leurs réponses sont *sans objet* ou *non pertinent*. En revanche, la présence d'un environnement stressant dans la majorité des clusters objectifs n'était pas attendue ; ce résultat fait écho à d'autres travaux montrant l'augmentation des facteurs de stress au cours des années, en particulier en lien à la "transformation numérique" des entreprises [Datchary.C, 2011].

2.3.2 Autonomie et clusters

Les différents clusters identifiés et leurs intersections peuvent être utilisés pour approfondir les liens entre QVT et d'autres facteurs tel que l'autonomie au travail des individus. Pour réaliser cette étude, nous avons défini un score d'autonomie à partir de 4 questions. Le score sur dix est une somme pondérée en fonction de la réponse aux questions, dont les détails sont donnés Table 2.5. Afin d'obtenir un score homogène, nous ne considérons que les personnes ayant répondu aux quatre questions.

Les quatre questions utilisées sont :

1. COMMENT : Les indications données par vos supérieurs hiérarchiques vous disent ce qu'il faut faire. En général, est-ce que...
 - (a) ils vous disent aussi comment faire
 - (b) ils indiquent plutôt l'objectif du travail et vous choisissez vous-mêmes la façon d'y arriver.
2. STARK : Vous recevez des ordres, des consignes, des modes d'emploi. Pour faire votre travail correctement, est-ce que ...
 - (a) vous appliquez strictement les consignes
 - (b) dans certains cas, vous faites autrement
 - (c) la plupart du temps vous faites autrement
 - (d) sans objet (pas de consignes)
3. INCIDENT : Quand au cours de votre travail, il se produit quelque chose d'anormal, est-ce que...
 - (a) la plupart du temps, vous réglez personnellement l'incident
 - (b) vous réglez personnellement l'incident mais dans des cas bien précis, prévus d'avance
 - (c) vous faites généralement appel à d'autres (un supérieur, un collègue, un service spécialisé)
4. REPETE : Votre travail consiste-t-il à répéter continuellement une même série de gestes ou d'opérations ?
 - (a) Oui
 - (b) Non

Les scores d'autonomie aux intersections des différents clusters sont représentés Figure 2.10. Les clusters d'indépendants n'ont pas été représentés car i) les questions relatives à l'autonomie ne sont pas toujours pertinentes pour eux (pas de supérieur) et ii) l'intersection avec les autres cluster est presque vide, ce qui rend toute moyenne et comparaison non significative.

L'analyse des résultats permet de tirer plusieurs enseignements :

- Indépendamment des clusters subjectifs, l'autonomie des cadres apparaît logiquement bien plus élevée que celle des autres clusters, par contre il existe peu de différence entre public et

	Réponse			
	(a)	(b)	(c)	(d)
COMMENT	0	3	-	-
STARK	0	1	2	3
INCIDENT	3	1	0	-
REPETE	0	1	-	-

TABLE 2.5: Pondération des réponses aux questions pour le calcul du score d'autonomie

privé. Le cluster de la Santé est quand à lui le groupe avec l'autonomie la plus faible.

- En étudiant le lien avec les clusters subjectif (ordonnés selon une QVT approximativement décroissante de HEUR. à MALH.), l'autonomie apparaît comme décroissante pour tous les groupes objectifs (les clusters avec une faible autonomie sont ceux ayant une faible QVT). De façon plus détaillée, l'absence d'autonomie est très fortement liée aux groupes MALH. et dans une moindre mesure STRESS(et ce pour tous les clusters objectifs). Le groupe HEUR. n'est par contre pas toujours caractérisée par l'autonomie la plus élevée (qui est souvent atteinte par RAS).

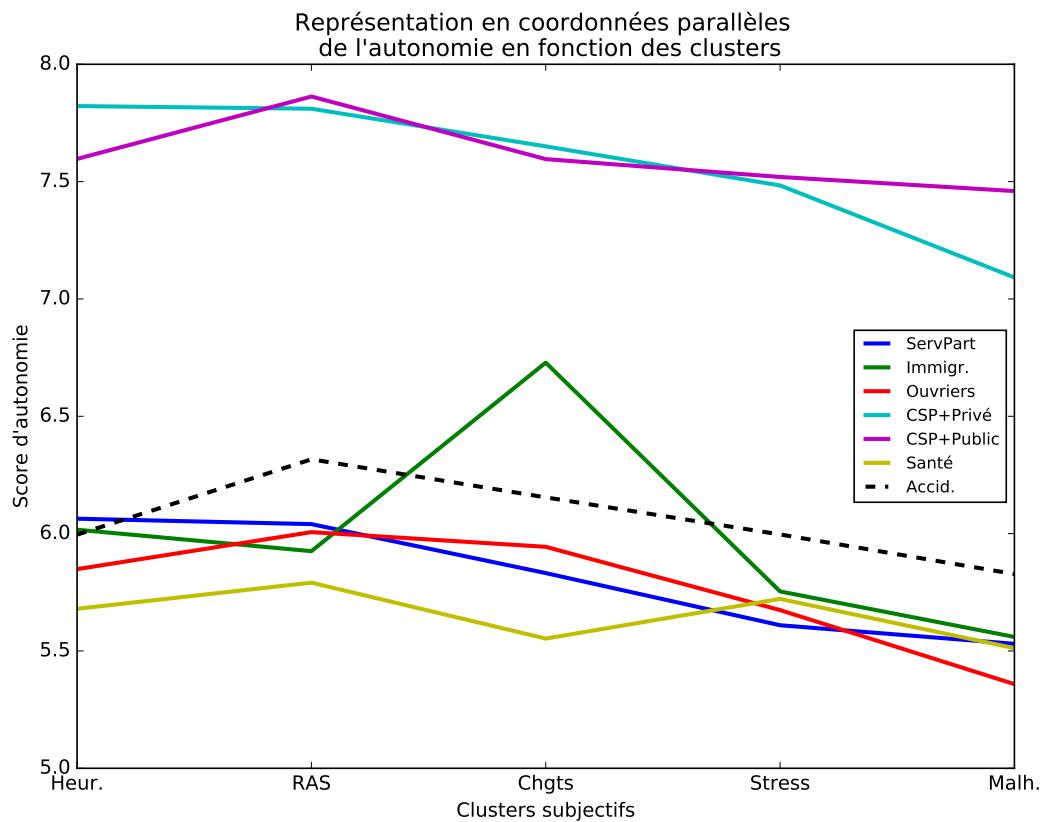
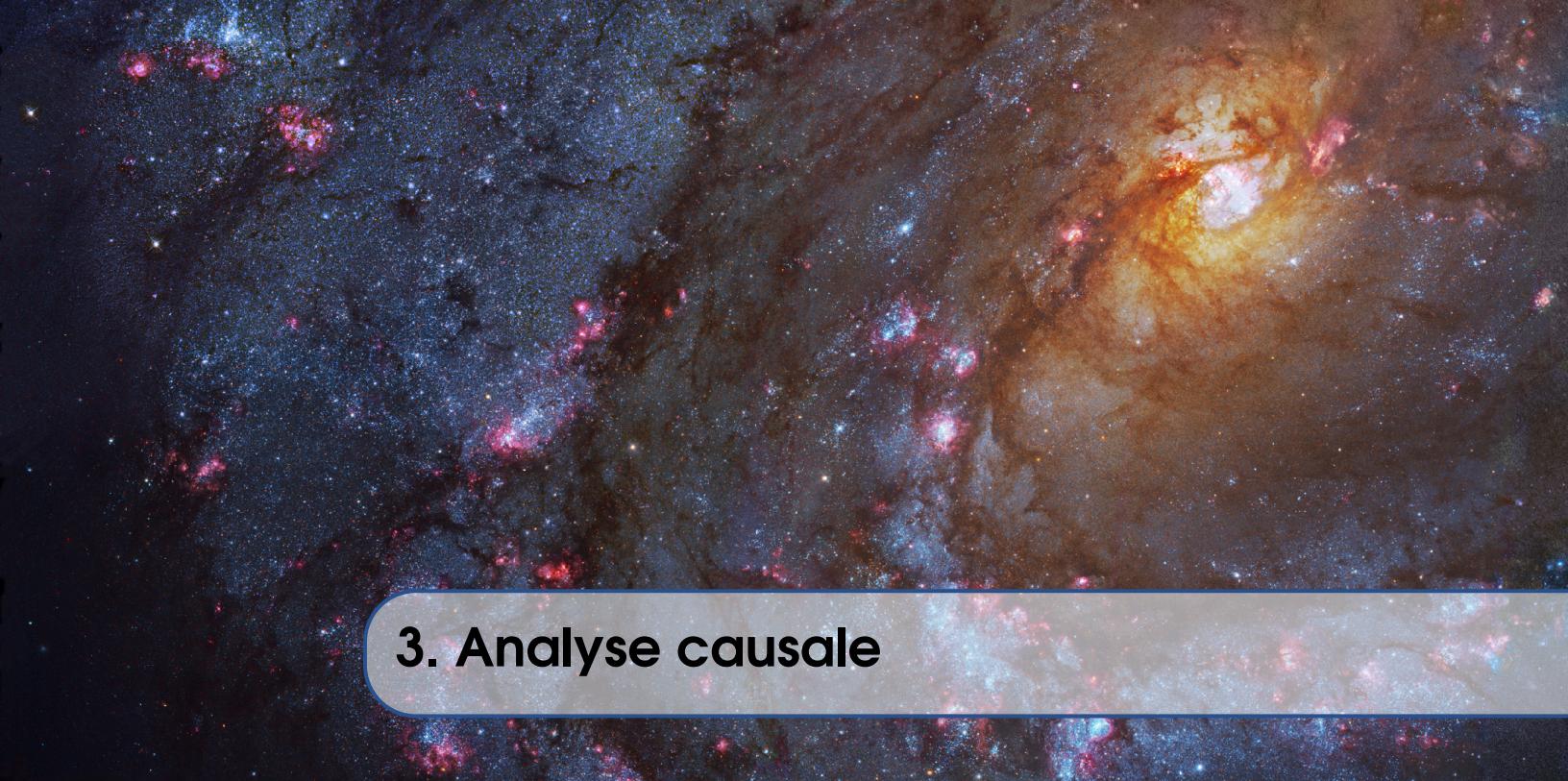


FIGURE 2.10: Représentation en coordonnées parallèles du score d'autonomie en fonction des clusters. Par exemple, les individus à la fois dans le cluster objectif Santé (ligne la plus basse - jaune) et dans le cluster subjectif HEUR. (tout à gauche sur le graphique) ont un score d'autonomie moyen de 5,6.



3. Analyse causale

3.1 Méthodologie

Notre objectif est de pouvoir construire un graphe reliant causallement l'ensemble, ou un maximum de variables de notre questionnaire. Pour cela, on adopte une méthodologie employée en analyse de graphes causaux, c'est-à-dire :

1. Construire un graphe relationnel non dirigé en analysant les indépendances entre les variables
2. Repérer et ne conserver que les relations d'adjacences (retirer les liens indirects, déconvolution)
3. Orienter les relations restantes à l'aide du coefficient de causalité entre les deux variables

Mais plusieurs difficultés viennent s'ajouter à cette approche : l'hétérogénéité des données et la nature du coefficient de causalité.

3.1.1 Hétérogénéité des données

La méthode utilisée "habituellement" afin de construire le graphe relationnel consiste à calculer le coefficient de corrélation de Pearson sur l'ensemble des variables afin d'avoir une idée de la force des liens entre les différentes variables, ce qui va être employé par la suite pour appliquer la deuxième étape.

Toutefois, nous avons dans notre cas des variables hétérogènes de nature, un ensemble de variables numériques, catégorielles (ordonnées ou non) et booléennes (notamment les drapeaux). Les coefficients de corrélation de classiques n'ont donc plus de sens en face de ces données. Il existe tout de même des mesures de corrélation pour les différents types de variables ; mais les techniques de déconvolution utilisent principalement des relations matricielles sur les matrices de liens ou de corrélation, ce qui relève une autre difficulté : la nécessité d'obtenir une matrice homogène avant d'appliquer les algorithmes de déconvolution.

De plus, nous allons revenir sur la décomposition des variables catégorielles ; il n'est pas utile voire contre-intuitif de continuer de décomposer les variables catégorielles en variables booléennes :

- le lien entre les variables booléennes sera ignoré par l'étude d'indépendance des variables
- les algorithmes d'inférence sur la causalité ont une moins bonne performance sur les variables booléennes

3.1.2 Nécessité d'une déconvolution

La nécessité de retirer les liens indirects ne semble pas être évidente au premier abord, et conserver tous les liens pour avoir le graphe total. En effet, pour les graphes de petite taille, le graphe reste intelligible comme l'indique la Figure 3.1.

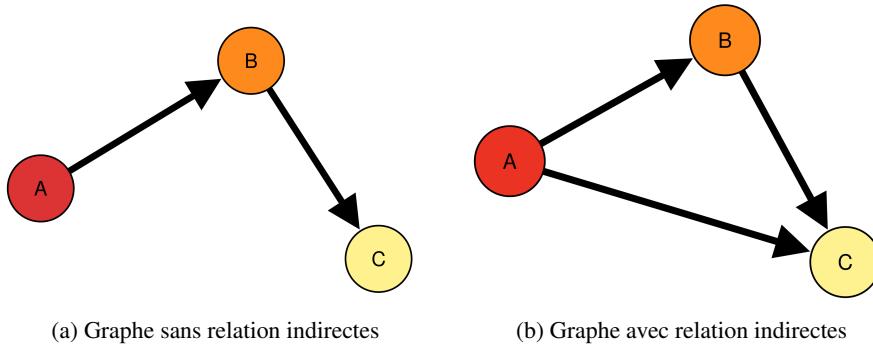


FIGURE 3.1: Graphe causal simple sans et avec la relation indirecte

Ainsi, on pourrait penser que notre étude se limiterait à étudier les coefficients de causalité, et d'interpréter les résultats obtenus par la suite. Toutefois, avec notre volume de variables, il est impossible d'interpréter les résultats sans appliquer une déconvolution, comme le montre la Figure 3.2. En effet, avec 470 variables¹, il y a plus de 5000 liens entre toutes les variables.

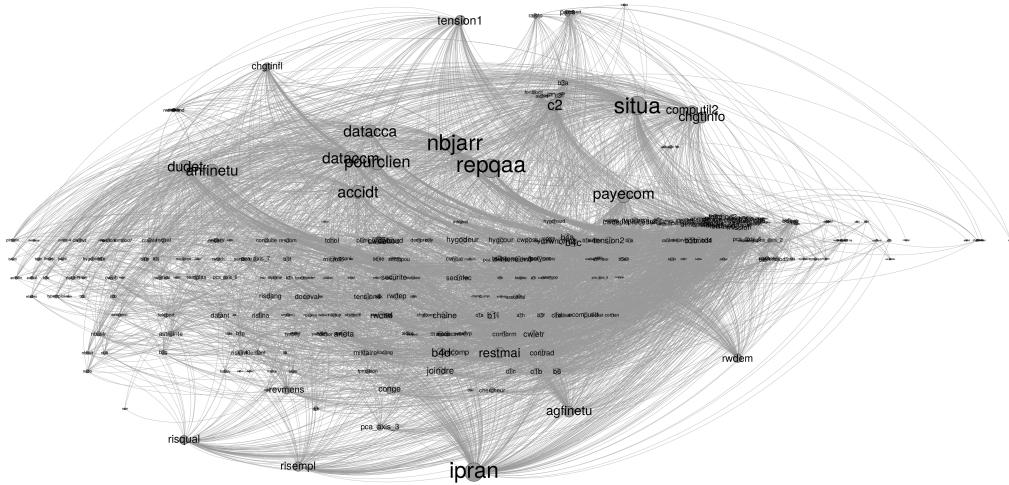


FIGURE 3.2: Graphe de causalité en se basant uniquement sur les résultats inférés par les paires deux à deux

1. Les variables drapeaux ont été retirées dans un premier temps.

3.1.3 Coefficient de causalité

Afin de pouvoir orienter nos graphes de causalité, on souhaite inférer la relation de causalité entre les variables deux à deux à la dernière étape de notre méthodologie. Pour cela, on fait appel aux résultats des compétitions *Kaggle*² et *Codalab*³ sur la causalité, notamment les algorithmes ayant obtenu de bon résultats : [Fonollosa, 2016] et [Lopez-Paz *et al.*, 2015]. Dans ces différents challenges, on a demandé de présenter les résultats sous la forme $target \in [-1; 1]$, $target$ étant la valeur de la prédiction de la causalité d'une paire A, B de variables. Ainsi, $target$ est interprétée de la manière suivante :

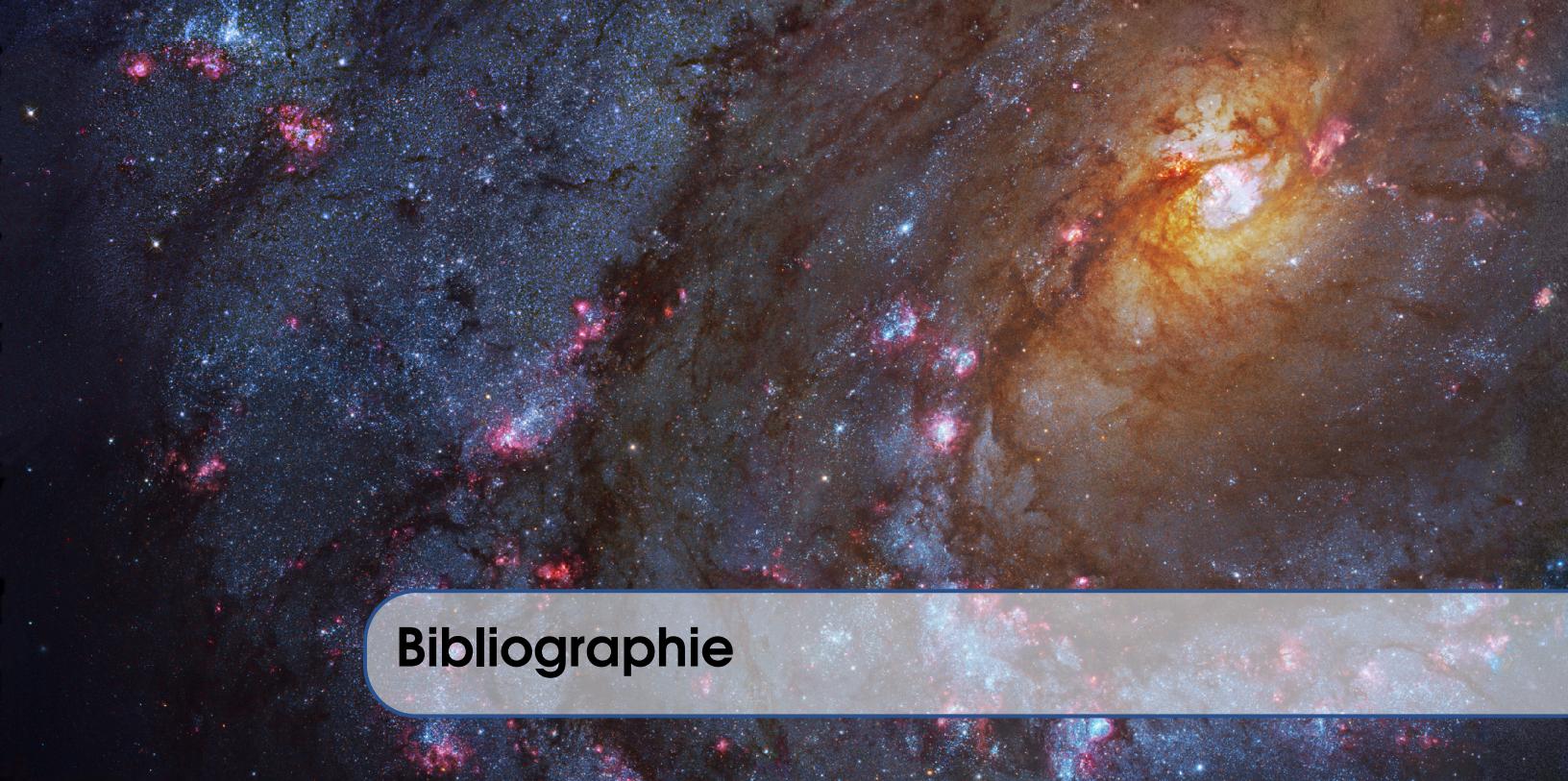
- Si $target = -1$, on a $A \leftarrow B$
- Si $target = 1$, on a $A \rightarrow B$
- $target = 0$ pour les autres cas⁴

Ce dernier point peut montrer les limites du modèle

2. kaggle.com/c/cause-effect-pairs

3. competitions.codalab.org/competitions/1381

4. Les cas de *cofounder* (existence d'une variable C causant A et B), *indépendance*, *cycle*, et de *contrainte*



Bibliographie

- [Arthur et Vassilvitskii, 2007] ARTHUR, D. et VASSILVITSKII, S. (2007). k-means++ : The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.
- [Datchary.C, 2011] DATCARY.C (2011). *La dispersion au travail*. Octarès Editions.
- [Fonollosa, 2016] FONOLLOSA, J. A. R. (2016). Conditional distribution variability measures for causality detection. *ArXiv e-prints*.
- [Granger, 1969] GRANGER, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.
- [Jones, 1972] JONES, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.
- [Lebart *et al.*, 2006] LEBART, L., MORINEAU, A. et PIRON, M. (2006). *Statistique exploratoire multidimensionnelle*. Dunod.
- [Lopez-Paz *et al.*, 2015] LOPEZ-PAZ, D., MUANDET, K., SCHÖLKOPF, B. et TOLSTIKHIN, I. (2015). Towards a Learning Theory of Cause-Effect Inference. *ArXiv e-prints*.
- [Meilă, 2006] MEILĂ, M. (2006). The uniqueness of a good optimum for k-means. In *Proceedings of the 23rd international conference on Machine learning*, pages 625–632. ACM.

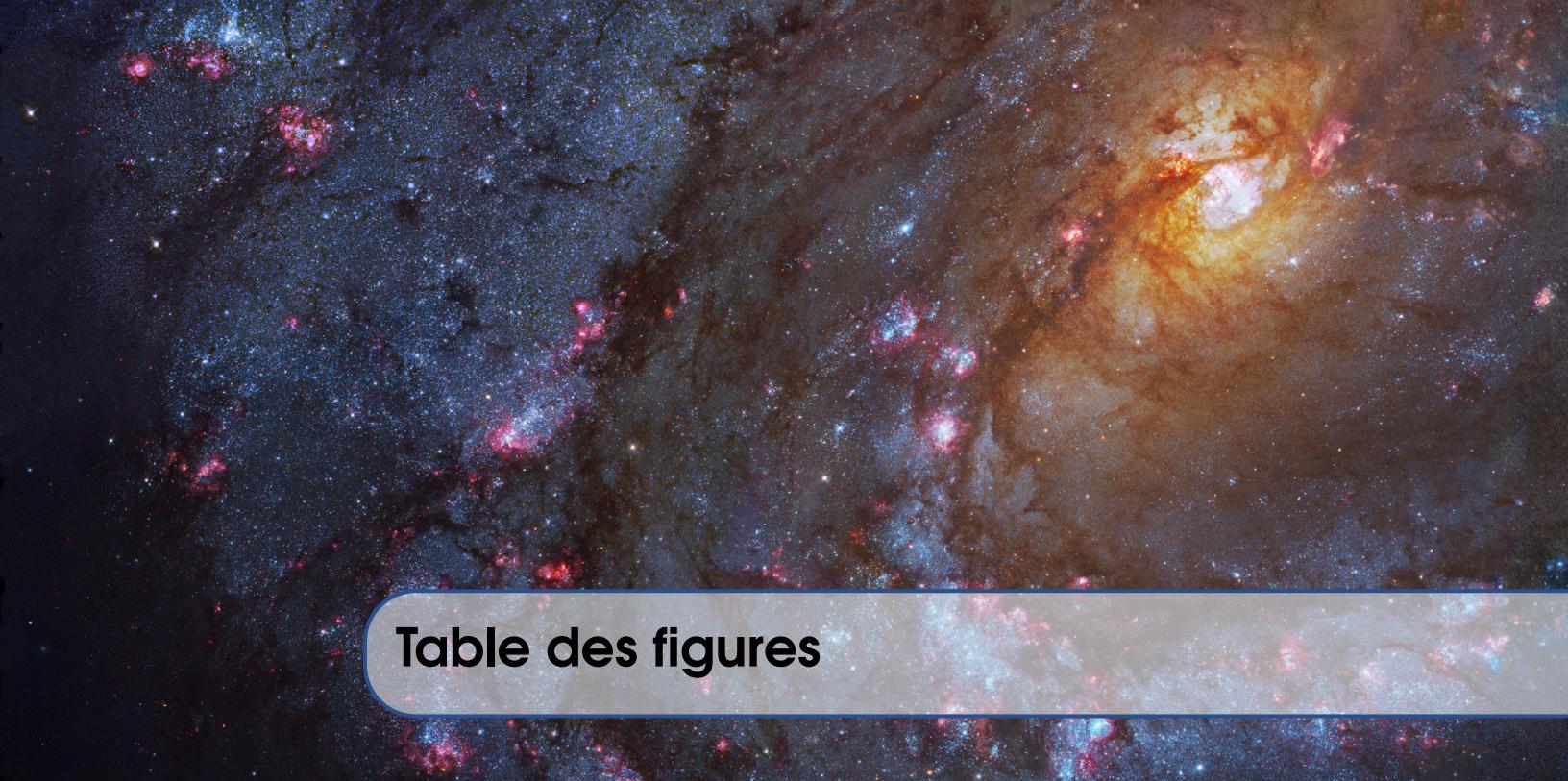


Table des figures

1.1	Répartition des types de questions en fonction des catégories	8
2.1	Données DARES, variables objectives : Spectre de la matrice de covariance.	12
2.2	Données DARES, variables subjectives : Spectre de la matrice de covariance.	12
2.3	Poids total des catégories de variables sur les nouveaux axes objectifs	13
2.4	Somme des valeurs absolues des poids des variables dans la définition de chaque axe subjectif, par catégorie de variable	13
2.5	Représentation en coordonnées parallèles des centres des clusters objectifs sur les axes de l'ACP	17
2.6	Valeurs V-test des clusters objectifs sur les codes NAF17	17
2.7	Représentation en coordonnées parallèles des centres des clusters subjectifs sur les axes de l'ACP	20
2.8	Valeurs V-test des clusters subjectifs sur les codes NAF17	20
2.9	Matrice de correspondance entre les clusters subjectifs et objectifs	22
2.10	Représentation en coordonnées parallèles du score d'autonomie en fonction des clusters. Par exemple, les individus à la fois dans le cluster objectif Santé (ligne la plus basse - jaune) et dans le cluster subjectif HEUR. (tout à gauche sur le graphique) ont un score d'autonomie moyen de 5,6.	25
3.1	Graphe causal simple sans et avec la relation indirecte	27
3.2	Graphe de causalité en se basant uniquement sur les résultats inférés par les paires deux à deux	27



Liste des tableaux

1.1	Catégories des questions du questionnaire de la Dares	7
2.1	Tableaux des principales contribution des variables pour les 8 premiers axes de l'ACP des variables objectives	14
2.2	Tableaux des principales contributions des variables pour les 5 premiers axes de l'ACP des variables subjectives	15
2.3	Identification des clusters objectifs	19
2.4	Identification des clusters subjectifs	21
2.5	Pondération des réponses aux questions pour le calcul du score d'autonomie	24