

Humanités causales

Analyse et causalités sur des données de qualité

de vie au travail

Diviyan Kalainathan (ISAE-ENSMA)

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 Démarche	7
1.1.1 Détermination des profils types	7
1.1.2 Causalité	7
2 Prétraitement des données	9
2.1 Présentation des données	9
2.2 Choix méthodologiques	9
3 Analyse descriptive des données	11
3.1 Réduction de la dimensionnalité des données	11
3.1.1 Principe	11
3.1.2 Obtention des nouveaux axes	12
3.1.3 Étude des axes	13
3.1.4 Interprétation des axes	13
3.2 Détermination des profils types	17
3.2.1 Méthode employée	17
3.2.2 Profils-types objectifs	18
3.2.3 Profils-types subjectifs	21
3.3 Correspondance entre clusters objectifs et subjectifs	23
3.3.1 Croisement des populations de clusters	23
3.3.2 Autonomie et clusters	26
3.4 Récapitulatif de l'analyse descriptive	27
4 Analyse causale	29
4.1 Motivation	29

4.2 Méthodologie	29
4.2.1 Coefficient de causalité	30
4.2.2 Hétérogénéité des données	30
4.2.3 Nécessité d'une déconvolution	30
4.3 Travaux futurs	31
Bibliographie	32
Table des figures	33
Liste des tableaux	34

REMERCIEMENTS

Je souhaite tout d'abord remercier Michèle Sebag, Phillippe Caillou, Isabelle Guyon et Paola Tubaro pour m'avoir accepté en tant que stagiaire à TAO, mais aussi pour m'avoir encadré et apporté de nombreux conseils qui m'ont permis de me développer et d'acquérir de nouvelles compétences.

Je remercie aussi Olivier Goudet, pour son esprit d'équipe et son aide précieuse. En travaillant aussi sur le projet AMIQAP il m'a apporté un autre point de vue à mon étude qui m'a été très bénéfique.

Je suis très reconnaissant envers les partenaires du projet AMIQAP, en particulier M.Thierry Weil et Émilie Bourdu de La Fabrique de l'industrie pour leur expertise dans le monde du travail.

Je souhaite également remercier toute l'équipe TAO pour son amabilité et sa convivialité.

PARTIE 1

INTRODUCTION

La qualité de vie au travail (QVT) a comme objectif de concilier les modalités de l'amélioration des conditions de travail et de vie pour les professionnels et la performance collective de l'entreprise, selon la Haute Autorité de Santé. De plus, la QVT a été identifiée comme un levier possible de compétitivité industrielle, selon une étude menée par *La Fabrique de l'industrie* [Bourdu *et al.*, 2016]. Toutefois, la notion de qualité de vie au travail est difficile à évaluer à cause de son caractère subjectif. Ainsi dans cette étude on souhaite étudier le lien entre travail et la satisfaction au travail pour faire apparaître la QVT, mais aussi les mécanismes de la QVT.

Cette étude fait partie du projet AMIQAP¹, qui a pour objet de définir une méthodologie opérationnelle pour analyser l'impact de la QVT sur la performance de l'entreprise, en caractérisant les relations entre d'une part les données liées à la qualité de la vie en entreprise et d'autre part les indicateurs de performance des entreprises. Les partenaires de ce projet sont :

- Michèle Sebag, Philippe Caillou, Paola Tubaro, Isabelle Guyon, Diviyan Kalainathan, Olivier Goudet : TAO, CNRS - INRIA - LRI, Univ. Paris-Sud, Univ. Paris-Saclay
- Jean-Luc Bazet & Ahmed Bounfour : RITM (Réseaux Innovation Territoires et Mondialisation), Université Paris-Sud
- Emilie Bourdu-Szwedek & Thierry Weil : La Fabrique de l'Industrie
- Valérie Fernandez & Valérie Beaudouin : SES, Telecom-ParisTech & Institut Interdisciplinaire de l'Innovation, CNRS UMR 9217

Nous disposons d'une quantité importante de données : les réponses de 33673 personnes sur un questionnaire effectué par la DARES² en collaboration avec l'INSEE³, 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 afin de pouvoir en tirer des interprétations sur les mécanismes à l'origine des différences entre la satisfaction au travail des enquêtés et leur situation.

Ce projet de fin d'études est un sujet multi-disciplinaire combinant sociologie et 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. À l'aide de ces réponses et l'analyse de celles-ci par le biais

1. Analyse multi-variée des impacts de la qualité de vie au travail sur la performance de l'entreprise
2. 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.
3. Institut national de la statistique et des études économiques collecte, produit, analyse et diffuse des informations sur l'économie et la société françaises

de techniques et d'algorithmes issus de la recherche en informatique, il s'agit d'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 de permettre de tester la validité des techniques sur un ensemble de données avec les connaissances des sociologues.

1.1 Démarche

Après une première phase de prétraitement des données, l'étude est divisée en deux : 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.

1.1.1 Détermination des profils types

Les données pré-traitées comportent un nombre important de variables, rendant l'étude complexe : regrouper les individus pour former des profils types nécessite de prendre en compte toutes les variables. La solution employée est l'analyse en composantes principales, qui 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 latent des axes, chaque individu est un vecteur de \mathbb{R}^d . On utilise la classification (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 de la mesure statistique valeur-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'OMS⁴ ; mais aussi étudier le croisement des populations entre les groupes objectifs et subjectifs est une analyse qui permet mettre en évidence le lien entre la situation de l'enquêté et son ressenti de sa situation. Cette analyse a été appuyée par le feedback d'experts sur l'interprétation des résultats d'analyse. La méthodologie est illustrée à la Fig.1.1.

1.1.2 Causalité

La deuxième partie de l'étude consiste à approfondir l'étude en étudiant 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 variables⁵. L'étude de la causalité, se basant sur des techniques complexes et variées, entre prédiction 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

4. cf www.euro.who.int.

5. 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.

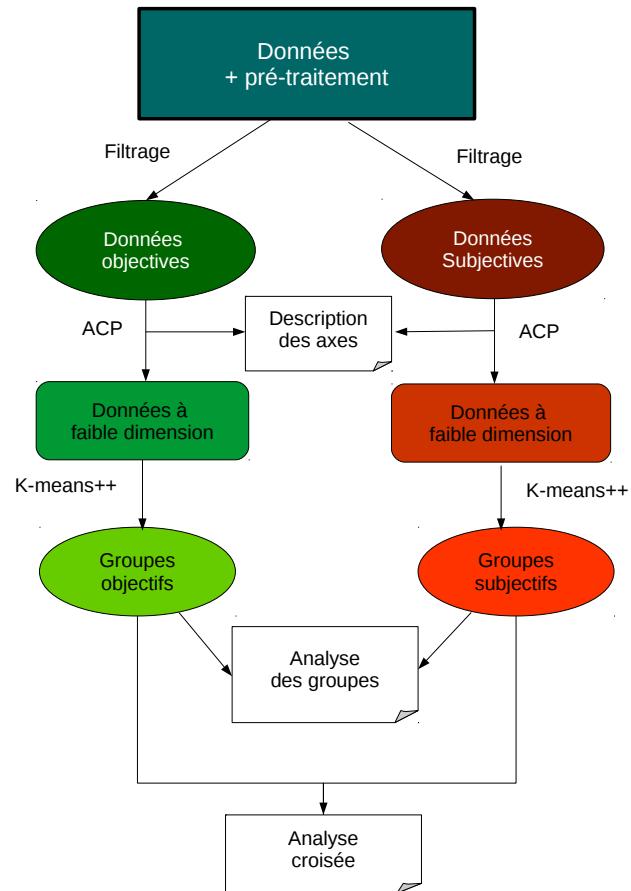


FIGURE 1.1: Méthodologie de l’analyse descriptive des données

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.

Le but étant de pouvoir faire des recommandations aux managers afin d’améliorer la qualité de vie au travail, les enjeux que représentent cette étude sont mis en valeur par le fait que la causalité à l’aide d’une approche d’analyse de données observationnelles n’a pas été étudiée jusqu’à aujourd’hui en sociologie. En effet, plusieurs méthodes permettent de déterminer les relations causales entre les différentes variables, mais ne sont pas applicables dans notre cas : premièrement, il s’agit d’effectuer des expériences contrôlées sur le système étudié en faisant varier des paramètres afin d’en déduire les liens causaux. Toutefois, cette approche ne peut pas être appliquée par faisabilité ou éthique (P.ex. licencier des employés pour analyser leur impact sur le marché du travail ou encore provoquer des accidents au travail pour étudier l’impact sur leur ressenti du travail). Deuxièmement, les expériences d’économie en laboratoire, permettent d’étudier la causalité, mais leur transposition au monde réel est difficile et est donc peu applicable. C’est donc une nouvelle approche d’étudier les liens causaux entre les variables avec seulement des données d’observation.

PARTIE 2

PRÉTRAITEMENT DES DONNÉES

2.1 Présentation des données

Les données se composent des réponses de 33673 personnes sur un questionnaire de 520 questions dans le cadre de l'enquête Conditions de travail 2013¹ réalisée par la DARES. 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 2.1). 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 2.1: Catégories des questions du questionnaire de la Dares

2.2 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)². Une question

1. cf. dares.travail-emploi.gouv.fr/

2. 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

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 question³). 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 continue⁴, nous avons choisi de partitionner les questions en deux groupes, correspondant respectivement aux éléments factuels (questions objectives) 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 2.1.

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.

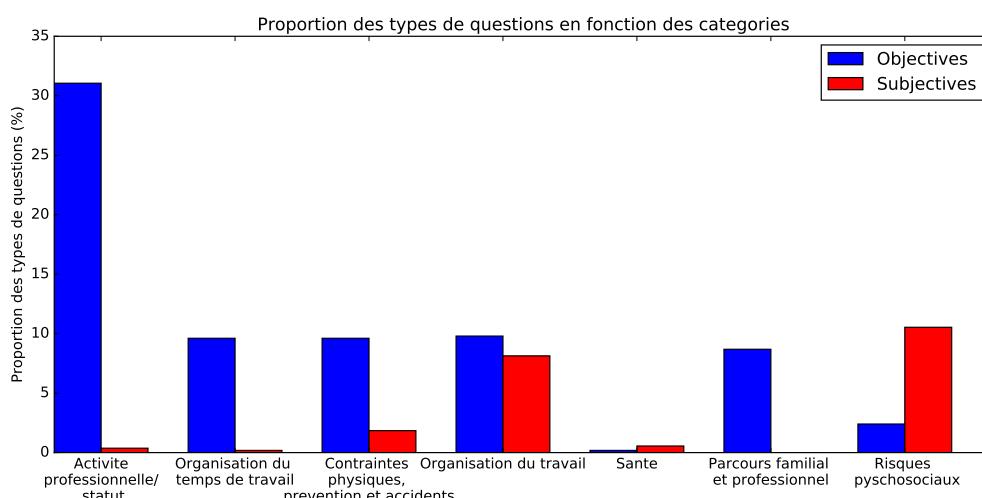


FIGURE 2.1: Répartition des types de questions en fonction des catégories

la valeur *opt* et faux sinon.

3. 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).

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

PARTIE 3

ANALYSE DESCRIPTIVE DES DONNÉES

Les données pré-traitées sont représentées par une matrice de $2463 \text{ variables} \times 31112 \text{ enquêtés}$. Un nombre de variables élevé ne permet pas d'analyser de manière simple les individus, dans le but d'établir des profils types. Notre démarche serait de réduire la dimensionnalité de nos données, puis d'établir des profils-types et enfin analyser les résultats obtenus (Fig.1.1).

3.1 Réduction de la dimensionnalité des données

3.1.1 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 3.1.1 — 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
    D ← Données $[i,:]$  // Vecteur de données pour la variable  $i$ 
    M $[i,:] \leftarrow \frac{D - \text{moyenne}(D)}{\text{variance}(D)}$ 
fin
M ← matrice_covariance(M)
W ← vecteurs_propres(M)
W ← W $[:p,:]$  // On tronque les  $p$  vecteurs propres qu'on a choisis
R ← M × Données $^T$ 
retourner 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 .

3.1.2 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.3.1a et Fig.3.2a. 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.3.1a. 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.

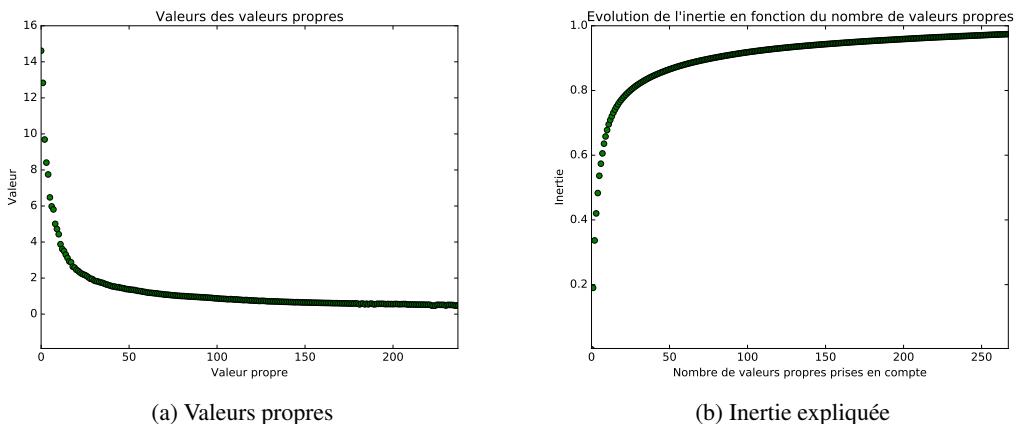


FIGURE 3.1: Données DARES, variables objectives : Spectre de la matrice de covariance.

ACP sur les variables subjectives

Dans le cas des variables subjectives (environ 20% des variables), le spectre est représenté Fig.3.2a. Le fait de retenir les 5 premiers vecteurs propres permet de capturer 80% de l'inertie des données.

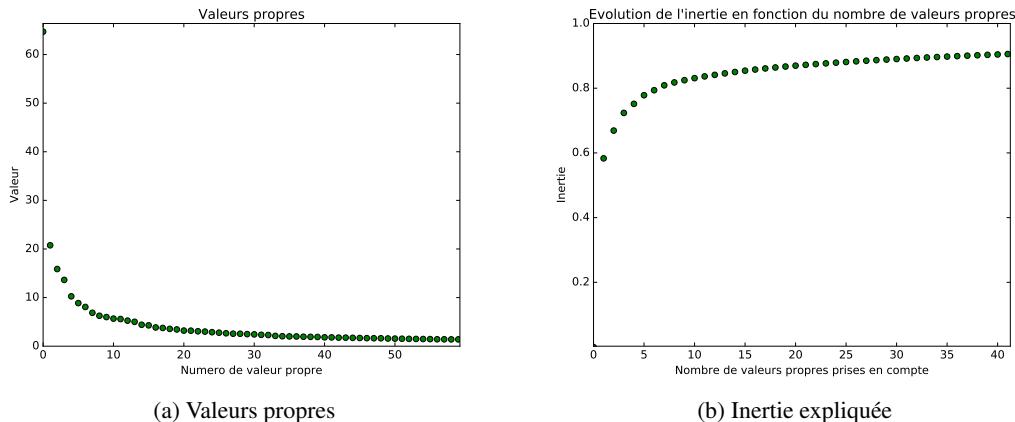


FIGURE 3.2: Données DARES, variables subjectives : Spectre de la matrice de covariance.

3.1.3 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 est un bon indicateur, qui est représenté sur la Fig.3.3 pour les axes objectifs et sur la Fig.3.4 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 Fig.2.1 ce qui réduit la variance expliquée par ces catégories. Ainsi, les poids des catégories de variables sur les nouveaux axes subjectifs ressemblent à un graphe complémentaire sauf pour la "Santé", car elle ne comporte que 4 questions dans le questionnaire. Cette première étude ne nous permet pas de caractériser les axes de l'ACP.

3.1.4 Interprétation des axes

Afin de pouvoir déterminer précisément la nature des axes, il est nécessaire d'étudier la composition de ceux-ci, c'est-à-dire la contribution des variables aux axes. Ainsi, observer les variables les plus corrélées à l'axe, positivement et négativement, nous permet d'inférer sur les caractéristiques de l'axe. Les résultats de l'analyse sont décrits aux Tables 3.1 et 3.2.

1. 8 axes objectifs et 5 axes subjectifs

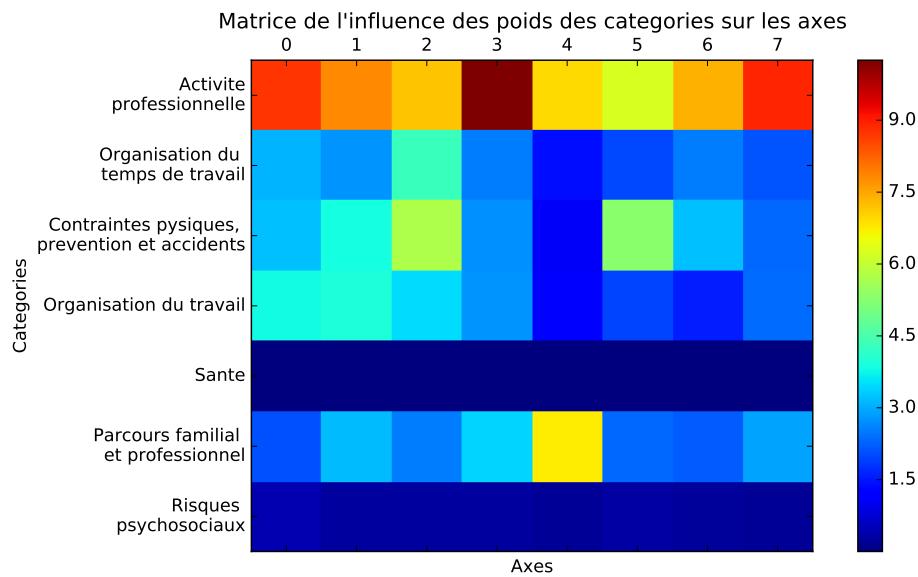


FIGURE 3.3: Poids total des catégories de variables sur les nouveaux axes objectifs

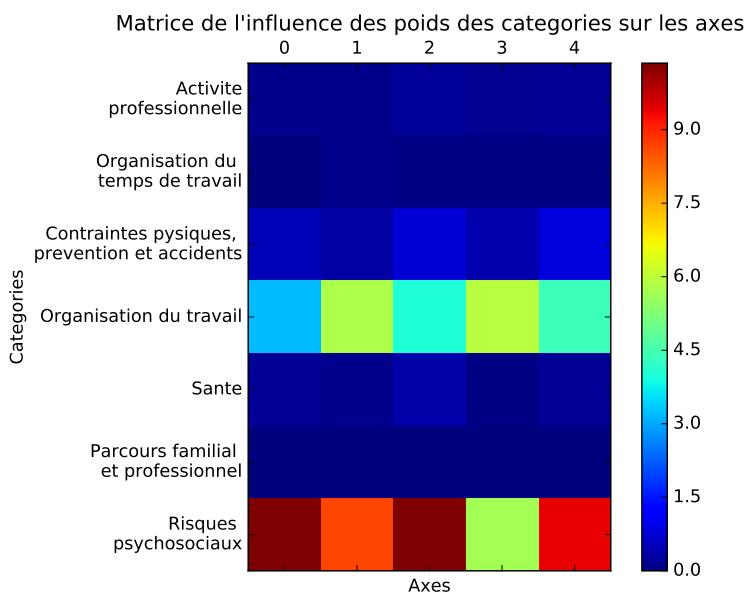


FIGURE 3.4: Somme des valeurs absolues des poids des variables dans la définition de chaque axe subjectif, par catégorie de variable

Axes de l'ACP objectif	Inertie	Variables corrélées positivement	Variables corrélées négativement
Axe 1 : Taille de l'entreprise employant l'enquêté	20%	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	16%	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	6%	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	6%	Salarié du privé Entreprise de grande taille Cadres d'entreprise	Employé d'administration publique, enseignement, santé, social Salarié de l'État
Axe 5 : Immigration	5%	Père/Mère nés en France Pas de lien à la migration	Mère/Père immigré(e) Immigré Naturalisé ou étranger
Axe 6 : Accidents du travail	5%	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	3%	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	3%	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 3.1: Tableaux des principales contribution des variables pour les 8 premiers axes de l'ACP des variables objectives

Axes de l'ACP subjectif	Inertie	Variables corrélées positivement	Variables corrélées négativement
Axe 1 : Risques psychosociaux	59%	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	8%	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	5%	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	4%	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	3%	Bonne humeur Frais et disposé, calme et tranquille Pas de pression Fier du travail	Pas de collègues Pas de supérieurs

TABLE 3.2: Tableaux des principales contributions des variables pour les 5 premiers axes de l'ACP des variables subjectives

3.2 Détermination des profils types

3.2.1 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). La projection est effectuée avec les $d^{\text{èmes}}$ vecteurs propres normalisés fournis par la PCA, en les multipliant à notre matrice de données, comme l'indique l'algorithme 1. On a utilisé les vecteurs propres normalisés pour ne pas associer plus d'importance aux premiers axes de données lors de notre classification. En effet, si l'on les vecteurs de norme la valeur propre, les premiers axes séparentont plus les données sur leur dimension associée, à cause de leur norme. Ainsi, les données seraient plus dispersées selon les premiers axes, qui détermineront de cette manière les groupes associés à la classification, alors que l'on souhaite que toutes les caractéristiques liées aux axes séparent les données avec une contribution égale.

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], est décrit à l'algorithme 2. On obtient ainsi 8 groupes objectifs et 6 groupes subjectifs³. 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 : cette baisse de performance peut être interprétée comme les clusters de situation et de ressenti qui se superposent, ce qui impacte la stabilité des groupes. Parmi 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. L'implémentation utilisée est celle de la librairie *Scikit-Learn*.

3. Le nombre de clusters est choisi à l'aide d'un compromis entre stabilité au sens de [Meilă, 2006], la dispersion minimale et un faible nombre de clusters

Algorithme 3.2.1 — Kmeans++

Données : Données de taille $m_{dimensions} \times n_{exemples}$,
hyper-paramètre k : nombre de clusters, r : nombre de runs

Résultat : Classification des données dans des groupes : Table $n_{exemples} \times 1$

```

pour  $j \leftarrow 1$  à  $r$  faire
    // Initialisation : KMeans++
    pour  $i \leftarrow 1$  à  $k$  faire
        cluster_center[i] ← rand((1,nexemples)) // Centre choisi  

aléatoirement
        pour  $x \leftarrow 1$  à  $n_{exemples}$  faire
             $| D(x) \leftarrow \|x - cluster\_center[i]\|_2$ 
        fin
        cluster_center[i] ← rand((1,nexemples),weights = D(x)^2 // Centre choisi  

aléatoirement, avec une distribution de probabilité  

pondérée de D(x)^2
    fin
    // K-means clustering
    tant que not converged faire
        pour  $x \leftarrow 1$  à  $n_{exemples}$  faire
             $| C[X] \leftarrow min_D(x,cluster\_center)$ 
            // Trouver/assigner le cluster le plus proche
        fin
        pour  $c \leftarrow cluster\_center[1]$  à  $cluster\_center[k]$  faire
             $| c \leftarrow \frac{1}{|S_i^{(cluster(c))}|} \sum_{x_j \in S_i^{(cluster(c))}} x_j$ 
            // Mise à jour des centres des clusters
        fin
    fin
     $R[j] \leftarrow C$ 
fin
retourner  $min_{dispersion}(R)$ 

```

3.2.2 Profils-types objectifs

Chaque cluster est interprété en fonction de son centre (représenté en coordonnées parallèles en fonction des axes de l'ACP⁴ à la Fig.3.5), 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 NAF17⁵ (Fig.3.6) 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 3.3.

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

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

5. Nomenclature d'Activités Française en 17 classes, cf. recherche-naf.insee.fr

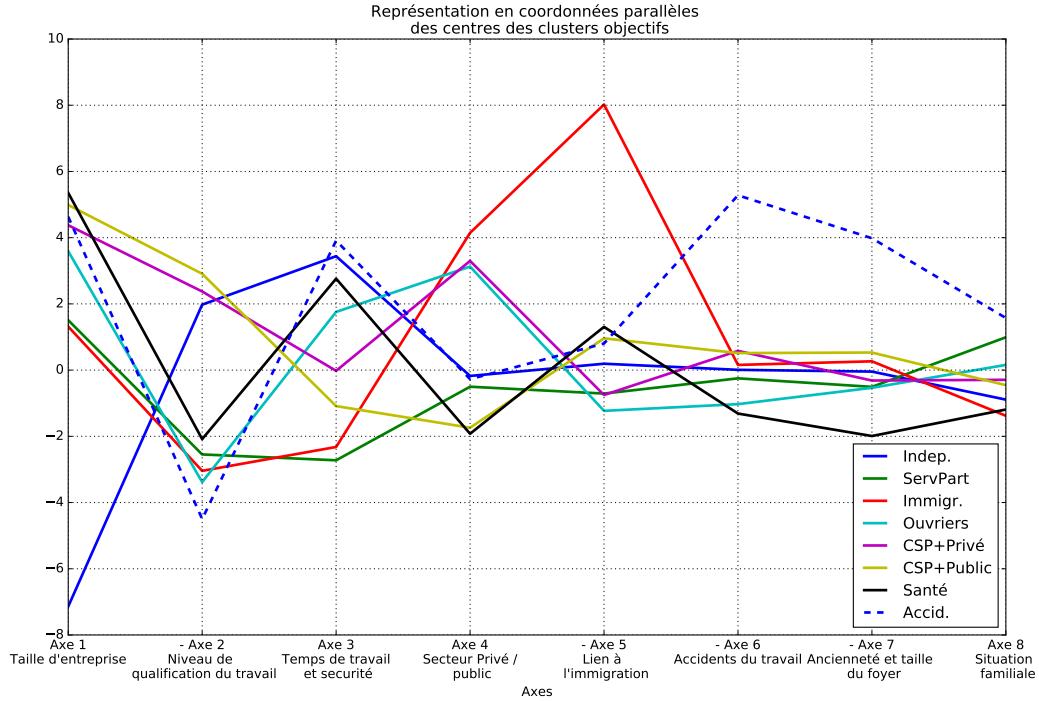


FIGURE 3.5: Représentation en coordonnées parallèles des centres des clusters objectifs sur les axes de l'ACP. Le groupe INDEP. est caractérisé par sa faible valeur sur l'axe 1, car les indépendants ont une taille d'entreprise peu importante.

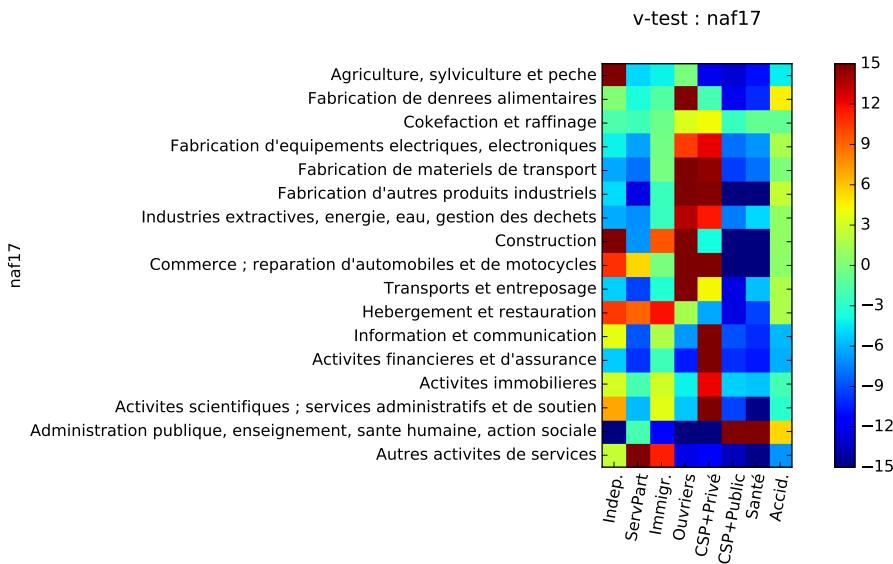


FIGURE 3.6: Valeurs V-test des clusters objectifs sur les codes NAF17

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 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.3.5, 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.3.5.

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 est caractérisé par 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 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)

Le profil type de ce groupe est similaire à celle du cluster 4, à l'exception du niveau de qualification qui est élevé et des secteurs d'activité (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 3.3: Identification des clusters objectifs

3.2.3 Profils-types 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.3.7 et de la répartition des v-test avec le code NAF17 à la Figure 3.8. Le résumé de ces analyses se retrouve à la Table 3.4.

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.

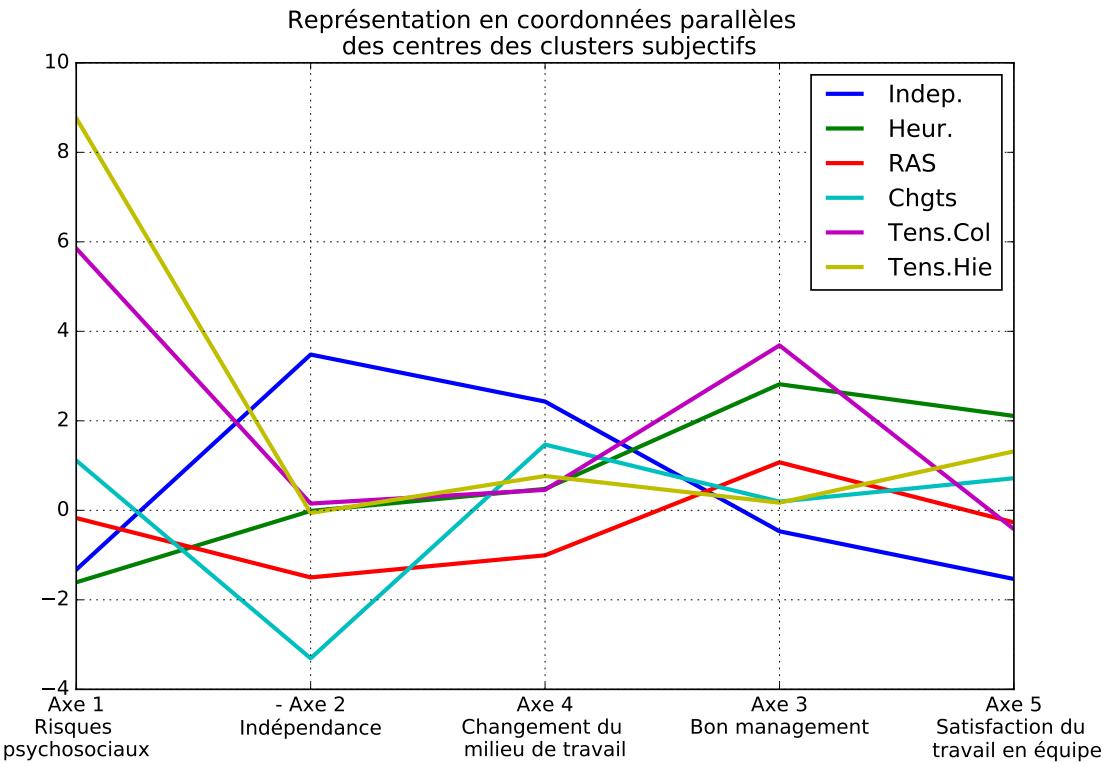


FIGURE 3.7: Représentation en coordonnées parallèles des centres des clusters subjectifs sur les axes de l'ACP. Ici, les enquêtés du groupe HEUR. ont une valeur assez élevée sur l'axe 5, traduisant une bonne satisfaction du travail en équipe.

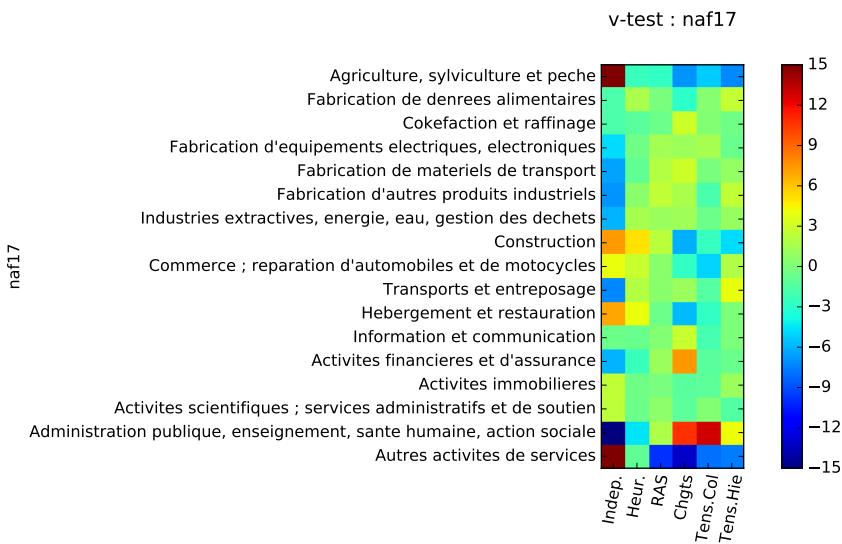


FIGURE 3.8: Valeurs V-test des clusters subjectifs sur les codes NAF17

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.

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 socio-professionnelle. 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 contre 1877€/mois en moyenne) 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 contre 1877€/mois en moyenne.

Groupe 5 : Tension avec les collègues (TENS.COL)

Cette communauté représente ceux qui sont satisfaits du management, mais ont des problèmes liés à leur collègues/équipe qui leur sont, selon leur point de vue, néfastes et sources 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 : Tension avec la hiérarchie (TENS.HIE)

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.

3.3 Correspondance entre clusters objectifs et subjectifs

3.3.1 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

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	TENS.COL	22.5%	Tension avec les collègues
6	TENS.HIE	13.0%	Tension avec la hiérarchie

TABLE 3.4: Identification des clusters subjectifs

de leur situation. La Figure 3.9 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 matrice de la Fig.3.9 vaut 1.

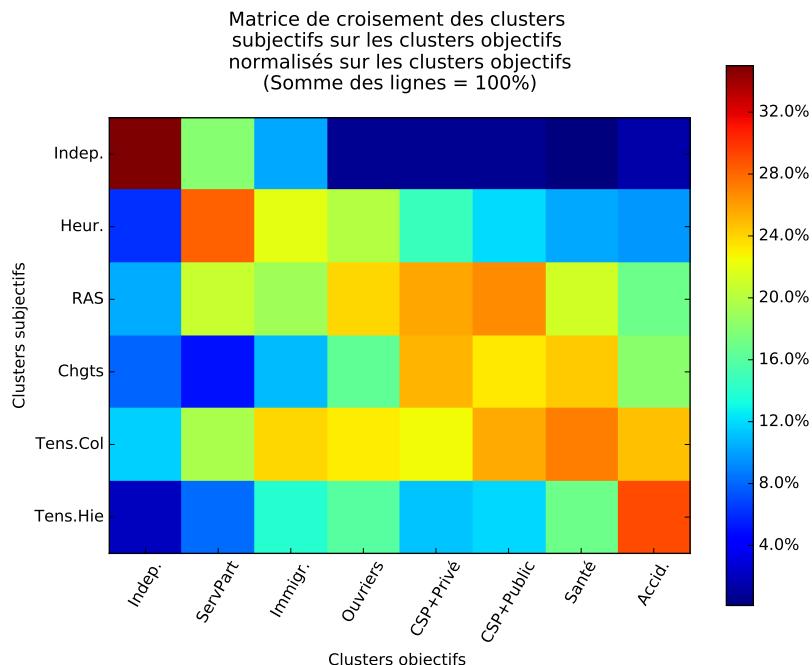


FIGURE 3.9: Correspondance entre les clusters objectifs (en colonnes) et les clusters subjectifs (en ligne) : La case (i, j) indique le pourcentage du cluster objectif i appartenant au cluster subjectif j . Par exemple, 30% des enquêtés du groupe ACCID. sont du groupe TENS.HIE et 25% des ACCID. sont dans le cluster TENS.COL .

Le premier élément qui apparaît de manière claire est le recouplement 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 SERV.PART est ventilé en quatre clusters subjectifs : INDEP. (17%), HEUR. (28%), RAS(20%), et TENS.COL (18%).

Les clusters objectifs OUVRIERS et IMMIGR. se répartissent de manière similaire entre les clusters subjectifs HEUR. & RAS, et TENS.COL. Les clusters CSP+PRIVÉ et CSP+PUBLIC de même se répartissent sur les clusters subjectifs RAS, CHGTS et TENS.COL. Enfin, le cluster ACCID. recoupe essentiellement les clusters subjectifs TENS.HIE (30%) et TENS.COL(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*. Par ailleurs, on remarque que les ACCID. ont une grande proportion de salariés soit en situation de tension(56%), une situation comparable au groupe SANTE ce qui traduit un environnement de travail qui peut perturber la productivité des employés. De plus, nous pouvons noter que les plus heureux au travail ne sont pas les groupes CSP+PRIVÉ ou CSP+PUBLIC (qui sont plutôt dans le groupe RAS), mais dans les clusters SERV.PART, IMMIGR. et OUVRIERS ; ce qui traduirait qu'une situation de travail qui paraît satisfaisante n'est pas celle qui mène au bonheur au travail. En revanche, la présence d'un environnement stressant et de tension 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].

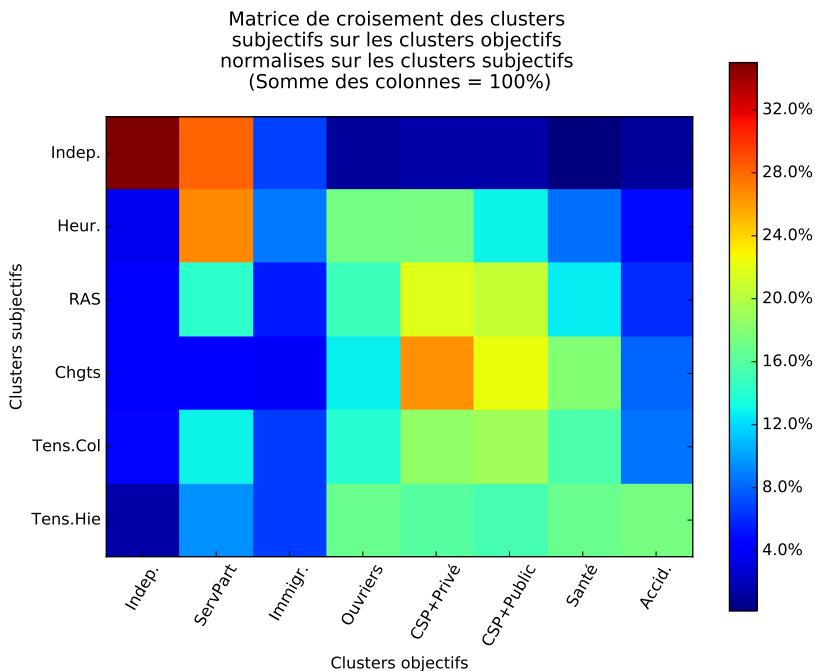


FIGURE 3.10: Correspondance entre les clusters objectifs (en colonnes) et les clusters subjectifs (en ligne) : La case (i, j) indique le pourcentage du cluster subjectif j appartenant au cluster objectif i . Par exemple, 25% des enquêtés du groupe CHGTS font partie du groupe CSP+PRIVÉ .

La Fig.3.10 permet d'ajouter des éléments pour nuancer les interprétations précédentes. En effet, en représentant la répartition des populations en fonction des clusters subjectifs, on remarque que les IMMIGR. sont sous-représentés dans tous les clusters : ils représentent en effet une population peu importante (6.2%), on note aussi que les HEUR.sont quand même assez représentés

dans les clusters CSP+PRIVÉ(17%) et CSP+PUBLIC(14%). Une grande proportion d'enquêtés ayant connu des changements dans l'environnement de travail font partie des CSP+PRIVÉ(26%) et CSP+PUBLIC(22%). Enfin, on remarque que les situations de TENS.HIE ou de CHGT sont plutôt réparties dans multiples groupes de situation à 16%.

3.3.2 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, un aspect de la QVT très étudié récemment. 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 données Table 3.5. 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

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

TABLE 3.5: Pondération des réponses aux questions pour le calcul du score d'autonomie

Les scores d'autonomie aux intersections des différents clusters sont représentés Figure 3.11. 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 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 TENS.HIE), 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 TENS.HIE et dans une moindre mesure TENS.COL (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).

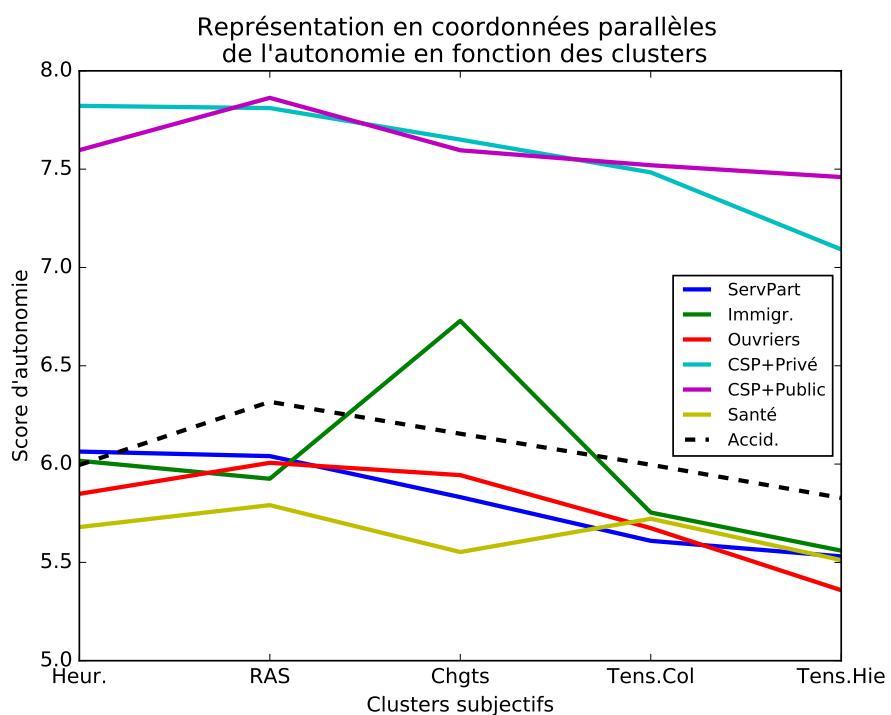


FIGURE 3.11: Représentation en coordonnées parallèles du score d'autonomie en fonction des clusters. Par exemple, les individus à la fois dans le cluster objectif Santé (ligne la plus basse - jaune) et dans le cluster subjectif HEUR. (tout à gauche sur le graphique) ont un score d'autonomie moyen de 5,6.

3.4 Récapitulatif de l'analyse descriptive

L'analyse descriptive nous a permis d'établir des profils-types à partir d'un volume important de données et hétérogènes. La méthodologie employée est intéressante dans le sens où tous les résultats ont été obtenus par des algorithmes traditionnels de regroupement et de réduction de dimension ; et que nous n'avons eu aucun contrôle sur la nature des résultats. Toutefois, nous avons quand même obtenu des résultats cohérents que ce soit en étude de situation ou de ressenti, sauf pour les groupes des indépendants qui ont été plutôt mis à l'écart à cause du questionnaire à branches et sur les multiples questions sur la hiérarchie et les collègues.

Les résultats que nous pouvons retenir de cette première partie sont le fait que l'autonomie est majoritairement supérieure pour les enquêtés qui sont plutôt satisfaits de leur travail comparé aux enquêtés victimes de malheur au travail(Fig.3.11), mais que les plus autonomes ne sont pas les plus heureux. De plus, on remarque que les groupes les plus heureux sont ceux ayant le moins de qualifications (Fig.3.9), que le secteur de la santé connaît une quantité importante d'enquêtés mécontents de leur travail, et que les accidentés au travail sont très mécontents de la hiérarchie. Ce dernier comportement est lié au manque de réaction de la hiérarchie face à l'accident qu'à subi l'enquêté.

Ainsi nous avons pu analyser de manière assez précise les données, mais cette première étude ne nous permet pas de déterminer les mécanismes à l'origine de la satisfaction au travail, c'est-à-dire si la situation détermine la satisfaction, ou si d'autres facteurs la déterminent.

PARTIE 4

ANALYSE CAUSALE

L'analyse descriptive nous permet d'identifier les phénomènes liés à la satisfaction au travail, mais ne nous permet pas de déterminer les causes derrière ces phénomènes [Pearl, 2000]. L'enjeu représenté par cette étude est important : elle nous permettrait de faire des recommandations aux managers afin de pouvoir améliorer la qualité de vie au travail.

4.1 Motivation

La corrélation caractérise le fait que deux variables ont des évolutions statistiques similaires, mais n'implique pas l'existence d'une relation de causalité entre les deux variables¹. La définition de la causalité pour cette étude est celle de [Statnikov, 2012] : "A est une cause de B ($A \rightarrow B$) si la loi de probabilité de B change pour une manipulation expérimentale de A."

Les approches classiques pour étudier la causalité sont tout d'abord l'expérience contrôlée² mais qui ont leur inconvénients : coûteux, non éthiques ou non réalisables, ou encore l'économie expérimentale, mais qui est dure à transposer à la réalité, mais aussi les expériences naturelles³ sur lesquelles les chercheurs n'ont aucun contrôle. Une nouvelle approche est d'inférer sur les relations causales avec seulement des données d'observations. Sur cette approche, plusieurs méthodes sont mises en évidence, telles que la classification avec les outils d'apprentissage machine ou encore une approche sur la complexité des modèles [Stegle *et al.*, 2010]. Afin de rassembler les connaissances et comparer les approches en inférence de causalité, plusieurs challenges ont été organisés par Isabelle Guyon. Nous allons nous baser sur les meilleurs algorithmes de ces compétitions pour lancer notre étude.

4.2 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

1. Telles que la corrélation entre les dépenses des États-Unis pour la science et le nombre de suicides

2. P.ex. les tests cliniques

3. P.ex. l'évolution du travail dans les restaurants Mc Donald's dans les différents états des États-Unis

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 fiabilité du graphe de causalité.

4.2.1 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*⁴ et *Codalab*⁵ 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 cas⁶

Ce dernier point peut montrer les limites du modèle : de nombreux cas ont été intégrés dans le cas $target = 0$, et l'on ne s'intéresse qu'aux relations de causalité directes.

4.2.2 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

4.2.3 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 4.1.

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

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

5. competitions.codalab.org/competitions/1381

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

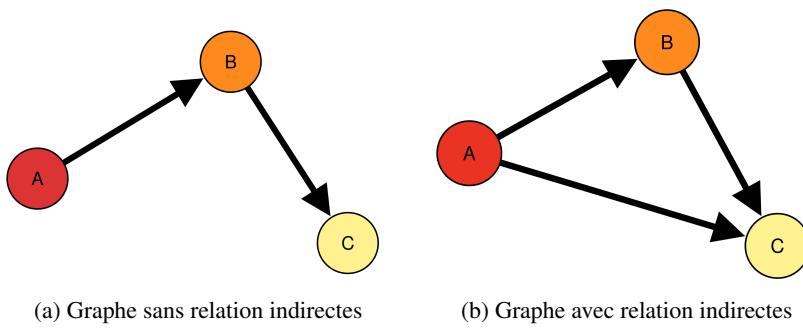


FIGURE 4.1: Graphe causal simple sans et avec la relation indirecte

est impossible d'interpréter les résultats sans appliquer une déconvolution, comme le montre la Figure 4.2. En effet, avec 470 variables⁷, il y a plus de 5000 liens entre toutes les variables.

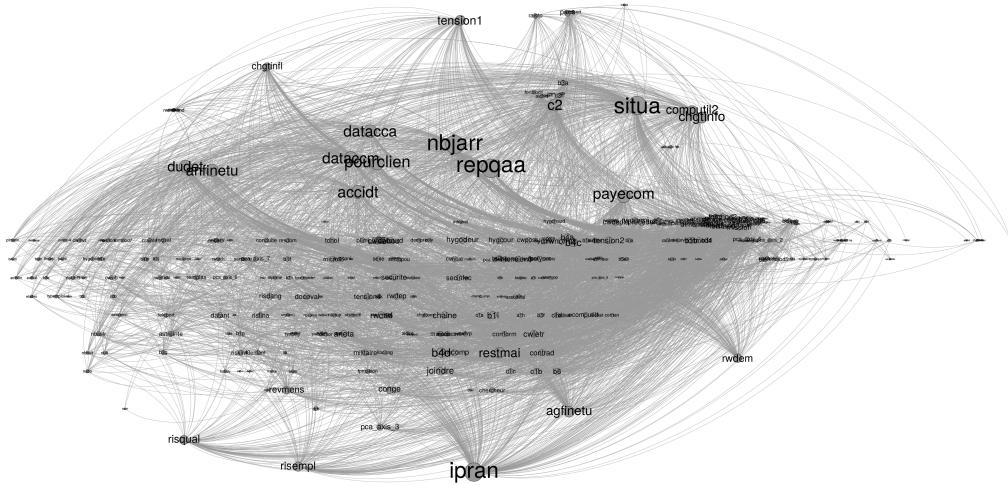


FIGURE 4.2: Graphe de causalité en se basant uniquement sur les résultats inférés par les paires deux à deux

4.3 Travaux futurs

Afin de pouvoir prendre en compte l'hétérogénéité des données, il s'agit de comparer un ensemble de critères d'indépendance. En créant un ensemble de paires de variables que l'on sait indépendantes⁸, on compte le nombre de fausses paires détectées par le modèle en fonction de la confiance du critère. De cette manière, on souhaite tester de multiples critères tels que (Pearson, χ^2 , FSIC, information mutuelle, Cramer's V) et ainsi déterminer le critère le plus adapté à nos données.

Pour inférer sur la structure du graphe ou encore vérifier la structure du graphe, nous pourra utiliser les travaux de [Aliferis *et al.*, 2010], qui cherchent à déterminer la couverture de Markov dans un graphe causal.

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

8. On génère des paires indépendantes à partir d'une vraie paire en mélangeant aléatoirement l'ordre des valeurs d'une des variables de la paire

BIBLIOGRAPHIE

- [Aliferis *et al.*, 2010] ALIFERIS, C. F., STATNIKOV, A., TSAMARDINOS, I., MANI, S. et KOUTSOUKOS, X. D. (2010). Local causal and markov blanket induction for causal discovery and feature selection for classification part i : Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(Jan):171–234.
- [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.
- [Bourdu *et al.*, 2016] BOURDU, E., PÉRETIÉ, M.-M. et RICHER, M. (2016). *La qualité de vie au travail : un levier de compétitivité*. La Fabrique de l’industrie.
- [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 PIROU, 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.
- [Pearl, 2000] PEARL, J. (2000). Causal inference without counterfactuals : Comment. *Journal of the American Statistical Association*, 95(450):428–431.
- [Statnikov, 2012] STATNIKOV, A. (2012). New methods for separating causes from effects in genomics data. *BMC genomics*, 13(8):1.
- [Stegle *et al.*, 2010] STEGLE, O., JANZING, D., ZHANG, K., MOOIJ, J. M. et SCHÖLKOPF, B. (2010). Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems*, pages 1687–1695.

TABLE DES FIGURES

1.1	Méthodologie de l'analyse descriptive des données	8
2.1	Répartition des types de questions en fonction des catégories	10
3.1	Données DARES, variables objectives : Spectre de la matrice de covariance.	12
3.2	Données DARES, variables subjectives : Spectre de la matrice de covariance.	13
3.3	Poids total des catégories de variables sur les nouveaux axes objectifs	14
3.4	Somme des valeurs absolues des poids des variables dans la définition de chaque axe subjectif, par catégorie de variable	14
3.5	Représentation en coordonnées parallèles des centres des clusters objectifs sur les axes de l'ACP. Le groupe INDEP. est caractérisé par sa faible valeur sur l'axe 1, car les indépendants ont une taille d'entreprise peu importante.	19
3.6	Valeurs V-test des clusters objectifs sur les codes NAF17	19
3.7	Représentation en coordonnées parallèles des centres des clusters subjectifs sur les axes de l'ACP. Ici, les enquêtés du groupe HEUR. ont une valeur assez élevée sur l'axe 5, traduisant une bonne satisfaction du travail en équipe.	22
3.8	Valeurs V-test des clusters subjectifs sur les codes NAF17	22
3.9	Correspondance entre les clusters objectifs (en colonnes) et les clusters subjectifs (en ligne) : La case (i, j) indique le pourcentage du cluster objectif i appartenant au cluster subjectif j . Par exemple, 30% des enquêtés du groupe ACCID. sont du groupe TENS.HIE et 25% des ACCID. sont dans le cluster TENS.COL	24
3.10	Correspondance entre les clusters objectifs (en colonnes) et les clusters subjectifs (en ligne) : La case (i, j) indique le pourcentage du cluster subjectif j appartenant au cluster objectif i . Par exemple, 25% des enquêtés du groupe CHGTS font partie du groupe CSP+PRIVÉ	25
3.11	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.	27
4.1	Graphe causal simple sans et avec la relation indirecte	31
4.2	Graphe de causalité en se basant uniquement sur les résultats inférés par les paires deux à deux	31

LISTE DES TABLEAUX

2.1	Catégories des questions du questionnaire de la Dares	9
3.1	Tableaux des principales contribution des variables pour les 8 premiers axes de l'ACP des variables objectives	15
3.2	Tableaux des principales contributions des variables pour les 5 premiers axes de l'ACP des variables subjectives	16
3.3	Identification des clusters objectifs	21
3.4	Identification des clusters subjectifs	24
3.5	Pondération des réponses aux questions pour le calcul du score d'autonomie	26