

Humanités causales

Analyse et causalités sur des données de qualité

de vie au travail

Diviyan Kalainathan (ISAE-ENSMA)

Encadré par M. Sebag, 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	24
3.3.1 Croisement des populations de clusters	24
3.3.2 Relations avec l'autonomie au travail	26
3.4 Récapitulatif de l'analyse descriptive	27
4 Analyse causale	29
4.1 Motivation	29

4.2	Méthodologie	30
4.2.1	Score de causalité	30
4.2.2	Score de corrélation et hétérogénéité des données	30
4.2.3	Phase d'élagage (déconvolution)	31
5	Perspectives	32
	Bibliographie	33
	Table des figures	34
	Liste des tableaux	35

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) est vue comme le résultat d'un compromis entre les conditions de travail et de vie pour les professionnels d'une part et la performance collective de l'entreprise d'autre part, selon la Haute Autorité de Santé. 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.

Notre stage s'inscrit dans le cadre du projet AMIQAP (*Analyse multi-variée des impacts de la qualité de vie au travail sur la performance de l'entreprise*), dont les partenaires 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

L'objectif général d'AMIQAP est 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. Dans ce cadre, l'objectif de notre stage concerne le lien entre les modalités du travail et la satisfaction au travail, dans le but de caractériser la QVT, ses modes et ses mécanismes.

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 aspects de la vie des enquêtés dans le cadre de l'enquête Conditions de travail 2013. Nous allons donc étudier ces données pour identifier et interpréter les relations entre la satisfaction au travail des enquêtés, et leur situation.

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

Ce projet de fin d'études est un sujet pluridisciplinaire, au confluent de la science des données et de la sociologie. Un aspect essentiel du travail consiste à proposer un modèle des données disponibles, qui fournit des éléments de réponse à notre problématique ; *in fine*, l'objectif est de constituer une méthodologie opérationnelle et statistiquement bien fondée pour analyser l'impact de la QVT sur la performance de l'entreprise. Mais un autre aspect consiste à valider cette méthodologie et les modèles proposés, en les confrontant aux connaissances et aux attendus des sociologues.

1.1 Démarche

Une première phase concerne le prétraitement des données. La seconde phase effectue une analyse descriptive approfondie des données, en distinguant les variables objectives et subjectives du questionnaire. La troisième phase enfin s'intéresse à l'analyse causale et aux relations d'implication entre les variables du questionnaire. L'ensemble des algorithmes de traitement est disponible à github.com/Diviyan-Kalainathan/causal-humans

1.1.1 Détermination des profils types

Les données 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, définies 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 en fonction de 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). Des clusters sont construits en considérant indépendamment les variables objectives (la situation de l'enquêté) et les variables subjectives (le ressenti). Les clusters sont étudiés par rapport à des variables choisies, telles que le revenu ou le score de bien-être défini par l'OMS³. Nous étudierons aussi les liens entre situation des enquêtés et ressenti, par analyse croisée des clusters objectifs et subjectifs. La méthodologie est illustrée à la Fig.1.1. Cette étude et l'interprétation des résultats ont bénéficié du retour des partenaires d'AMIQAP.

1.1.2 Causalité

La deuxième partie de l'étude s'intéresse aux relations causales entre variables, au sens de la causalité de Granger [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 uniquement la "ressemblance entre deux courbes", et ne permet pas de conclure à l'existence d'un réel lien entre les deux variables⁴. L'étude de la causalité se fonde sur des techniques sophistiquées allant de la prédiction en apprentissage statistique à

3. cf www.euro.who.int.

4. 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 liées causalement.

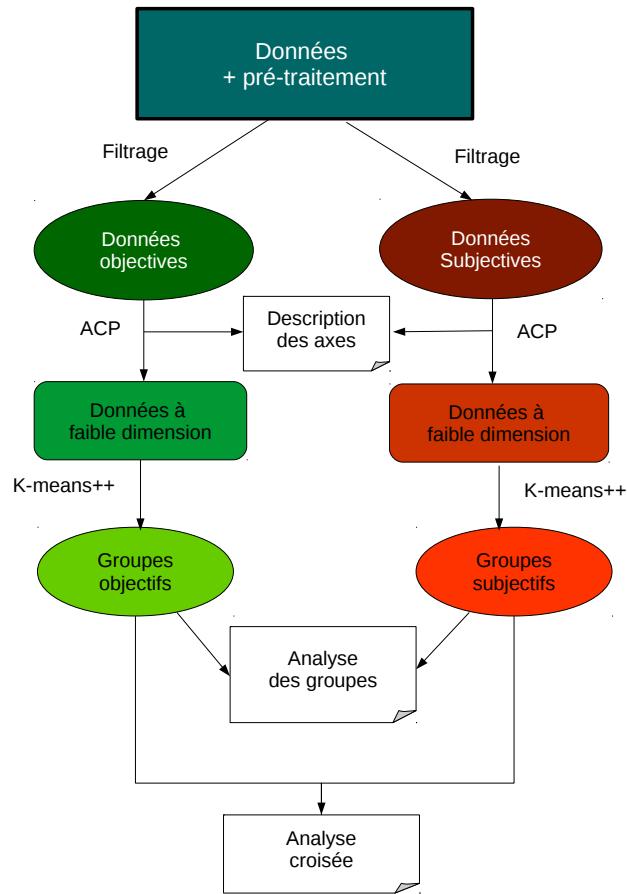


FIGURE 1.1: Méthodologie de l’analyse descriptive des données

l’inférence des distributions de probabilités jointes ; elle permet de déterminer la présence ou non d’une relation de causalité, et le sens de cette relation si elle existe.

Cette partie est la plus novatrice de notre étude. Elle se justifie dans la mesure où l’un des objectifs finaux consiste à émettre des recommandations aux managers afin d’améliorer la qualité de vie au travail de leurs employés. Or seul un modèle causal permet de fonder des recommandations. Cette étude et la recherche de relations causales fondées sur des données observationnelles est aussi novatrice du point de vue sociologique. En effet, plusieurs méthodes utilisées en sociologie permettent effectivement de déterminer les relations causales entre différentes variables ; cependant ces méthodes ne sont pas applicables dans notre cas. Ainsi une première méthode consiste à 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 faute de faisabilité ou pour des raisons éthiques (par exemple, 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). Une seconde approche est celle des expériences d’économie en laboratoire ; cependant si celles-ci permettent d’étudier la causalité, leur transposition au monde réel est difficile.

In fine, notre objectif est d’établir un graphe (aussi complet et fiable que possible) des variables et de leurs liens causaux, afin de comprendre les phénomènes moteurs observés par le questionnaire et commandant le bien-être au travail.

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. Seuls les 31112 enquêtés actifs, i.e. occupant un emploi à la date du questionnaire, seront considérés dans la suite de ce document. Les questions portant sur les aspects de la vie de l'enquêté sont regroupées en 7 rubriques (Table 2.1). Notons que les données issues du questionnaire sont hétérogènes, répondant à des questions numériques (*Quel est le montant de votre revenu ?*) ou catégorielles (questions à choix multiples, QCM). Les données comportent aussi une quantité importante de données manquantes (21%) ; d'une part les enquêtés avaient le choix de refuser de répondre à une question, ou de dire qu'ils ne connaissaient pas la réponse ; d'autre part le questionnaire est à multiples branchements.

Nom	Nombre de questions	% de questions objectives
1. Activité professionnelle/Statut	102	100%
2. Organisation du temps de travail	116	97.3%
3. Contraintes physiques, prévention et accidents	63	84.1%
4. Organisation du travail	97	54.6%
5. Santé	4	25%
6. Parcours familial et professionnel	64	79.7%
7. Auto-questionnaire sur les risques psychosociaux	70	18.6%

TABLE 2.1: Questionnaire de la Dares : catégories de questions, nombre de questions par catégorie, fraction de questions objectives par catégorie (voir texte).

2.2 Choix méthodologiques

Dans un premier temps, les variables sont recodées pour éviter les biais sur les variables catégorielles². Une question catégorielle comprenant N options est ainsi représentée par $N + 1$

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

2. Une option codée '3' (= pays du Maghreb) n'a pas de relation d'ordre avec 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

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 la nature catégorielle ou continue des questions (qui n'intervient que dans la phase de pré-traitement), 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 Table 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.

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).

PARTIE 3

ANALYSE DESCRIPTIVE DES DONNÉES

Les données pré-traitées sont représentées par une matrice de 2463 variables \times 31112 enquêtés. Le nombre élevé de variables ne permet pas d'analyser de manière simple les individus. Notre démarche sera 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. Les valeurs propres associées à ces nouvelles variables permettent de mettre en évidence la dimension intrinsèque des données ; en pratique, on retient les premières variables ACP, appelées axes dans la suite, qui capturent une fraction convenable de l'inertie (la variance) des données. Ces nouvelles variables, en nombre réduit, permettent de représenter les individus dans un espace intelligible et de remédier à la redondance des variables initiales. 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
    |    $\mathbf{D} \leftarrow$  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
 $\mathbf{M} \leftarrow \text{matrice\_covariance}(\mathbf{M})$ 
 $\mathbf{W} \leftarrow \text{vecteurs\_propres}(\mathbf{M})$ 
 $\mathbf{W} \leftarrow \mathbf{W}[:,p,:]$  // On tronque aux  $p$  premiers vecteurs propres
 $\mathbf{R} \leftarrow \mathbf{M} \times \text{Données}^T$ 
retourner  $\mathbf{R}$ 
```

Le choix de p réalise un compromis entre l'inertie expliquée et le bruit. En effet, la perte d'information diminue quand on rajoute des dimensions ; mais en contrepartie le bruit augmente.

Cette approche doit être adaptée toutefois compte tenu des questions catégorielles : en effet, les variables booléennes représentant les différentes options des questions catégorielles ont une variance très faible (voire nulle si cette option n'a jamais été choisie). La normalisation peut donc conduire de telles variables à avoir une grande amplitude, faussant 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 :

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

avec les notations de l'algorithme 2.1.1 ; en ayant de plus $\text{occurences}(\mathbf{D}_i) = \text{somme}(\mathbf{D})$ 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.

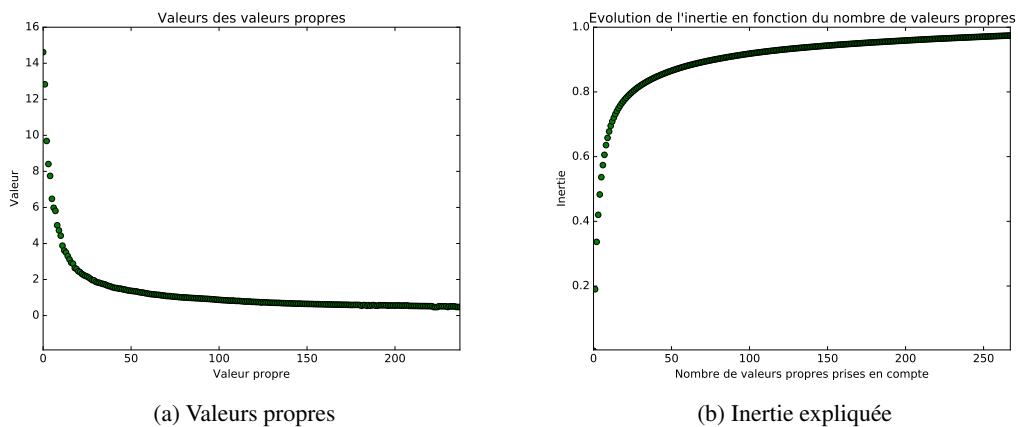


FIGURE 3.1: Données DARES, variables objectives : Spectre de la matrice de covariance.

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.

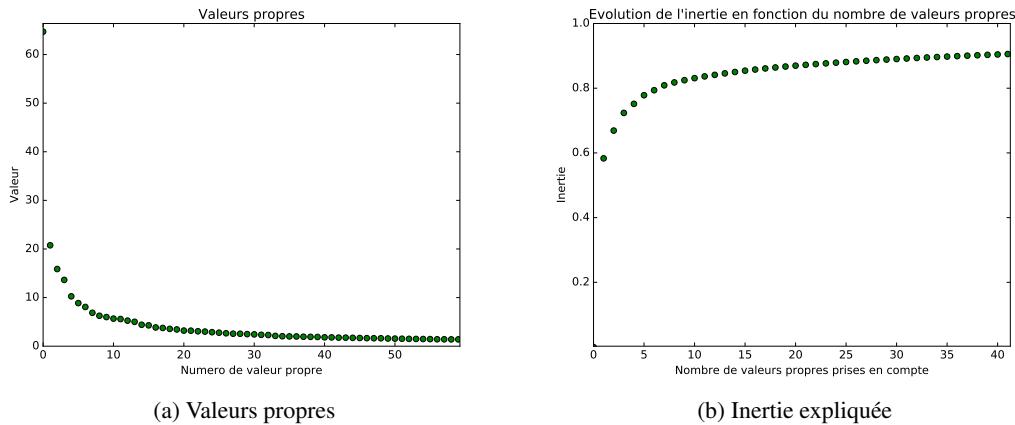


FIGURE 3.2: Données DARES, variables subjectives : Spectre de la matrice de covariance.

3.1.3 Étude des axes

13 nouvelles variables¹ sont ainsi obtenues, dont nous allons étudier les caractéristiques. Chacune de ces nouvelles variables est une combinaison linéaire des variables initiales. Une première interprétation des axes est faite en considérant les poids associés des variables initiales (Fig.3.3 pour les axes objectifs et Fig.3.4 pour les axes subjectifs).

Les risques psychosociaux, l'organisation du travail et la santé n'apparaissent pas clairement dans les axes objectifs : ceci est du au faible nombre de questions objectives dans ces catégories ce qui réduit la variance expliquée par ces catégories. Ainsi, ces figures ne nous permettent pas d'identifier précisément la nature des axes ; mais nous permettent toutefois de noter la provenance des poids des variables composant les différents axes de l'ACP.

3.1.4 Interprétation des axes

Une seconde interprétation des axes est effectuée en identifiant les variables les plus corrélées à l'axe, positivement et négativement (Tables 3.1 et 3.2).

1. 8 axes objectifs et 5 axes subjectifs

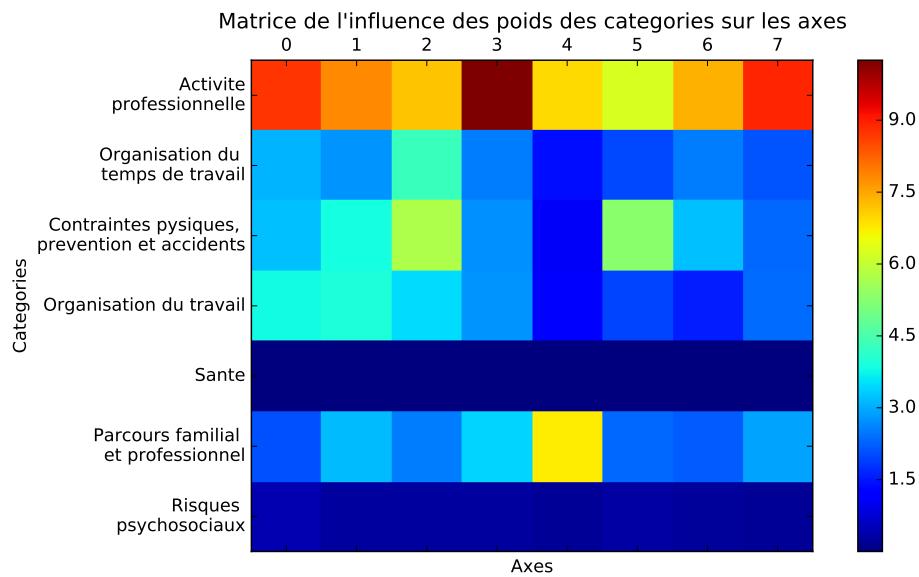


FIGURE 3.3: Poids total des catégories de variables sur les nouveaux axes objectifs

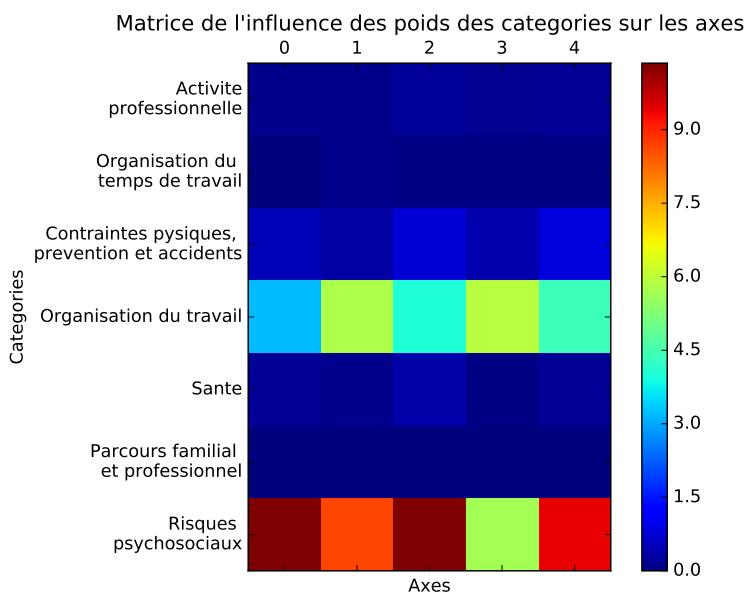


FIGURE 3.4: Somme des valeurs absolues des poids des variables dans la définition de chaque axe subjectif, par catégorie de variable

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 sur les $d^{\text{ièmes}}$ vecteurs propres normés de l'ACP (par multiplication avec la matrice de données initiales, comme indiqué par l'algorithme 1). On a utilisé les vecteurs propres normés (sans faire intervenir la valeur propre associée dans la distance) pour que 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 euclidienne 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 au sens de [Meilă, 2006], par rapport aux clusters obtenus à partir de l'ensemble des variables ; ceci est interprété comme le résultat des interférences entre situation objective et ressenti. Parmi nos outils d'analyse nous avons utilisé la valeur-test (v-test) de [Lebart *et al.*, 2006], définie 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
         $CC[i] \leftarrow \text{Données}[rand((1, n_{exemples}))]$  // Centre choisi aléatoirement
        // CC : Centres des clusters
        pour  $x \leftarrow 1$  à  $n_{exemples}$  faire
            |  $D(x) \leftarrow \|x - CC[i]\|_2$ 
        fin
         $CC[i] \leftarrow rand((1, n_{exemples}), weights = D(x)^2)$  // 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
            |  $T[X] \leftarrow \min_{c \in CC}(x, c)$ 
            // Trouver/assigner le cluster le plus proche
        fin
        pour  $l \leftarrow 1$  à  $k$  faire
            |  $CC[l] \leftarrow \frac{1}{|S_i^{(l)}|} \sum_{x_j \in S_i^{(l)}} x_j$ 
            // Mise à jour des centres des clusters
        fin
    fin
     $R[j] \leftarrow T$ 
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.

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

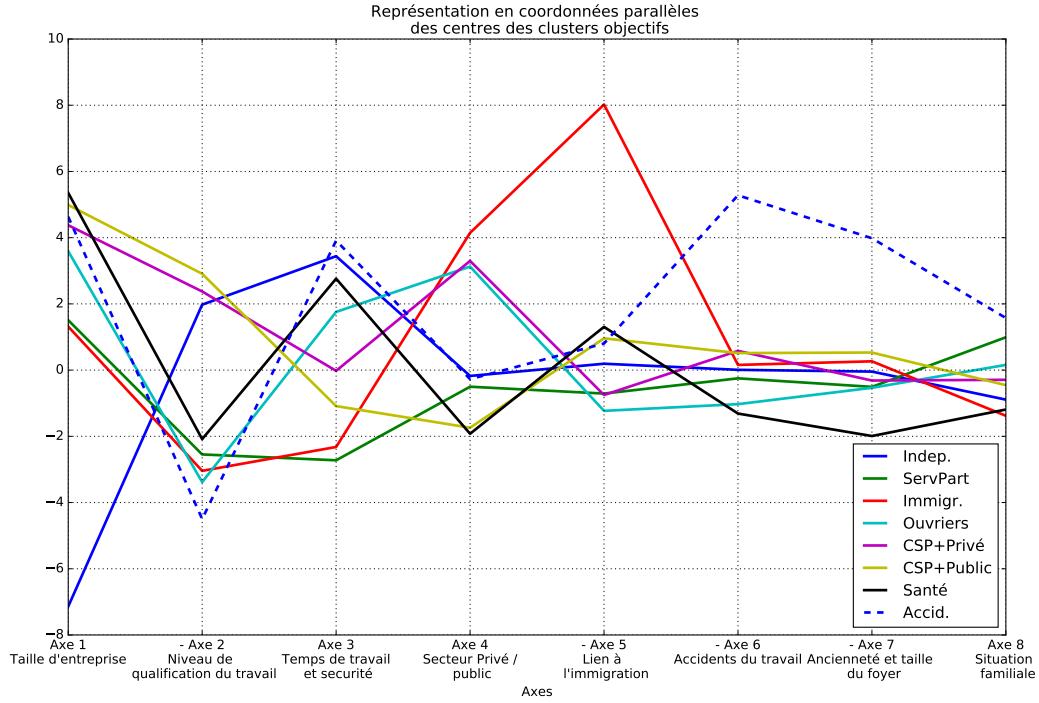


FIGURE 3.5: Représentation en coordonnées parallèles des centres des clusters objectifs sur les axes de l'ACP. Le groupe INDEP. est caractérisé par sa faible valeur sur l'axe 1, car les indépendants ont une taille d'entreprise peu importante.

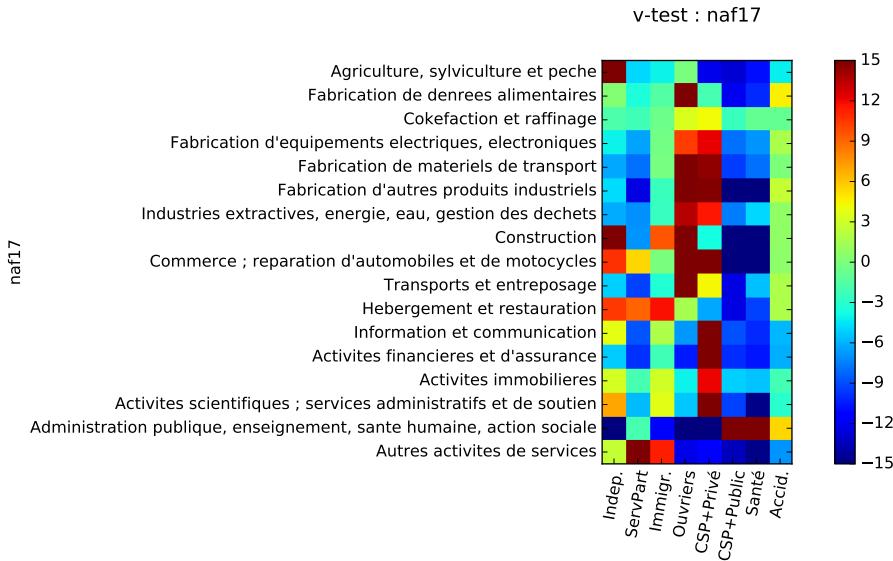


FIGURE 3.6: Valeurs V-test des clusters objectifs sur les codes NAF17

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 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. D'après les moyennes calculées sur le cluster, ils travaillent principalement dans le secteur privé (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 la pénibilité des tâches 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)

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	Abréviaction	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

Nous analysons les clusters subjectifs de la même manière, avant de mettre en relation 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

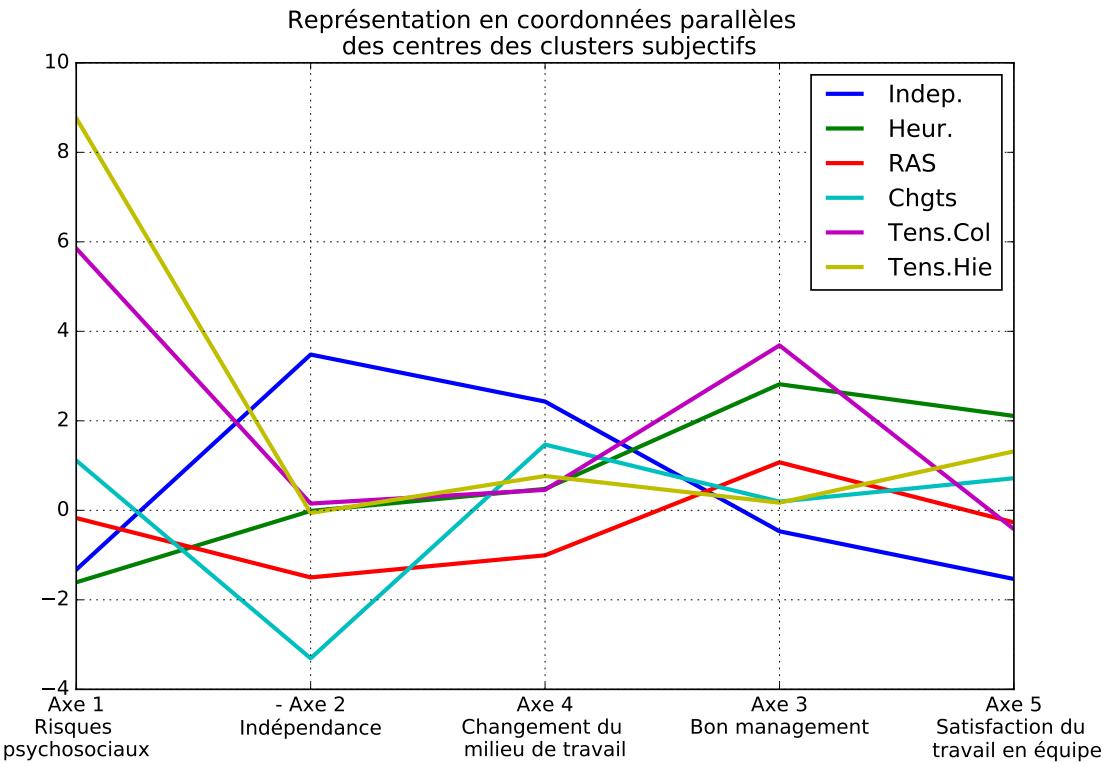


FIGURE 3.7: Représentation en coordonnées parallèles des centres des clusters subjectifs sur les axes de l'ACP. Ici, les enquêtés du groupe HEUR. ont une valeur assez élevée sur l'axe 5, traduisant une bonne satisfaction du travail en équipe.

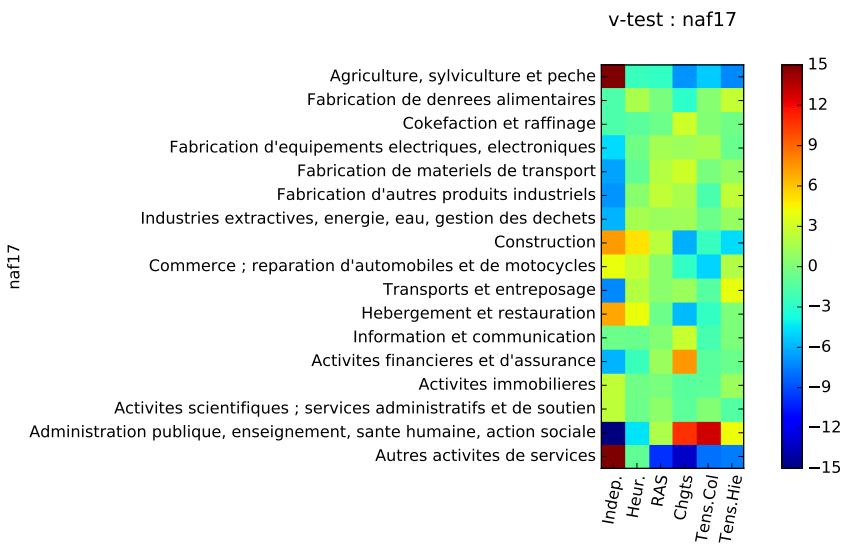


FIGURE 3.8: Valeurs V-test des clusters subjectifs sur les codes NAF17

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 ne 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 de toutes les catégories socioprofessionnelles. Ce cluster représente donc les personnes qui n'ont rien à signaler de particulier et sont 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)

Ce cluster se caractérise par le fait que ses membres proviennent de grandes organisations, et qu'ils ont 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.

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

3.3 Correspondance entre clusters objectifs et subjectifs

3.3.1 Croisement des populations de clusters

L'étape suivante consiste à mettre en relation la situation professionnelle des enquêtés et le ressenti de leur situation. La Figure 3.9 représente le recouplement des deux ensembles de clusters ; plus précisément, la case $M_{i,j}$ indique la fraction des personnes appartenant au cluster objectif i , qui appartient au cluster subjectif j :

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

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.3.9 vaut 1.

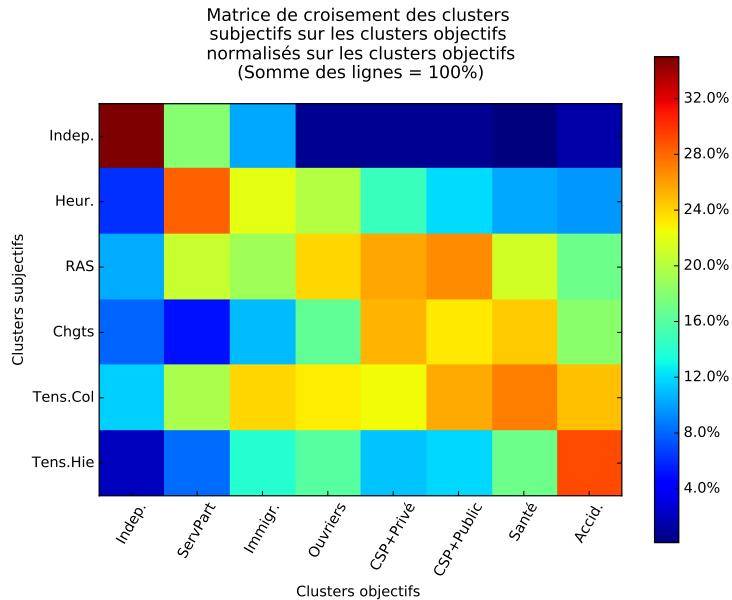


FIGURE 3.9: Correspondance entre les clusters objectifs (en colonnes) et les clusters subjectifs (en ligne) : La case (i, j) indique le pourcentage du cluster objectif i appartenant au cluster subjectif j . Par exemple, 30% des enquêtés du groupe ACCID. sont du groupe TENS.HIE et 25% des ACCID. sont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

Ce recouplement montre les faits suivants. En premier lieu, la spécificité des indépendants est attribuée au 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 en situation de tension (56%, une situation comparable à celle du groupe SANTE), ce qui traduit un environnement de travail qui peut perturber la productivité des employés. De manière surprenante (pour nous), les personnes les plus satisfaites au travail ne sont pas dans les clusters CSP+PRIVÉ ou CSP+PUBLIC (qui sont plutôt dans le groupe RAS), mais dans les clusters SERV PART, IMMIGR. et OUVRIERS ; ce qui confirme la complexité de la notion de satisfaction au travail.

Enfin, 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].

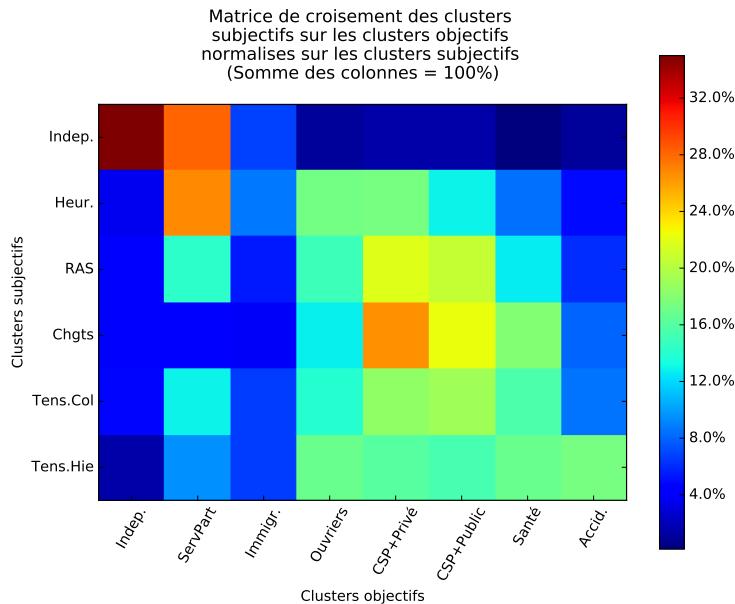


FIGURE 3.10: Correspondance entre les clusters objectifs (en colonnes) et les clusters subjectifs (en ligne) : La case (i, j) indique le pourcentage du cluster subjectif j appartenant au cluster objectif i . Par exemple, 25% des enquêtés du groupe CHGTS font partie du groupe CSP+PRIVÉ .

La Fig.3.10 permet de nuancer les remarques 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 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 CHGTS sont plutôt réparties dans multiples groupes de situation à 16%.

3.3.2 Relations avec l'autonomie au travail

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 récemment⁶. Pour réaliser cette étude, nous avons défini manuellement un score d'autonomie, comme une somme pondérée des réponses à 4 questions (détalée en 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

Les scores d'autonomie des clusters objectifs et subjectifs sont représentés Figure 3.11. Les clusters d'indépendants sont omis 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 clusters 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 SANTE est quant à lui le groupe avec l'autonomie la plus faible.
- En étudiant le lien avec les clusters subjectifs (ordonnés selon une QVT approximativement décroissante de HEUR. à TENS.HIE), l'autonomie apparaît comme décroissante pour tous

6. cf. anact.fr/

	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 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 caractérisé par l'autonomie la plus élevée (qui est souvent atteinte par RAS).

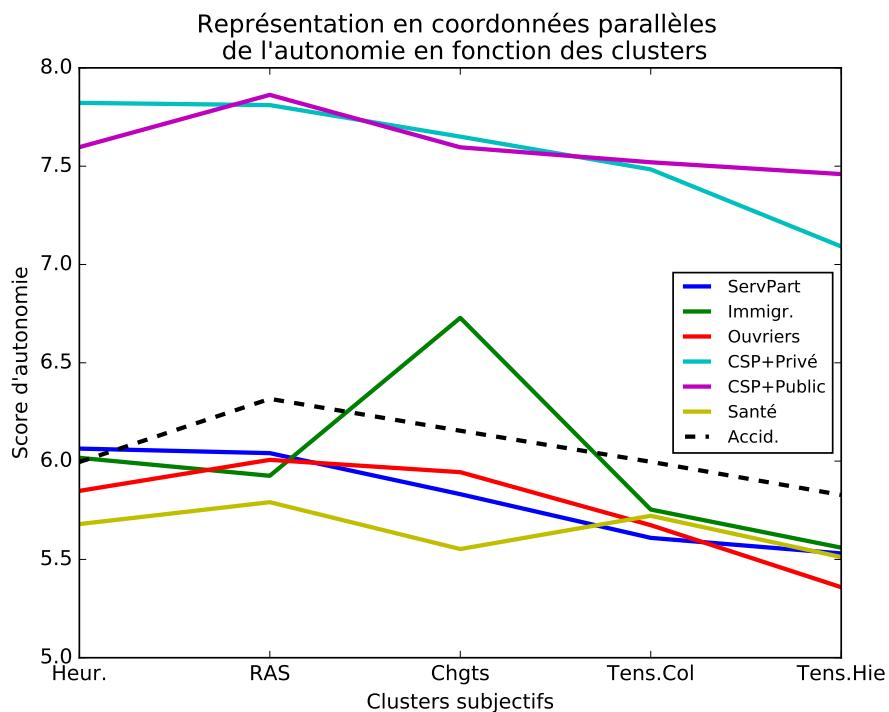


FIGURE 3.11: Représentation en coordonnées parallèles du score d'autonomie en fonction des clusters. Par exemple, les individus à la fois dans le cluster objectif Santé (ligne la plus basse - jaune) et dans le cluster subjectif HEUR. (tout à gauche sur le graphique) ont un score d'autonomie moyen de 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 hétérogènes. La méthodologie employée est intéressante dans le sens où tous les résultats ont été obtenus par des algorithmes classiques (clustering et réduction de dimension),

sans a priori sur la nature des résultats. Cependant les résultats obtenus sont considérés comme cohérents par les experts, tant pour les clusters objectifs que subjectifs ; la mise à l'écart des INDEP. est due à la structure du questionnaire.

Les résultats 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 ressenti semble lié au manque de réaction de la hiérarchie face à l'accident qu'a subi l'enquêté.

Ainsi cette première analyse établit la relation entre situation, autonomie et satisfaction au travail, dans certaines limites ; mais elle ne nous permet pas de déterminer la nature de ces relations et 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'un des enjeux de la présente étude est d'émettre des recommandations aux managers dans le but d'améliorer la qualité de vie au travail. Or, si l'analyse descriptive nous permet d'identifier les phénomènes liés à la satisfaction au travail, elle ne nous permet pas de déterminer les causes de ces phénomènes [Pearl, 2000]. La modélisation causale constitue ainsi l'un des objectifs majeurs en science des données : alors que des modélisations fondées sur des corrélations (ce qui est le cas usuel) permettent d'effectuer des prédictions, des modélisations causales sont *nécessaires* pour émettre des recommandations.

La phase de modélisation causale a été abordée dans les dernières semaines de ce stage ; les résultats obtenus sont donc préliminaires, et l'étude se poursuivra ultérieurement dans le cadre d'une thèse.

4.1 Motivation

La corrélation caractérise le fait que deux variables A et B ont des évolutions statistiques similaires ; cependant la corrélation n'implique pas l'existence d'une relation de causalité entre A et B (e.g. les dépenses des États-Unis pour la science et le nombre de suicides sont corrélés sans qu'une relation causale n'existe). La définition de la causalité retenue dans le cadre de cette étude est celle de [Statnikov, 2012] : " A est une cause de B ($A \rightarrow B$) si la loi de probabilité de B change sous l'effet d'une manipulation expérimentale de A .

Les approches classiques pour étudier la causalité, récapitulées ci-après, ne sont généralement pas accessibles pour des raisons de coût, d'impossibilité éthique, ou de faisabilité. Ainsi les expériences contrôlées (par exemple les tests cliniques) ne sont pas éthiquement admissibles dans le cadre du travail. L'économie expérimentale est dure à transposer à la réalité. Les expériences naturelles (par exemple l'évolution du travail dans les restaurants Mc Donald's dans les différents états des États-Unis, qui sont soumis à des régulations du travail différentes) ne sont pas sous le contrôle des chercheurs.

Une méthodologie nouvelle, datant des années 2010, consiste à inférer les relations causales à partir de données d'observation uniquement. Cette méthodologie repose sur plusieurs approches : la classification avec les algorithmes d'apprentissage statistique (la donnée de B aide-t-elle à prédire la valeur de A ?), ou l'évaluation de la complexité des distributions jointes [Stegle *et al.*, 2010]. Plusieurs challenges internationaux ont été organisés par I. Guyon (le challenge *Kaggle*¹ en

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

2013 et le challenge *Codalab*² en 2014) dans le but d'évaluer les approches existantes et leurs limites. Notre étude va exploiter cet état de l'art, et les meilleurs algorithmes au sens de ces compétitions.

4.2 Méthodologie

L'objectif est de construire un graphe reliant l'ensemble (ou une partie) des variables du questionnaire en fonction de leurs liens de causalité. La méthode suivie est la méthode usuelle en modélisation causale, c'est-à-dire :

1. Construire un graphe relationnel non dirigé excluant les arcs entre variables indépendantes ;
2. Identifier les relations secondaires (liens indirects, déconvolution) et les élaguer ;
3. Orienter les relations restantes à l'aide du coefficient de causalité entre les deux variables.

La première difficulté rencontrée concerne l'hétérogénéité des variables considérées. La seconde est la fiabilité des scores de causalité au niveau des paires de variables, affectant la structure même du graphe.

4.2.1 Score de causalité

Construire un score de causalité a fait l'objet de deux challenges, où l'objectif était d'associer à toute paire de variables (A, B) un score à valeurs dans $[-1, 1]$] tel que :

- $score = -1$ si $A \leftarrow B$
- $score = 1$ si $A \rightarrow B$
- $score = 0$ dans les autres cas : indépendance de A et de B , mais aussi "cofounder" (i.e. existence d'une variable C causant A et B), ou présence de *cycle* dans le graphe de causalité (A cause B qui cause C qui cause A), ou *contrainte*(e.g. les variables de la relation physique PV=nRT sont liées par contrainte).

Un tel score est approprié compte tenu de notre objectif d'identification des relations directes de causalité.

L'approche de [Fonollosa, 2016] s'appuie sur les données des challenges ; les distributions jointes de chaque paire de variables A et B étant connue (sous forme d'un échantillon), un grand nombre d'attributs (features) sur les distributions jointes et marginales sont définis et ces attributs sont utilisés dans le cadre d'une méthode d'apprentissage classique, les forêts aléatoires [Breiman, 2001], pour apprendre le score cherché. L'approche de [Lopez-Paz *et al.*, 2015] est très comparable, à la différence que les attributs sont définis à l'aide des méthodes de noyau.

4.2.2 Score de corrélation et hétérogénéité des données

La méthode classique s'appuie sur le score de corrélation de Pearson associé à toute paire de variables. Ce score permet d'évaluer la force des liens entre les différentes variables, dans le but de passer à l'étape d'élagage des liens.

La difficulté que nous rencontrons ici est due à l'hétérogénéité des variables : numériques, catégorielles (ordonnées ou non) et booléennes (notamment les variables drapeaux, indiquant si la réponse à une question est présente ou manquante). Les coefficients de corrélation classiques ne sont pas applicables directement. Certes il existe des mesures de corrélation entre paires de variables d'un même type, booléen ou continu. La difficulté est que ces mesures ne sont pas comparables d'un type à l'autre. Or les techniques de déconvolution utilisées pour la phase d'élagage, exploitant la matrice des scores des paires (A_i, A_j), requièrent que les scores soient homogènes.

2. competitions.codalab.org/competitions/1381

Les scores faisant intervenir des variables catégorielles doivent également être reconsidérés. En effet, la décomposition des variables catégorielles en variables booléennes a deux effets indésirables : i) le lien entre les variables booléennes est ignoré par l'étude d'indépendance des variables (et devrait donc être ré-introduit par la suite) ; ii) mais surtout, les scores de causalité ont de moins bonnes performances sur les variables booléennes.

4.2.3 Phase d'élagage (déconvolution)

Pourquoi retirer les liens indirects ? Le fait de préserver le graphe total semble en effet utile, sans nuire à l'intelligibilité des graphes de petite taille (Fig. 4.1). Dans le cas de grands graphes par contre, l'élagage est nécessaire pour une visualisation raisonnable du graphe ainsi que le montre la Figure 4.2 considérant 470 variables ; les variables drapeaux ont été retirées dans un premier temps, mais le nombre de liens (> 5000) nuit à la lisibilité. La déconvolution (retranchant à l'influence de A sur B tout ce qui peut être expliqué par l'influence de A sur C et de C sur B) est donc requise.

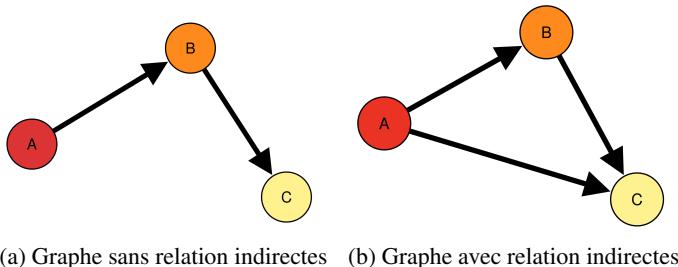


FIGURE 4.1: Graphe causal simple sans et avec la relation indirecte

Hors la phase de déconvolution, seules des relations très simples peuvent être établies aisément. Par exemple, nous avons noté que la variable associée à la question "Compte tenu du travail que vous réalisez, diriez-vous que vous êtes bien payé ?" (variable *Payecom*) pouvait être considérée comme une cause majeure (en relation causale avec de nombreuses autres variables).

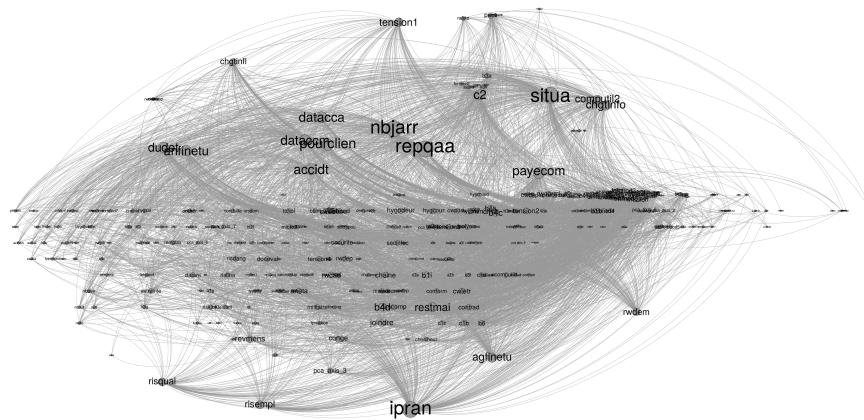


FIGURE 4.2: Graphe de causalité en se basant uniquement sur les scores associés aux paires de variables.

On remarque la variable *Payecom* mais aussi *Situa* qui traduit la situation professionnelle, *nbjarr* traduisant le nombre de jours d'arrêt de travail, ou encore *ipran*, un indicateur de revenus.

PARTIE 5

PERSPECTIVES

La perspective essentielle de notre travail, et l'objet de notre recherche future, concerne la construction de graphes causaux à partir de données hétérogènes, et les applications au problème de la QVT et plus généralement, dans le cadre des Humanités Numériques.

Deux directions de recherche ont été définies : la première concerne la définition d'un score hétérogène robuste, à même de construire une matrice de causalité homogène à partir de variables de nature différentes et de données réelles (e.g. avec valeurs manquantes), et de supporter une opération de déconvolution effective. Dans ce but, nous comparerons les critères existants sur des données générées par lésion (en introduisant des variables non-pertinentes, obtenues par permutation aléatoire de variables existantes). Les scores (e.g. Pearson, χ^2 , FSIC, information mutuelle, Cramer's V) obtenus par les paires faisant intervenir des variables non-pertinentes, aussi appelées probes, nous permettront tout d'abord de comparer les critères en termes de précision/rappel (taux de paires non-pertinentes classées dans les x% premières paires). En second lieu, ces scores nous permettront de calibrer les valeurs obtenues par les meilleurs scores, en alignant les scores obtenus sur les paires numériques et sur les paires catégorielles.

La seconde concerne la validation du graphe ainsi obtenu. Une première piste de recherche concerne la comparaison du graphe final ainsi obtenu aux approches fondées sur la couverture de Markov [Aliferis *et al.*, 2010].

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.
- [Breiman, 2001] BREIMAN, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- [Datchary.C, 2011] DATCHARY.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
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 scores associés aux paires de variables.	31

LISTE DES TABLEAUX

2.1	Questionnaire de la Dares : catégories de questions, nombre de questions par catégorie, fraction de questions objectives par catégorie (voir texte).	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	27