L'APPORT DE L'ANALYSE TEXTUELLE A LA STATISTIQUE D'ENTREPRISE : L'EXPLOITATION DE DIX ANNEES DE VISITES D'ENTREPRISES PAR LES DIRECCTE

Nicolas CAVALLO[1] (*)

(*) Insee, Direction de la statistique d'entreprise

Résumé

La base de données ISIS, qui rassemble l'information issue des 60 000 visites d'entreprises réalisées en dix ans par les Directions régionales des entreprises, de la concurrence, de la consommation, du travail et de l'emploi (Direccte), les antennes régionales de la Direction Générale des Entreprises (DGE), constitue une matière idéale pour mettre en œuvres les méthodes de la statistique textuelle avec, à la clé, des applications importantes aussi bien pour la statistique d'entreprise elle-même que pour les décideurs économiques grâce à un éclairage original porté sur les entreprises.

Deux catégories de méthodes du Data Mining ont été mises en œuvre avec les données sur les entreprises de la base ISIS afin d'éclairer des domaines stratégiques de l'analyse économique : la conjoncture et la connaissance des filières.

Une première analyse exploite le vocabulaire et les champs lexicaux de 18 000 projets d'entreprises décrits dans la base ISIS. Elle permet d'apprécier la situation conjoncturelle et son évolution telle qu'elle est perçue par les entreprises. Plus précisément, une dizaine d'indicateurs conjoncturels ont été construits sur la période 2008-2013, correspondant à autant de thématiques clés, comme l'activité, l'investissement, le développement à l'international, etc. Ces thématiques clés ont été identifiées par l'utilisation de méthodes d'analyse factorielle et de classification appliquées aux données textuelles relatives à la description des projets d'entreprises. Les indicateurs reflètent la fréquence d'apparition de ces thématiques clés. Grâce à l'enrichissement quotidien d'ISIS, cette première analyse constitue la base d'un nouvel outil d'analyse conjoncturelle dont pourrait se doter la DGE et les Direccte.

Une seconde analyse, réalisée à partir de l'information - des données textuelles également – issue de plus de 25 000 visites d'entreprises, permet de déterminer à quelle filière appartient une

[1] nicolas.cavallo@insee.fr

entreprise à partir des données textuelles disponibles. L'exercice a été mené dans le cas de la filière automobile. Il permet de mesurer, par le calcul d'un score, la probabilité qu'une entreprise appartienne à cette filière. Le calcul de ce score est établi grâce à un *modèle d'apprentissage supervisé* exploitant le pouvoir de caractérisation des entreprises plus fort et surtout plus ouvert des variables textuelles. Ce modèle apporte une connaissance appréciable des filières, placées au cœur de la politique industrielle actuelle, la statistique d'entreprise, sectorielle, n'offrant qu'un éclairage partiel sur ce sujet.

Abstract

Public statistic would benefit from the whole of textual data available on the net but also internally within numerous underused bases. The Study of 10 years of businesses visits provides an overview of the issues of textual data for business statistics. A quantitative analysis of the vocabulary and the lexical fields from 18 000 comments allows to understand the effects of the economic crisis on business projects. A second analysis of 25 000 visits using a supervised learning approach shows that textual variables have a great predictive power.

Mots-clés

- classification hiérarchique, statistique textuelle, données textuelles, statistique d'entreprise
- Hierarchical clustering, textmining, textual data, business statistics

Table des matières

Introduction

La statistique d'entreprise a pour objectif principal de représenter l'économie en observant l'activité des entreprises. Comme toute statistique, elle essaie de s'adapter à un monde en perpétuel changement. L'actuel débat qui anime tous les statisticiens d'entreprises porte sur la définition même de l'entreprise et sur la forme de l'unité statistique à observer. Mais un autre sujet devrait prendre de plus en plus d'importance. En effet, les constantes améliorations des méthodes statistiques cumulées à celles des puissances de calculs devraient amener le statisticien à repenser son approche pour mieux comprendre les entreprises.

Jusqu'à présent, afin d'appréhender certains phénomènes relatifs aux entreprises, la statistique d'entreprise a toujours eu deux approches : d'un côté, l'utilisation d'enquêtes quantitatives pour mesurer un effet connu, de l'autre l'utilisation d'entretiens dit qualitatifs pour appréhender certains phénomènes encore inconnus ou du moins mal-compris par le statisticien. Ces derniers offrent, à travers les informations recueillies, une capacité de prospective que les enquêtes quantitatives ne permettent pas vraiment. En effet, contrairement aux enquêtes quantitatives qui limitent l'interprétation d'un phénomène à un choix prédéfini de modalités, les enquêtes qualitatives permettent de récupérer une information « brute ».

La limite de l'approche qualitative réside dans son coût. Alors que la réalisation et la saisie d'un entretien qualitatif peut prendre plusieurs heures à un enquêteur, cela ne lui prendra que quelques minutes pour un questionnaire quantitatif. On comprend donc bien que ce surcoût a pour conséquence de restreindre le nombre d'individus enquêtés. Ce qui de ce fait, peut rendre impossible toute généralisation des résultats. Cependant, lorsque cette approche qualitative provient d'opérations importantes et durables dans le temps, ce nombre d'individus enquêtés n'est plus aussi limité et s'approche plus particulièrement d'échantillons rencontrés dans le cadre d'enquêtes quantitatives. Dans la plupart des cas, ces opérations répondaient à des besoins métiers qui n'avaient pas prévu que les données textuelles collectées pouvaient former des bases de données exploitables à grande échelle. Pourtant, les statisticiens disposent aujourd'hui de méthodes extrêmement puissantes pour faire parler les chiffres, mais aussi les lettres. Celles-ci permettent ainsi de structurer toutes ces données textuelles afin d'en dégager les informations essentielles.

Cette étude a comme objectif de mesurer pour la statistique d'entreprise les enjeux de l'exploitation quantitative des données textuelles. Pour cela, nous nous intéressons à 10 ans de visites d'entreprises réalisées par les antennes régionales de la Direction Générale des Entreprises (DGE). Chacune de ces visites permet de collecter un nombre très important de données textuelles portant sur des informations stratégiques très détaillées. Ainsi, alors que l'objectif principal de ces visites portait sur l'accompagnement des entreprises, nous observerons que, grâce à l'accumulation de toutes ces données collectées, elles permettent également de comprendre les entreprises à travers une toute nouvelle approche.

Nous verrons à cet effet, au sein d'une première exploitation, que ces données enrichies quotidiennement constituent un enjeu majeur pour comprendre les entreprises et leur conjoncture économique. Dans ce cadre-là, nous ferons parler les données textuelles d'elles-mêmes grâce à un modèle d'apprentissage non-supervisé. Enfin, le pouvoir de caractérisation des entreprises qui réside dans ces données textuelles sera analysé à travers un modèle d'apprentissage supervisé. Cette deuxième exploitation montrera que les données textuelles peuvent constituer, dans le cadre de modèles prédictifs, de nouvelles variables explicatives très utiles.

Avant toute analyse, nous nous attarderons à présenter dans une première partie la base de données issue des visites d'entreprises, afin d'appréhender la structure des entreprises concernées et des données collectées. Puis, nous détaillerons dans une seconde partie la méthodologie préalable à la statistique textuelle.

1. 10 ans de visites d'entreprises pour comprendre l'entreprise

La Direction Générale des Entreprises (DGE), placée sous l'autorité du ministre de l'économie, du redressement productif et du numérique, a pour mission de développer la compétitivité et la croissance des entreprises. La DGE dispose d'antennes régionales (les DIRECCTE) qui ont pour mission, d'une part d'accompagner les entreprises à chaque étape de leur évolution et, d'autre part, d'anticiper et d'accompagner les mutations économiques. C'est dans le cadre de cette mission que les DIRECCTE réalisent, grâce à près de 400 chargés de mission, jusqu'à 7 000 visites d'entreprises chaque année. Au total, c'est près de 60 000 visites d'entreprises, réalisées depuis 2004 sur tout le territoire français, qui apportent une quantité considérable d'informations sur les entreprises. Cependant, avant d'exploiter ces données, il est important de comprendre d'une part la structure des entreprises visitées et le poids qu'elles représentent dans l'économie française, et d'autre part, de comprendre les informations collectées à la suite de ces visites.

1.1 Quelles sont les entreprises visitées ?

1.1.1 Selon le secteur : Des visites principalement dans l'industrie

Historiquement, la DGE a toujours porté une attention très marquée au secteur de l'industrie. C'est pourquoi, la majorité des visites porte sur des entreprises de ce secteur. Ce poids de l'industrie a cependant tendance à diminuer. En effet, il est passé de 72 % des entreprises visitées en 2004 à 57 % en 2013 (Figure 1). Au vu du poids de l'industrie beaucoup plus faible dans la démographie d'entreprises française (7 % environ), certains secteurs tel que le commerce ou la construction sont fortement sous représentés au sein des entreprises visitées.

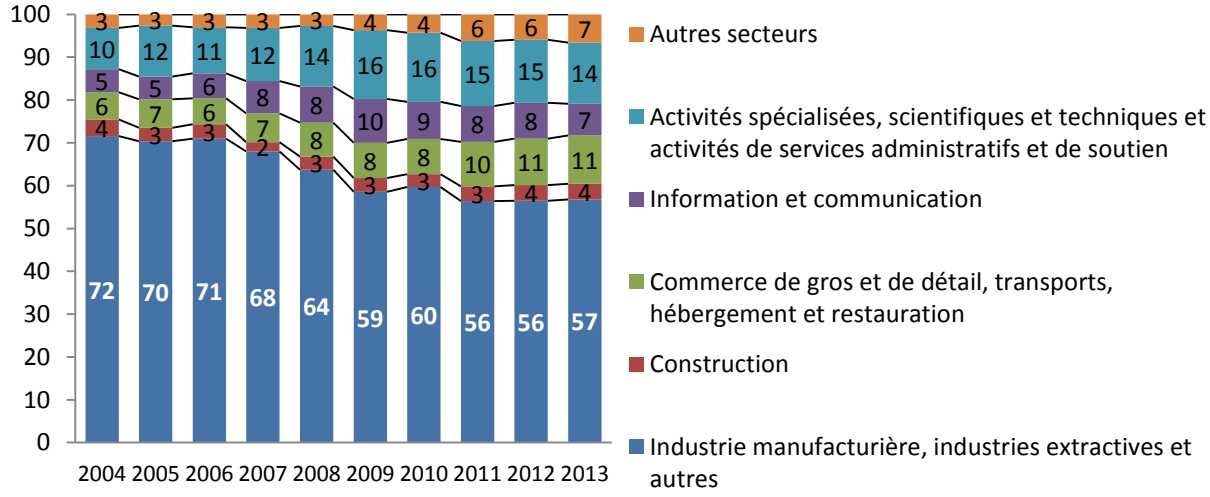


Figure 1 : répartition des entreprises visitées selon le secteur

Les entreprises visitées génèrent chaque année un peu plus de 5 % du chiffre d'affaires de l'économie française (7 % en 2011) et près de 15 % des exportations françaises. Cette bonne représentation des

exportations françaises est principalement due au fait que l'industrie, qui concerne la moitié des exportations françaises, est surreprésentée : chaque année, les entreprises industrielles visitées représentent environ un cinquième de l'activité du secteur et un quart de ses exportations.

1.1.2 Selon la taille : Des visites qui ciblent principalement les grandes entreprises (ou de taille intermédiaire)

La loi de modernisation de l'économie (LME) a défini en 2008 quatre catégories d'entreprises qui dessinent un partage relativement équilibré de l'emploi et de la valeur ajoutée : les grandes entreprises (GE), les entreprises de taille intermédiaire (ETI), les petites et moyennes entreprises (PME) et, au sein de cette catégorie, les micro-entreprises. Cette nouvelle définition de l'entreprise nécessite de prendre en compte l'activité économique de l'entreprise et les liaisons financières qui peuvent exister entre les différentes unités de production qui la composent. Le temps d'obtenir ces données et la difficulté à déterminer ces liaisons nous imposent de travailler sur la période 2009-2011 seulement (Figure 2).

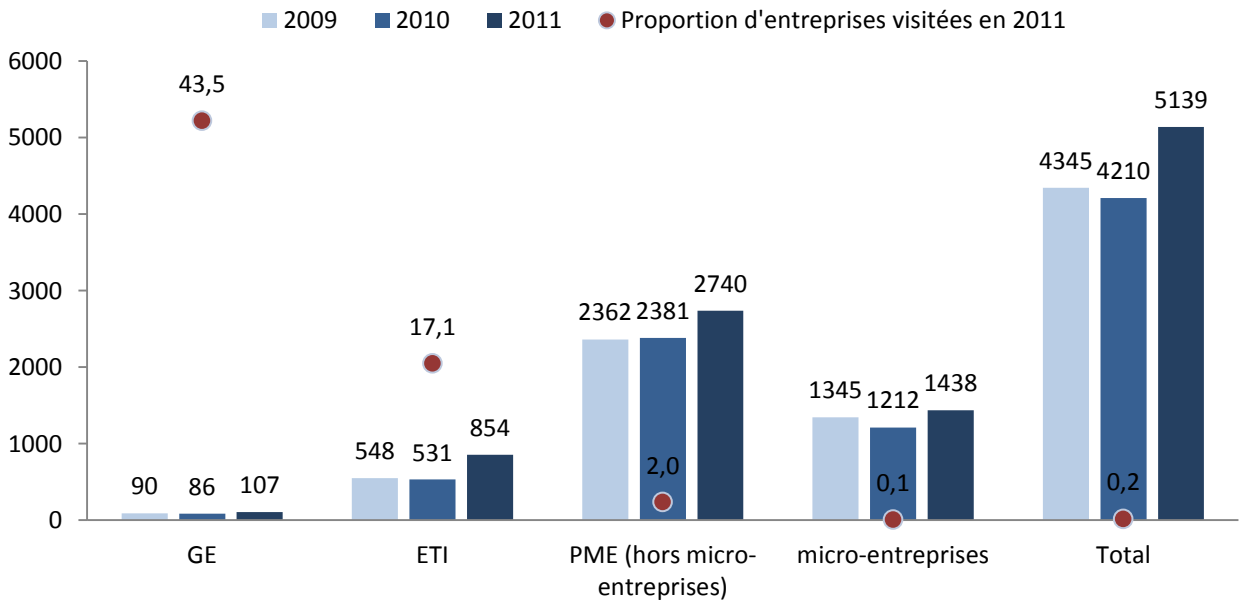


Figure 2 : Nombre d'entreprises au sens LME visitées chaque année

Les visites d'entreprises ciblent principalement les grandes entreprises ou les ETI : 44 % des Grande entreprises et 17 % des ETI ont étés concernées par des visites d'entreprises en 2011, contre 2 % des PME (hors micro-entreprises) et 0,1 % des micro-entreprises.

Pour les grandes entreprises ou les ETI, on ne peut cependant pas considérer que la globalité de l'activité est concernée pour chaque visite. En effet, seules quelques unités de production appartenant à ces grandes entreprises sont visitées. Pour appréhender le poids de l'activité visitée dans l'activité totale, il est donc judicieux de ne comptabiliser que l'activité générée par ces unités de production visitées. On observe que pour les ETI et les GE, les unités de production visitées représentent environ

8 % du chiffre d'affaires de ces entreprises chaque année contre 3 % pour les PME (hors micro-entreprises).

Les exportations et le déploiement à l'international sont pratiquement limités aux grandes entreprises et à celles de taille intermédiaire. C'est pourquoi, la bonne observation de celles-ci cumulée à la surreprésentation du secteur de l'industrie explique en grande partie la bonne observation chaque année de l'activité à l'exportation française.

1.2 Quelles sont les données collectées ?

Les visites d'entreprises peuvent être assimilées à des entretiens semi-directifs réalisés la plupart du temps auprès du dirigeant de l'entreprise. A la suite de cette visite d'entreprise, les chargés de mission remplissent au sein d'un logiciel informatique (appelé « ISIS ») les informations qu'ils ont réussi à collecter. Cet outil métier permet de normaliser, d'ordonner et de guider la saisie d'information pour le chargé de mission. Mais outre la compréhension du tissu local, un des principaux enjeux de ces visites est d'accompagner les entreprises. C'est pourquoi cet outil métier permet également une saisie quasi exhaustive des informations collectées au cours de ces visites d'entreprises. Ainsi, il offre la possibilité au chargé de mission de saisir un nombre très important de données textuelles.

Tableau 1 : Les variables textuelles issues des visites d'entreprises

Différentes Variables	taille max en nombre de caractères	nombre de commentaires non vides	Nombre d'occurrences total *(en million)*	renseignée dans "ISIS" depuis	Les formes « pleines » les plus employées
Conclusion générale	32 000	53 640	29,3	2004	*entreprise, société, activité, marché, chiffre_affaires, client, projet, production, produit, groupe*
Projet de développement	200	17 851	0,3	2008	*entreprise, projet, développement, marché, activité, nouveau, développer, produit, export, souhaiter*
Clients principaux	200	25 045	0,3	2008	*chiffre_affaires, client, grand, entreprise, particulier, collectivité, automobile, secteur, industrie, groupe*
Grands donneurs d'ordre	200	13 205	0,1	2008	*grand, client, psa, chiffre_affaires, renault, groupe, automobile, entreprise, airbus, industrie*
Commentaire sur un sous-critère spécifique	4 000	120 170	2,5	2008	*entreprise, client, produit, site, société, marché, production, activité, permettre,*

Entre 2004 et 2013, 59 000 visites d'entreprises ont été stockées au sein de l'application « ISIS ». Les possibilités de stockage ont évolué au cours du temps. Elles sont passées d'un simple compte-rendu à un ensemble d'informations très détaillées (projet, innovation, ressources humaines, etc.). Nous nous intéressons ici uniquement aux seules données textuelles présentées dans le Tableau 1 :

- La « conclusion générale » est en quelque sorte le résumé de la visite. On y trouve des informations très détaillées et très variées sur l'entreprise visitée (son histoire, son activité, les difficultés rencontrées, les projets, ses besoins, etc.).
- La variable « projet de développement », comme son nom l'indique porte sur le projet de développement de l'entreprise. Elle est limitée à 200 caractères et impose donc au chargé de mission d'être concis.
- La variable « clients principaux » éclaire sur la composition de la clientèle de l'entreprise. On y trouve souvent les noms des grandes entreprises françaises voire mondiales et le poids qu'elles représentent en termes de chiffre d'affaires.
- La variable «grands donneurs d'ordre » concerne principalement les entreprises de la sous-traitance. Elle est, au vu des formes les plus employées, très liée avec la variable « clients principaux ».
- La dernière variable provient d'un commentaire que les chargés de missions peuvent laisser pour chacun des 46 sous-critères listés dans le Tableau 2. Plus de 14 000 visites d'entreprises comportent au moins un de ces commentaires (en moyenne 8 sous-critères commentés par visite). Ces commentaires détaillent les points positifs ou négatifs de l'entreprise pour chacun des 46 sous-critères.

Tableau 2 : Répartition des commentaires laissés selon les 46 sous-critères

Critère	Sous-critère	Nombre de commentaires
COMMERCIAL	Action commerciale - marketing	4 555
	Connaissance concurrence	3 872
	Dépendance client	5 191
	Dépendance fournisseur	2 775
	Innovation commerciale	1 193
	Niveau technologique produits	2 328
	Positionnement concurrentiel	4 756
	Protection marques - stratégie PI	2 595
	Présence internationale	5 063
	Réseau distribution	2 263
	Savoir faire - renommée	3 548
	Service - écoute clients	1 629
ENVIRONNEMENT DE L'ENTREPRISE	Adéquation site - développement	5 433
	Implication vie locale	2 942
	Patrimoine immobilier	2 811
	Qualité infrastructures transport	2 115
	Réseaux TIC	1 029
	Veille sur son environnement (techno, concurrence, règlementaire)	2 068
INNOVATION	Ecoconception	1 559
	Innovation organisationnelle	1 595
	Innovation procédés	2 299
	Innovation produits design	1 774
	Innovation produits techno	4 459
	Innovation services	2 604
MOYENS PRODUCTION	Adaptation procédés - besoins	2 021
	Gestion flux, stock	1 787
	Maintenance industrielle	749
	Niveau technologique	2 207
	Organisation production	2 556
	Outil production - gestion matériel	2 336
	Respect normes environnementales	2 145
QUALITE	Management de la qualité	4 281
	Niveau qualité produits	4 105
	Qualité délais	1 912
RESSOURCES HUMAINES	Climat social	2 861
	Facilités recrutement	3 759
	Formation - gestion compétences	2 356
	Niveau qualification	2 596
	Pyramide âges	2 104
	Qualité équipe dirigeante	3 877
	Recours intérim	1 226
USAGES TIC	Achats en ligne Ventes en ligne Services eadministration	1 476
	Conception (CAO, PLM.)	1 794
	Gestion interne de l'entreprise (PGI, ERP.)	1 842
	Outils de travail collaboratifs (groupware, vidéo conférence, etc..)	883
	Relation client (CRM.)	841

1.3 Quels sont les enjeux de ces visites ?

Une des grandes richesses de ces visites d'entreprises réside dans les différentes données textuelles récoltées depuis 10 ans. Alors que l'intérêt principal de ces données est de conserver des informations détaillées (fonctionnement, caractéristiques, spécificités, projets de développement, etc.) sur une entreprise afin d'en assurer son suivi, elles peuvent maintenant, grâce à l'accumulation de toutes ces visites, être utilisées de façon plus quantitative. En effet, les méthodes d'analyse textuelle cumulées à des outils informatiques performants permettent une toute nouvelle exploitation de celles-ci. Ces données peuvent répondre à un des enjeux majeurs de la statistique d'entreprise : comprendre les entreprises tout en comprenant l'entreprise.

Le premier intérêt de ces visites d'entreprises provient de la qualité des données collectées sur chaque entreprise. Ces données principalement textuelles s'adaptent aux spécificités des entreprises et permettent de les caractériser. Ce pouvoir de caractérisation des données textuelles est étudié en partie IV.

Le deuxième intérêt qui apparait dans ces données réside dans sa réactivité. En effet, les chargés de mission peuvent renseigner dans l'outil métier « ISIS » les informations de la visite ayant eu lieu la veille voire le jour même. Au total, les informations d'une vingtaine de visites d'entreprises sont renseignées quotidiennement dans l'outil métier « ISIS ». Ces informations conjoncturelles, qui sont bien évidemment les plus intéressantes pour mesurer le pouls de l'économie, sont habituellement les plus compliquées à obtenir. Cette réactivité dans la collecte cumulée à la qualité et la quantité d'informations collectées sur l'entreprise laissent donc entrevoir des possibilités intéressantes d'un point de vue conjoncturel. Afin d'appréhender ces possibilités, les résultats d'une analyse textuelle sur les projets des entreprises sont présentés en partie III.

2. L’analyse textuelle, comment l’utiliser ?

Le domaine des analyses de données textuelles, également appelé Text Mining, réunit l’ensemble des techniques et méthodes qui permet le traitement automatique de données textuelles, afin de les structurer et de faire ressortir des thèmes ou toutes informations cachées. La grande difficulté dans l’analyse de données textuelles est de déterminer quelle unité statistique utiliser. En effet, on peut analyser les paragraphes, les phrases, les mots ou directement les caractères.

Nous nous intéressons dans cette partie à l’approche lexicale de l’analyse textuelle. L’analyse lexicale consiste à étudier statistiquement l’usage des mots. Bien que cette analyse limite le contenu à un ensemble de mots-clés, elle permet de dégager une structure claire de l’organisation des textes étudiés et comprendre les thèmes abordés. Il existe plusieurs méthodes d’analyse textuelle basées sur la lexicométrie. Nous détaillerons ici les étapes communes qui précèdent toutes ces analyses textuelles. En effet, la première étape consiste à créer le lexique et à le simplifier. Cette simplification va permettre de créer un tableau lexical concis sur lequel nous pourrons appliquer toute sorte de méthodes d’analyses statistiques multivariées. En ce qui concerne ces analyses possibles, nous nous attarderons particulièrement sur la méthode de classification proposée par Reinert (1983).

2.1 Première étape : Le lexique

2.1.1 Créer le lexique

La première étape essentielle à toute analyse lexicale consiste à créer le lexique issu du corpus à étudier. Pour cela, on analyse toutes les formes employées dans le corpus étudié que l’on présente principalement selon leur ordre de fréquence décroissante d’apparition. Le lexique permet de visualiser rapidement le contenu. Il donne du coup des informations utiles pour comprendre et se familiariser avec la thématique étudiée. Cependant, dans le cas de gros corpus alliant taille importante et vocabulaire riche, il est intéressant voire obligatoire de simplifier le lexique. En effet, le nombre de formes différentes peut facilement atteindre des centaines de milliers, ce qui *de facto* risque de noyer l’information essentielle.

2.1.2 Simplifier le lexique

L’étude de gros corpus nécessite de réduire l’étendu du vocabulaire à étudier. Pour cela, on réalise une opération de lemmatisation sur le corpus qui consiste à remplacer les mots du corpus par leur forme « racine » : on ramène les verbes à l'infinitif, les noms au singulier et les adjectifs au masculin singulier.

Un exercice préalable à la lemmatisation consiste à gérer les expressions spécifiques à la thématique du corpus étudié. En effet, certaines expressions risquent tout simplement de disparaitre par la lemmatisation (Par exemple, « Etats-Unis » qui deviendrait « état unir »). De plus, si plusieurs

expressions ont le même sens (« CA », « chiffre d'affaires »), en l'absence d'unification, l'analyse lexicale risque de réduire leur poids global dans le corpus. Afin de garder le sens unique de certaines expressions, il est important d'unir les formes verbales qui composent celles-ci. Une partie des expressions relatives à la thématique des données étudiées se dévoilent au fur et à mesure de l'étude du corpus. Il faut par conséquent mettre à jour continuellement la liste des expressions pour améliorer l'analyse.

Il est possible de simplifier encore davantage l'information en distinguant les formes « pleines » des formes « outils ». On considère comme forme « outils » tous les pronoms, prépositions, conjonctions, auxiliaires ou les nombres, et comme formes « pleines » tous les autres (noms, verbes, adjectifs et adverbes). Cette classification crée une sorte de hiérarchie sémantique qui a pour objectif de limiter l'analyse aux mots les plus pourvus de sens.

2.1.3 Se familiariser avec le lexique

L'étude de la fréquence des mots permet de comprendre très rapidement la thématique du corpus étudié. L'hypothèse sous-jacente est qu'il est possible de cerner un certain nombre de facteurs en fonction de la fréquence d'apparition des occurrences. Dans le Tableau 3 présentant les formes « pleines » les plus utilisées dans le corpus des « conclusions générales » (voir partie I), on remarque que les étapes précédentes ont permis de supprimer les mots les moins porteurs de sens. Ce tableau permet ainsi de se familiariser avec le contenu du corpus avant d'entreprendre des analyses plus approfondies.

Tableau 3 : Effet de la simplification du lexique sur les 20 formes les plus fréquentes de la « conclusion générale »

Lexique initial		
Forme	fréquence	type
de	1 795 481	pre
la	757 955	art_def
l	687 689	art_def
et	678 280	con
en	569 854	pre
d	566 989	pre
le	549 512	art_def
à	538 778	pre
des	503 222	art_ind
les	437 332	art_def
est	330 322	aux
du	319 290	art_def
pour	309 261	pre
un	308 340	art_ind
une	272 404	art_ind
a	257 069	aux
entreprise	240 917	nom
sur	232 343	pre
par	213 087	pre
dans	201 250	pre

Apres gestion des expressions, lemmatisation et distinction entre formes « pleines » et formes « outils »		
Forme "pleine"	fréquence	type
entreprise	273 446	nom
société	115 339	nom
activité	108 562	nom
marché	99 350	nom
chiffre_affaires	98 939	nom
client	95 810	nom
projet	91 466	nom
production	84 168	nom
produit	80 528	nom
groupe	71 566	nom
site	68 650	nom
développement	64 886	nom
mettre	55 337	ver
commercial	52 651	adj
permettre	52 297	ver
réaliser	50 997	ver
nouveau	48 474	adj
france	43 181	nr
fabrication	42 388	nom
grand	39 300	adj

2.2 Deuxième étape : Le tableau lexical entier

Si l'on veut étudier le lien entre les mots, il va tout d'abord falloir déterminer la taille d'un segment de texte pour laquelle une utilisation simultanée de certains mots ait un sens. Dans le cadre de textes trop longs il faudra procéder à un découpage de celui-ci en un ensemble de segments de texte. Ce découpage dépend du corpus à étudier et peut porter sur une suite d'occurrences, de phrases ou de quelques paragraphes.

Une fois cette unité textuelle déterminée, il est nécessaire de construire un tableau disjonctif complet dont les lignes sont constituées par les segments de texte et les colonnes par les différentes formes issues du lexique construit précédemment. Ce tableau, appelé tableau lexical entier (TLE), a pour valeur en $i^{ème}$ ligne et $j^{ème}$colonne le nombre de fois que la $j^{ème}$ forme est présente dans le $i^{ème}$ segment de texte du corpus. Chaque segment de texte ne comportant qu'une partie infime des mots du lexique, ce tableau est principalement composé de zéros.

Ce TLE peut être réduit à un tableau d'absence-présence (Tableau 4) où seule la présence d'un mot dans un segment de texte est prise en compte. Ce choix de résumer l'information fréquentielle à une information de présence-absence se justifie par le fait qu'un mot est très rarement utilisé par plus de 50% de la population et par conséquent que sa présence à elle seule constitue l'information essentielle (Reinert, 1983).

Tableau 4 : Tableau Lexical Entier

	Mot n°1	**Mot n°2**	**Mot n°3**	.	.	.	.	.	....	.	.	.	.	.	**Mot n°M**
Segment de texte n°1	1	1	0	0	0	0	1	0		0	0	0	0	0	0
Segment de texte n°2	0	0	1	0	0	0	0	0		0	0	1	0	0	0
Segment de texte n°3	1	0	0	0	0	0	0	0		0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Segment de texte n°N	0	0	0	0	1	0	0	0		0	1	0	0	0	0

2.3 Troisième étape : Analyser le tableau

L'information lexicale du corpus a donc été en quelque sorte résumée dans un tableau sur lequel nous pouvons réaliser des méthodes classiques d'analyse de données. L'analyse textuelle se base particulièrement d'une part sur les méthodes factorielles pour dégager la principale structuration du corpus selon les axes factoriels et, d'autre part, sur les méthodes de classification automatique pour représenter les proximités entre les différents éléments du TLE. Nous nous intéressons premièrement à l'analyse des correspondances qui apparaît comme la méthode factorielle la plus adaptée à l'analyse textuelle, puis nous détaillerons la classification descendante hiérarchique proposée par Reinert en 1983 qui est une méthode de classification efficace pour traiter ce type de tableau.

2.3.1 Analyse factorielle des correspondances (Benzécri 1973)

La méthode factorielle la plus adaptée à l'analyse du TLE est l'Analyse Factorielle des Correspondances (AFC). En effet, cette analyse présentée par Benzécri (1973) permet de décrire efficacement les tableaux de contingence en représentant graphiquement l'association entre les lignes et les colonnes. Cette association est déterminée à partir de la distance du Khi^2 qui permet de corriger le fait que les mots les plus fréquents auront mécaniquement un poids plus important. L'AFC permet de représenter sur un plan les principales proximités et les oppositions (au sens de la distance du Khi^2) qui se dégagent entre les différents mots employés dans le corpus.

2.3.2 Classification descendante hiérarchique (Reinert 1983)

Il existe plusieurs méthodes de classification mais nous nous intéressons particulièrement à la classification descendante hiérarchique (CDH) que Reinert a proposée en 1983. En effet, cette

méthode de classification est très bien adaptée à l'exploitation de tableaux creux (plein de zéros) telle que le TLE. Cette classification exploite le TLE réduit au tableau d'absence-présence (Tableau 4).

2.3.2.1 Fonctionnement de la CDH

L'objectif de la CDH est de chercher les ensembles de segments de texte les plus contrastés du point de vue de leur vocabulaire (Tableau 5). La CDH part de l'ensemble entier de tous les segments de texte ; celui-ci est scindé en deux parties qui sont à leur tour scindées en deux, jusqu'à ce que tous les sous-ensembles obtenus soient réduits à un segment de texte unique. Cependant, la scission en deux d'une classe à n segments de texte demande l'examen de $2^{n-1} - 1$ bipartitions, ce qui peut demander un temps de calcul considérable. Reinert a eu l'idée d'opérer la scission selon le 1er axe factoriel issue de l'AFC (réalisée sur l'ensemble à scinder), afin d'obtenir de bons résultats tout en réduisant le nombre de bipartitions à observer. Le nombre de bipartitions possibles est ainsi réduit à $n - 1$, et la coupure intervient pour celle qui maximise l'inertie interclasse, c'est-à-dire pour celle où les barycentres des deux classes résultantes sont les plus éloignés.

Tableau 5 : Principe de la séparation de deux ensembles de segments de texte en fonction du vocabulaire

		Vocabulaire 1	**Vocabulaire 2**
		Mot n°. . . .. Mot n°.	**Mot n°. . . .. Mot n°. Mot n°. . . .. Mot n°.**
Classe 1	**ST n°X**	**1 0 1 .... 1 0 1**	0 0 0 .... 0 1 0 0 0 0 .... 0 1 0
	⋮	**Beaucoup de 1**	Beaucoup de 0
	ST n°XX		
Classe 2	**ST n°X**	0 0 0 **....** 1 0 0	**1 1 0 .... 1 1 0 1 1 0 .... 1 1 0**
	⋮	Beaucoup de 0	**Beaucoup de 1**
	ST n°X		

Trois variables doivent être paramétrées afin de réaliser efficacement une CDH. Premièrement, le nombre de bipartitions totales (ou phases terminales) est à déterminer afin de se retrouver au final avec un nombre de classes exploitables. Les deux autres paramètres permettent notamment de pallier une des faiblesses de la CDH. En effet, la CDH entraine la création de classes de petits effectifs qui sont souvent créées par la présence de mots de faible fréquence. Ces classes pouvant être créées lors des premières bipartitions, le premier paramètre ne permet pas de les éliminer. Il est donc important de déterminer un seuil à partir duquel les classes seront considérées comme exploitables : les classes comportant un nombre de segments de texte inférieur à ce seuil ne seront pas analysées. Ce seuil est souvent égal au nombre moyen de segments de texte par classes (nombre total de segments de texte / nombre de phases terminales). Le dernier paramètre porte sur le nombre de formes « pleines » exploitées. Reinert avait observé que les mots de faible fréquence jouaient par leur nombre un rôle non négligeable dans la classification. Il avait donc préconisé de garder toutes les formes « pleines » employées au moins trois fois. Cela permettait de réduire considérablement les temps de calcul tout en gardant l'information contenue dans les formes « pleines » de faible fréquence. L'étude de Ratinaud & Marchand (2012) a cependant montré que sur

de gros corpus, l'information contenue dans les formes « pleines » les plus fréquentes était très proche de celle contenue dans les formes « pleines » légèrement moins fréquentes. C'est pourquoi, ce seuil peut sans doute être augmenté dans le cas de gros corpus.

2.3.2.2 Interprétation des résultats

Les classes en sortie de la CHD sont définies par un ensemble de formes « pleines » et représentent, à travers ce vocabulaire, une thématique ou des thématiques « significativement » présentes dans le corpus. Des analyses complémentaires permettent d'interpréter ces classes. À chaque couple (formes « pleines », classe) correspond une valeur de Khi^2 mesurant l'intensité de leur association. De plus, les classes peuvent être représentées graphiquement à partir d'une analyse factorielle des correspondances portant sur le tableau qui croise les formes « pleines » analysées et les classes. Au final, l'organisation de ces classes appelées *mondes lexicaux* par Reinert (1993), permettent de structurer l'information contenue dans le corpus.

2.4 Les logiciels utilisées

De nombreux logiciels sont disponibles pour réaliser des analyses textuelles. Chacun possède ses avantages en termes d'automatisation, de méthodes implémentées, de facilité d'utilisation ou de qualité dans la représentation des résultats. Nous nous intéressons particulièrement aux deux logiciels IRaMuTeQ et SAS TextMiner.

2.4.1 IRaMuTeQ

IRaMuTeQ est un logiciel libre, basé sur R et le langage Python. Ce logiciel permet de réaliser efficacement toutes les opérations consistant à simplifier le lexique (gestion des expressions, lemmatisation, distinction entre formes « pleines » et formes « outils »). Son intérêt majeur réside cependant dans le fait qu'il permet de reproduire la méthode de classification de Reinert. Pour cette raison, nous utiliserons ce logiciel dans la partie III.

2.4.2 SAS Text Miner

SAS Text Miner est un module de SAS Enterprise Miner. Il permet de réaliser toutes les exploitations préalables à l'analyse textuelle. Il offre un grand nombre de méthodes de modélisation. Dans le cadre d'un modèle d'apprentissage supervisé, il permet notamment de convertir une information textuelle en une probabilité aidant à la prédiction d'une variable cible. Nous utiliserons à cet effet SAS Text Miner dans la partie IV. Globalement, l'atout principal de ce module réside dans sa complémentarité avec SAS Enterprise Miner. Il est ainsi possible de réunir dans un environnement commun les données structurées (quantitatives) avec les informations textuelles (non structurées).

3. L'analyse textuelle pour comprendre : L'exemple des projets des entreprises.

Depuis 2008, les chargés de missions peuvent renseigner dans « ISIS » des informations portant sur les projets de développement des entreprises. Il est proposé au chargé de mission deux moyens complémentaires de coder l'information : une variable qualitative avec quelques modalités proposées et une variable textuelle limitée à 200 caractères.

La difficulté dans l'exploitation de la variable qualitative provient de l'insertion au cours du temps de nouvelles modalités. Cette variable ne renseignant que sur le principal projet (une seule modalité possible par visite), l'ajout de nouvelles modalités impacte fortement le choix du chargé de mission. La Figure 3 montre que l'ajout à partir de mi-2012 des modalités « croissance d'activité », « croissance interne» et « diversifications marchés » semblent fortement impacter les modalités «différenciation produits, spécialisation », « croissance externe » et dans une moindre mesure la modalité « développement à l'export ». Ces nouvelles modalités représentent en 2013 déjà plus de 40 % des principaux projets de développement des entreprises visitées. C'est pourquoi, l'exploitation de cette variable est limitée à la qualité et à la stabilité des modalités proposées aux chargés de mission.

On comprend donc ici tout l'intérêt des variables textuelles. En effet, le chargé de mission n'est pas obligé de simplifier voire déformer les informations qu'il a pu collecter à la suite de sa visite d'entreprise. Les données textuelles sont des données « brutes ». Elles ne sont pas biaisées par des hypothèses que l'on aurait pu faire sur les informations attendues et qui auraient restreint l'information collectée à un ensemble de modalités déterminées en amont.

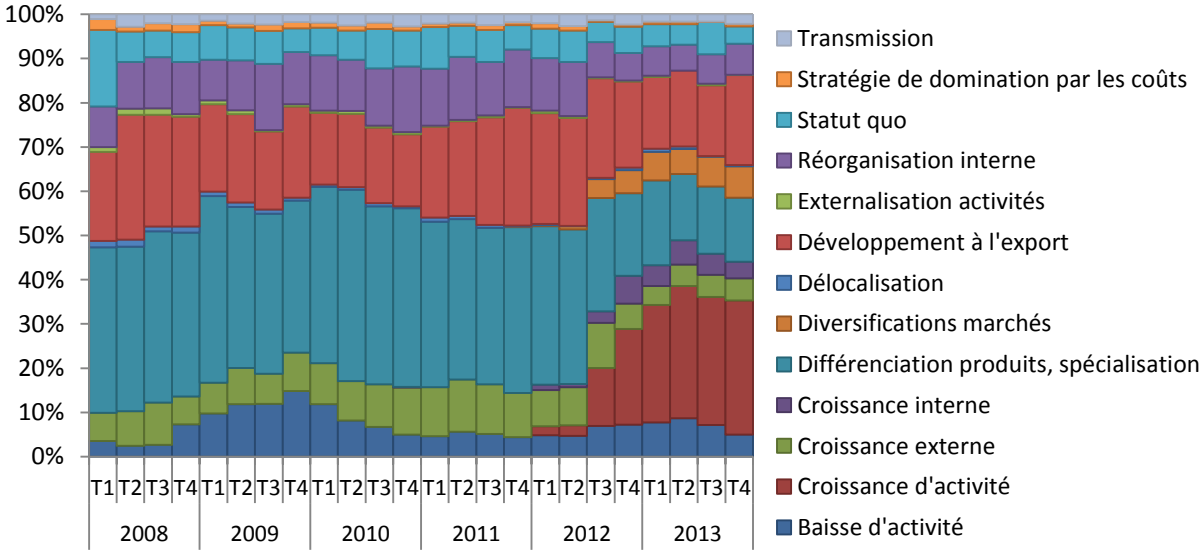


Figure 3 : Répartition trimestrielle des projets de développement renseignés

Dans cette partie, nous faisons parler les données textuelles d'elles-mêmes : les thématiques principales relatives aux projets des entreprises émergent automatiquement à partir du vocabulaire employé. Pour cela, nous réalisons plusieurs classifications de Reinert sur le corpus des projets de développement.

La première classification a pour objectif d'étudier à un niveau agrégé les différentes thématiques des projets. Elle montrera que les données textuelles peuvent apporter des informations nouvelles et aider par la même occasion à la compréhension des entreprises. La deuxième classification servira à étudier l'intérêt de travailler à un niveau plus fin. Cela permettra de justifier et compléter les résultats observés grâce à la 1ère classification. Enfin, la troisième classification utilisera des données plus récentes afin d'observer, d'une part, la stabilité dans le temps des résultats et, d'autre part, d'avoir un aperçu des possibilités en termes d'analyse économique conjoncturelle qu'offre l'étude des projets des entreprises. Les problèmes de représentativité des résultats seront étudiés dans une dernière partie qui se conclura par une exploitation des résultats des trois classifications au champ de l'industrie seulement.

3.1 Préparation de la base

Entre 2008 et 2013, les chargés de mission ont renseigné 17 851 projets de développement. Chacun de ces projets est associé à une visite d'entreprise et donc à une entreprise. Une même entreprise a beau pouvoir être visitée plusieurs fois, il est plus intéressant de travailler au niveau de la visite d'entreprise. En effet, les projets d'une même entreprise peuvent changer au cours d'une même année, et ce serait donc une perte d'information que de réunir au sein d'une même observation ces données.

La classification de Reinert ne prend en compte que le contenu de la variable textuelle étudiée. C'est ce contenu à lui tout seul qui permet de diviser l'ensemble des projets des entreprises en sous-ensembles plus homogènes. Cette homogénéité est calculée à partir du vocabulaire employé au sein des projets. Afin de s'assurer de la cohérence de ces sous-ensembles, nous rajoutons en tant que variables de contrôle la variable « projet de développement» à 13 modalités (Figure 3). Nous rajoutons également des variables supplémentaires pour étudier l'évolution de ces projets selon les caractéristiques principales de l'entreprise : année et trimestre du projet, activité de l'entreprise, taille de l'entreprise et région de l'entreprise.

Pour réaliser toutes les étapes de la classification de Reinert, nous utilisons le logiciel IRaMuTeQ. Le corpus en entrée doit être présenté selon une structure particulière (Figure 4) : chaque texte doit être introduit par une première ligne comportant toutes les caractéristiques du projet (la modalité d'une variable s'écrit de la forme **variable_modalité*). Une fois le corpus renseigné dans le logiciel, il ne reste plus qu'à régler les différents paramètres (détaillé en partie II.3.2) de la classification en fonction des résultats souhaités.

```
**** *a_2008 *ac_industrie *eff_ms20sal *proj_devexport *at_2008t2 *reg_11
        L'entreprise a modifié sa gamme depuis deux ans dans la perspective de son développement
        à l'export. Le succès de cette stratégie implique un renforcement des effectifs.

**** *a_2008 *ac_industrie *eff_20a100sal *proj_reorga_interne *at_2008t1 *reg_82
        augmentation de la productivité par la modernisation du parc à grumes
```

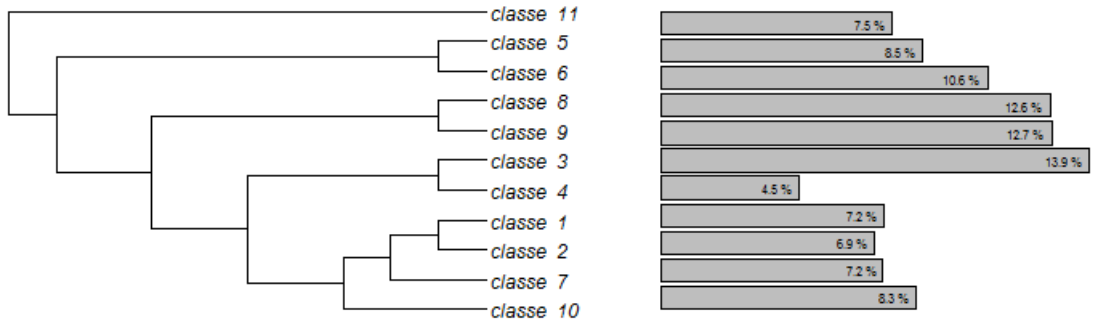
Figure 4 : Exemple de mise en forme des projets pour IRaMuTeQ

3.2 Une première classification pour comprendre les projets des entreprises

Cette première classification a pour objectif de montrer que de grandes thématiques se dégagent des projets des entreprises. Pour cela, nous nous contenterons d'une classification à 30 phases terminales que nous appliquons au TLE limité aux seules formes « pleines » dont la fréquence d'apparition est supérieure ou égale à 6, soit 2 469 formes. Par la suite, les classes provenant de cette classification seront étudiées afin de déterminer les principales thématiques des projets. Enfin, l'évolution sur la période 2008-2013 du poids de ces différentes thématiques permettra d'appréhender les effets de la crise sur les projets des entreprises françaises.

3.2.1 Organisation des projets de développement

Comme nous l'avons vu précédemment, un des paramètres de la classification de Reinert permet de conserver dans l'exploitation finale uniquement les classes suffisamment représentées. En ne conservant que les classes dont le nombre de segments de texte est supérieur à la moyenne, la classification de Reinert ne fait apparaître que 11 classes lexicales distinctes (comportant au moins 17 851 / 30= 596 segments de texte) dans le corpus « projets de développements des entreprises ». Au final, ces 11 classes représentent tout de même 15 819 projets de développement, soit 89 % des projets.



Légende : les histogrammes indiquent la proportion de commentaires de projet de développement représentés par la classe

Figure 5 : Les mondes lexicaux du corpus « projets de développements » (11 classes).

La Figure 5 permet de représenter ce que Reinert appelle les « mondes lexicaux » pour le corpus des projets de développement. Elle illustre clairement les proximités qu'il peut y avoir entre les différentes classes. Cette proximité provient bien évidemment de similitudes au sein du vocabulaire employé. Le Tableau 6 présente les formes spécifiques de chaque classe et permet de confirmer ces proximités à travers les thématiques qui se dégagent.

Tableau 6 : Description des classes de la 1ère classification

Thème de chaque classe		Traits lexicaux typiques par classe[1]
1	Diversification clientèle	client / diversifier / développer / souhaiter / grand / chercher / marché / taille / part / série / clientèle / portefeuille / rester / donneur / dépendance / petit / se / ses / répondre / gagner / élargir / recherche / étranger / travailler / entreprise / compte
2	Leader mondial	mondial / marché / société / leader / racheter / devenir / américain / bureau / français / allemand / concurrent / souhaiter / national / référence / positionner / acteur / étude / distributeur / pénétrer / présenter / stratégique / européen / *proj_export / rang / constituer / le
3	Diversification produit	produit / gamme / *proj_diversificationproduit / développement / produire / haut / ajouter / propre / valeur / innovants / nouveau / innovation / matériau / bio / design / innovant / conception / diversification / commercialisation / composite / poursuite / durable / différencier / niche / technique / procédé
4	diversification marché	diversification / énergie / secteur / renouvelables / véhicule / électrique / éolien / industrie / aéronautique / médical / ferroviaire / solaire / nucléaire / aéro / vers / *proj_diversificationmarches / marine / chimie / spatial / *proj_diversificationproduit / automobile / air / santé / photovoltaïque / naval / domaine
5	cessation transmission	*proj_transmission / judiciaire / redressement / transmission / reprise / cession / liquidation / plan / fils / repreneur / pse / retraite / céder / salarié / dirigeant / suite / restructuration / rj / transmettre / observation / départ / juillet / actionnaire / continuation / fermeture / lever
6	baisse d'activité	baisser / *proj_baisseactivité / crise / difficile / commande / difficulté / situation / chiffre_affaires / charger / carnet / activité / perte / 2009 / année / subir / économique / trésorerie / *reg_43 / financier / partiel / retrouver / conjoncture / chômage / rentabilité / 2008 / face
7	objectif commercial	commercial / stratégie / terme / réseau / démarche / force / moyen / courir / ambition / marketing / structurer / cohérent / renforcer / structuration / qualité / définir / poursuivre / certification / réel / renforcement / potentiel / continuer / consolider / mener / prospects / effort
8	Investissement immobilier	site / extension / bâtiment / déménagement / local / création / projet / *proj_reorga_interne / construction / agrandissement / immobilier / usine / m2 / construire / prévoir / emploi / atelier / déménager / regrouper / commun / transfert / saint / neuf / centre / st / logistique
9	investissement matériel	investissement / production / ligne / machine / unité / modernisation / matériel / nouvelle / projet / capacité / traitement / usinage / outil / achat / *ac_industrie / d / acquisition / investir / peinture / découper / productivité / automatiser / bois / automatisation / installation / fabrication
10	développement offre de service	service / offrir / logiciel / *ac_informationcommunication / proposer / solution / *aceff_info_commms20sal / mobile / gestion / web / informatique / donnée / communication / internet / *proj_diversificationproduit / plateforme / apporter / éditeur / open / tic / analyse / virtuel / *reg_11 / source / mode / global
11	Prospection de marchés à l'international	*proj_export / coface / allemagne / prospection / chine / brésil / pays / assurance / russie / amérique / japon / canada / etats_unis / espagne / inde / unir / italie / ap / suisse / user / afrique / asie / europe / royaume / belgique / maghreb

[1] *Sont également renseignées les modalités surreprésentées dans la classe. Elles s'écrivent de la forme « *variable_modalité »*

Les projets des classes n°1 et n°2 tournent autour de la notion de « clientèle ». La première concerne plutôt une problématique de diversification globale de la clientèle, tandis que la seconde concerne des projets relatifs à l'augmentation de sa clientèle internationale dans l'objectif de devenir un leader mondial (dans le domaine d'activité de l'entreprise en question).

Les classes n°3 et n°4, représentant respectivement 13,9 % et 4,5 % des projets classés, sont assez proches car elles portent sur des projets de « diversification » :

- La classe n°3 regroupe les projets de diversifications tournés autour du produit. Elle concerne les entreprises qui souhaitent créer des produits innovants ou tourner leurs produits vers le haut de gamme, le bio ou vers des marchés de niche.
- La classe n°4 concerne également des projets de diversifications mais cette fois sur la problématique des secteurs d'activités de l'entreprise. Les projets de cette classe évoquent les nouveaux marchés que l'entreprise veut atteindre ou, au contraire, ceux qu'elle veut abandonner. C'est pourquoi, les différents secteurs d'activité de l'économie y sont surreprésentés (automobile, aéronautique, énergie, ferroviaire, etc.).

Les projets des classes n°5 et n°6 font références à des projets assez pessimistes :

- La classe n°5 porte sur des projets plutôt liés à la fin de vie de l'entreprise. D'un côté, on observe en quelque sorte la fin de vie naturelle de l'entreprise avec des projets de transmission d'entreprises (on parle également de dirigeant âgé qui cherchent à partir en retraite). De l'autre, on parle de fin de vie de l'entreprise (ou d'une partie de l'activité) liée à la conjoncture économique : redressement, liquidation judiciaire, fermeture.
- La classe n°6 met en avant les difficultés économiques rencontrées par les entreprises et les impacts sur leurs projets. Dans ces projets, on parle de baisse d'activité ou de carnet de commandes, de situation difficile ou directement de la crise économique. On voit que les projets des entreprises de la région Franche comté (*reg_43) sont surreprésentés dans cette classe. Ce qui s'explique sans doute par le poids très important dans la région de la filière automobile (également surreprésenté dans cette classe) qui a fortement subi la crise économique.

La classe n°7, qui représente 7,2 % des projets classés concerne principalement des projets commerciaux. L'entreprise évoque la mise en place d'une stratégie commerciale ou marketing pour renforcer (ou consolider) ses ventes et sa présence à l'international notamment.

Les classes n°8 et n°9 représentent légèrement plus du quart des projets classés. Elles ont un vocabulaire en commun car elles rassemblent toutes les deux des projets d'investissement :

- La classe n°8 concerne plutôt les investissements de type immobilier. On y parle d'extension, d'agrandissement voire de construction de bâtiments ou de site de production.
- La classe n°9 concerne quant à elle l'investissement matériel. Les projets de cette classe parlent de modernisation ou d'acquisition de machines dans un objectif d'automatisation de la

production. On remarque que cette classe concerne principalement les entreprises industrielles (**ac_industrie*).

La classe n°10 porte sur des projets qui évoquent des solutions ou des offres de services. Cette classe concerne principalement les entreprises non-industrielles et particulièrement celles de l'information et de la communication.

Pour la classe n°11, il est évident au vu des formes surreprésentées dans cette classe qu'on parle ici de projets à l'exportation. En effet, on retrouve la quasi-totalité des pays (hormis la France) ou continents ainsi que les formes « *prospections* » ou « *export* ». La variable qualitative «principale projet de développement » confirme également cette thématique d'exportation car, pour 83 % des projets de la classe n°11, cette variable a pour valeur la modalité « projet à l'exportation » contre 18 % en moyenne.

3.2.2 Les effets de la crise sur les projets des entreprises (premiers résultats)

Nous regarderons en partie III.5 tous les problèmes de représentativité relatifs à l'étude de ces entreprises visitées. Pour le moment, et afin de limiter en partie le biais de sélection qui peut sembler important dans le cadre de ces données, nous n'allons observer que l'évolution dans le temps du poids des 11 classes détaillées précédemment. Nous supposons donc qu'au cours de la période 2008-2013 la répartition des entreprises visitées est restée la même et qu'aucune nouvelle mesure de la DGE n'est venue entre temps interférer sur la sélection des entreprises. Nous supposons également que les évolutions dans le temps des projets de développement observées au sein des entreprises visitées s'approchent des évolutions réelles qui aurait pu être observées au sein des entreprises françaises. Sous ces hypothèses, nous étudions l'évolution annuelle puis trimestrielle des onze classes mises en avant par la CHD afin d'observer les effets de la crise sur les projets des entreprises entre 2008 et 2013.

3.2.2.1 Année par année :

Pour étudier les évolutions annuelles des projets, nous regardons année par année la surreprésentation (ou sous-représentation) de chacune des 11 classes par rapport à sa moyenne sur la période 2008-2013. Nous étudions cette différence à la moyenne grâce à la distance du Khi^2. Pour cela, nous calculons les Khi^2 signés à un degré de liberté sur les tableaux de contingence croisant la variable année et les 11 classes de la CHD (Figure 6). Pour rappel, le Khi^2 à un degré de liberté et sa version signée sont calculés de la manière suivante :

- $Khi^2 = \frac{(eff_{observé} - eff_{théorique})^2}{eff_{théorique}}$
- $Khi^2 signé = signe(eff_{observé} - eff_{théorique}) \frac{(eff_{observé} - eff_{théorique})^2}{eff_{théorique}}$

Le Khi^2 signé permet de distinguer les éléments sous-représentés (Khi^2 signé inférieur à zéro) des éléments surreprésentés (Khi^2 signé supérieur à zéro). Plus la valeur du Khi^2 est élevée, plus la différence entre l'effectif observé et l'effectif théorique est significative sur le plan statistique. Si le Khi^2 calculé pour une année et une classe donnée est supérieur à $2{,}71$, on peut affirmer avec un risque d'erreur de 10 % que cette classe est sur ou sous-représentée cette année-là.

Par exemple, pour la classe « baisse d'activité » en 2011, le nombre total (sur la période 2008-2013) de segments de texte appartenant à la classe « baisse d'activité » étant de 1 678 et le nombre de segments de texte classés de 15 819 (dont 2 819 en 2011), on en déduit un $eff_{théorique}$ de $\frac{1\,678}{15\,819} * 2\,819 = 299$. Or, nous observons cette année-là 251 segments de texte dans cette classe, le Khi^2 vaut donc $\frac{(251-299)^2}{299} =$ 7,7 et le Khi^2 signé vaut -7,7.

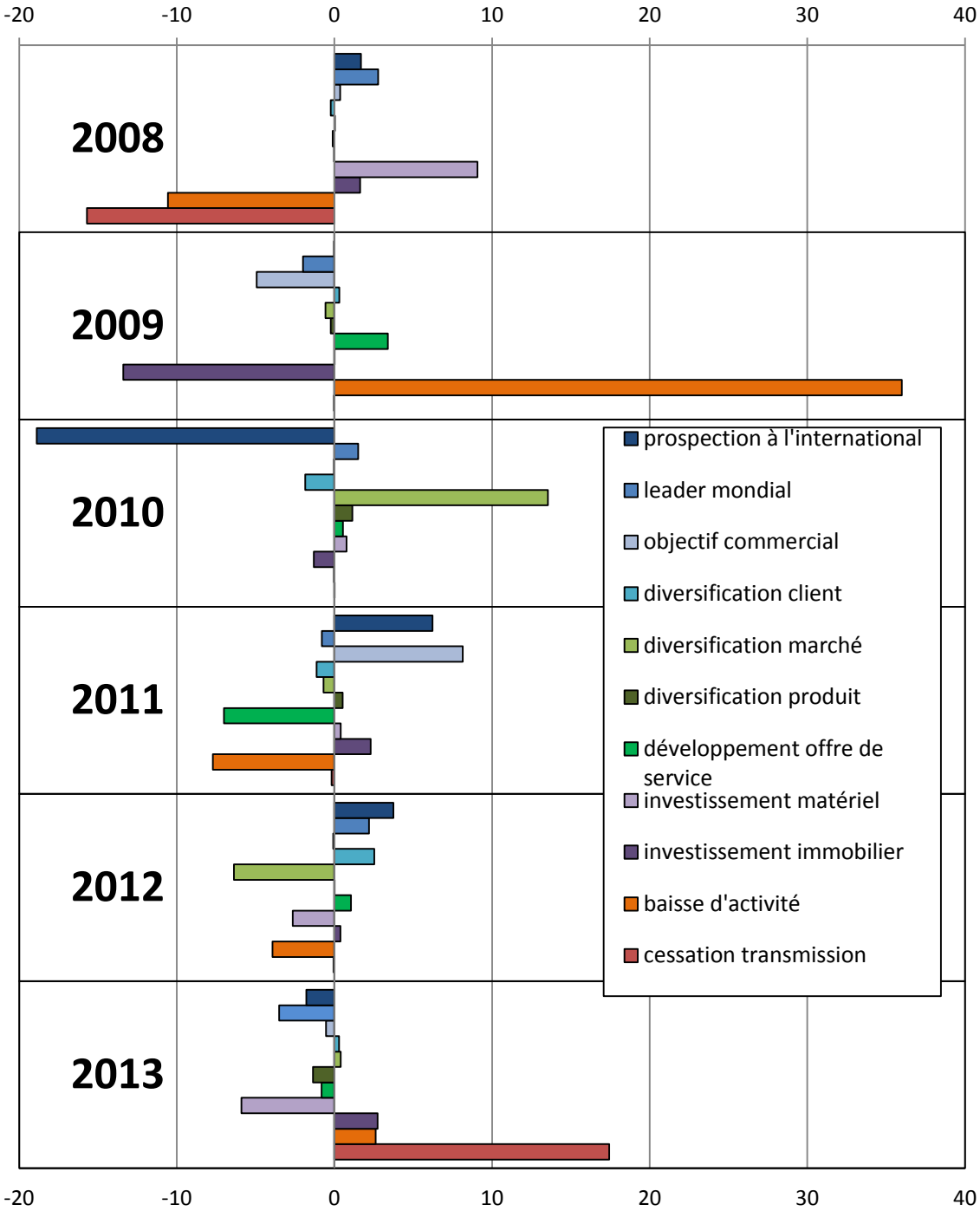


Figure 6 : Evolution année par année du poids des 11 classes (Khi^2 signé)

2008 : LA CRISE N'IMPACTE PAS ENCORE LES PROJETS DES ENTREPRISES

La crise ne semble pas encore impacter les projets des entreprises en 2008. En effet, on parle beaucoup moins de « baisse d'activité » ou de « cessation transmission » d'entreprise cette année que sur la période 2009-2013 (Khi^2 signé de respectivement – 10,6 et – 15,7). Les projets d'investissements matériels sont au contraire un peu plus souvent évoqués que la moyenne (Khi^2 signé de + 9,1).

2009 : ON PARLE BAISSE D'ACTIVITE !

En 2009, la crise touche les entreprises. On parle beaucoup plus de baisse d'activité dans les projets des entreprises (Khi^2 signé de + 36).

2010 : LA CRISE EST MONDIALE, L'ENTREPRISE VEUT SE DIVERSIFIER !

En 2010, la crise est mondiale, tous les pays sont touchés et on parle beaucoup moins de projets à l'export (Khi^2 signé de - 18,9). L'entreprise se retourne vers sa production et ce qu'elle produit. Elle souhaite se diversifier et changer de secteur d'activité (Khi^2 signé pour la classe « diversification marché » de + 13,5). Cette diversification de marchés provient (sans doute) des fortes difficultés économiques rencontrées dans la filière automobile.

2011 : ON IMAGINE LA FIN DE LA CRISE. ON REPARLE EXPORT, PROJETS COMMERCIAUX…

On imagine la fin de la crise. Les entreprises parlent beaucoup moins de baisse d'activité (Khi^2 signé de – 7,7). On reparle de projets à l'export et de projets commerciaux (Khi^2 signé de respectivement + 6,2 et + 8,1).

2012 : LES PROJETS A L'EXPORT SONT TOUJOURS FAVORABLEMENT ORIENTES… MAIS L'INVESTISSEMENT COMMENCE A ETRE MOINS PRESENT.

Dans la continuité avec l'année 2011, on parle encore légèrement plus d'international (Khi^2 signé de + 3,7 pour la classe « prospection à l'international » et + 2,2 pour la classe « leader mondial »). Les projets de baisse d'activité sont toujours moins présents qu'en moyenne sur la période 2008-2013 (- 3,9). On constate cependant un début de baisse d'investissement (Khi^2 signé de – 2,6 pour la classe « investissement matériel »). De même, on parle un peu moins de « diversification marchés », une tendance qui pourrait être la conséquence d'une sur-diversification deux ans auparavant. Globalement, l'année 2012 apparait comme une année moyenne sur la période 2008-2013 : aucun Khi^2 ne dépasse la valeur de 7.

2013 : LA CRISE IMPACTE DE NOUVEAU FORTEMENT LES ENTREPRISES : REDRESSEMENT JUDICIAIRE, BAISSE D'ACTIVITE ET DEMENAGEMENT DE LA PRODUCTION…

L'année 2013 marque le retour de la crise. On parle beaucoup plus de cessions d'entreprises redressement ou liquidation judiciaire (Khi^2 signé + 17,4). On parle également dans une moindre mesure un peu plus d'investissement immobilier (Khi^2 signé de + 2,7). Ce qui n'est également pas bon signe car au vu des mots surreprésentés dans cette classe, les projets de cette classe parlent beaucoup de déménagement de site (délocalisation) de réorganisation interne. Les projets d'investissements matériels semblent encore moins nombreux cette année (Khi^2 signé de – 5,9), ce qui confirme la tendance baissière observée en 2012.

On peut également étudier le premier plan de l'AFC réalisé sur le tableau de contingence croisant les formes « pleines » avec les 11 classes résultant de la CHD. Cette AFC est disponible dans le logiciel IRaMuTeQ. La Figure 7 représente les formes sur le 1er plan factoriel en fonction de leur classe d'appartenance. Comme pouvait le faire apparaitre l'organisation des mondes lexicaux (Figure 5), on remarque sur ce 1er plan que les deux classes « baisse d'activité » et « cessation transmission » semblent assez proches et font en grande partie allusion à un vocabulaire relatif à des difficultés

économiques. Seules trois classes contribuent fortement à la création du premier plan factoriel (Figure 8) : la classe n°5 « cessation transmission », la classe n°6 « baisse d'activité » et la classe n°11 « prospection à l'international ».

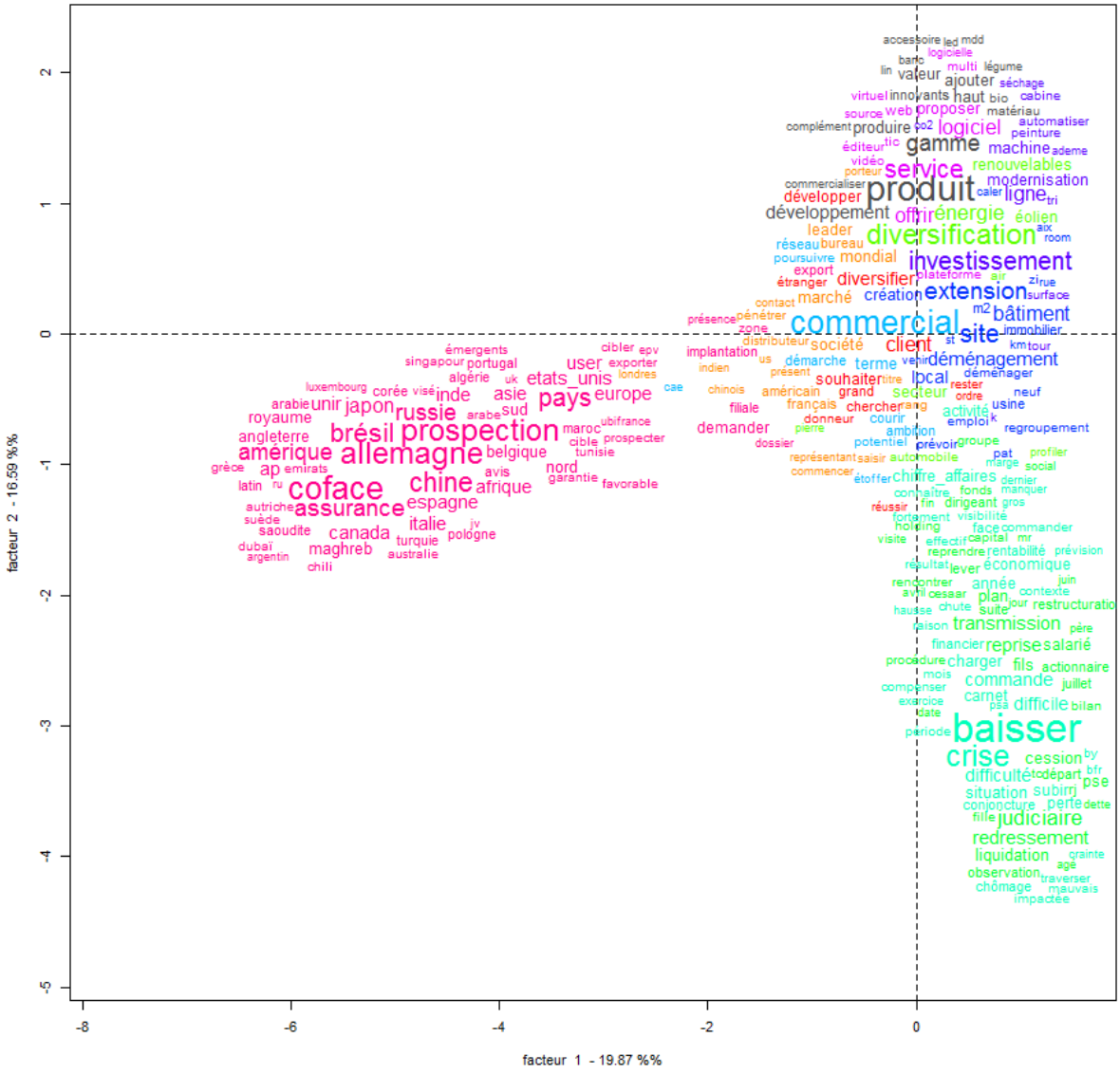


Figure 7 : Représentation des formes « pleines » sur le 1er plan factoriel

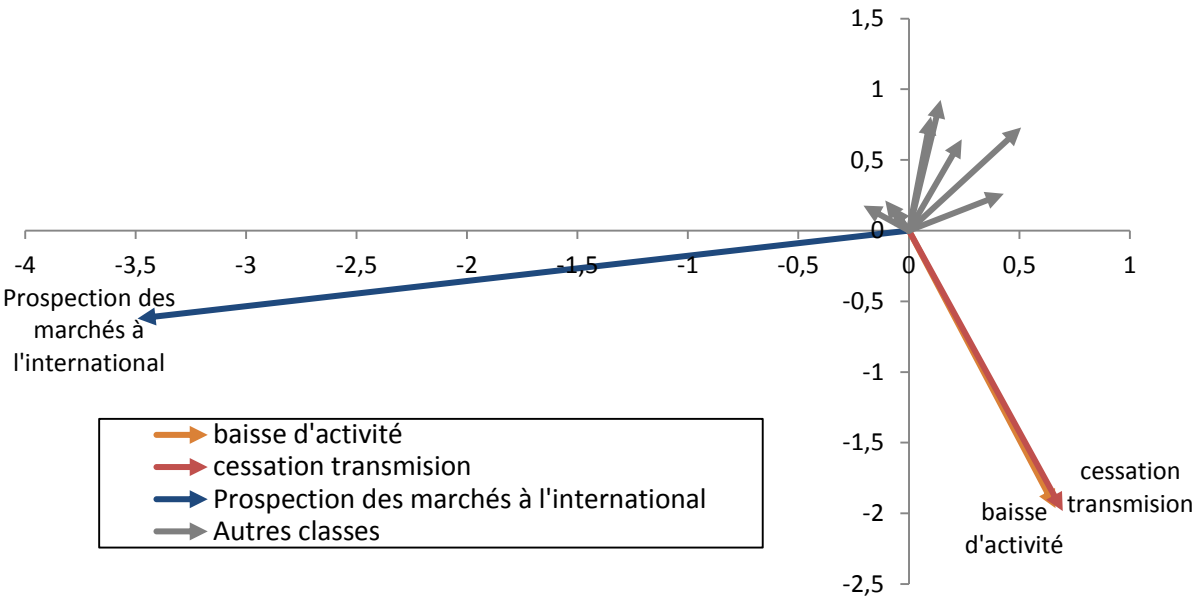


Figure 8 : Représentation des classes sur le 1er plan factoriel

Nous pouvons ensuite regarder comment des variables supplémentaires se comportent sur le 1er plan factoriel. Si les différentes modalités sont suffisamment bien représentées sur le 1er plan factoriel (sommes des qualités de représentation sur les deux premiers axes du plan proche de 1), alors nous pouvons interpréter les proximités entre ces modalités (quelles proviennent d'une même variable ou de variables différentes). Si nous nous intéressons à la variable relative à l'année du projet de développement, nous pouvons remarquer que ce 1er plan factoriel représente un résumé de ce que nous avons observé précédemment (ce qui est l'objectif même du 1er plan factoriel). En effet, la Figure 9 montre que nous sommes passés d'une année 2008 où les difficultés semblaient moins importantes à une année 2009 où les difficultés économiques étaient beaucoup plus présentes. L'année 2010, est surtout marquée par une baisse des projets à l'exportation (coordonnée très positive sur l'axe 1). Les années 2011 et 2012 repassent de l'autre côté signe d'exportations qui redémarrent. L'année 2013 redevient proche de l'année 2009, signe cette fois de difficultés qui impactent de nouveau les entreprises.

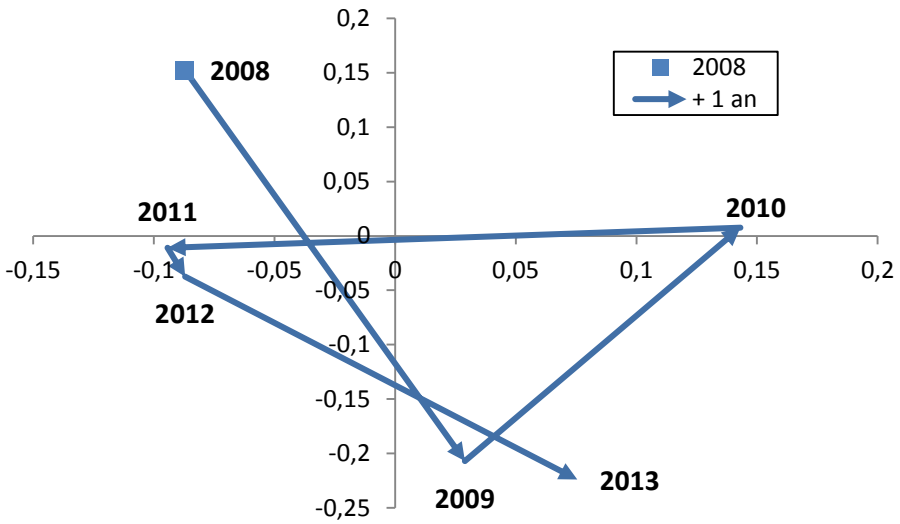


Figure 9 : Représentation des années sur le 1er plan factoriel

3.2.2.2 Trimestre par trimestre :

Dans la continuité de ce que nous venons d'observer, nous pouvons essayer d'analyser les projets des entreprises au niveau trimestriel. Pour cela, nous allons d'abord regarder comment se caractérisent les différents trimestres intervenus entre 2008 et 2013 sur le 1er plan factoriel. La Figure 10 montre clairement un lien entre deux trimestres consécutifs. On observe que les difficultés économiques liées à la crise débutent au 4ème trimestre de 2008 pour atteindre un pic au 4ème trimestre 2009, trimestre marqué par de très fortes difficultés économiques d'un côté et des projets à l'export plus rares de l'autre. Les projets à l'export sont restés plus rares tout au long de 2010, et n'ont vraiment repris qu'à partir du 2ème trimestre 2011 (passage sur le 1er axe d'une coordonnée positive de + 0,10 au T1 à une coordonnée négative de – 0,14 au T2). Les difficultés économiques semblent également moins impactées les projets des entreprises sur cette période. Cette légère reprise apparait de courte durée. En effet, dès le 2ème trimestre 2012, tous les trimestres qui suivent semblent marqués par des difficultés économiques de plus en plus présentes.

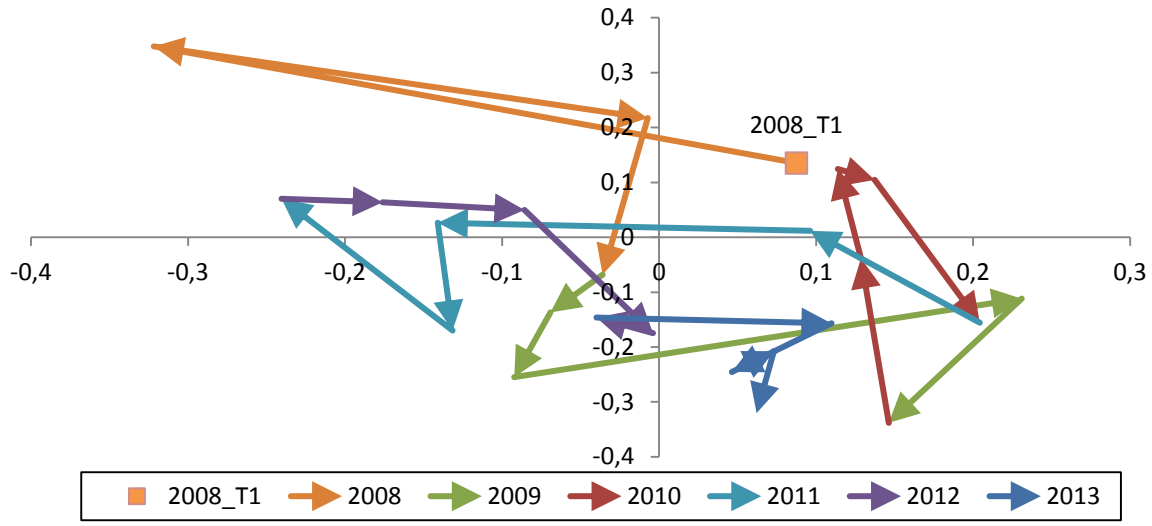


Figure 10 : Représentation des trimestres de 2008 à 2013 sur le 1er plan factoriel.

Au vu des seules trois classes contributives et bien représentées sur le premier axe factoriel nous pouvons directement regarder l'évolution trimestrielle du poids de ces trois classes dans les projets des entreprises. On perd ainsi les défauts de l'interprétation d'un 1er plan factoriel qui aurait nécessité de vérifier la représentativité de chaque trimestre selon les deux premiers axes. Cette analyse complémentaire (Figure 11) permet d'une part de conforter l'analyse trimestrielle faite du 1er plan factoriel et, d'autre part, d'aller plus loin dans l'interprétation. En effet, elle permet de dissocier la classe « cessation transmission » de la classe « baisse d'activité ». Nous observons d'ailleurs que les conséquences des difficultés économiques rencontrées depuis fin 2012 ne sont pas les mêmes que celles rencontrées en 2009. Les cessations, redressements et liquidation judiciaires semblent ainsi beaucoup plus impactées les entreprises en 2013 (Khi^2 signé supérieur à 2 chaque trimestre et égal à + 11,5 au 2ème trimestre).

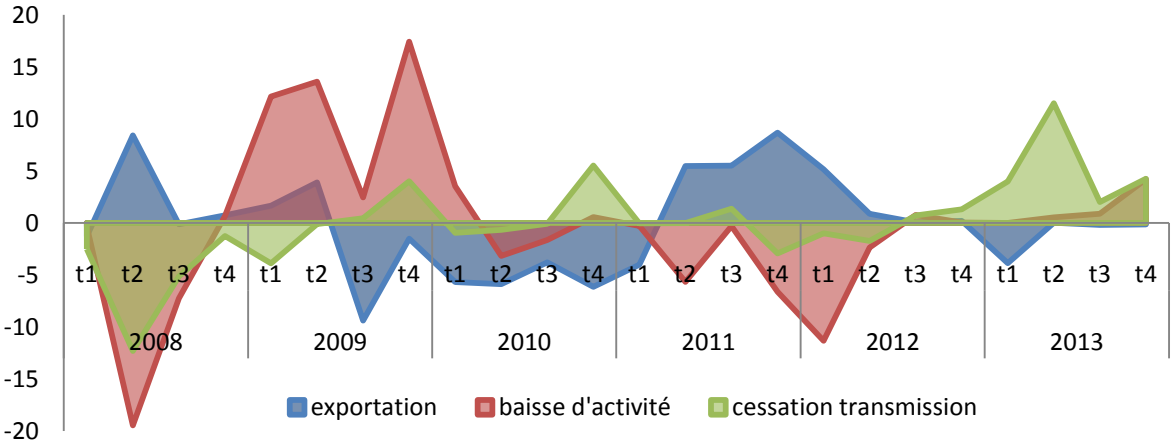


Figure 11 : Evolution trimestrielle du poids des classes n°5, 6 et 11 (Khi^2 signé)

3.3 Une deuxième classification pour aller à un niveau plus fin.

L'organisation des projets de développement des entreprises qui apparaît à travers les onze classes précédentes apporte des informations très intéressantes. En étudiant les évolutions annuelles puis

trimestrielles des segments de texte selon ces onze thématiques, on a pu constater les effets de la crise sur les projets (baisse d'activité, réorientation à l'export, diversification, redressement judiciaire, etc.).

On remarque cependant que certaines classes peuvent comporter différentes thématiques que l'on souhaiterait dissocier (la classe « cessation transmission » par exemple). C'est pourquoi, nous allons ici augmenter le nombre de phase terminale de la classification à 100 et analyser toutes les classes comportant un nombre de segments de texte supérieur à la moyenne (17851/100 = 179 commentaires). Nous élargissons également l'analyse aux formes pleines dont la fréquence est supérieure ou égale à 3 (contre >=6 précédemment). Nous arrivons cette fois à un total de 32 classes à analyser (Figure 12).

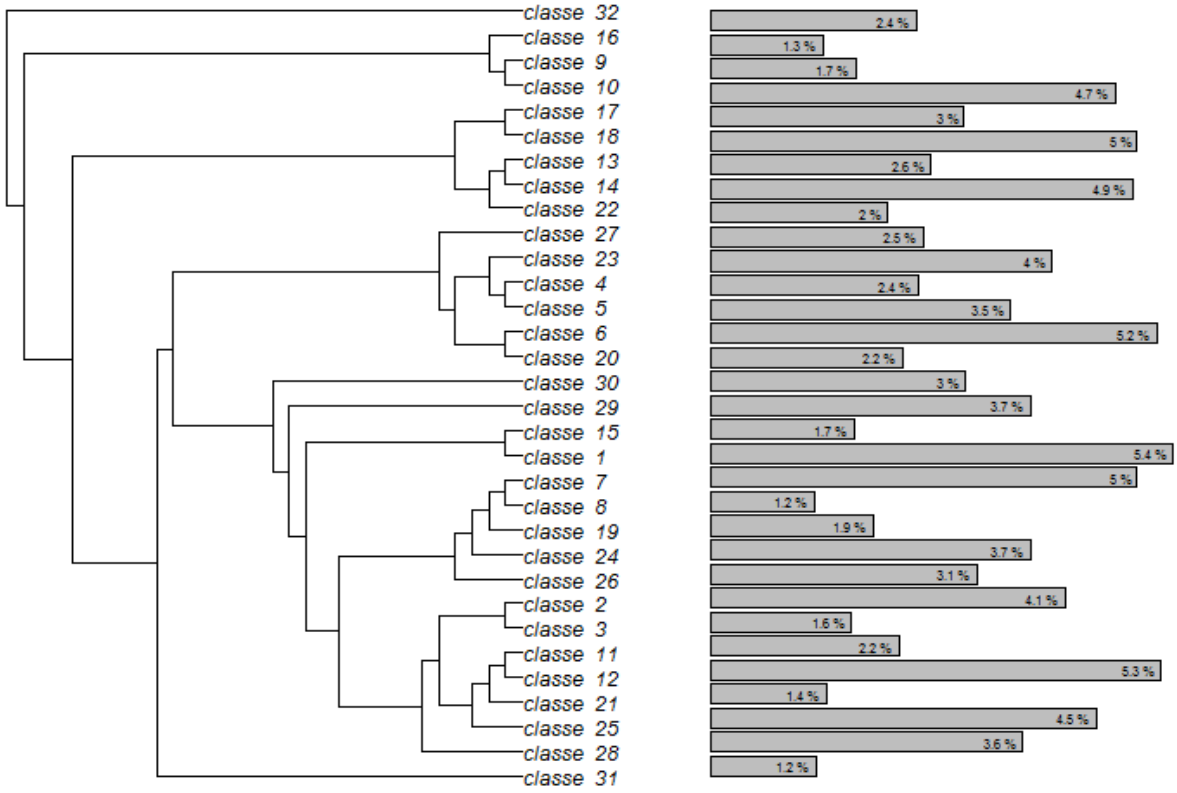


Figure 12 : Les 32 mondes lexicaux du corpus « projets de développements ».

Le principe même de la CHD est de partitionner la plus grosse classe à chaque étape. Ce qui explique qu'aucune d'entre elles ne dépasse les 6 % de commentaires classées alors que dans la classification précédente certaines classes représentaient près de 14 % des segments de texte classés. Malgré que nous ayons augmenté le nombre de formes pleines analysées, le Tableau 7 montre que ces nouvelles classes sont étroitement liées avec celles de la 1ère classification à 11 classes.

Tableau 7 : Répartition des commentaires selon les deux modèles.

		Anciennes classes (1ère classification)												
		Non classée	1	2	3	4	5	6	7	8	9	10	11	Ensemble
Nouvelles classes (2ème classification)	non classée	24%	3%	3%	14%	4%	2%	10%	3%	20%	9%	6%	2%	100%
	1	4%	2%	4%	13%	4%	0%	0%	2%	2%	5%	63%	0%	100%
	2	1%	24%	14%	5%	2%	1%	0%	28%	6%	2%	17%	0%	100%
	3	4%	19%	36%	1%	1%	1%	1%	23%	5%	4%	5%	0%	100%
	4	1%	2%	1%	2%	1%	1%	0%	1%	6%	86%	0%	0%	100%
	5	2%	0%	0%	2%	1%	1%	1%	0%	24%	68%	0%	0%	100%
	6	4%	1%	0%	1%	0%	1%	0%	1%	83%	7%	1%	0%	100%
	7	6%	2%	2%	65%	3%	0%	0%	6%	4%	3%	7%	0%	100%
	8	7%	9%	3%	50%	7%	1%	0%	8%	5%	5%	7%	0%	100%
	9	0%	0%	4%	1%	0%	0%	0%	2%	1%	2%	1%	89%	100%
	10	1%	2%	3%	1%	0%	0%	0%	2%	1%	0%	1%	89%	100%
	11	1%	21%	20%	4%	2%	0%	0%	14%	5%	4%	29%	0%	100%
	12	1%	23%	13%	3%	2%	0%	1%	41%	3%	1%	10%	0%	100%
	13	0%	1%	1%	0%	0%	3%	91%	0%	1%	1%	1%	0%	100%
	14	0%	1%	0%	1%	1%	1%	94%	0%	0%	0%	1%	0%	100%
	15	5%	2%	4%	30%	22%	1%	1%	1%	6%	21%	8%	0%	100%
	16	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	98%	100%
	17	1%	1%	0%	1%	1%	90%	1%	1%	2%	0%	0%	0%	100%
	18	1%	1%	1%	0%	0%	89%	4%	0%	3%	0%	0%	0%	100%
	19	3%	2%	4%	70%	5%	1%	0%	4%	2%	7%	3%	0%	100%
	20	2%	1%	2%	1%	1%	1%	1%	1%	76%	13%	2%	0%	100%
	21	1%	13%	14%	4%	2%	0%	1%	42%	4%	4%	16%	0%	100%
	22	0%	1%	1%	0%	0%	14%	82%	0%	1%	0%	1%	0%	100%
	23	2%	4%	1%	1%	1%	0%	1%	2%	21%	66%	1%	0%	100%
	24	1%	4%	3%	77%	2%	0%	0%	4%	1%	3%	5%	0%	100%
	25	2%	12%	46%	5%	2%	1%	1%	10%	10%	3%	9%	0%	100%
	26	1%	2%	1%	33%	55%	0%	1%	1%	2%	1%	2%	0%	100%
	27	4%	2%	3%	12%	5%	1%	0%	2%	6%	63%	2%	0%	100%
	28	4%	45%	13%	2%	15%	1%	1%	5%	3%	2%	8%	0%	100%
	29	78%	1%	2%	5%	1%	1%	2%	3%	4%	2%	1%	0%	100%
	30	69%	2%	2%	5%	1%	1%	0%	3%	6%	8%	3%	0%	100%
	31	91%	2%	0%	1%	0%	0%	0%	0%	3%	1%	0%	0%	100%
	32	99%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	100%

Le fait d'augmenter le nombre d'étape dans la classification va permettre de répartir l'information dans un nombre plus important de classes et donc permettre de décomposer certains phénomènes observés à partir du modèle à 11 classes. De plus, le fait d'avoir diminué le seuil minimum de segments texte dans la prise en compte d'une classe laisse apparaître des classes qui avaient été considérées comme non-classées dans la classification précédente.

3.3.1 Pour comprendre plus en détail les classes précédentes.

Au vu du

Tableau 7, nous remarquons que le vocabulaire spécifique des classes n°5, n°6, n°9 et n°11 de la première classification peut être en partie décomposé grâce à l'analyse de cette deuxième classification. Nous nous intéresserons particulièrement à l'interprétation de la décomposition de ces deux premières classes (la classe n°5 « cessation transmission » et la classe n°6 « baisse d'activité »).

Le même travail consistant à synthétiser le vocabulaire dominant par classe a été effectué et est disponible en annexe. Nous allons regarder comment évolue dans le temps (année par année) ces nouvelles composantes. L'objectif est de détailler les évolutions observées précédemment, et par la même occasion de vérifier que les conclusions précédentes restent cohérentes.

3.3.1.1 Décomposition de la classe « cessation transmission »

Une des faiblesses de l'analyse de la classe « cessation transmission » (classe n°5 de la première classification) est le fait qu'elle porte sur deux phénomènes assez différents. D'un côté on parle de transmission d'entreprises, de l'autre, on parle de fin de vie de l'entreprise (ou d'une partie de l'activité) liée à la conjoncture économique : redressement judiciaire, pse, fermeture. Cette 2ème classification plus fine permet de répondre en partie à cette problématique vu qu'elle permet de mettre de côté tout ce qui est lié à la transmission naturelle de l'entreprise. On remarque d'ailleurs sur la Figure 13 que ces projets de transmission ne semblent pas trop influencés par la conjoncture. Au contraire, les projets où l'on parle de fermeture et de redressements judiciaires sont beaucoup plus présents en 2013 (Khi^2 signé de + 21,6) qu'ils ne l'étaient sur la période 2008-2012. Ce qui conforte l'interprétation faite en partie III.2.2 sur l'évolution dans le temps de la classe « cessation transmission ».

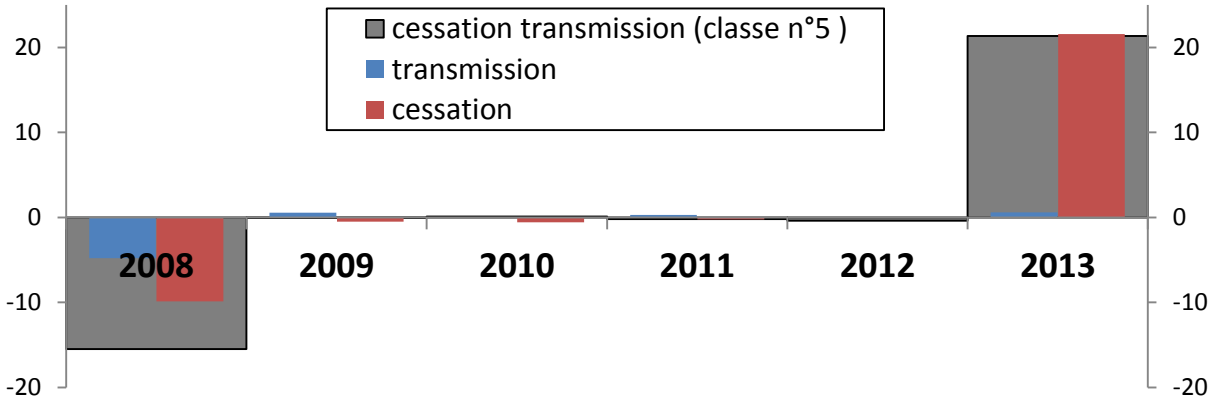


Figure 13 : Evolution annuelle du poids (Khi^2 signé) des composantes de la classe « cessation transmission »

3.3.1.2 Décomposition de la classe « baisse d'activité »

La classe « baisse d'activité » (classe n°5 de la 1ère classification) est décomposée en trois classes. Une première traite des difficultés rencontrées dans l'année (ou dans le trimestre) en termes d'activité de résultats ou de carnet de commande. On remarque sur la Figure 14 que cette classe « activité difficile » n'évolue pas ou très peu dans le temps. La deuxième classe se réfère aux mêmes difficultés

économiques mais la cause est cette fois identifiée. On parle de la crise économique que l'entreprise a subit et à laquelle elle doit faire face. C'est cette classe qui explique le pic de 2009 que nous avions observé lors de la première classification (Khi^2 signé de + 62,2 pour la classe « crise économique »). La dernière classe qui parle principalement de difficultés financières liées à des problèmes de trésorerie, de besoin de fonds de roulement (bfr) ou de paiement, permet d'expliquer en partie la dégradation de la conjoncture économique en 2013. En effet, on remarque que les projets évoquant des difficultés financières sont bien plus fréquents en 2013 (Khi^2 signé de + 16,1).

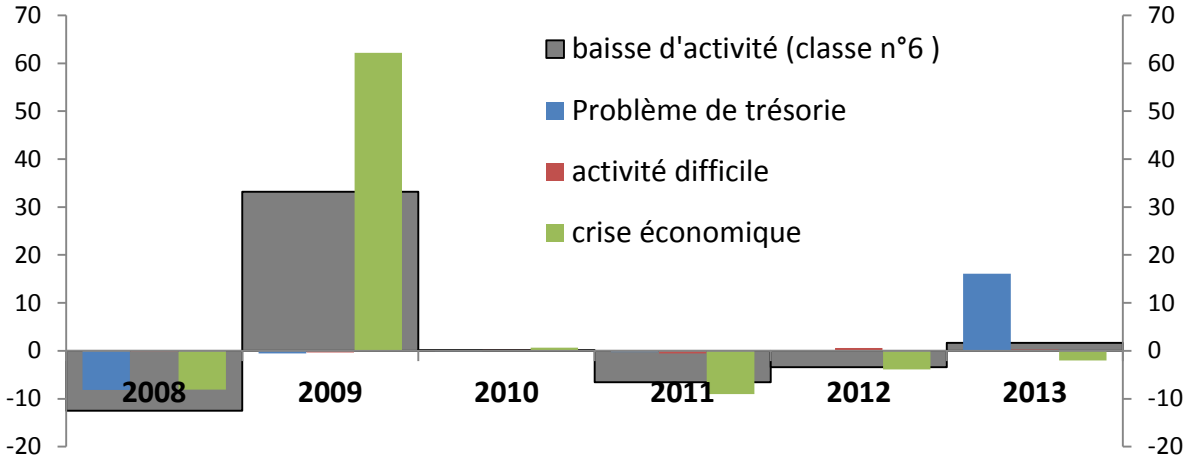


Figure 14 : Evolution annuelle du poids (Khi^2 signé) des composantes de la classe « baisse d'activité ».

3.3.2 Pour faire apparaître des nouvelles thématiques

L'autre intérêt de cette 2ème classification est de faire apparaître de nouvelles thématiques non prises en compte dans la 1ère classification. En effet, dans cette 2ème classification nous étudions les classes de plus de 179 segments de texte contre plus de 596 segments de texte pour la 1ère classification. Cette modification du seuil minimum fait donc apparaître des classes qui n'étaient pas analysées dans la 1ère classification. C'est le cas par exemple des classes n°29 à 32. La classe n°30 met en avant les projets qui sont liés à des « actions collectives ». Le dispositif « actions collectives » déployé par les DIRECCTE a pour objectif d'accompagner des actions communes à plusieurs entreprises, dans le but de renforcer le tissu industriel, artisanal et commercial d'un territoire et d'en améliorer les performances. Ce dispositif qui représentait un budget assez important pour les DIRECCTE a diminué considérablement depuis 2009. C'est ce qui semble d'ailleurs apparaître en étudiant l'évolution dans le temps du poids de cette classe : la classe n°30 représentait 5,1 % des projets classés en 2009, 2,9 % en 2010 et 1,6 % en 2013.

Le vocabulaire représentatif de la classe n°32 (« voir », « compte_rendu », « conclusion », « général », etc.) ne semble pas être lié à une thématique de projet d'entreprise. On en comprend rapidement le contenu en lisant quelques segments de texte caractéristique de la classe : « voir conclusion générale », « cf. compte rendu ». Il est clair que les projets qui ont été classés ici ne comportent aucune information utile puisqu'ils signalent seulement que l'information est ailleurs. Après discussion avec certains chargés de missions, ce choix de reporter l'information dans la conclusion

générale de la visite d'entreprise provient principalement de la limitation à 200 caractères imposée pour le commentaire de projets de développement. La limitation du nombre caractères peut donc avoir des conséquences non souhaitées sur le contenu renseigné, surtout si on laisse la possibilité au chargé de mission de reporter l'information ailleurs. L'avantage de cette 2ème classification plus détaillée est qu'elle a permis de repérer ces segments de texte mal renseignés. On peut donc aisément les enlever de l'analyse. Ce qui a d'ailleurs déjà été fait dans l'analyse de la 1ère classification puisque ces segments de texte faisaient partie des non-classés. Ils feront également partie des non-classées dans la 3ème classification.

3.4 Une troisième classification avec des données plus récentes

L'objectif de cette 3ème classification est de montrer que nous pouvons nous servir de cette analyse pour constituer un outil d'analyse conjoncturel efficace. En effet, nous avons remarqué que l'analyse au niveau trimestriel, bien que plus instable, garde tout son intérêt et permet de dégager des tendances. Le logiciel IRaMuTeQ ne permet malheureusement pas d'appliquer le modèle créé ultérieurement à de nouveaux projets. Nous sommes donc obligés de relancer l'analyse sur un corpus complété des nouveaux projets recueillis à la date du 30 mai 2014 (soit 887 projets supplémentaires). Afin de rester sur un champ en partie comparable, nous réalisons une CHD avec un nombre de phases terminales permettant d'obtenir 11 classes à analyser.

Tableau 8 : Répartition des segments de texte selon la 1ère et la 3ème classification.

		Anciennes classes (1ère classification)											
		non classée	1	2	3	4	5	6	7	8	9	10	11
nouvelles classes (3ème classification)	non classée	63%	1%	2%	6%	4%	0%	0%	3%	3%	16%	2%	0%
	1	2%	4%	4%	71%	3%	0%	0%	3%	3%	5%	4%	0%
	2	1%	14%	47%	5%	1%	0%	0%	11%	8%	3%	10%	0%
	3	2%	23%	14%	5%	1%	0%	0%	35%	5%	2%	11%	0%
	4	1%	3%	2%	1%	1%	3%	86%	2%	0%	1%	0%	0%
	5	1%	1%	2%	1%	0%	87%	6%	1%	2%	0%	0%	0%
	6	1%	3%	1%	2%	1%	0%	0%	1%	18%	70%	1%	0%
	7	3%	1%	1%	1%	0%	1%	0%	1%	77%	12%	1%	0%
	8	3%	4%	5%	8%	4%	0%	0%	1%	2%	3%	70%	0%
	9	3%	19%	9%	19%	40%	0%	0%	3%	2%	1%	3%	0%
	10	72%	2%	2%	3%	1%	1%	0%	2%	5%	9%	3%	0%
	11	1%	1%	2%	1%	0%	0%	0%	1%	1%	1%	0%	93%
	Ensemble	11%	6%	6%	12%	4%	8%	9%	6%	11%	11%	7%	7%

Champ : seuls sont comptabilisés les projets de développement (ou segments de texte) communs entre les deux classifications. Les projets de développement (ceux de début 2014) rajoutés dans la 3ème classification sont donc omis.

Ces onze nouvelles classes restent assez proches des onze classes précédentes. Si on observe les 17 851 segments de texte utilisés à la fois dans 1ère et dans la 3ème classification, trois des nouvelles classes apparaissent comme quasiment identiques avec une des anciennes classes (Tableau 8). En effet, la nouvelle classe n°4 a 86 % de ses segments de texte en commun avec l'ancienne classe « baisse d'activité », la classe n°5 en a 87 % avec l'ancienne classe n°5 « cessation transmission », et La classe n°11 en a 93 % avec l'ancienne classe n°11 « prospection à l'international ». Les huit autres classes sont légèrement plus difficiles à expliquer à partir des anciennes classes : elles ont au maximum 77 % de leurs segments de texte en commun avec une ancienne classe.

Devant le travail nécessaire à l'interprétation de ces nouvelles classes, nous allons nous limiter à l'interprétation de l'analyse factorielle des correspondances qui résulte de cette nouvelle classification. En effet, d'une part, les formes les plus contributives à l'axe 1 et à l'axe 2 sont restées les mêmes. D'autre part, les trois classes les plus contributives aux deux premiers axes (Figure 15) sont celles qui ressemblent très fortement aux trois classes les plus contributives de l'AFC résultant de la 1ère classification (Figure 8).

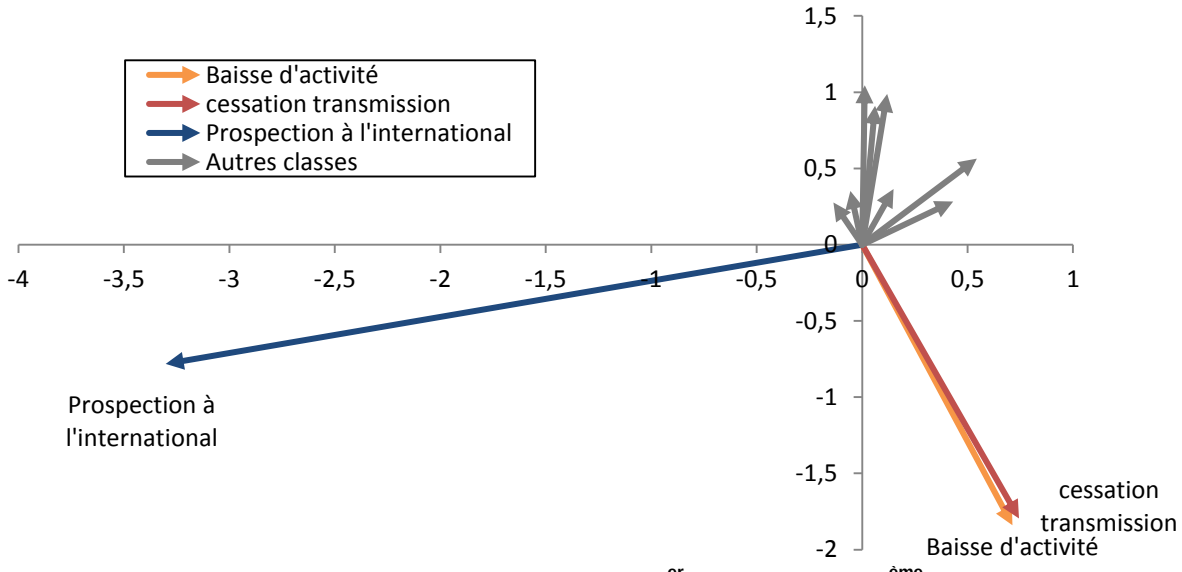


Figure 15 : Représentation des classes sur le 1er plan factoriel (3ème classification)

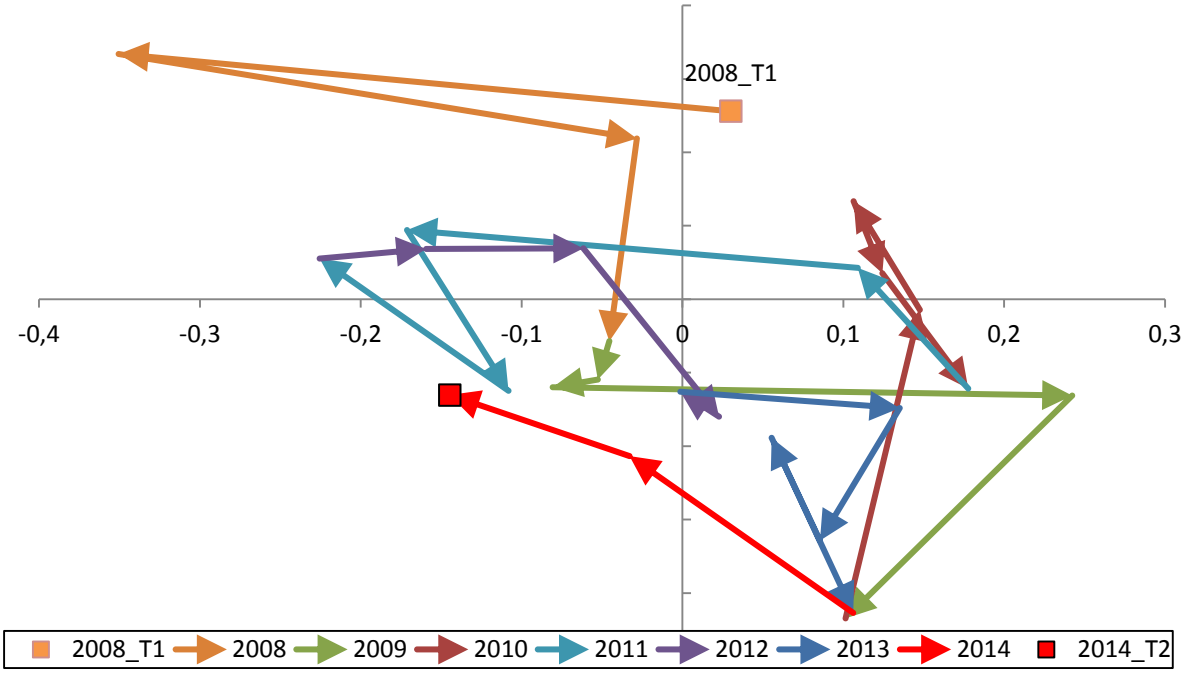


Figure 16 : Représentation sur le 1er plan factoriel du 1er trimestre 2008 au 2ème 2014 (3ème classification)

La construction des deux premiers axes est à peu près équivalente à celle des deux premiers axes de l'AFC réalisée lors de la 1ère classification. Il est donc normal que la représentation sur le 1er plan factoriel des trimestres de 2008 à 2013 (Figure 16) soient à peu près analogue à celle observée en partie III.2.2 (Figure 10). On peut remarquer que la conjoncture semble s'améliorer au vue de la position des deux premiers trimestres 2014 (le deuxième trimestre 2014 n'est pas définitif vu qu'il manque tous les projets de juin 2014). En effet, la conjoncture économique en ce début d'année 2014 semble se rapprocher de celle « légèrement optimiste » de fin d'année 2011.

Si l'on analyse l'évolution temporelle des trois classes n°4,5 et 11 et qu'on les compare aux trois classes correspondantes de la 1ère classification, on remarque que les évolutions sur la période 2008-2013 sont quasiment les mêmes (Figure 17, Figure 18 et Figure 19). Cette similitude peut justifier l'interprétation des tendances « optimistes » qu'on observe en ce début d'année 2014 : un poids des cessations et redressements judiciaires qui se réduit et retrouve un niveau normal, et un léger frémissement des exportations. Cette similitude confirme également qu'il existe une certaine stabilité sur la définition de ces trois classes, et qu'il est donc possible de les interpréter assez rapidement à chaque fois qu'on relance une classification avec de nouveaux segments de texte.

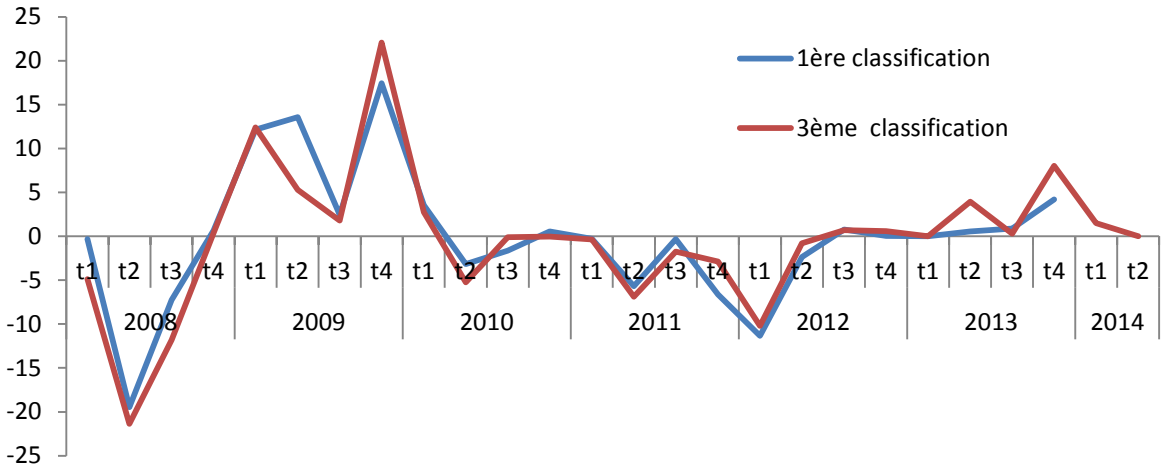


Figure 17 : Comparaison trimestrielle entre la 1ère et de la 3eme classification du poids de la classe « baisse d'activité » (Khi² signé)

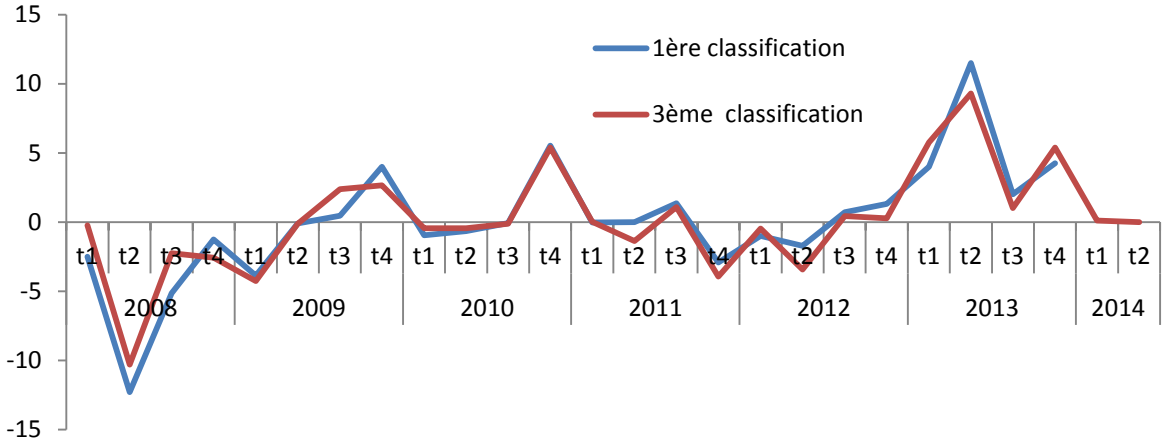


Figure 18 : Comparaison trimestrielle entre la 1ère et de la 3éme classification du poids de la classe « cessation transmission » (Khi² signé)

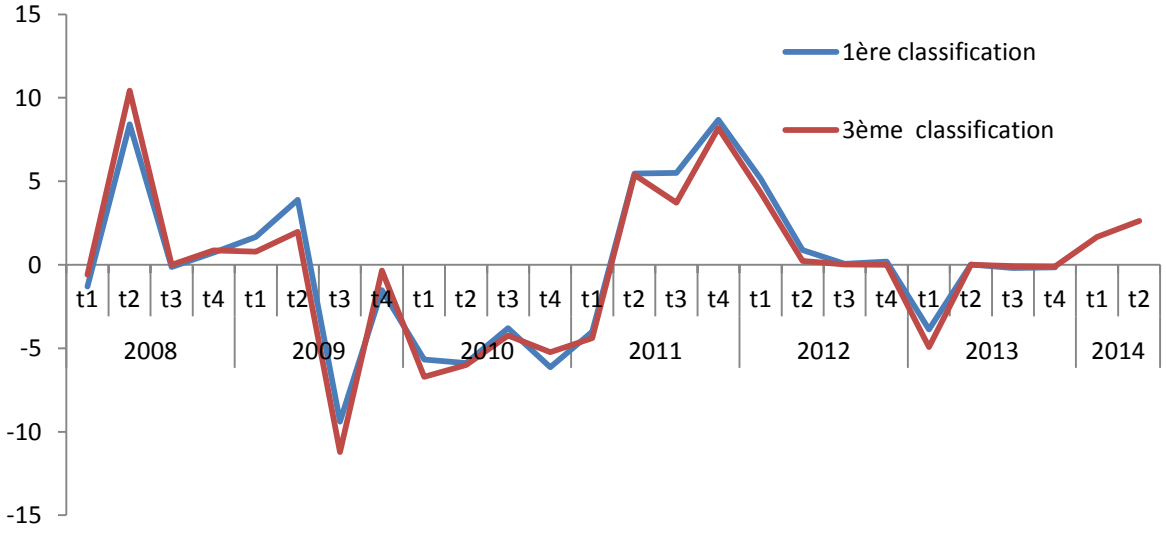


Figure 19 : Comparaison trimestrielle entre la 1ère et de la 3éme classification du poids de la classe « prospection à l'international » (Khi² signé)

3.5 Interprétation des résultats

3.5.1 La problématique de la représentativité des résultats

Une des critiques des résultats mis en avant jusqu'à présent réside dans la représentativité des entreprises visitées. Nous avons supposé jusqu'à présent que les entreprises visitées étaient représentatives des entreprises françaises. Or, l'industrie représente bien plus de la moitié des entreprises visitées alors que, pour rappel, les entreprises industrielles ne représentent que 7 % des entreprises françaises. Il y a donc bien évidemment un problème de représentativité des résultats au vu de la structure des entreprises visitées.

Afin que les entreprises visitées représentent l'ensemble de l'économie française, on pourrait affecter un poids en fonction de la taille ou de l'activité. Malheureusement, la surreprésentation de l'industrie dans les entreprises visitées est telle, que le poids des autres entreprises deviendrait trop important par rapport à ceux de l'industrie (un facteur 43 entre la construction et l'industrie). C'est pourquoi, il est sans doute plus intéressant de regarder les évolutions par secteur voire de se limiter au seul secteur de l'industrie qui représente bien plus de la moitié des projets classés. Enfin, au sein même d'un secteur, les entreprises visitées ne sont pas représentatives des entreprises françaises en termes de taille. En effet, plus de la moitié des entreprises visitées ont plus de 20 salariés contre seulement 3 % pour l'ensemble des entreprises françaises. De la même façon que précédemment, la création de pondérations pour être représentatif des entreprises française en termes de nombre donnerait un poids trop faible à toutes les grandes entreprises. Ce qui n'est pas intéressant au vu du poids qu'elles représentent en termes d'activité. On pourrait d'ailleurs essayer de créer une pondération pour être représentatif en termes de chiffre d'affaires ou de valeur ajoutée mais on serait confronté à un nouveau problème : les données fiscales sont disponibles avec trois ans de retard et peuvent fortement variées d'une année sur l'autre (surtout en période de crise).

On comprend donc que l'étape envisagée de calage des marges sur les entreprises visitées n'est pas simple et risque surtout, au vu de la structure particulière des entreprises visitées, de donner à certaines entreprises un poids trop important comparativement à d'autres. Ce qui aura pour conséquence de perturber fortement les analyses fines qu'on pourrait faire des données, notamment au niveau trimestriel.

L'exploitation de ces données textuelles sur les projets des entreprises n'a peut-être pas vocation, pour le moment du moins, à fournir des résultats exacts mais elle a le mérite de dégager des tendances générales. Ce qui semble le plus efficace pour le moment dans l'exploitation de ces données est de faire une analyse secteur par secteur, plus particulièrement sur l'industrie seulement. Ce choix de ne pas pondérer par la taille s'explique en partie par une structure des projets qui n'est pas si différente selon la taille des entreprises. En effet, la Figure 20 montre que pour le secteur de l'industrie, les 11 classes (1ère classification) ont chacune à peu près le même poids quelle que soit la taille de l'entreprise (seuls les projets d'investissements semblent légèrement plus caractéristiques des plus grandes entreprises). C'est pourquoi, les légères évolutions dans le temps de la structure en

termes de tailles ne devraient pas trop impacter les résultats. Il faudra tout de même vérifier que cette évolution de structure reste légère. Au final, Il sera intéressant de compléter l'analyse globale du secteur par une analyse par tranche d'effectifs.

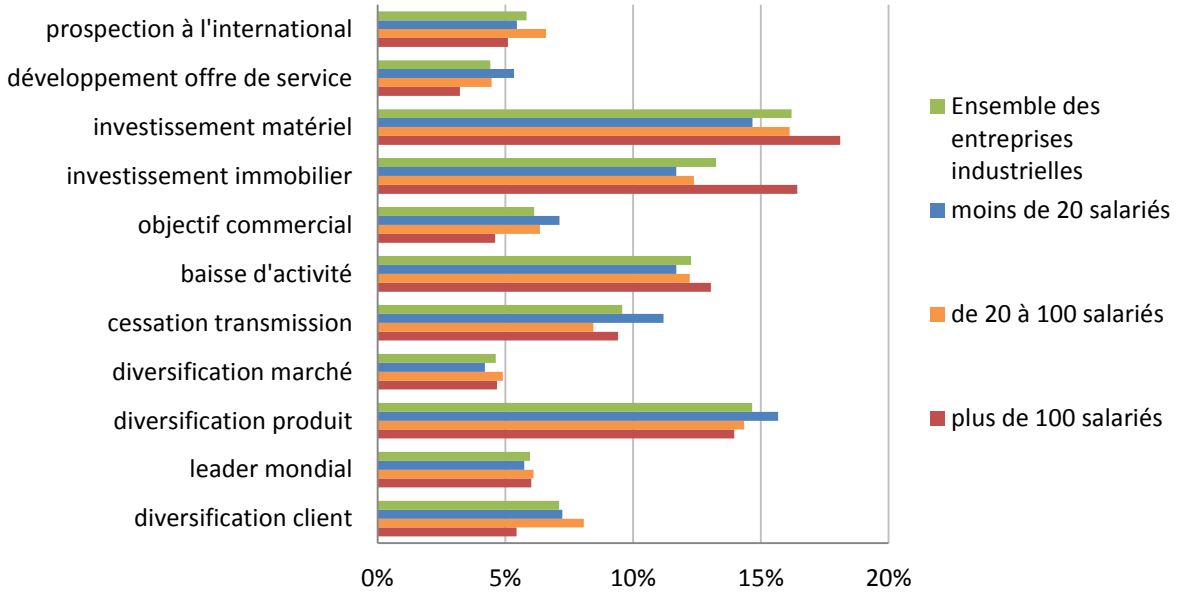


Figure 20 : Répartition des projets classés des entreprises industrielles selon la taille

3.5.2 Pour résumer : Les projets dans l'industrie pendant la crise

3.5.2.1 Le secteur dans son ensemble

Les entreprises industrielles représentent plus de la moitié des entreprises visitées. Il est donc évident que les résultats observés au sein de ce secteur seront très similaires à ceux observés pour l'ensemble des entreprises visitées.

A travers de nombreuses études ou d'enquêtes de conjoncture, nous savons déjà que l'industrie a été un secteur fortement touché par la crise. C'est pourquoi, nous allons ici en plus de retracer cette situation économique, essayer de comprendre comment les dirigeants d'entreprises industrielles ont essayé de faire face à une des plus grandes crises économiques.

Il est important de rappeler que nous étudions ici les projets de développement des entreprises industrielles. Assez souvent les dirigeants décident de modifier leur projet à la suite de difficultés économiques qu'ils constatent. C'est pourquoi, les difficultés économiques observées sur les projets peuvent arriver plus tardivement que celles observées sur des indicateurs économiques plus court-termistes telle que la demande globale dans l'industrie calculée par l'Insee (Figure 21).

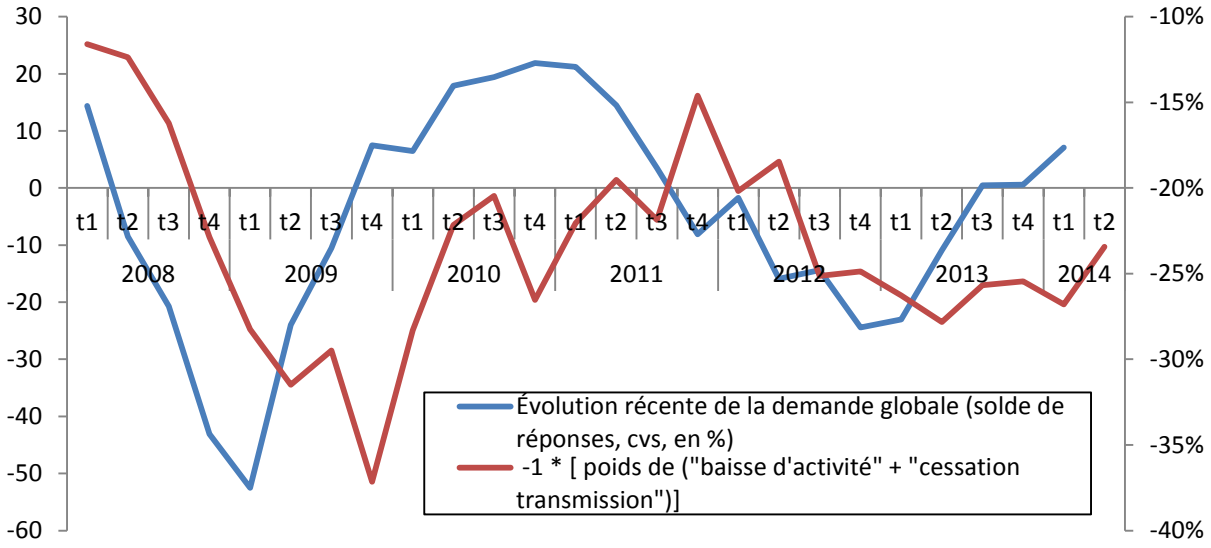


Figure 21 : Comparaison entre l'évolution des difficultés économiques observées sur les projets des entreprises et celle des perspectives de demandes globales

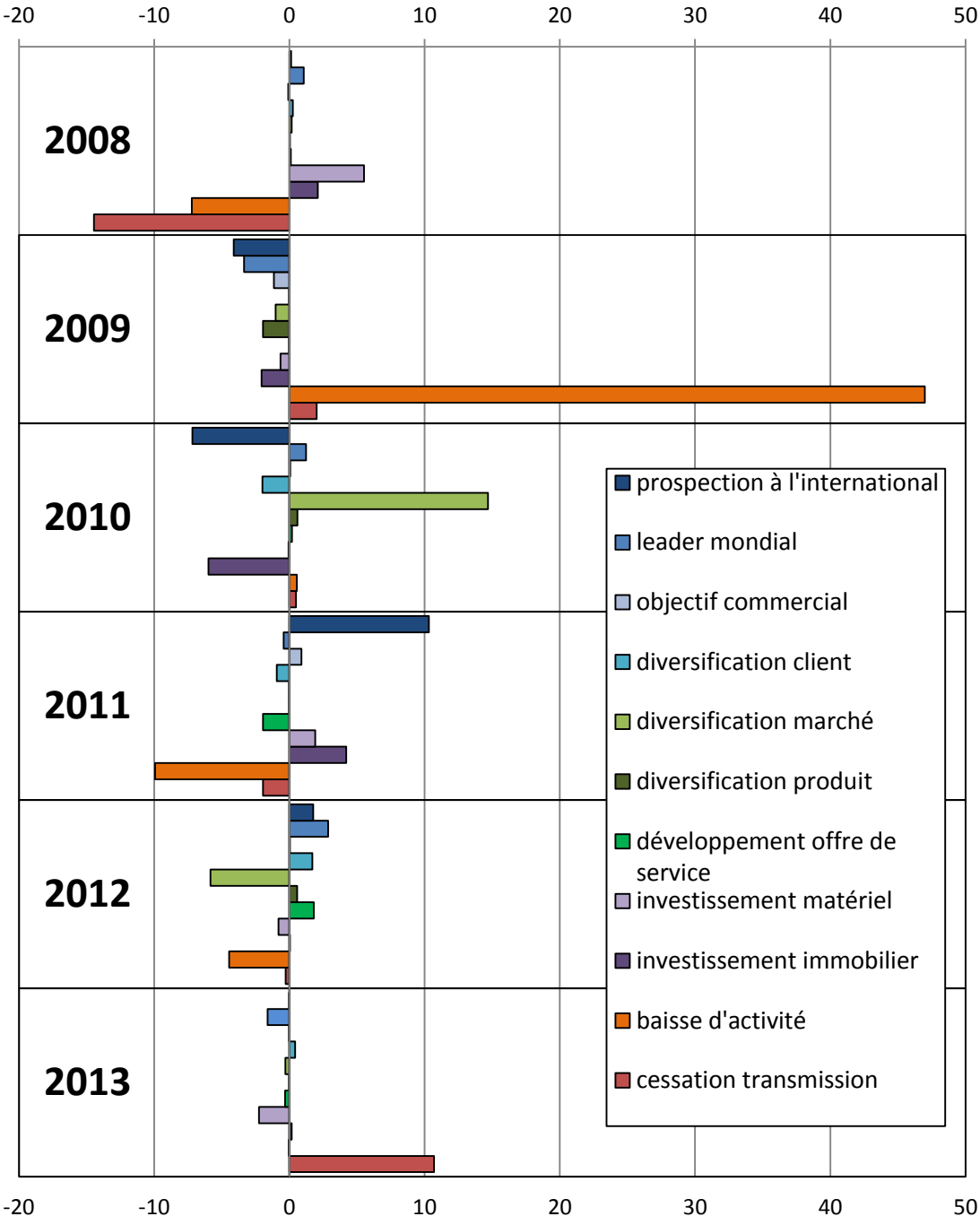


Figure 22 : Evolution annuelle du poids (Khi^2 signé) de chaque classe de la 1ère classification pour les entreprises industrielles

L'année 2008 se caractérise pour l'industrie par des investissements matériels encore très présents (Figure 22 - Khi^2 signé de + 5,5) et des difficultés économiques impactant beaucoup moins les entreprises que sur les cinq années qui vont suivre (Khi^2 signé de – 7,2 pour la classe « baisse d'activité » et de – 14,5 pour la classe « cessation transmission »). Cependant, le dernier trimestre de 2008 laissent déjà apparaitre une situation économique dégradée (Tableau 9 – les deux premiers projets).

En 2009, la crise économique s'est installée dans les projets des entreprises industrielles. Des difficultés économiques apparaissent dans le tiers des projets (classées) de 2009 alors que ce n'était le cas que pour un dixième des projets au début de 2008. Toutes les branches de l'industrie n'ont pas subi la crise avec la même intensité. D'après l'Insee, la branche automobile a été la plus touchée et sa production en volume a reculé de 24 % en 2009 après une baisse de 7 % en 2008.

Afin de faire face à ces difficultés sectorielles, les dirigeants ont donc entamé un travail de diversification vers les secteurs porteurs et moins impactés par la crise (Tableau 9 – quatre derniers projets). C'est d'ailleurs cette diversification de marchés qui caractérise principalement l'année 2010 (Figure 22 / Khi^2 signé de + 14,7).

Tableau 9 : Exemple de projets de développement entre 2008 et 2010 d'entreprise industrielles

Date du projet	Commentaire du projet
4ème trimestre 2008	*L'entreprise connaît une baisse d'activité généralisée depuis octobre liée à la crise. Réduction du carnet de commandes.*
1er trimestre 2009	*L'entreprise connaît depuis octobre 2008 une forte baisse de son activité avec une situation financière difficile.*
1er trimestre 2009	*Baisse de charge aujourd'hui dans l'automobile mais bonnes perspectives sur d'autres secteurs*
1er trimestre 2009	*Baisse d'activité due à la crise de 30%. Rebond envisagé sur secteur aéronautique et énergie*
2ème trimestre 2010	*Les projets de développement de l'entreprise visent à réduire la dépendance de l'entreprise du secteur de l'automobile pour prendre de nouvelles parts de marchés dans l'aéro et l'éolien: Projet en cours avec succès*
4ème trimestre 2010	*L'entreprise cherche à se diversifier pour moins dépendre du secteur de l'automobile, elle s'oriente vers les énergies renouvelables (éolien).*

L'année 2011 est signe de redémarrage pour les entreprises industrielles. Les projets à l'international, qui avaient été reportés à cause des difficultés économiques et des incertitudes mondiales qui pesaient, semblent revenir (Khi^2 signé de + 10,3 en 2011 après - 7,2 en 2012). De même, les investissements reprennent une place plus importante dans les projets des entreprises. Dans la continuité de 2011, l'année 2012 reste légèrement orientée à l'international. Cependant, des signaux négatifs commencent à apparaitre au sein des projets avec notamment un début de baisse des investissements matériels. Ces signaux étaient malheureusement fondés puisque cette baisse des investissements s'est légèrement accentuée en 2013 (Khi^2 signé de - 2,4). Les dernières enquêtes trimestrielles de l'Insee (janvier et avril 2014) sur les investissements dans l'industrie font également état d'une baisse des investissements en 2013 (- 7 %). Plus particulièrement, l'année 2013 apparait comme une année très difficile au vu du poids de la classe « cessation transmission » (Khi^2 signé de + 10,7). Les entreprises industrielles commencent à subir les contrecoups des difficultés rencontrées

en 2008-2009 (Tableau 10 – 1er et 3ème projets), ce qui remet en cause la viabilité de l'entreprise (redressement, liquidation judiciaire). Enfin, l'analyse de la 2ème classification semble également mettre en avant une hausse en 2013 des difficultés de trésorerie (Khi^2 signé de + 19,6).

Tableau 10 : Exemple de projets de développement entre 2013 et 2014 d'entreprise industrielles

Date du projet	Commentaire du projet
1er trimestre 2013	*Entreprise en situation de cessation de paiement. Placée en redressement judiciaire assorti d'une période d'observation de six mois. L'entreprise a deux mois pour présenter un plan de redressement.*
1er trimestre 2013	*baisse de la commande publique et saisonnalité ont mis l'entreprise dans de grosses difficultés financières.*
3ème trimestre 2013	*La contraction des marges pousse le dirigeant à une vision très pessimiste de l'évolution de la société, voir même la liquidation judiciaire*
1er trimestre 2014	*Développement important affiché mais problèmes de trésorerie et difficulté à lever des fonds*
2ème trimestre 2014	*Croissance d'activité entrainant un BFR conséquent et des difficultés de trésorerie.*

Pour l'année 2014, l'étude de la 3ème classification (avec des projets sur les 5 premiers mois de l'année) laisse apparaitre cette fois des signaux positifs. Après deux années de baisse, les investissements matériels se stabilisent (Khi^2 signé de + 0 pour le début 2014 après - 2,4 en 2013). Les projets orientés à l'international semblent également plus présents (+ 1,9 au début 2014 après - 0 en 2013). Les projets relatifs à des cessions, redressements ou liquidations judiciaires retrouvent un niveau normal (+ 0 après + 9,3). Un seul point négatif demeure (obtenu grâce à l'étude d'une 4ème classification non renseignée ici). En effet, les difficultés financières (problèmes de trésorerie, besoin de fond de roulement, etc.) semblent en ce début d'année 2014 toujours impactées les entreprises industrielles. Ces difficultés sont cependant étroitement liées à des perspectives de croissance légèrement plus optimistes (Tableau 10 – deux derniers projets).

3.5.2.2 Selon la taille

Afin de comparer les entreprises industrielles selon la taille, nous allons étudier l'évolution dans le temps des quatre classes « baisse d'activité », « cessation transmission », « prospection à l'international » et « investissement matériel » selon les tranches d'effectifs « moins de 20 salariés » (2800 projets classées), « de 20 à 100 salariés » (3700 projets classées) et « plus de 100 salariés » (2400 projets classées).

Les entreprises industrielles ont subi la crise économique, et ce quelle que soit la taille (Figure 23). Le poids de la classe « baisse d'activité » a connu un pic en 2009, il s'est réduit pendant les deux années qui ont suivi avant de repartir légèrement à la hausse en 2012. Cependant, la situation économique a

recommencé à se dégrader qu'en 2013 pour les entreprises industrielles de moins de 20 salariés (TPE industrielles).

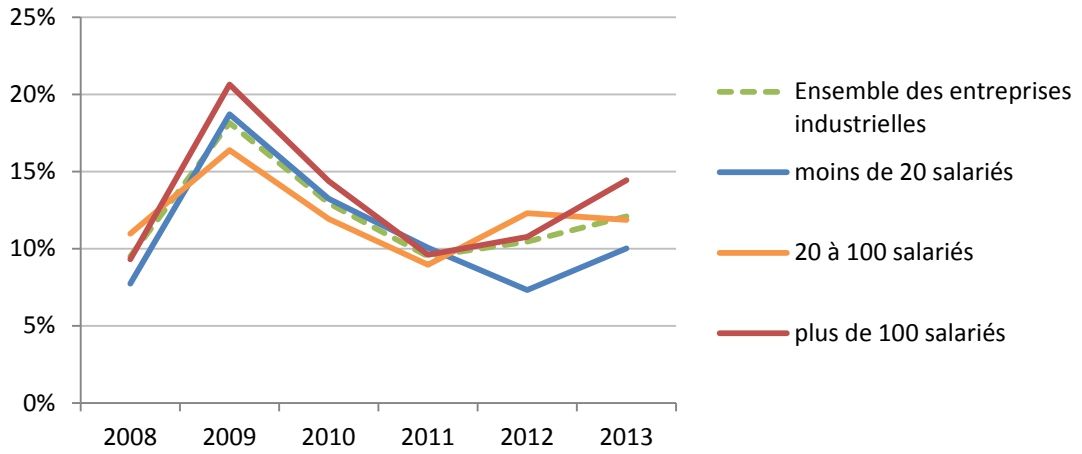


Figure 23 : Evolution annuelle du poids de la classe « baisse d'activité» selon la taille

L'étude de la classe « cessation transmission » montre également un léger retard sur l'évolution des projets des TPE industrielles (Figure 24). En effet, les TPE industrielles ont connu jusqu'à 2010 de plus en plus en difficultés amenant à des cessations, redressements ou vente de l'entreprise, alors que ce pic a été atteint en 2009 pour les autres entreprises industrielles. L'année 2013, bien que difficile en terme de redressements et liquidations judiciaires pour toute les entreprises industrielles, semblent avoir beaucoup plus marqués les plus grandes entreprises, et plus particulièrement celles de plus de 100 salariés dont le poids de la classe « cessation transmission » a presque doublé en un an (13 % des projets classés en 2013 contre 7 % en 2012).

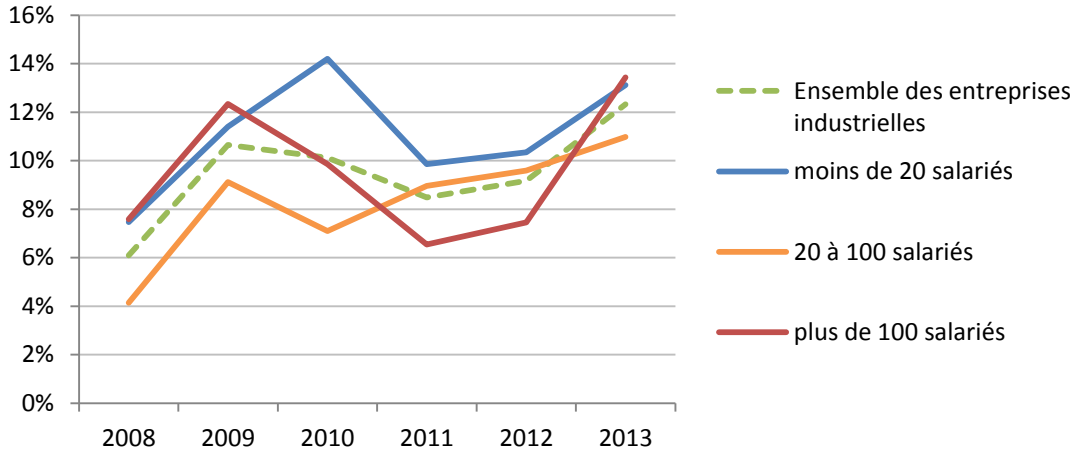


Figure 24 : Evolution annuelle du poids de la classe « cessation transmission» selon la taille

Quelle que soit la taille, les dirigeants d'entreprises industrielles ont dû reporter des projets à l'international à cause de la crise. Cependant, alors que les entreprises industrielles de plus de 20 salariés ont redémarré leurs projets à l'international dès 2010, les TPE industrielles (moins de 20 salariés) semblent avoir beaucoup plus réduit ou reporté leurs projets à l'international. En effet, le poids de ces projets a connu deux années de baisse et ne représentait plus que 2 % des projets

classées en 2010 contre 9 % en 2008. Encore une fois, on observe donc un léger décalage entre les TPE industrielles et les autres entreprises industrielles. Les projets à l'export connaissent cependant quelle que soit la taille deux années de baisse depuis 2011.

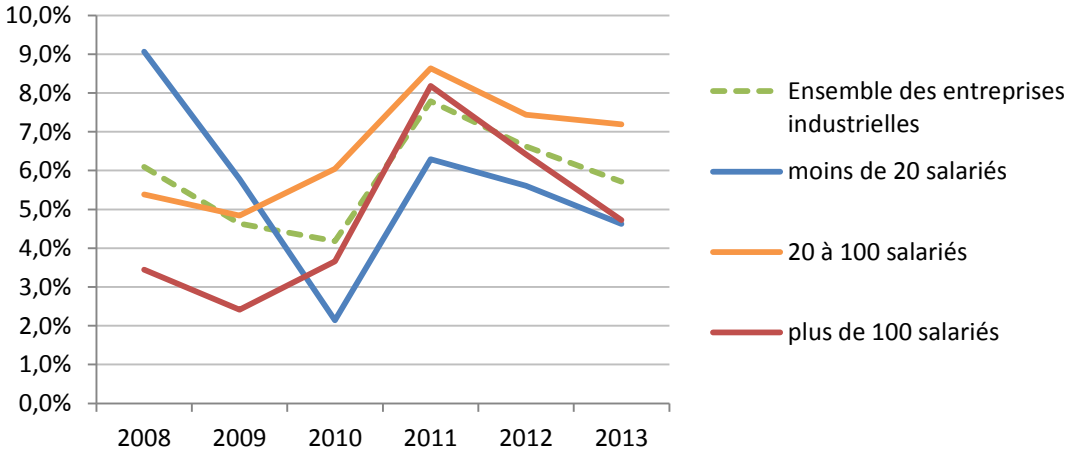


Figure 25 : Evolution annuelle du poids de la classe « Export» selon la taille des entreprises industrielles

L'étude de la classe « investissement matériel » laisse apparaitre un effet très négatif de la crise sur l'investissement des entreprises industrielles. Les très grandes entreprises industrielles semblent cependant avoir bénéficié de leur taille pour limiter cette baisse d'investissement. En effet, alors que le poids de la classe « investissement matériel » avait en 2008 un poids à peu près équivalent quelle que soit la taille des entreprises industrielles, ce poids a diminué en cinq ans d'un tiers pour celles de moins de 100 salariés tout en restant stable pour celles de plus de 100 salariés (Figure 26). Les grandes entreprises industrielles (plus de 100 salariés) qui ont bien évidemment reporté en 2009 des projets d'investissement comme les autres, sont les seules avoir su les relancer. Or, l'investissement est une dépense qui engage l'avenir et est essentiel à la survie de l'entreprise. Un poids des investissements matériels de moins en moins important dans le projet de ces plus petites entreprises industrielles est donc un mauvais signal quant à la confiance en l'avenir de ces dirigeants.

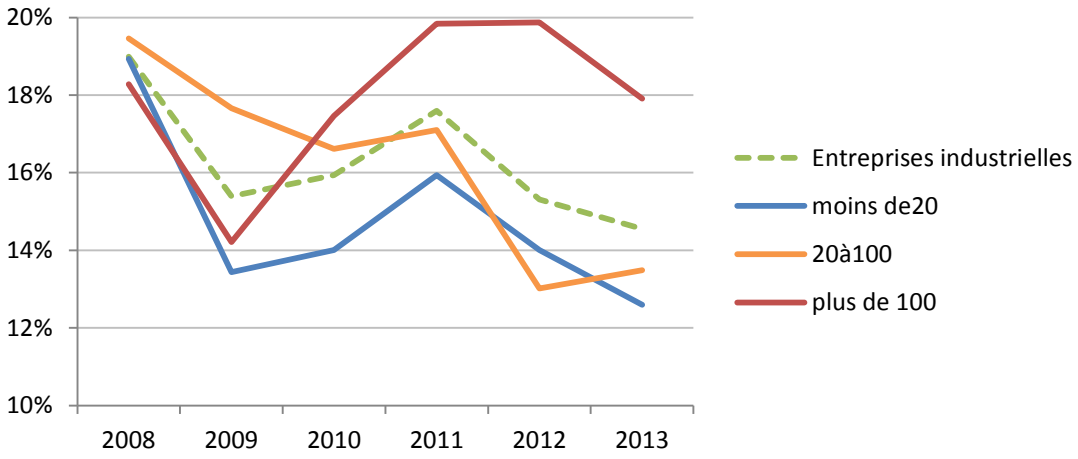


Figure 26 : Evolution annuelle du poids de la classe « investissement matériel» selon la taille

Globalement les dirigeants des entreprises industrielles, quelle que soit leur taille, ont tous modifié leurs projets à la suite de cette crise économique. Ces modifications semblent cependant intervenir légèrement plus tard pour les plus petites entreprises industrielles. Il est possible que la structure de l'industrie caractérisée par de grands donneurs d'ordre d'un côté et une myriade de sous-traitants de l'autres explique en partie ce retard. En effet, les sous-traitants industriels sont principalement dépendants de leurs grands donneurs d'ordre, et particulièrement de toute modification de leurs projets de développement. On aurait donc en premier lieu des grands donneurs d'ordre, qui très dépendants de la situation économique globale, ont reporté ou modifié des projets. Ces reports ont dans les mois qui suivent impacté les sous-traitants industriels qui ont dû modifier leurs projets en conséquence, d'où un léger décalage observé.

4. L'analyse textuelle pour prédire : L'exemple de l'appartenance à la filière automobile

Dans la partie III, nous avons fait parler les données textuelles d'elles-mêmes afin de dégager les thématiques du corpus étudié. Aucune variable extérieure n'est intervenue dans la création de ces thématiques. Dans cette partie, nous allons cette fois nous placer dans le cas d'un modèle d'apprentissage supervisé afin de mettre en avant le pouvoir prédictif que peut contenir une variable textuelle. Pour cela, nous utilisons SAS *Text miner* pour convertir les différentes données textuelles en variable continue. Cette conversion n'est bien évidemment pas effectuée aléatoirement. Elle a pour but de quantifier le pouvoir prédictif des différentes données textuelles, et dépend donc du phénomène que l'on souhaite expliquer.

La DGE mène depuis un certain nombre d'années une politique économique guidée principalement par une approche filière. Or, les bases de données de la statistique publique d'entreprise sont structurées selon une approche « secteur d'activité ». C'est pourquoi, de nombreuses enquêtes ont été réalisées pour essayer d'appréhender la situation économique des principales filières économiques. Pour chacune de ces enquêtes, les statisticiens ont été confrontés à la difficulté de bien cibler les entreprises concernées. En effet, la seule variable disponible permettant d'orienter ce ciblage est l'activité principale de l'entreprise (APE) qui est codée selon la Nomenclature d'Activité Française (NAF). Cette variable est donc codée selon une approche « secteur d'activité ».

Depuis 2008 les chargés de mission de la DGE peuvent renseigner dans ISIS les différents marchés pour lesquels l'entreprise visitée exerce. Nous nous intéressons dans cette partie au fait d'appartenir à la filière automobile. Dans cette étude, les entreprises qui produisent, vendent ou rendent des services pour le marché automobile seront considérées comme appartenant à la filière automobile. Connaissant l'appartenance ou la non-appartenance à la filière automobile pour un nombre assez important d'entreprises, nous allons montrer que le ciblage très limité effectué jusqu'à présent gagnerait en précision à prendre en compte toute les données textuelles récoltées sur les entreprises. Pour cela, nous comparons 5 modèles. Les trois premiers modèles mettent en avant le pouvoir prédictif que peut posséder une variable textuelle. Quant aux deux derniers modèles, ils montrent que l'accumulation de différentes variables textuelles améliore davantage encore ce pouvoir prédictif.

4.1 Création de la base

Les chargés de mission de la DGE ont la possibilité de renseigner la filière de l'entreprise sans que cela ne soit obligatoire. C'est pourquoi afin de ne pas considérer à tort des entreprises comme non-appartenant à la filière automobile, nous nous limiterons aux visites d'entreprises dont au moins une information sur les marchés de l'entreprise a été renseignée, soit un total de 25 232 visites entre 2008 et 2013.

Tout comme dans la partie III, nous continuerons de travailler au niveau de la visite d'entreprise. C'est pourquoi, une même entreprise peut se retrouver plusieurs fois dans la base. Cette surreprésentation de quelques entreprises peut créer un biais dans la création des différents modèles mais encore une fois nous ferons fi de ce biais. En effet, ces 25 000 visites ont été réalisées sur une période de 6 ans. Or, une entreprise peut changer de filière au cours du temps (ce qui est notamment arrivé dans la filière automobile au vu des difficultés économiques rencontrées dans cette filière au cours de la crise) et il est donc intéressant d'avoir une même entreprise plusieurs fois dans le temps. De plus, une analyse complémentaire ne gardant qu'une visite par entreprise montre que les résultats présentés par la suite sont tout aussi probants. Enfin, l'objectif ici n'est pas de créer un modèle parfait pour caractériser une entreprise mais plutôt de trouver un modèle permettant de cibler efficacement les entreprises française selon la problématique souhaitée (ici l'appartenance à la filière automobile). Nous considérons donc que l'analyse porte sur 25 232 entreprises. Parmi ces entreprises, 17 % d'entre-elles appartiennent à la filière automobile. Afin d'analyser les résultats sur des données différentes de celles utilisées pour la création du modèle, nous partitionnons l'échantillon en deux : l'échantillon d'apprentissage (70 % de l'échantillon) et l'échantillon de validation (30 % de l'échantillon). Ce partitionnement est réalisé selon une méthode d'échantillonnage aléatoire simple afin de préserver la même proportion de cible dans chaque échantillon.

4.2 La création des différents modèles

4.2.1 Modèle n°1 : Sans les données textuelles d'ISIS

Ce premier modèle qui n'en est peut-être pas un au vu de la simplicité de celui-ci, est celui qui a souvent été utilisé pour les anciennes enquêtes de la filière automobile. En effet, la seule information pour trouver les entreprises de la filière automobile est la variable « APE » codée selon la nomenclature NAF. Un travail de reconnaissance des codes NAF *a priori* en lien avec l'industrie automobile a déjà été réalisé et a permis de distribuer les différents codes NAF selon 4 classes.

La première classe comportant 4 codes NAF représente le noyau de la filière automobile. La deuxième classe, appelée « partielle » et constituée de 11 codes NAF industrielles, représente les activités où la filière automobile semble avoir un poids très important. La troisième classe réunissant 20 codes NAF représente les activités où il est possible de trouver des entreprises de la filière automobile, d'où le nom de « potentielle ». Enfin, une dernière classe englobe tous les 697 codes NAF restants.

Le Tableau 11 réalisé sur les 7 600 visites de l'échantillon de validation (afin d'être comparable avec les autres modèles) montre clairement que cette classification des codes NAF est tout à fait justifiée. On observe que 85,4 % des entreprises du « noyau » appartiennent à la filière automobile contre 30,4 % pour les entreprises de la classe « potentielle » et 9,5 % pour la classe « autres codes NAF ». Ce qu'il est cependant intéressant de regarder est que seul 26,9 % des entreprises de la filière automobile appartiennent aux classes « noyau » ou « partielle ». Or, ce sont ces deux seules classes

qui ont été utilisées pour cibler les entreprises de la filière automobile pour la dernière enquête de filière. Près des trois quart des entreprises de la filière ont donc été écartées du ciblage. On comprend bien que malgré des activités qui peuvent paraître plus éloignées, ces entreprises écartées représentent une composante importante de la filière automobile qu'il ne faut pas oublier si l'on veut mesurer efficacement la situation économique de la filière.

Tableau 11 : Répartition des entreprises selon l'appartenance à la filière et la classification NAF.

	Répartition en ligne			**Répartition en colonne**	
	n'appartient pas à la filière auto	appartient à la filière auto		n'appartient pas à la filière auto	appartient à la filière auto
Noyau	14,7	85,4		0,4	10,4
Partielle	47,9	52,1		3,1	16,5
Potentielle	69,7	30,4		14,5	30,9
autres codes NAF	90,5	9,5		82,0	42,2

Champ : Entreprises visitées de l'échantillon de validation

4.2.2 Modèle n°2 : Avec la « Conclusion générale » Seulement.

Ce deuxième modèle va également contenir qu'une seule variable. Cette variable va cependant être construite à partir de la variable textuelle « conclusion générale » (cf partie I.2). Nous utilisons SAS *Text Miner* sur l'échantillon d'apprentissage pour établir un lien entre cette variable et le fait d'appartenir à la filière automobile. Pour cela, SAS Text Miner génère automatiquement un ensemble de règles booléennes basées sur les termes employés (dans la « conclusion générale ») et ordonnées selon la probabilité de prédire l'appartenance à la filière automobile. Cet ensemble de règle permet de convertir la « conclusion générale » en une probabilité d'appartenir à la filière. C'est cette probabilité qui est utilisée dans ce 2ème modèle.

4.2.3 Modèle n°3 : Modèle n°2 + Modèle n°1.

Le 3ème modèle a pour objectif de mesurer l'apport de la « conclusion générale » au modèle utilisé jusqu'à présent pour cibler les entreprises de la filière automobile. Pour cela, on rajoute au modèle n°1 la probabilité (issue de la « conclusion générale ») utilisée dans le modèle n°2. Une régression logistique est ensuite réalisée avec ces deux variables explicatives.

4.2.4 Modèle n°4 : Avec toutes les données textuelles d'ISIS

Dans le modèle n°2 nous n'avons utilisé que la variable textuelle « conclusion générale » alors qu'il existe dans ISIS de nombreuses données textuelles (cf. partie I.2). L'objectif de ce modèle est donc de montrer que nous avons intérêt à utiliser la totalité des données textuelles disponibles. Pour cela, nous ajoutons à la variable textuelle « conclusion générale » dix nouvelles variables textuelles : les

projets de développements, les grands donneurs d'ordre, les clients principaux et sept variables textuelles issues des 46 sous-critères énumérés en partie 1.2.

Pour rappel, les chargés de missions peuvent laisser un commentaire pour 46 sous-critères. Ces 46 sous-critères sont classés selon sept critères (Commercial, environnement de l'entreprise, innovation, moyens de production, qualité, ressources humaines et usages TIC). Pour simplifier l'exploitation de ces 46 variables tout en conservant la totalité des données textuelles renseignées, nous créons une variable textuelle pour chacun des sept critères. Chacune des sept variables textuelles est définie par la concaténation des variables textuelles laissées pour les sous-critères appartenant au critère en question. Par exemple, si une visite d'entreprise contient un commentaire pour les sous-critères « management de la qualité », « niveau qualité produit » et « qualité délais » alors la variable « qualité » sera la concaténation de ces trois commentaires.

De la même façon que pour le modèle n°2, nous utilisons SAS Text Miner sur chacune de ces dix nouvelles variables pour les convertir en une probabilité d'appartenir à la filière automobile. Nous nous retrouvons donc avec 11 variables explicatives sur lesquelles nous pouvons réaliser une régression logistique pour prédire l'appartenance à la filière automobile. La question à se poser avant de commencer toute analyse est de regarder la colinéarité entre les variables. En effet, au vu de la création de ces différentes variables, qui représentent une probabilité d'appartenance à la filière, la colinéarité entre elles peut être très forte.

Les différentes probabilités peuvent être considérées comme des variables continues mais sont très éloignées d'une loi gaussienne (test de Kolmogorov-Smirnov). C'est pourquoi nous nous intéresserons principalement au coefficient de corrélation de Spearman. Les deux variables les plus liées sont d'une part la probabilité créée à partir du commentaire laissé sur les clients principaux de l'entreprise, d'autre part celle créée à partir du commentaire relatif aux grands donneurs d'ordre. Le coefficient de colinéarité entre ces deux variables reste cependant assez faible (0,36). L'étude de la multicolinéarité (à travers l'étude de la « tolérance ») entre ces 11 variables amène au même résultat. Ainsi, la régression logistique effectuée sur ces 11 variables ne devrait pas être impactée par des problèmes de colinéarité.

La sélection du meilleur modèle a consisté à chercher parmi les 11 variables celles qui « expliquent le mieux » le fait d'appartenir à la filière automobile. Le meilleur sous-ensemble a été sélectionné grâce à une méthode descendante (« backward » sous SAS) qui a enlevé la variable « usages TIC ». Cette variable apparaît comme peu explicative du fait d'appartenir à la filière automobile. Au final, le modèle n°4 proviendra de la régression logistique réalisée sur l'ensemble des 10 autres variables seulement.

4.2.5 Modèle n°5 : modèle n°4 + modèle n°1

Le dernier modèle a pour objectif de créer le meilleur modèle susceptible de prédire le fait d'appartenir à la filière automobile. Nous nous permettons donc d'ajouter aux 10 variables utilisées dans le modèle n°4 l'information issue de l'APE de l'entreprise. Le dernier modèle est le résultat de la régression logistique réalisée sur ces 11 variables.

4.3 La comparaison des modèles

Pour comparer ces différents modèles, nous les appliquons sur l'échantillon de validation. Nous disposons de plusieurs moyens pour mesurer le pouvoir discriminant des différents modèles. Nous nous contenterons ici de la courbe ROC (Figure 27) et de la courbe LIFT (Figure 28).

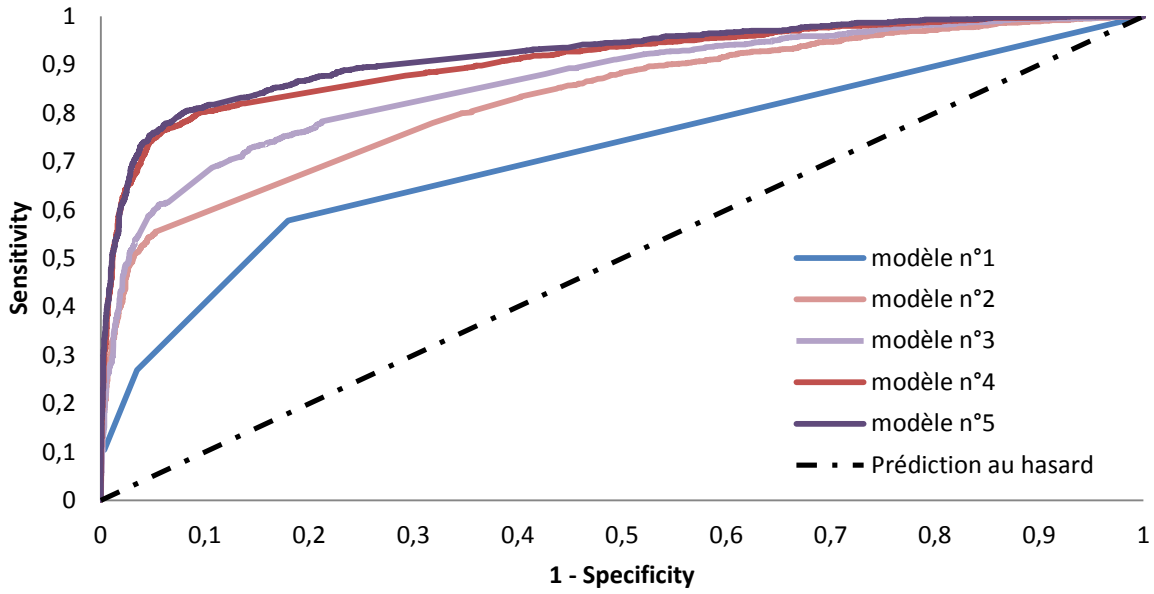


Figure 27 : Courbe ROC des 5 modèles

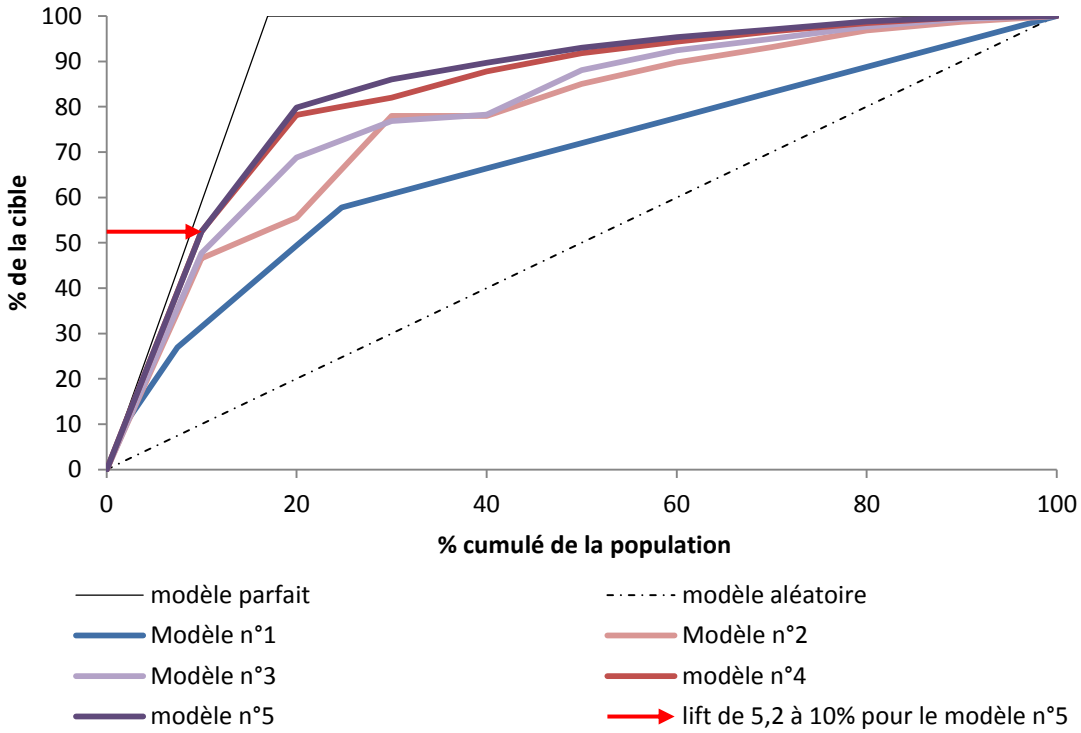


Figure 28 : Courbe lift des 5 modèles

Les 5 modèles créées donnent à chacune des entreprises un score explicatif du fait d'appartenir à la filière automobile. La courbe ROC représente, pour un score supérieur à un seuil s, en ordonné la

probabilité de bien détecter une entreprise appartenant à la filière automobile, et en abscisse la proportion d'entreprise n'appartenant pas à la filière mais détecter comme y appartenant. Un modèle est donc d'autant plus performant que son air sous la courbe est élevée (max=1). Le Tableau 12 montre clairement que les données textuelles à elles seules possèdent un pouvoir de prédiction très fort. En effet, l'aire sous la courbe du modèle n°2 est bien plus élevée que celle du modèle n°1. On remarque également que l'utilisation de l'ensemble des données textuelles permet la création d'un modèle plus efficace encore (Aire de 0,908 pour le modèle n°4 contre 0,827 pour le modèle n°2). Enfin, ces données textuelles apportent une information complémentaire à celles déjà disponibles. On observe ainsi une amélioration du modèle en cumulant les données textuelles et la variable APE (Aire de 0,919 pour le modèle n°5 contre 0,908 pour le modèle n°4). Globalement le modèle n°5 affiche des résultats tout à fait intéressants. En enquêtant 10 % des entreprises ayant le score le plus élevée, on cible avec ce modèle plus de la moitié (52,5 %) des entreprises de la filière (contre 31,5 % pour le modèle n°1). De ce fait, la proportion d'entreprise de cet échantillon appartenant à la filière automobile atteint 89 % (contre 52 % pour le modèle n°1).

Tableau 12 : Comparaison de la performance des 5 modèles

	Modèle				
	n°1	**n°2**	**n°3**	**n°4**	**n°5**
Aire sous la courbe ROC	0,715	0,827	0,863	0,908	0,919
lift à 10 %	3,15	4,65	4,77	5,25	5,25

Lecture : le lift à 10 % du modèle n°5 signale que pour ce modèle les 10 % d'entreprises ayant le score le plus élevé concentre 52,5 % des entreprises appartenant à la filière automobile.

4.4 L'exploitation des modèles

Ces modèles vont permettre de trouver davantage d'entreprises susceptibles d'appartenir à la filière automobile. D'une part en appliquant le modèle n°3 sur près de 18 000 visites d'entreprises réalisées avant 2008 (dont la conclusion générale est non vide). D'autre part, en appliquant le modèle n°5 sur près de 13 000 visites d'entreprises où aucune information sur la filière n'est renseignée. Si la proportion d'entreprises appartenant à la filière automobile est équivalente à celle observée sur les 25 000 utilisées jusqu'à présent, on peut s'attendre à trouver près de 3 500 visites d'entreprises portant sur une entreprise de la filière l'automobile. Ce score pourra donc servir à la DGE pour améliorer le ciblage de futures enquêtes filières, mais il pourra également servir directement aux DIRECCTE qui ont pour objectif d'accompagner les filières régionales. Les chargés de mission en région ont en effet besoin d'avoir une vision la plus élargie possible sur les entreprises de la filière ou susceptible d'y appartenir.

Conclusion

Dans cette étude, nous nous sommes évertués à présenter les intérêts de l'analyse textuelle pour la statistique d'entreprise. Ce travail a été axé sur l'exploitation d'une base de données textuelles issues de visites d'entreprises. Cette base, par sa taille, son ancienneté et son enrichissement quotidien constitue l'exemple parfait des enjeux que représente l'utilisation des données textuelles. Depuis plus de 10 ans des centaines de chargés de missions renseignent des informations très détaillées sur les entreprises sans que celles-ci ne soient utilisées à un niveau agrégé. Or, les résultats que nous avons obtenus montrent que ces données textuelles sont d'une richesse impressionnante.

En laissant parler les données textuelles d'elles-mêmes, nous avons été capables de caractériser et de quantifier les évolutions des projets des entreprises pendant la crise. Plus particulièrement, nous avons pu expliquer et comprendre ces évolutions. Ce que des variables quantitatives ou qualitatives n'auraient sans doute par permis. L'analyse textuelle a donc permis d'allier la connaissance qualitative de l'entreprise à une connaissance globale des entreprises. De plus, en essayant de tirer parti de l'enrichissement quotidien de cette base de données textuelles, nous pouvons aisément envisager de créer de nouveaux indicateurs conjoncturels susceptibles d'éclairer le débat économique.

Nous avons également montré dans la dernière partie que les données textuelles ont un pouvoir de caractérisation de l'entreprise plus fort et surtout plus ouvert que celui de certaines variables quantitatives ou qualitatives. En effet, nous pouvons réitérer l'analyse pour prédire l'appartenance à d'autres filières afin d'aider au ciblage de futures enquêtes de filières (santé, aéronautique, logiciels et services informatiques, etc.). Nous pouvons également l'utiliser pour améliorer la prédiction d'autres phénomènes : type d'innovation, utilisation de la R&D, situation économique de l'entreprise, etc. Il suffit pour cela de disposer d'une base d'apprentissage suffisamment grande alliant les données textuelles et le phénomène à expliquer. Cette base peut aisément être construite en récupérant les informations issues des très nombreuses enquêtes « entreprises » déjà réalisées.

Le principal avantage des données textuelles est de permettre la collecte d'une information que l'on peut quasiment qualifier de « brute ». L'information collectée n'est ainsi pas convertie en un ensemble de variables fermées. Il n'y a donc pas de perte d'informations inhérente à toute conversion. La richesse des données textuelles risque de modifier complètement la façon de faire de la statistique d'entreprise. Actuellement, on part d'hypothèses sur le fonctionnement des entreprises qu'on tente de valider et de quantifier en procédant à des enquêtes. Cependant, c'est souvent après avoir analysé les résultats de l'enquête que de nouvelles questions émergent. Ces nouvelles questions nécessitent alors la réalisation de nouvelles enquêtes. Or, les données textuelles, parce qu'elles collectent une information quasi exhaustive sur l'entreprise, peuvent permettre de répondre en partie à ces nouvelles questions. Le recours à davantage de questions ouvertes au sein des enquêtes « entreprises » voire à davantage d'entretiens va donc se poser.

Cependant, deux difficultés majeures entourent toujours l'utilisation des données textuelles. La première difficulté provient évidemment de la collecte de ces données. L'utilisation de questionnaires numériques étant encore rarement exploitée au sein des enquêtes de la statistique d'entreprise, la transcription des questions ouvertes a encore un coût très élevé. De même, dans le cadre des visites d'entreprises, la saisie de l'intégralité des informations obtenues du dirigeant occupe un temps non-négligeable des chargés de missions. La deuxième difficulté découle de l'exploitation que l'on peut faire de ces données. En effet, l'analyse lexicale, bien que présentant des résultats intéressants à un niveau agrégé, présente quelques faiblesses à niveau plus fin. Ces faiblesses proviennent d'erreurs dans l'analyse du texte : non prise en compte de la négation, fautes d'orthographe, style atypique de l'auteur, etc. C'est pourquoi, il faudra veiller à réduire ce bruit et à vérifier qu'il n'impacte pas excessivement les résultats.

Plus généralement, on comprend bien que le « Big Data » qui englobe toutes les données publiques disponibles sur le net et toutes les données (structurées ou non) collectées par des entreprises privées voire publiques, représente une mine d'or d'informations pour comprendre les entreprises. Certaines en font même leur business. L'entreprise Data Publica a ainsi lancé il y a plus de trois ans le projet « c-radar ». Ce projet permet d'exploiter la quantité faramineuse d'informations sur les entreprises disponible sur le net afin de créer une nouvelle base de données sur les entreprises françaises. Cette base de données caractérisée, d'une part, par la richesse dans les informations collectées, et d'autre part, par une mise à jour en temps réel des informations, intéressent les institutions publiques et particulièrement la DGE. Le projet « c-radar » a d'ailleurs été récompensé par l'État dans le cadre du concours innovation 2030 (catégorie Big Data).

« Le Big Data n'est pas en passe de concurrencer toutes les sources de la statistique publique, qui sont variées» a déclaré Jean Luc TAVERNIER (directeur générale de l'INSEE) lors de la conférence débat « *statistique et démocratie : à quoi servent les chiffres ?* » co-organisée par le Conseil national de l'information statistique (Cnis) et le Conseil économique, social et environnemental (CESE). Il est évident que les bases de données administratives ou les bases de données fiscales continueront d'être l'essence même de la statistique d'entreprise. Cependant, il reste indéniable que si le « Big Data » permet de réaliser des études en se passant d'enquêtes souvent très lourdes, il doit être également vu comme une source supplémentaire d'informations. Il faudra juste, comme pour toute source de données, s'assurer d'une qualité minimum. C'est d'ailleurs, cette labellisation de la qualité qui va fortement occuper dans les années à venir les statisticiens publiques ou privés. En effet, éclaircir le flou inhérent au « Big Data » est la dernière étape pour atteindre et profiter pleinement de toutes ces nouvelles mines d'or.

Bibliographie

Benzecri, J.-P. (1973). *L'analyse des Données* (Vol. 1 et 2). Paris: DUNOD.

Bestgen, Y. (2012). *Analyse des différences lexicales entre des corpus : test ou distance du Khi-2.* JADT.

Couvreur, A., & Lehuede, F. (2002). *Essai de comparaison de méthodes quantitatives et qualitatives à partir d'un exemple : le passage à l'euro vécu par les consommateurs.* CREDOC.

della Ratta, F., & Morrone, A. (2000). *Du texte aux variables : les contributions de l'analyse textuelle des questions ouvertes à l'analyse traditionnelle des données.* JADT.

Fallery, B., & Rodhain, F. (2007). Quatre approches pour l'analyse de données textuelles : lexicale, linguistique, cognitive, thématique. Dans *les Actes de la XVIème Conférence internationale de management stratégique* (pp. 1-27). Montréal.

Garnier, B., & Guérin-Pace, F. (2010). *Appliquer les méthodes de la statistique textuelle.* Paris: CEPED (Les clefs pou).

Guérin-Pace, F. (1997). La statistique textuelle. Un outil exploratoire en sciences. Dans *Population,52e année, n°4* (pp. 865-887).

INSEE. (1989). *Observer et représenter un monde de plus en plus complexe : un défi pour la statistique d'entreprise.* Paris: INSEE méthodes n°54.

Lahlou, S. (1994). L'analyse lexicale. Dans *Variance* (pp. 13-24).

Lebart, L., & Salem, A. (1994). *Statistique textuelle.* Paris: Dunod.

Ratinaud, P. (2014). *Visualisation chronologique des analyses ALCESTE : application à Twitter avec l'exemple du hashtag #mariagepourtous.* Paris: JADT.

Ratinaud, P., & Marchand, P. (2012). *Application de la méthode ALCESTE aux « gros » corpus et stabilité des « mondes lexicaux » : analyse du « CableGate » avec IRAMUTEQ.* JADT.

Reinert. (1983). Une méthode de classification descendante hiérarchique. Dans *Cahiers de l'Analyse des Données* (pp. 187-198).

Reinert, M. (1990). Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval. Dans *Bulletin de méthodologie sociologique* (pp. 24-54).

Reinert, M. (1993). Les "mondes lexicaux" et leur 'logique" à travers l'analyse statistique d'un corpus de récits de cauchemars. Dans *Langage et société* (pp. 5-39).

Reinert, M. (1999). Quelques interrogations à propos de l'"objet" d'une analyse de discours de type statistique et de la réponse "Alceste". Dans *Langage et société* (pp. 57-70).

Reinert, M. (2001). Alceste, une méthode statistique et sémiotique d'analyse de discours : Application aux "Rêveries du promeneur solitaire". Dans *La Revue Française de Psychiatrie et de Psychologie Médicale* (pp. 32-36).

Reinert, M. (2008). *Mondes lexicaux stabilisés et analyse statistique de discours.* JADT.

Liste des abréviations

AFC : Analyse Factorielle des correspondances

APE : Activité principale de l'entreprise

CHD : Classification Descendante Hiérarchique

DGE : Direction Générale des Entreprises

DIRECCTE : Direction Régionale des Entreprises, de la Concurrence, de la Consommation, du Travail et de l'Emploi.

ETI : Entreprise de Taille Intermédiaire

GE : Grande entreprise

NAF : Nomenclature d'Activité Française

PME : Petite et Moyenne Entreprise

ROC : « Receiver Operatig Characteristic »

ST : Segment de Texte

TPE : Très Petite Entreprise

Glossaire

Analyse factorielle : famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à extraire des "facteurs" résumant approximativement par quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ.

Analyse factorielle des correspondances : méthode d'analyse factorielle s'appliquant à l'étude de tableaux à double entrée composés de nombres positifs. Elle est caractérisée par l'emploi d'une distance (ou métrique) particulière dite distance du Khi^2.

Caractère délimiteur : Distinction opérée sur l'ensemble des caractères qui entrent dans la composition du texte, permettant aux procédures informatisées de segmenter le texte en occurrences.

Classification : technique statistique permettant de regrouper des observations ou des individus entre lesquels a été définie une distance.

Classification hiérarchique : technique particulière de classification produisant par agglomération progressive des classes ayant la propriété d'être, pour deux quelconques d'entre-elles, soit disjointes, soit incluses.

Corpus : ensemble de textes réunis à des fins de comparaison; servant de base à une étude quantitative.

Cooccurrence : présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus etc.) des occurrences de deux formes données.

Distance du Khi^2 : Il s'agit d'une distance entre des valeurs observées et des valeurs théoriques. Cette différence entre observée et théorique est élevée au carré puis rapportée à cette même valeur théorique.

Segment de Texte : représente la plus petite unité sémantique du texte

Forme : archétype correspondant aux occurrences identiques dans un corpus de textes, c'est-à-dire aux occurrences composées strictement des mêmes caractères non-délimiteurs d'occurrence.

Forme canonique : Voir lemme.

Forme « outils » : Désigne les formes canoniques considérées comme moins porteuses de sens (pronoms, prépositions, conjonctions, auxiliaires, nombres).

Forme « pleine » : Désigne les formes canoniques qui ne sont pas des formes « outils ».

Lemmatisation : regroupement sous une forme canonique (en général à partir d'un dictionnaire) des occurrences du texte. En français, ce regroupement se pratique en général de la manière suivante :

- les formes verbales à l'infinitif,
- les substantifs au singulier,
- les adjectifs au masculin singulier,
- les formes élidées à la forme sans élision.

Lemme : Désigne la forme de référence d'un mot, c'est-à-dire la forme du mot sans les marques (dites marques de flexion) qui l'actualisent dans le discours. En d'autres termes, il s'agit de la forme infinitive pour un verbe, masculin singulier pour un adjectif, singulier pour un nom, etc. Généralement, les entrées des dictionnaires (que ce soit les dictionnaires sur papiers ou les dictionnaires électroniques) sont des lemmes. On parle aussi de forme canonique.

Lexicométrie : Ensemble de méthodes permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire d'un corpus de textes.

Lexique : Ensemble des éléments ou signes constitutifs d'un texte donné. Le lexique comprend donc les mots ou groupes de mots, signes typographiques, chiffres et caractères spéciaux reconnus dans le texte par le système d'analyse de texte. C'est donc l'ensemble des formes, présentés généralement sous forme de liste, avec leur fréquence d'apparition dans le texte.

Occurrence : suite de caractères non-délimiteurs bornée à ses extrémités par deux caractères délimiteurs de forme.

Segments répétés : suites d'occurrences non séparées par un caractère délimiteur

Tableau de contingence : tableau dont les lignes et les colonnes représentent respectivement les modalités de deux questions (ou deux variables nominales), et dont le terme général représente le nombre d'individus correspondant à chaque couple de modalités.

Tableau lexical entier (TLE) : tableau disjonctif complet dont les lignes sont constituées par les segments de texte et les colonnes par les différentes formes issues du corpus. Le terme générique k(i,j) du TLE est égal au nombre de fois que la forme j est présente dans la partie i du corpus. Le tableau d'absence-présence où k(i,j) est égale à 1 si la forme j est présente dans la partie i du corpus et 0 sinon est également un TLE.

Texte : collection de mots définie selon le découpage naturel du corpus

ANNEXE : Description des classes de la 2ème classification

Thème de chaque classe		Traits lexicaux typiques par classe
1	**Projet informatique**	logiciel / *ac_info_comm / solution / *aceff_info_commms20sal / mobile / numérique / web / gestion / système / donnée / informatique / éditeur / plateforme / plate_forme / open / *proj_diffprod / détection / vidéo / source / virtuel / application / robotique / service / information / impression / jeu / *reg_11 / business / cancer / serveur / multi / analyse / tic / hébergement / réalité
2	**vision court moyen terme**	terme / moyen / courir / petit / humain / apporter / enjeu / série / rester / acquérir / stratégique / métier / définir / artisanal / stratégie / dimension / quantité / qualité / alpha / imposer / adaptation / évolution / environnemental / principal / connaissance / créneau / taille / dynamique / objectif / nombreux / clair / opportuniste / centrer / coeur / rapide
3	**Devenir leader mondial**	leader / avance / atteindre / devenir / mondial / eti / ambition / conserver / critique / stratégique / compétitivité / taille / concurrentiel / position / bénéficier / porter / longueur / copier / autonomie / avantage / niveau / adopter / innover / puissance / traiteur / hauteur / garder / conteneur / crèpes / monsanto / pasta / sandwitch / sud_ouest / transducteur / aborder
4	**usinage**	ligne / nouvelle / peinture / usinage / cabine / automatiser / conditionnement / unité / grenaillage / pièce / assemblage / production / fonderie / cuve / une / soudage / emballage / 1ère / mettre / m3 / troisième / investissement / four / acide / internalisation / connectique / polissage / décapage / verrerie / achat / acquisition / forger / chrome / centre / conformité
5	**investissement matériel**	investissement / machine / matériel / achat / immobilier / parc / investir / découper / cov / performant / automatique / k / acquisition / projet / tour / euro / million / programme / laser / d / ctp / oenotourisme / obsolète / inauguration / plaque / *ac_industrie / couleur / béton / 12m / laminoir / productique / lourd / presse / pliage / fnrt
6	**investissement immobilier**	extension / bâtiment / déménagement / construction / local / site / déménager / agrandissement / m2 / construire / agrandir / prévoir / terrain / atelier / stockage / installer / projet / neuf / recrutement / nouveau / *proj_reorga_interne / commun / nouvel / foncier / supplémentaire / usine / un / moderne / étroit / surface / côté / ingénieur / vaste / démanagement / inadapté
7	**innovation**	produit / développement / innovation / innovants / propre / durable / r / poursuite / *proj_diffprod / axe / reposer / poursuivre / stratégie / matière / bio / axer / baser / formulation / développé / spécifique / nouveau / polymères / semence / developpement / huile / technique / lancer / orienter / procédé / maîtrise / respectueux / miser / lutte / politique / process
8	**projet de recherche**	recherche / éco / permanent / laboratoire / collaboration / scientifique / association / impliquer / éclairage / collaboratifs / fabriquer / essentiellement / designer / solaire / signature / concevoir / produit / issu / prochainement / concours / identique / biscuiterie / fourniture / transformation / travailler / algue / génie / visiter / mousse / agro_alimentaire / tropical / inscrire / photovoltaïque / fiabilité / déployer
9	**Implantation pays émergents**	filiale / implantation / arabie / saoudite / *proj_devexport / chine / brésil / tchèque / tunisie / unir / dubaï / algérie / maroc / emirats / koweit / bulgarie / jv / république / etats_unis / inde / pologne / jordanie / succursale / arabe / venture / mexique / création / ouverture / canada / joint_venture / représentation / roumanie / royaume / hongrie / argentin
10	**Exportation pays développés**	*proj_devexport / allemagne / pays / amérique / japon / chine / russie / italie / asie / moyen_orient / suisse / europe / afrique / brésil / user / sud / espagne / nord / etats_unis / belgique / maghreb / prospecter / canada / inde / export / royaume / corée / prospection / latin / angleterre / unir / portugal / cible / exporter / autriche

Thème de chaque classe		Traits lexicaux typiques par classe
11	**classe 11**	offrir / public / appel / privé / partenaire / disposer / proposer / conseil / positionner / complet / acteur / idée / global / mener / obtenir / dévelopement / aider / biophotoniques / soutien / institutionnel / molécule / école / opérateur / accompagner / agrément / serious / informer / télécom / labo / régional / capable / différent / chantier / notoriété / pharmaceutique
12	**projet commercial**	commercial / développer / souhaiter / force / renforcer / réseau / recruter / part / structurer / société / démarche / conseil / france / fonction / continuer / son / vente / également / particulier / démarcher / priorité / concentrer / assoir / technique / renforcement / assurer / *ac_info_comm / prescripteurs / représenter / partie / *aceff_info_commms20sal / entendre / offrir / équipe / effort
13	**activité difficile**	année / chiffre_affaires / difficile / résultat / trimestre / commande / exercice / stable / enregistrer / 2009 / dernier / *proj_baisseact / prévisionnel / négatif / crainte / rentabilité / effectif / mensuel / équilibre / semestre / déficitaire / espérer / mauvais / net / carnet / 2010 / chiffre / mort / prévision / tassement / seuil / baisser / affaire / correct / redémarrage
14	**crise économique**	baisser / crise / *proj_baisseact / subir / situation / activité / charger / économique / retrouver / conjoncture / impactée / face / perte / carnet / commande / difficile / souffrir / *reg_43 / contexte / fouet / chute / 2008 / toucher / ralentissement / compenser / commander / la / entraîner / 2009 / fortement / conséquence / depuis / raison / tenir / stabiliser
15	**classe 15**	électrique / véhicule / oseo / banc / brevet / essai / moteur / avion / hydrogène / chirurgie / léger / radar / ulm / faisabilité / hybride / voiture / motorisation / scooter / biplace / variateurs / validation / dga / hélice / soutenir / instrument / pile / élaboration / conception / recharger / *proj_diffprod / kit / futur / antenne / fibre / dépôt
16	**demande Coface**	coface / assurance / prospection / demander / ap / avis / *proj_devexport / garantie / favorable / dossier / déposer / guinée / propection / renouvellement / cibler / *obj_aide / caution / sénégal / *reg_11 / solliciter / auprès_de / emirats / etats_unis / maroc / arabe / durablement / enveloppe / mali / pays / allemagne / unir / canada / ubifrance / brésil / accompagnement
17	**Transmission naturelle**	*proj_transmission / transmission / fils / retraite / céder / transmettre / dirigeant / lever / fille / repreneur / succession / fonds / départ / capital / père / reprendre / agé / actionnaire / familial / pdg / an / cession / existence / recapitalisation / 62 / gérant / michel / investisseur / entreprise / régler / âgé / âge / succéder / viellard / voici
18	**Cessation - redressement liquidation**	judiciaire / redressement / liquidation / reprise / pse / plan / rj / suite / fermeture / observation / *proj_baisseact / cession / restructuration / juillet / cesaar / salarié / continuation / procédure / bilan / tribunal / sauvegarde / janvier / social / mars / octobre / *obj_demande / tc / convention / crp / pat / arrêt / cessation / mai / dette / fin
19	**innovation industrialisation**	produire / innovant / commercialisation / industrialisation / prototype / conducteur / lancement / critère / phare / semi / breveté / encre / nano / pré / lancer / piscine / oséo / propre / phase / *proj_diffprod / transport / anti / eau / animal / cstb / mouton / sandwich / injection / finaliser / un / série / finalisation / petit / air / filtre
20	**Regroupement site**	site / regroupement / saint / transfert / usine / regrouper / rapatriement / seul / relocaliser / situer / denis / distant / st / fermer / rapatrier / km / caler / rassemblement / exclusivité / ajout / *proj_reorga_interne / idf / dompierre / loubès / condé / réaménagement / actuellement / groupe / center / juridique / commun / unité / cylindre / sur / méthanisation
21	**classe 21**	place / mettre / cohérent / potentiel / réel / prospects / conquérir / drce / espace / accompagnement / partager / structuration / biologique / prestataire / vendre / simple / prospective / commercial / oeuvre / valider / viser / souche / accompagner / en / profit / plv / accélerer / learning / service / final / assistance / bombardier / canadien / effectuer / conséquent

Thème de chaque classe		Traits lexicaux typiques par classe
22	**difficultés de trésorerie**	difficulté / trésorerie / financier / problème / confronter / bfr / cause / tendu / apprentissage / fragile / *proj_baisseact / rencontre / freiner / manquer / engendrer / frein / pb / réglementation / pbs / refuser / déloyal / connait / mais / désengagement / lier / faute / structurelles / gros / grave / paiement / recrutement / cash / qualifié / trouver / accorder
23	**Augmentation des capacités de productions**	production / capacité / productivité / outil / modernisation / réorganisation / augmentation / flux / doubler / augmenter / réorganiser / *proj_reorga_interne / automatisation / gain / améliorer / interne / atelier / compétitif / amélioration / investissement / robotisation / rationnalisation / gagner / autorisation / stockage / compétivité / oxyde / travail / *ac_industrie / sous_traiter / changement / compétitivité / erp / moderniser / coût
24	**diversification produit : haut de gamme**	gamme / haut / valeur / ajouter / produit / marque / *proj_diffprod / luxe / design / accent / décoration / article / différencier / intérieur / restauration / boutique / mdd / réorientation / accessoire / spécial / historique / monter / fruit / bio / benne / plume / voilier / gms / contemporain / distribution / magasin / élaboré / légume / sport / élargir
25	**classe 25**	racheter / créer / bureau / territoire / français / ouvrir / agence / venir / allemand / national / américain / région / présenter / étude / société / normandie / chinois / pari / us / consortium / etp / partenariat / constituer / direct / gérer / étendre / saturer / japonais / grouper / mailler / cluster / touristique / parisien / spécialiser / concession
26	**Diversification marchés : filière**	diversification / industrie / aéronautique / branche / secteur / vers / chimie / implant / ferroviaire / santé / médical / hors / agro / *proj_diversifmarches / automobile / usage / piste / aéro / négliger / *proj_diffprod / textile / réussite / luminaire / défense / nautisme / cosmétique / manutention / pharmacie / armement / micro / composite / tertiaire / pierre / dvpt / civil
27	**recyclage / environnement**	bois / panneau / valorisation / photovoltaïques / déchet / approvisionnement / recyclage / isolation / toiture / métal / ossature / cintrage / pv / biomasse / traitement / tri / broyage / guadeloupe / sédiment / poutre / rabotage / massif / installation / polystyrène / maison / fabrication / reconversion / sciage / parachèvement / tube / chaufferie / triage / palette / terre / construction
28	**diversifier clientèle car dépendance de grands donneurs d'ordre**	diversifier / donneur / secteur / dépendance / ordre / nucléaire / rang / eurocopter / naval / portefeuille / civil / maintenance / aéronautique / auprès_des / se / aeronautique / iter / ferroviaire / chercher / spatial / dépendant / client / réussir / automobile / airbus / *proj_diversifmarches / réparation / équipementiers / areva / réduire / cea / intégrateur / clientèle / défense / diminuer
29	**croissance externe**	croissance / externe / interne / *proj_croissexter / organique / rachat / consolidation / par / mature / ext / safran / fmea / relais / bourse / duplication / cheops / erma / op / saturation / rationalisation / belge / armature / infogérance / leadership / opportunité / opération / corep / finance / eads / structuration / acquisition / doper / miniaturisation / obligatoire / pilier
30	**Action collective**	collectif / participation / action / lean / up / participer / dinamic / pm / *obj_actionco / manufacturing / start / portée / salon / environnement / communication / groupement / collaboratif / drire / aco / marketing / université / pôle / management / acamas / excellence / performance / partenariats / messier / connaitre / technologique / identifier / *reg_22 / boucle / palier / niort
31	**stratégie diversification**	différenciation / spécialisation / diversifications / *reg_26 / domination / statu_quo / status / quo / externalisation / produit / coût / débouché / réorganisation / externe / éolien / croissance / strategie / interne / export / *ac_industrie / ampoule / diamètre / entrerpise / entretenir / médecine / mériter / préciser / reconquérir / résidentiel / sapin / californie / comptable / dynamiser / exceptionnel / hard

Thème de chaque classe		Traits lexicaux typiques par classe
32	**Projet non communiqué ici**	conclusion / voir / général / generale / cr / cf / commentaire / joint / annexe / remarque / visite / piece / compte_rendu / *reg_11 / annexer / ficher / champ / note / *reg_31 / doc / perception / ci_dessous / *obj_etudsecto / *reg_41 / *aceff_commercesup100sal / *at_2008t2 / *obj_demande / *proj_croissexter / document / rubrique / *obj_aide / bp / dessous / mécanorex / synthèse

Liste des figures

Liste des tableaux