L'apport de l'analyse textuelle à la statistique d'entreprise :

L'exploitation de 10 ans de visites d'entreprises par les DIRECCTE

STATISTIQUE D'ENTREPRISE

Nicolas CAVALLO

JMS 2015
SESSION 3 : ANALYSE DES DONNÉES
1er avril 2015

le cnam

DGE DIRECTION GÉNÉRALE DES ENTREPRISES

Sommaire

Les données textuelles

- 10 ans de visites d'entreprises

Statistique Textuelle

La conjoncture

- Comprendre les effets de la crise sur les projets des entreprises

Les filières

- Cibler les entreprises appartenant à la filière automobile

Les données textuelles

Les données textuelles

- 10 ans de visites d'entreprises

La Direction Générale des Entreprises

DIRECTION GÉNÉRALE DES ENTREPRISES

a pour objectif d'accompagner les entreprises

- Plus de 400 chargés de missions visitent fréquemment des entreprises

A partir de 2004, mise en place d'un outil métier pour collecter les données issues de ces visites d'entreprises

Objectif initial	Nouvel objectif
• Assurer le suivi de l'entreprise visitée	• Exploiter ces données pour **comprendre les entreprises**

Les visites représentent …

En 10 ans,

- ✓ 60 000 visites d'entreprises
- ✓ Plus de 30 000 entreprises différentes visitées

Chaque année :

- ✓ Jusqu'à 6 000 entreprises différentes visitées
- ✓ Plus de 5% de l'activité française concernée
- ✓ Plus de 15% des exportations françaises

Les entreprises visitées sont …

Principalement des entreprises industrielles

- ✓ Plus de la moitié des entreprises visitées sont industrielles
- ✓ 1/5 de l'activité industrielle concernée par des visites chaque année
- ✓ 1/4 des exportations industrielles concernées par des visites chaque année

Principalement des grandes entreprises

- ✓ Chaque année, 44 % des grandes entreprises concernées
- ✓ 8 % de l'activité des ETI et GE
- ✓ Plus de 15 % des exportations des ETI et GE

Les données collectées sont …

Principalement des données textuelles

Données textuelles collectées pour une entreprise visitée

0-1

Conclusion générale
32 000 caractères

0-46

Points forts et faiblesses
4 000 caractères

0-1

Grands donneurs d'ordres
200 caractères

0-1

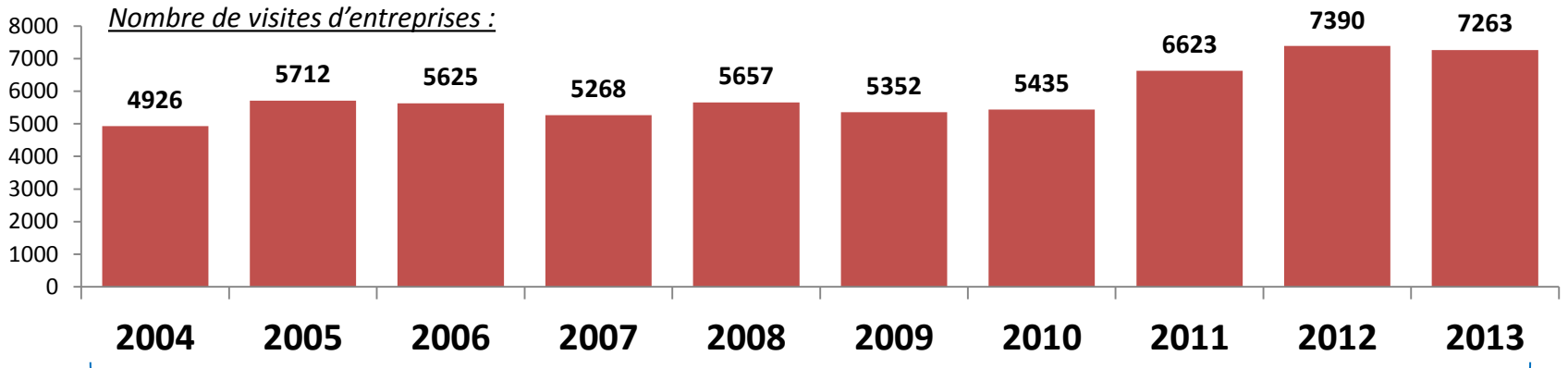
Clients principaux
200 caractères

0-1

Projets de développement
200 caractères

Les données collectées sont …

…Volumineuses



- 53 640 Conclusions Générales

≈ 30 millions d'occurrences

- 17 851 Projets de développement
- 25 045 Clients principaux
- 13 205 Grands donneurs d'ordre
- 120 170 Points forts et faiblesses

≈ 3 millions d'occurrences

La statistique textuelle

Les données textuelles

- 10 ans de visites d'entreprises

Statistique Textuelle

Création du lexique

Formes	fréquence
de	1 795 481
la	757 955
l	687 689
et	678 280
en	569 854
d	566 989
le	549 512
à	538 778
des	503 222
les	437 332
est	330 322
du	319 290
pour	309 261
un	308 340
une	272 404
a	257 069
entreprise	240 917
sur	232 343

- **Analyse des formes employées dans le corpus**
- **Présentées selon la fréquence d'apparition**
- **Problèmes :**
  - ✓ Les formes mises en avant ne sont pas porteuses de sens

Simplification du lexique

Formes	fréquence
de	1 795 481
la	757 955
l	687 689
et	678 280
en	569 854
d	566 989
le	549 512
à	538 778
des	503 222
les	437 332
est	330 322
du	319 290
pour	309 261
un	308 340
une	272 404
a	257 069
entreprise	240 917
sur	232 343

- ✓ Gérer les expressions
- ✓ Gérer les abréviations
- ✓ Lemmatisation
- ✓ Suppression des formes « outils » *(pronoms ,« être »,« avoir »,..)*

Formes « pleines »	fréquence
entreprise	273 446
société	115 339
activité	108 562
marché	99 350
chiffre_affaires	98 939
client	95 810
projet	91 466
production	84 168
produit	80 528
groupe	71 566
site	68 650
développement	64 886
mettre	55 337
commercial	52 651
permettre	52 297
réaliser	50 997
nouveau	48 474
france	43 181

Le tableau lexical entier

Les « formes pleines »

Les segments de texte

	Entreprise	Exporter	Investir	.	.	.	.	.	....	.	.	.	.	.	Mot n°M
ST n°1	1	1	0	0	0	0	1	0	....	0	0	0	0	0	0
ST n°2	0	0	1	0	0	0	0	0	....	0	0	1	0	0	0
ST n°3	1	0	0	0	0	0	0	0	....	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
ST n°N	0	0	0	0	1	0	0	0		0	1	0	0	0	0

- **Application sur ce tableau des méthodes d'analyses multivariées**

La conjoncture

Les données textuelles

- 10 ans de visites d'entreprises

Statistique Textuelle

La conjoncture

- Comprendre les effets de la crise sur les projets des entreprises

Analyse textuelle sur les projet des entreprises

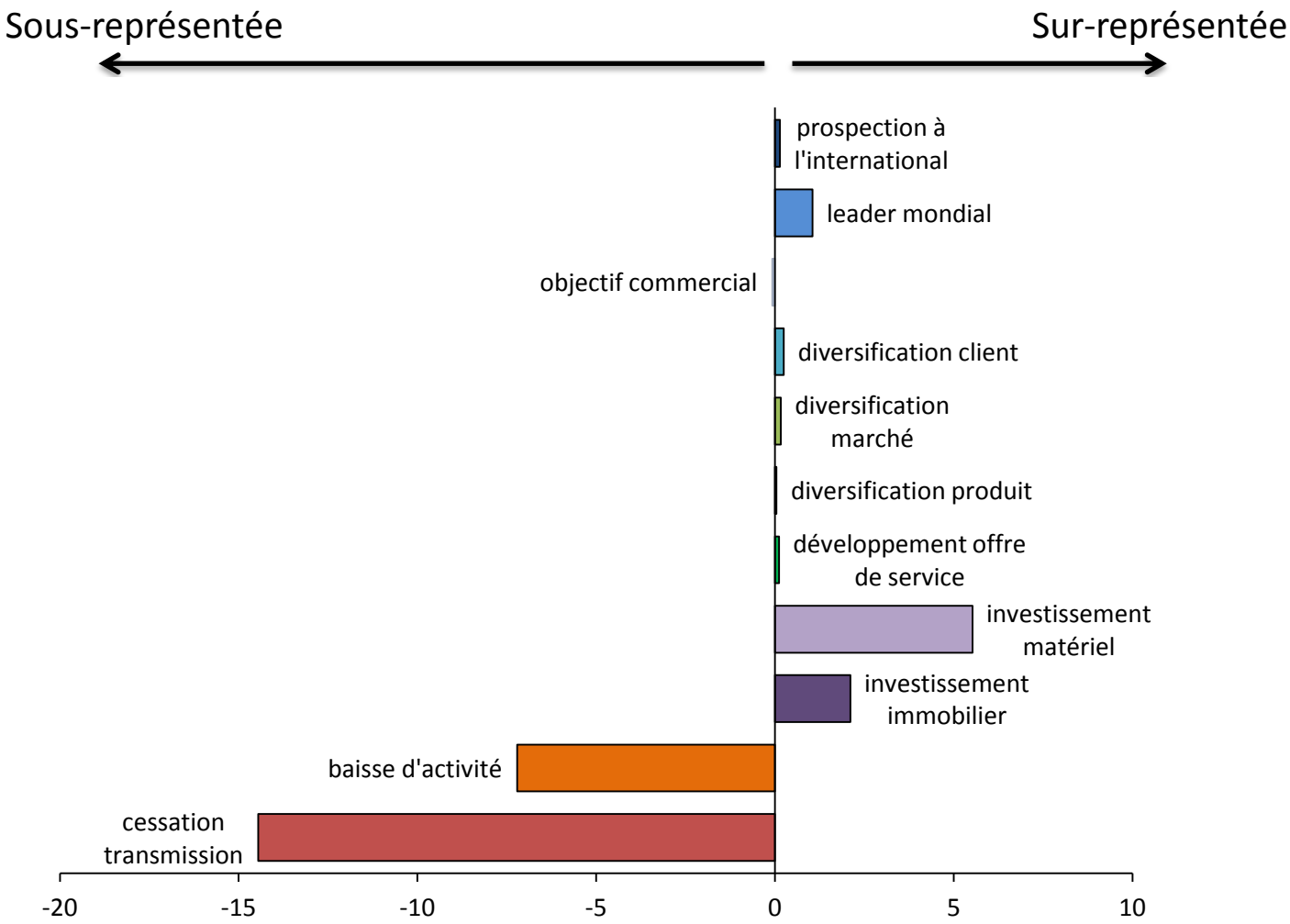
- Application de la méthode de classification ALCESTE (Reinert) aux 18 000 projets de développement renseignés entre 2008 et 2013
- **Création de 11 classes ou thématiques** principales relatives aux projets des entreprises

Exemple pour 3 classes:

Thème de la classe	Traits lexicaux typiques par classe
baisse d'activité	baisser / crise / difficile / commande / difficulté / situation / chiffre_affaires / charger / carnet / activité / perte / 2009 / année / subir / économique / trésorerie / financier / partiel / retrouver / conjoncture / chômage / rentabilité / 2008 / face
investissement matériel	investissement / production / ligne / machine / unité / modernisation / matériel / nouvelle / projet / capacité / traitement / usinage / outil / achat / acquisition / investir / peinture / découper / productivité / automatiser / bois / automatisation / installation / fabrication
Prospection de marchés à l'international	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 / export / implantation

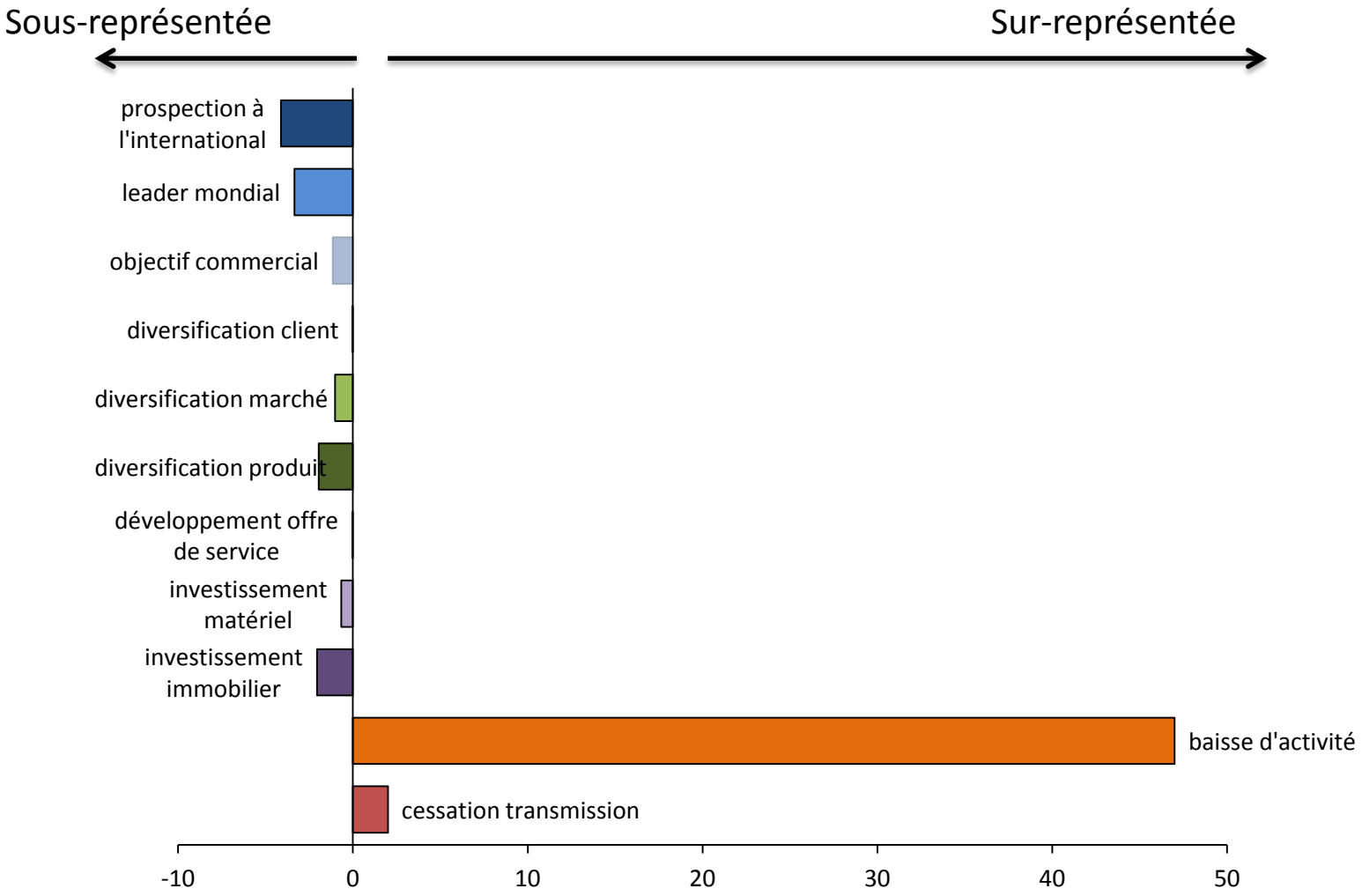
2008

La crise n'impacte pas encore les projets



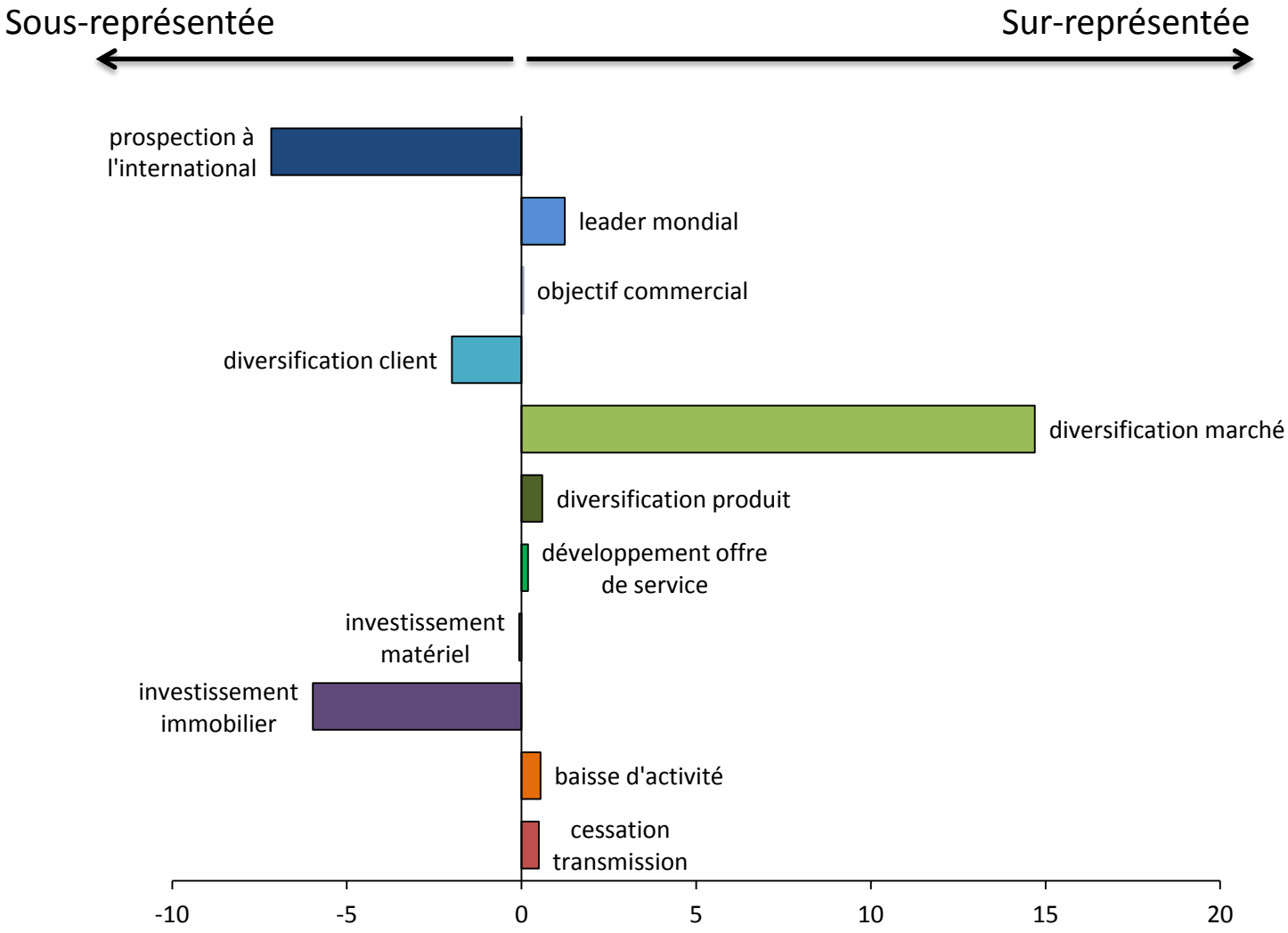
2009

Les entreprises subissent la crise… et prévoient une baisse d'activité



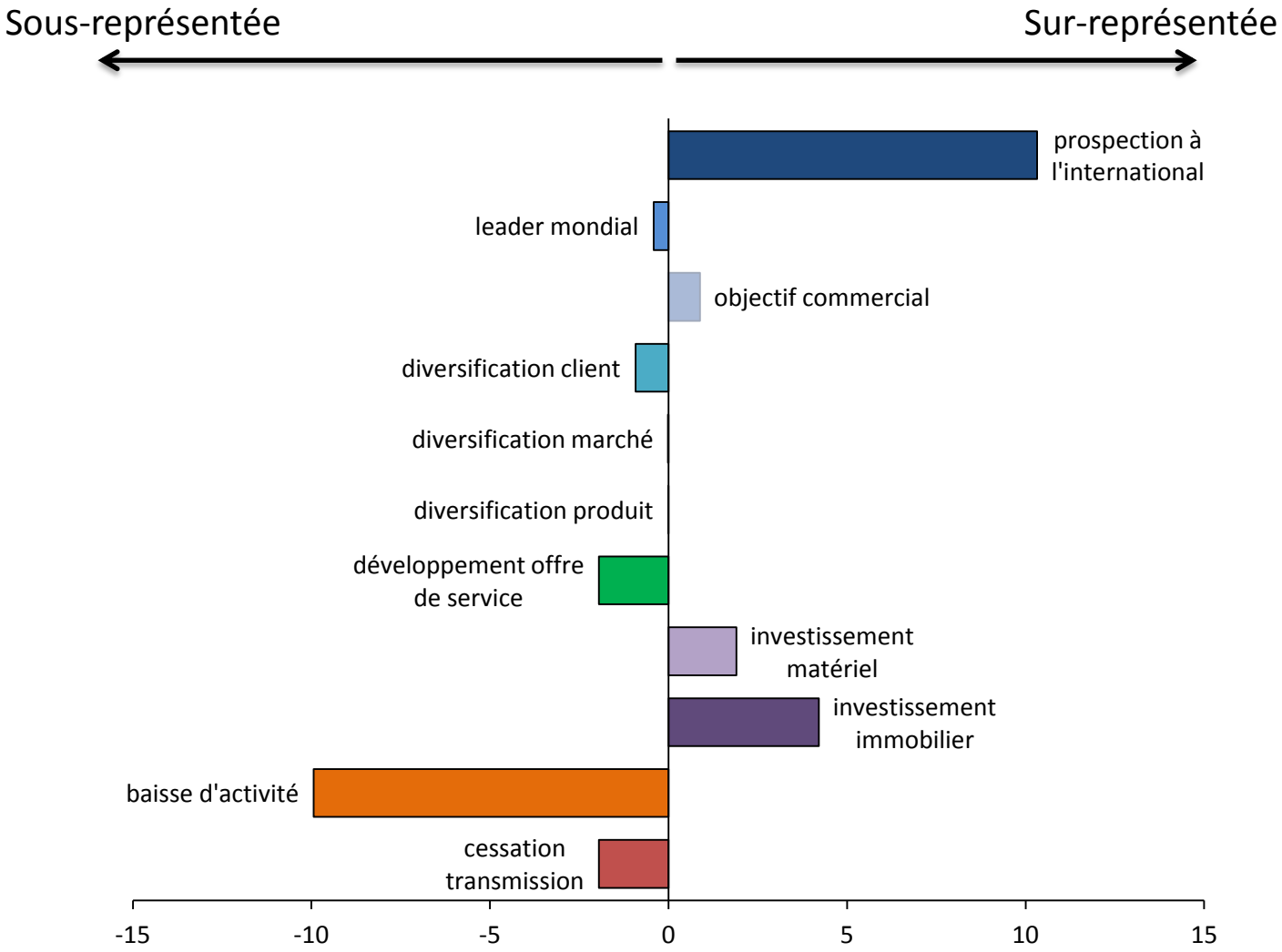
2010

La crise est mondiale… il faut se diversifier



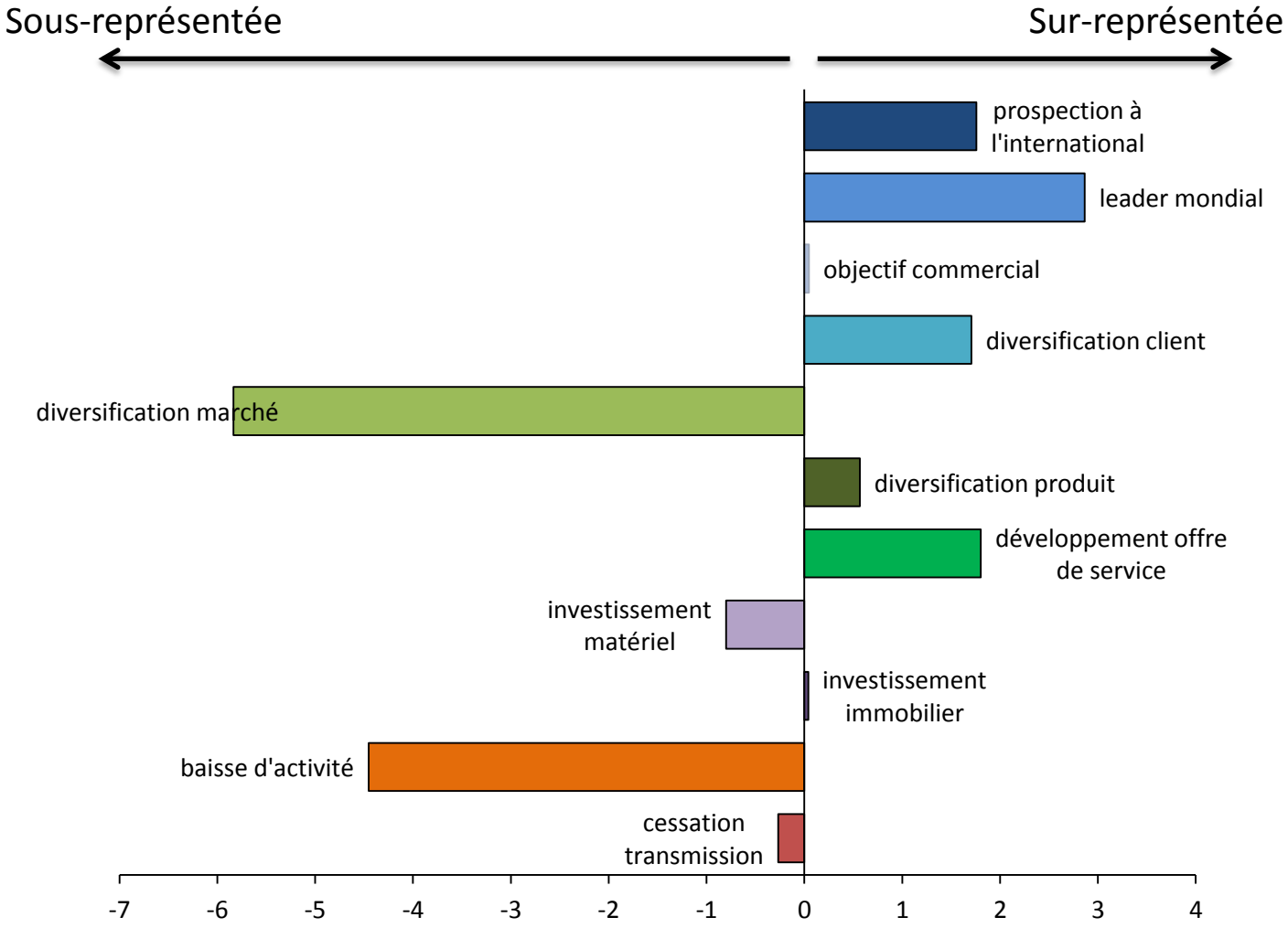
2011

Fin de la crise ?



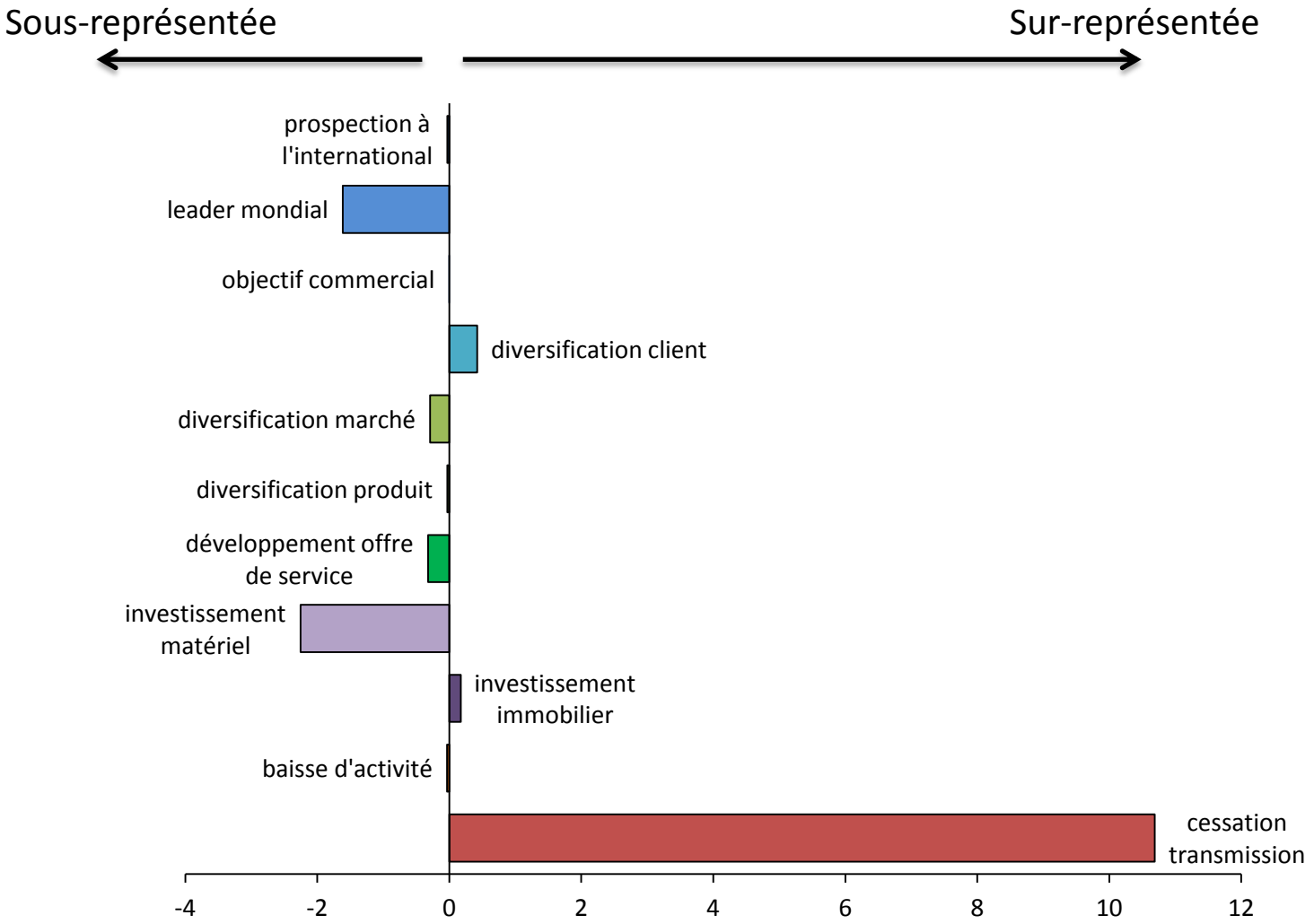
2012

L'international toujours dans la tête des dirigeants



2013

baisse des investissements et davantage de cessations…



Les filières

Les données textuelles

- 10 ans de visites d'entreprises

Statistique Textuelle

La conjoncture

- Comprendre les effets de la crise sur les projets des entreprises

Les filières

- Cibler les entreprises appartenant à la filière automobile

Prédire l'*appartenance à la filière automobile*

- **Problématique :** très peu d'information sur les filières car la statistique publique a une approche « secteur d'activité »

- **Pour le moment :** seule l'APE (code NAF) est disponible pour cibler les entreprises d'une filière

- **Solution :**
  - ✓ Utiliser les données textuelles comme variable explicative

Créer la base d'apprentissage

- **L'information sur les filières est disponible depuis 2008, soit pour 25 232 visites d'entreprises**
- **Pour évaluer le pouvoir prédictif des variables textuelles on partitionne :**

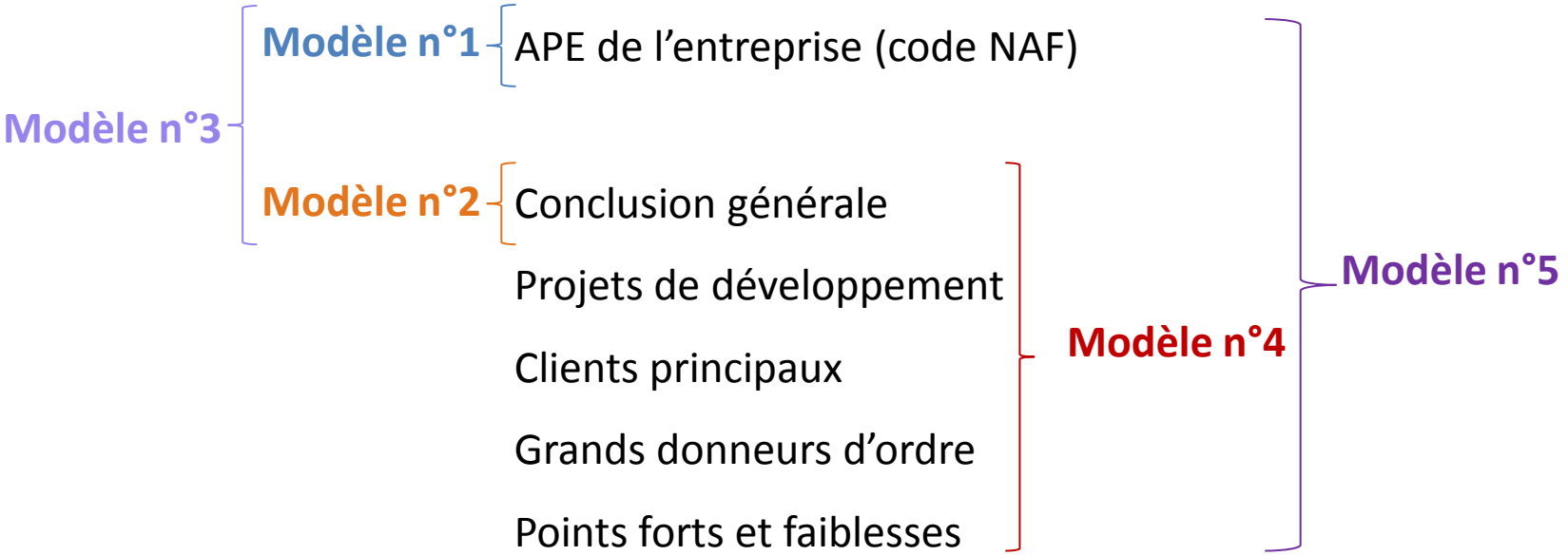
Base initiale		
Total	25 232	
Dans la filière auto	**4 283**	17%

70% *30%*

Base d'apprentissage		
Total	17 661	
Dans la filière auto	**2 998**	17%

Base de validation		
Total	7 571	
Dans la filière auto	**1 285**	17%

Création et comparaison de 5 modèles



	Modèle n°1	Modèle n°2	Modèle n°3	Modèle n°4	Modèle n°5
AUC	0,715	0,827	0,863	0,908	0,919
Lift à 10%	3,15	4,65	4,77	5,25	5,25

Conclusion sur les analyses textuelles

	A permis de …	Va permettre de…
1ère Analyse	• **Observer les effets de la crise** sur les projets des entreprises *(investissements, exportations, difficultés rencontrées)*	• **Comprendre plus en détail** les entreprises • Créer de **nouveaux indicateurs** pour disposer d'un nouvel **outil conjoncturel**
2ème Analyse	• **Créer un modèle efficace** pour cibler les entreprises appartenant à la filière automobile	• **Améliorer les répertoires de filières** *(+3500 nouvelles entreprises de la filière « automobile » envisagées)* • **Aider au ciblage** pour d'autres phénomènes *(situation économique de l'entreprise, ensemble des filières, innovation, etc..)*

Conclusion sur les données textuelles

Les données textuelles sont riches

Il faut tout de même mesurer les faiblesses de l'analyse textuelle à un niveau fin (non prise en compte de la négation, fautes d'orthographe, style atypique de l'auteur, etc)

Maintenant, l'objectif est de **trouver davantage de données textuelles**

Soit en les récupérant
(*WEB, Dans des bases déjà disponibles)*

Soit en les créant
(*+ de réponses ouvertes dans les enquêtes)*

Pour élargir le champ des données exploitées au sein de la statistique d'entreprise

Merci pour votre attention

STATISTIQUE D'ENTREPRISE

Nicolas CAVALLO

le cnam

DGE DIRECTION GÉNÉRALE DES ENTREPRISES