CAHIER DE RECHERCHE

FEVRIER 92 ■ N° 25

EFFETS CUMULES DE DIFFERENTS CRITERES SOCIO-DEMOGRAPHIQUES SUR LES REPONSES A UNE QUESTION D'OPINION :

Résultats empiriques commentés d'une segmentation, d'une régression logistique et d'une analyse discriminante sur coordonnées factorielles

Laurent Clerc
Jean-Luc Volatier

CREDOC

CAHIER DE RECHERCHE

EFFETS CUMULES DE DIFFERENTS CRITERES SOCIO-DEMOGRAPHIQUES SUR LES REPONSES A UNE QUESTION D'OPINION :

Résultats empiriques commentés d'une segmentation, d'une régression logistique et d'une analyse discriminante sur coordonnées factorielles

Laurent Clerc
Jean-Luc Volatier

FEVRIER 1992

Ce travail a bénéficié des conseils d'Anastasios Iliakopoulos du CREDOC

142, rue du Chevaleret
75013 - PARIS

DEPARTEMENT "CONDITIONS DE VIE ET ASPIRATIONS DES FRANCAIS"

Ce travail a bénéficié d'un financement au titre de la subvention recherche attribuée au CREDOC par le Commissariat Général du Plan.

La question analysée dans ce rapport a été financée par le Ministère du Travail, de l'Emploi et de la Formation Professionnelle (SES).

Le département "Conditions de vie et Aspirations des Français" est composé de :

. Georges Hatchuel (Directeur adjoint)
. Laurent Clerc, Catherine Duflos, Ariane Dufour, Françoise Gros, Lucette Laurent, Viviane Payet-Thouvenot, Jean-Luc Volatier.

C R E D O C

Président : Bernard Schaefer
Directeur : Robert Rochefort

SOMMAIRE

Introduction

Il est très fréquent, dans les enquêtes d'opinion, d'observer des effets conjoints ou cumulés de différentes caractéristiques socio-démographiques sur les opinions des individus[1]. On constate par exemple souvent des variations d'opinion selon simultanément le sexe, l'âge, le niveau d'études, le niveau d'urbanisation. Le plus fréquemment, l'effectif des personnes enquêtées (2000 pour l'enquête "Aspirations et conditions de vie" du CREDOC) ne permet cependant d'observer une variable que selon deux critères croisés.

Plusieurs méthodes différentes sont couramment utilisées pour contourner cette difficulté. Une méthode descriptive, la segmentation, consiste à "économiser" les effectifs d'enquêtés disponibles en sélectionnant les croisements les plus significatifs et en regroupant les modalités d'une même variable qui s'avèrent proches au regard de la question étudiée.

D'autre part, deux méthodes de régression, la régression logistique et l'analyse factorielle discriminante, font l'hypothèse de l'existence d'effets additifs[2] des variables socio-démographiques sur les variables de comportement ou d'opinion étudiées pour pouvoir mesurer ces effets.

L'objectif de cette note est de comparer, sur un cas concret, les résultats obtenus par ces différentes méthodes. En effet, si les avantages théoriques de l'analyse discriminante ont été démontrés dans les "bonnes" conditions[3], la supériorité pratique de l'une ou l'autre des deux méthodes de régression varie selon ces mêmes conditions d'application (normalité des variables explicatives, identité des matrices de variance intra des deux groupes étudiés)[4]. Dans notre cas, les circonstances sont plutôt favorables à la régression logistique, c'est le résultat que nous tenterons d'illustrer plus loin.

---

[1] L. Lebart, *"Sept ans de perceptions - Evolution et structure des opinions en France de 1978 à 1984"*, CREDOC, 1986.

[2] En régression logistique, on peut aussi prendre en compte des interactions entre variables explicatives.

[3] B. Efron, *"The efficiency of logistic regression compared to normal discriminant analysis"*, JASA 70, pp 892-898, 1975.

[4] J. J. Press and S. Wilson, *"Choosing between logistic regression and discriminant analysis"* JASA 73, pp 699-705, 1978.

Après avoir présenté plus en détail la question d'opinion étudiée, nous allons dans un premier temps décrire les associations statistiques entre les réponses à cette question et les variables socio-démographiques, au moyen d'une segmentation. Cette outil permettra de mettre en relief des effets additifs de variables socio-démographiques ou de signaler l'existence d'effets non additifs. Il se révèle à ce titre un bon préliminaire aux méthodes de régression.

Dans un second temps, nous allons appliquer l'analyse discriminante et la régression logistique et comparer les résultats des deux méthodes. Dans la dernière partie de cette note, on essaiera de contourner l'hypothèse d'additivité qui est sans doute ici trop forte, en réalisant plusieurs analyses discriminantes sur des sous-populations.

Mais auparavant, décrivons rapidement le sujet de l'étude qui sert ici d'application.

1. Le sujet d'étude : les raisons pour lesquelles les entreprises ont du mal à recruter

1.1 La variable étudiée

Pourquoi les entreprises ont-elles des difficultés à embaucher alors que les chômeurs sont nombreux sur le marché du travail ? les Français ont différentes interprétations du phénomène. Les uns attribuent la responsabilité de ce déséquilibre aux entreprises, les autres désignent les demandeurs d'emploi comme insuffisamment formés ou trop exigeants.

La question à laquelle nous souhaitons donner une réponse est la suivante : *selon quels clivages socio-démographiques la population se sépare-t-elle pour donner les deux types d'explications à ce déséquilibre du marché de l'emploi ? Les différences socio-culturelles (approchées par l'âge, le niveau d'études,...) jouent-elles plus fortement dans les réponses apportées que l'expérience personnelle du chômage ? L'effet de la catégorie socio-professionnelle est-il "propre" ou ne fait-il que refléter des effets d'âge et de diplôme ?*

La question étudiée dans cette note est issue de la vague d'automne 1990 de l'enquête "Aspirations et conditions de vie des Français"(*). En voici l'intitulé :

Les entreprises déclarent rencontrer de plus en plus de difficultés pour recruter les personnes dont elles ont besoin, notamment pour des emplois qualifiés. Selon vous, quelle en est la raison principale ?

(en %)

1. Les entreprises sont trop exigeantes	16
2. Il n'y a pas assez de personnes qualifiées ou compétentes	36
3. Les salaires proposés sont insuffisants	20
4. Les conditions de travail offertes sont pénibles	2
5. Les entreprises ne veulent pas payer la formation nécessaire	15
6. Les demandeurs d'emploi sont trop exigeants	7
7. Les emplois proposés sont souvent trop éloignés du domicile.	3
8. Ne sait pas	1
Ensemble	100

(*) - Cette question a été financée par le Ministère du Travail, de l'Emploi et de la Formation Professionnelle - SES.

Un recodage a été effectué pour traiter les réponses : les modalités font en effet appel soit à la responsabilité des entreprises (modalités 1, 3, 4 et 5), soit à la responsabilité des demandeurs d'emploi (modalités 2 et 6) ; la modalité 7 restante n'ayant été retenue que par 2% des enquêtés, nous l'avons assez arbitrairement affectée au groupe faisant appel à la responsabilité des chômeurs[5].

La nouvelle variable détermine ainsi deux groupes (respectivement 55% de la population des enquêtés pour la responsabilité des "entreprises" et 45% pour la responsabilité des "chômeurs"). Pour chaque enquêté, nous pouvons donc croiser le groupe auquel il appartient et l'ensemble de ses réponses aux questions de l'enquête "Aspirations et conditions de vie".

1.2 Les critères socio-démographiques les plus caractéristiques : âge, diplôme, PCS, vécu du chômage

Après un croisement systématique avec l'ensemble des variables, les critères socio-démographiques les plus liés aux réponses données à cette question s'avèrent être les suivants : la PCS, l'âge, le "vécu du chômage", c'est-à-dire le nombre de situations de chômage vécues au cours des dix années précédant l'enquête (cinq modalités), et le niveau de diplôme (en quatre modalités).

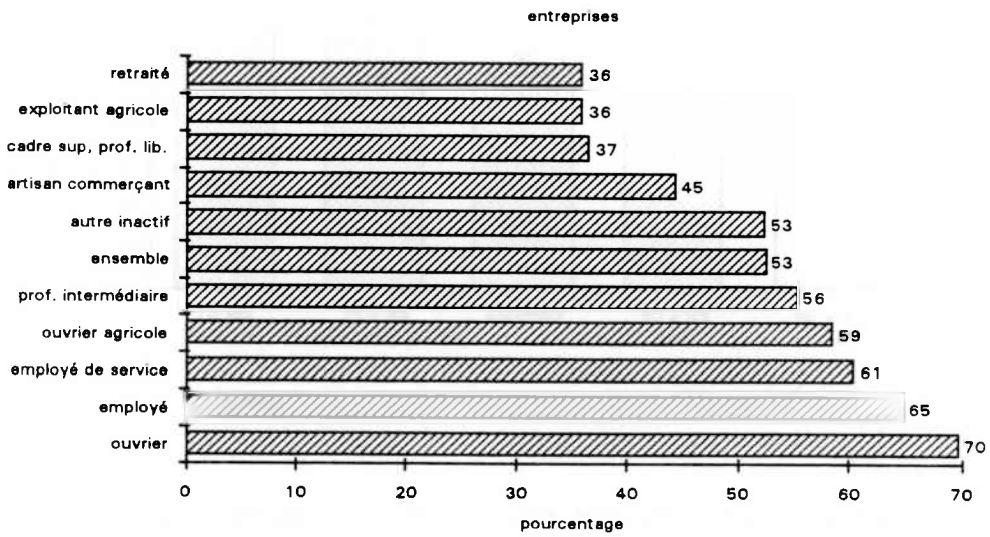
Ce sont les retraités, les indépendants et les cadres supérieurs, qui attribuent le moins souvent aux entreprises, et donc le plus souvent aux demandeurs d'emploi, la responsabilité des difficultés de recrutement des entreprises (graphique 1).

---

[5] Cette affectation est sans conséquence sur les résultats qui suivent : des analyses ont été effectuées sans tenir compte des individus ayant choisi cette modalité de réponse. Elles ont conduit à des résultats voisins. Cf. annexe.

Graphique 1

Part des Français qui attribuent la responsabilité des difficultés d'embauche aux entreprises selon la CSP

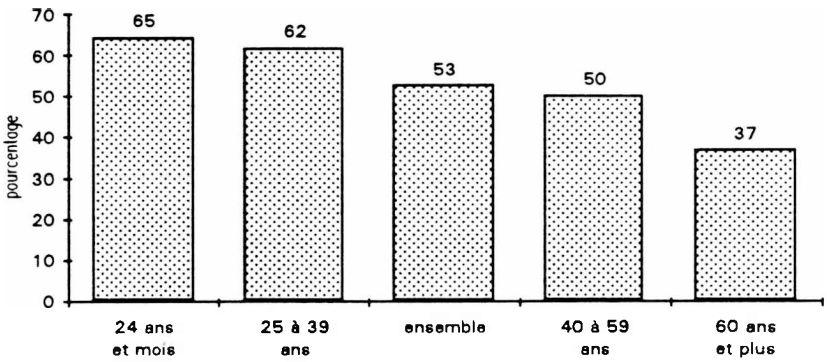


A l'opposé, ce sont les ouvriers qui sont les plus nombreux à attribuer cette responsabilité aux entreprises (70%).

Les opinions sont aussi partagées selon l'âge : les plus jeunes font plus souvent porter la responsabilité du déséquilibre sur les offreurs d'emploi que sur les demandeurs. La position des personnes âgées est proche de celle des retraités : ils considèrent plus souvent que ce sont le **manque de formation** et, dans une moindre mesure, **les exigences** des demandeurs d'emploi, qui expliquent les difficultés des entreprises à embaucher.

Graphique 2

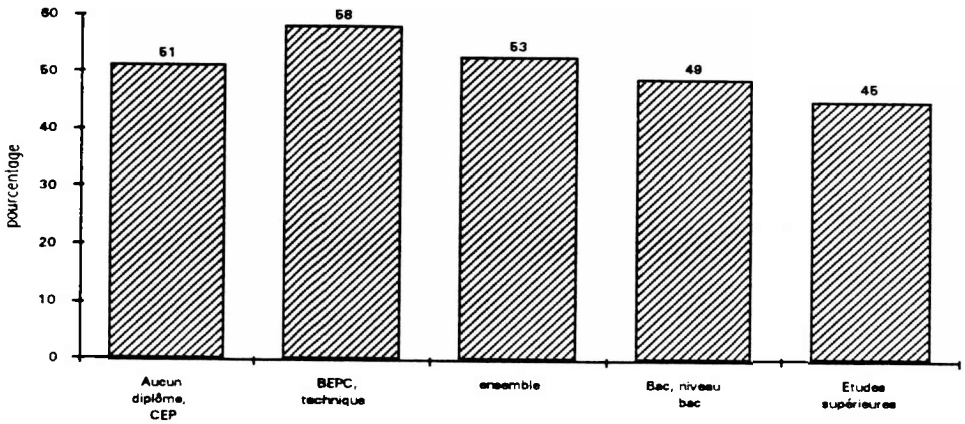
Les jeunes font plus porter la responsabilité des difficultés d'embauche aux entreprises



L'effet du diplôme semble aussi corrélé à celui de la CSP : les plus diplômés, tout comme les cadres supérieurs, citent moins souvent les entreprises comme responsables. Il faut donc souligner que l'effet du niveau d'études (plus on est diplômé, moins on désigne les entreprises) va à l'encontre de celui de l'âge (plus on est jeune, plus on désigne les entreprises) ; il y a en effet plus de diplômés parmi les jeunes que parmi les personnes âgées. Les effets "purs" de l'âge et du diplôme sont donc sans doute plus marqués que les variations unidimensionnelles ne le laissent penser. Un autre résultat nous conduit à nous poser une interrogation analogue : l'effet du diplôme ne semble pas *a priori* régulier. Les Français munis du seul CEP ou d'aucun diplôme ont une position proche de la moyenne. S'agit-il alors de la neutralisation des effets contraires de l'âge et du niveau d'études mentionnés plus haut, mais agissant ici dans d'autres catégories : les personnes âgées et les non-diplômés ?

Graphique 3

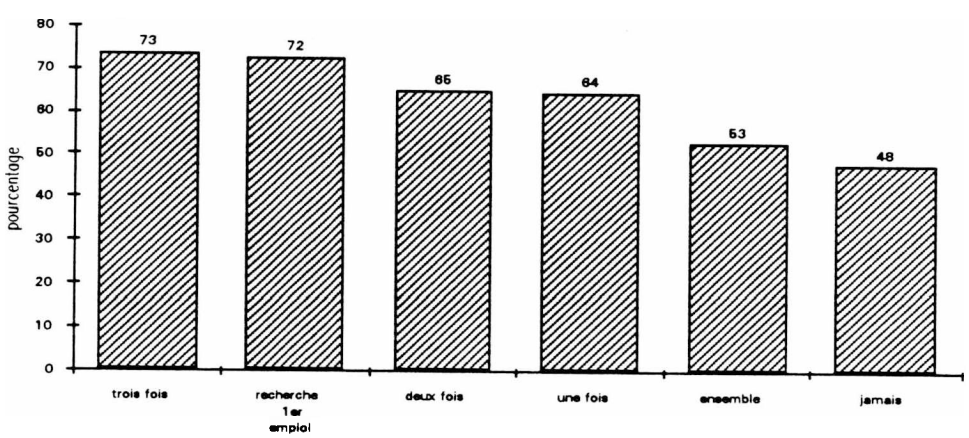
Les diplômés du secondaire attribuent la responsabilité du déséquilibre aux entreprises



Enfin, l'expérience du chômage induit assez clairement une plus grande compréhension vis-à-vis des demandeurs d'emploi et corrélativement une plus grande sévérité vis à vis des entreprises.

Graphique 4

Citation de la responsabilité des entreprises selon le nombre de périodes de chômage[6]



[6] Nombre de périodes de chômage dans les dix ans précédant l'enquête.

2. Une segmentation

Pour analyser les effets conjugués de l'âge, du niveau de diplôme, de la PCS et de l'expérience du chômage sur les opinions vis-à-vis du marché de l'emploi que nous étudions ici, nous avons dans un premier temps appliqué une méthode descriptive : la segmentation.

Appliqué à cette étude, le principe consiste à rechercher, dans un premier temps, la variable socio-démographique la plus liée statistiquement à la variable d'opinion à étudier, en comparant les niveaux de significativité des tests du chi2. On obtient ainsi deux catégories socio-démographiques (pour chacune des deux opinions retenues) et deux distributions bien différentes de la variable à étudier dans chacune de ces deux catégories ; on cherche ensuite dans chacun des deux sous-échantillons la meilleure variable explicative et ainsi de suite.

L'avantage de la segmentation est que son résultat se visualise par un arbre très suggestif (graphique 5)[7]. Dans notre exemple sur le chômage, les deux premières dichotomies se font selon la même variable, de la PCS en dix postes, et conduisent à une séparation de l'arbre en trois branches. On constate que c'est le risque vis-à-vis du chômage qui intervient ici : les catégories les moins concernées (agriculteurs, artisans-commerçants-chefs d'entreprises, cadres supérieurs et professions libérales, retraités) considèrent à 60% que la responsabilité de la situation paradoxale actuelle (les entreprises ont du mal à recruter des personnes qualifiées, tandis que le nombre de chômeurs ne cesse de s'accroître) repose sur les demandeurs d'emploi plutôt que sur les entreprises.

Elles s'opposent fortement aux catégories "à fort risque de chômage" (employés, ouvriers, ouvriers agricoles, personnels de service), qui font le raisonnement inverse, considérant à 69,5% que les entreprises sont responsables de cet état de fait. Entre ces deux groupes, les professions intermédiaires et les autres inactifs (peut-être influencés par la profession et la situation de leur conjoint) sont d'un avis plus partagé.

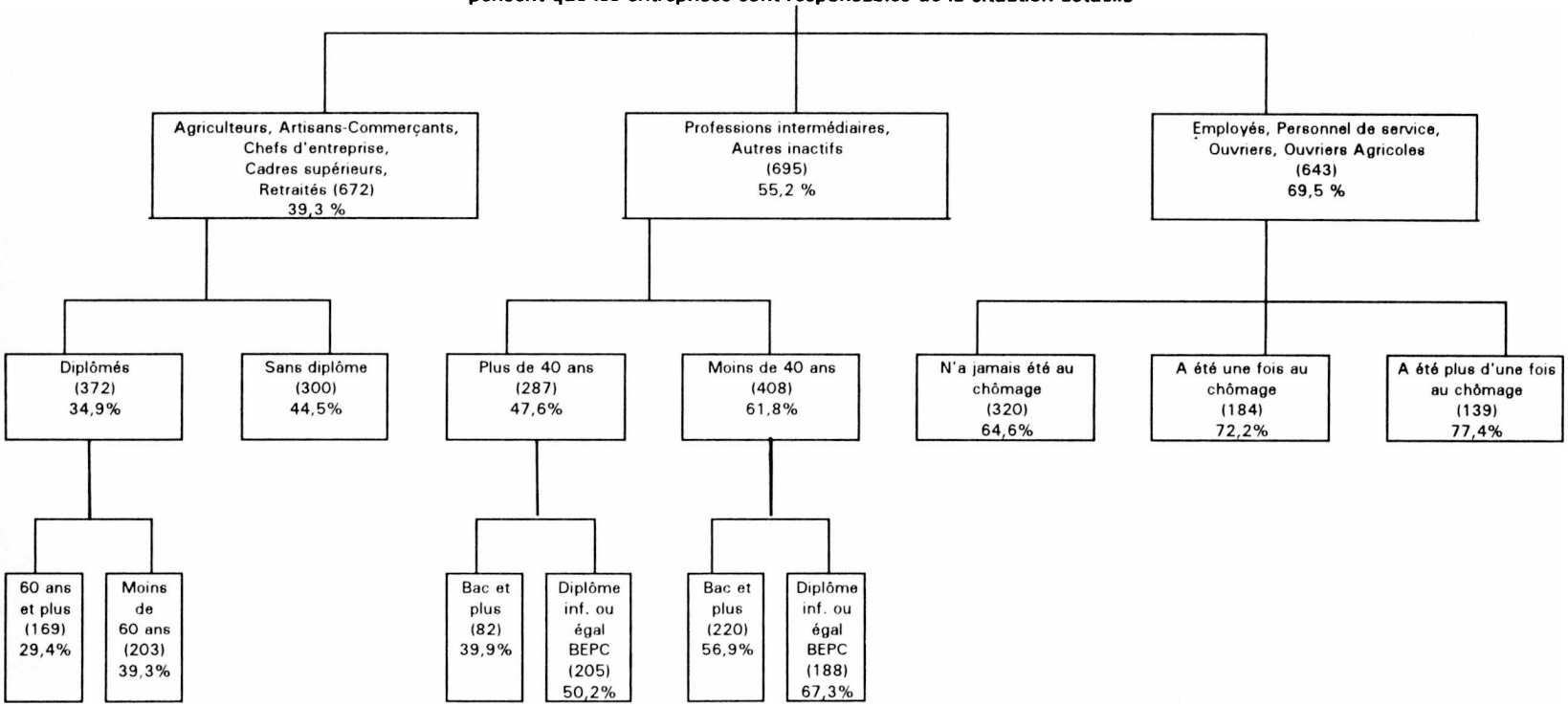
Au deuxième niveau, la variable "expliquant le mieux", en termes statistiques, la variable étudiée, en l'occurence l'opinion sur la responsabilité entreprise/chômeurs, n'est pas la même suivant les groupes déterminés au premier niveau.

---

[7] Les pourcentages calculés dans cette segmentation ne sont pas redressés du fait de l'utilisation d'un logiciel spécifique.

Graphique 5

Au sein de la population totale ((2010 individus), 54,7% (soit 1100 individus) pensent que les entreprises sont responsables de la situation actuelle



Répartition de l'ensemble de la population selon la proportion de ceux qui imputent aux entreprises les difficultés de recrutement actuelles

Pour les personnes les moins concernées par le risque de chômage (agriculteurs, artisans-commerçants-chefs d'entreprise, cadres supérieurs et professions libérales, retraités, professions intermédiaires et autres inactifs), ce sont les variables d'âge et de niveau de diplôme qui jouent ensuite, de façon symétrique. Ces variables opposent respectivement les plus jeunes aux plus âgés, les plus diplômés aux moins diplômés. On obtient ainsi des groupes estimant que la responsabilité de la situation est à imputer aux employeurs dans des proportions variant de 29,4% (agriculteurs, artisans-commerçants-chefs d'entreprise, cadres supérieurs et professions libérales supérieures, retraités, ayant plus de 60 ans et disposant d'un diplôme supérieur au CEP) à 67,3% (professions intermédiaires, autres inactifs, de moins de 40 ans et n'ayant que le CEP ou pas de diplôme).

Pour le groupe des personnes les plus exposées au risque de chômage, c'est la variable "vécu du chômage au cours des dix années précédant l'enquête" qui intervient au deuxième niveau, créant trois sous-groupes suivant que l'on n'a jamais connu le chômage, qu'on l'a connu une fois seulement, ou plusieurs fois. Les proportions de personnes jugeant les entreprises responsables de la situation actuelle s'échelonnent ici de 64,6% à 77,4%. Pour l'ensemble de ce groupe des personnes "à risque", c'est donc le vécu du chômage qui intervient le plus.

Il faut cependant se méfier de la puissance évocatrice de la segmentation qui n'est qu'un angle de description parmi d'autres des effets conjoints des variables socio-démographiques étudiées. Sans mettre en doute la valeur heuristique de cette méthode, ses résultats doivent être accueillis avec un regard critique et confrontés aux résultats obtenus au moyen d'autres outils. Dans cette étude, qui a pour but principal de comparer deux méthodes de régression, l'analyse discriminante et la régression logistique, l'emploi préliminaire de la segmentation a pour objectif de donner quelques informations initiales sur les conditions de cette comparaison. L'information principale est qu'il existe bien des effets additifs (par exemple de l'âge et du diplôme), mais ces effets sont parfois conditionnels : par exemple, l'âge et le diplôme jouent surtout chez les cadres et les indépendants.

La comparaison des résultats de l'analyse discriminante et de la régression logistique va donc s'effectuer sur des données qui ne sont pas parfaitement adaptées. On est donc bien loin d'une simulation, mais au contraire plus proche des conditions d'analyse de la plupart des questions d'opinion de l'enquête "Aspirations et conditions de vie". Pour nombre de sujets d'étude, plusieurs catégories d'individus caractérisées

par des approches différentes du sujet coexistent dans la population : les utilisateurs d'un service ou d'une institution et les non-utilisateurs, ceux qui connaissent le sujet et ceux qui le connaissent moins bien, etc. Par ailleurs, au sein de chacune de ces différentes catégories, les variations des réponses à la question étudiée obéissent à des logiques différentes.

3. Comparaison d'une analyse discriminante et d'une régression logistique

3.1 Les deux méthodes

A partir des quatre variables socio-démographiques, toutes nominales, l'objectif de l'analyse discriminante est de construire une fonction discriminante permettant d'affecter les enquêtés dans deux sous-populations, l'une faisant allusion à la responsabilité des demandeurs d'emploi, l'autre à celle des entreprises. Comme ces variables sont nominales, on réalise au préalable une analyse factorielle, une ACM dans le cas présent, et on effectue l'analyse factorielle discriminante à partir des coordonnées sur les axes de l'ACM. Il s'agit de l'application de la méthode Disqual, développée par G. Saporta[8].

L'objectif de la régression logistique est voisin : calculer des probabilités, pour l'interviewé, d'être dans chacune des deux sous-populations (responsabilité des demandeurs d'emploi, responsabilité des entreprises) sachant les caractéristiques socio-démographiques de cet interviewé. Cette probabilité prend par hypothèse la forme d'une fonction logistique :

$P = [1 / 1 + e (- bx)]$ où b est un vecteur de paramètres à estimer et x le vecteur ayant pour composantes les valeurs des différentes variables socio-démographiques explicatives.[9]

3.2 La régression logistique a obtenu un meilleur taux de bons classements

Pour les deux méthodes, on peut comparer les réponses réelles des individus et les réponses estimées. Par cette comparaison, on obtient, pour l'ensemble de la population, un pourcentage d'individus bien ou mal classés :

---

[8] De nombreux manuels d'Analyse des données présentent la théorie de l'analyse discriminante :
. M. Volle *"L'analyse des données"*, Ed. Economica.
. J.M. Bouroche et G. Saporta, *"L'analyse des données"*, collection Que sais-je ?
. G. Saporta, *"Probabilités, analyse des données et statistiques"*, Ed. Technip, 1990.

[9] On trouvera une présentation théorique de la régression logistique et de ses applications dans :
C. Gourieroux, *"Econométrie des variables qualitatives"*, Ed. Economica.

Pourcentages de classement dans chacun des groupes et au total

	Bien classés	Mal classés	Total
Analyse discriminante	61,7	38,3	100,0
Régression logistique	64,9	35,1	100,0

Ces pourcentages donnent une première estimation de la qualité de l'explication de la variable d'opinion par les quatre variables socio-démographiques explicatives. Dans notre exemple, le pourcentage de bien classés est légèrement supérieur dans le cas de la régression logistique (65 %).

3.3 Les effets des modalités des différentes variables explicatives n'ont pas toujours le même sens d'après les deux méthodes

3.3.1 L'analyse discriminante

L'analyse factorielle discriminante conduit à estimer les coefficients de la fonction discriminante de Fisher qui permet d'affecter les individus à chacun des deux groupes d'opinion étudiés selon les valeurs (positives ou négatives) de cette fonction. C'est une fonction linéaire des variables indicatrices d'appartenance aux différentes modalités des caractéristiques socio-démographiques. Plus l'effet d'une modalité socio-démographique sur la variable étudiée est fort, plus le coefficient de la fonction discriminante de Fisher correspondant à cette modalité est élevé en valeur absolue. Ainsi, il apparaît dans le tableau suivant que le fait de n'avoir jamais été au chômage (coefficient -1,23) joue fortement sur le partage des opinions étudiées ici. Le signe négatif signifie que le fait de n'avoir jamais été au chômage est lié au fait de ne pas attribuer la responsabilité du déséquilibre sur le marché de l'emploi aux entreprises, mais plutôt aux demandeurs d'emploi. Parallèlement, ce résultat indique que c'est principalement le fait d'avoir été au chômage qui conduit à considérer que ce sont les entreprises qui sont responsables des difficultés de recrutement.

Coefficients de la fonction discriminante

Variable	Coefficient
Une fois au chômage	0,20
Deux fois au chômage	0,09
Trois fois ou plus au chômage	0,09
Recherche 1er emploi *	0,01
Jamais au chômage	-1,23
Agriculteur	-0,04
Artisan-Comm-Chef d'entr.	-0,01
Cadre sup-Prof.lib. sup.	-0,09
Prof. intermédiaire	-0,06
Employé	0,16
Personnel de service	0,03
Ouvrier	0,26
Ouvrier agricole *	0,01
Retraité	-0,27
Autre inactif	0,01
Moins de 24 ans	0,15
25-40 ans	0,38
40-60 ans	-0,21
Plus de 60 ans	-0,30
Aucun diplôme-CEP	-0,19
Diplôme niveau BEPC	0,46
Bac ou équivalent	0,02
Diplôme supérieur	-0,21
Constante	0,01

* Effectif inférieur à 50.

Cette première méthode nous montre donc que, sous l'hypothèse d'un modèle additif, les modalités qui interviennent le plus sont les suivantes :

- "N'a jamais été au chômage au cours des dix années précédant l'enquête" (dans le sens "la responsabilité repose sur les demandeurs d'emploi");

- "Niveau de diplôme BEPC ou technique inférieur au Bac" (dans le sens "la responsabilité repose sur les entreprises");

- "Age 25 à 39 ans" (dans le sens "la responsabilité repose sur les entreprises").

D'une manière plus générale, si l'on regarde les variables ayant plusieurs modalités aux coefficients importants en valeur absolue, on s'aperçoit que le modèle linéaire fournit l'ordre suivant pour les variables explicatives : fréquence des situations de chômage, catégorie d'âge et niveau de diplôme[10], PCS.

Soulignons aussi que les effets des différentes modalités d'une variable "ordonnée" ne sont pas toujours classés dans le sens où l'analyse univariée initiale nous conduirait à l'attendre :

- Le fait d'avoir été plusieurs fois au chômage ne semble pas conduire à citer encore plus souvent les entreprises que le fait d'avoir été une fois au chômage.

- Les moins de 24 ans semblent moins incriminer les entreprises que les 25-39 ans.

- Les Français sans diplôme ou munis du seul CEP considèrent plutôt que la responsabilité revient aux demandeurs d'emploi. Ce résultat est assez surprenant : on pouvait suspecter que la position *a priori moyenne* des non-diplômés était due en partie à un effet d'âge qui aurait contrecarré une plus forte citation des entreprises chez les non-diplômés "toutes choses égales par ailleurs".

Utilisant les regroupements de modalités suggérées par la segmentation, on obtient par une seconde analyse discriminante les résultats suivants (la probabilité de classement n'est pas modifiée par ce recodage) :

---

[10] Pour juger de l'importance d'une variable, on a considéré la somme pondérée des valeurs absolues des coefficients de la variable (l'écart entre le plus fort et le plus faible coefficient d'une même variable donne des résultats analogues).

Variable	Coefficient
Déjà été au chômage	0,38
Jamais été au chômage	-0,99
PCS1	-0,89
PCS2	0,15
PCS3	0,73
Moins de 40 ans	0,89
Plus de 40 ans	-0,98
Diplôme < Bac	1,14
Bac et plus	-0,46
Constante	0,02

PCS1 = agriculteurs, artisans-commerçants-chefs d'entreprise, cadres supérieurs et professions libérales supérieures, retraités ;
PCS2 = professions intermédiaires et autres inactifs ;
PCS3 = employés et ouvriers.

On voit que les effets vont dans les mêmes sens que ceux indiqués par la segmentation. Le rôle du chômage apparaît cependant comparativement plus faible : il semble ici jouer un rôle équivalent à celui de l'âge ou du diplôme, alors qu'il jouait un rôle plus important dans l'analyse détaillée.

3.3.2 La régression logistique

La régression logistique utilise les mêmes variables que l'analyse factorielle discriminante, et les deux mêmes groupes responsabilité entreprise/responsabilité demandeurs d'emploi. Contrairement à l'analyse discriminante sur coordonnées factorielles, la régression logistique rend nécessaire la définition pour chaque variable d'une modalité de référence par rapport à laquelle les autres modalités sont comparées. On a choisi les personnes n'ayant jamais été au chômage, les autres inactifs et femmes au foyer, les 60 ans et plus, les diplômés d'études supérieures.

Rappelons enfin que plus l'effet d'une modalité est fort par rapport à la situation de référence, plus le coefficient est élevé. La probabilité dans le tableau suivant est associée au test de non nullité du coefficient : quand elle est supérieure à 5%, le coefficient n'est pas significativement non nul.

Variable	Coefficient	Probabilité
Une fois au chômage	0,29	0,03
Deux fois au chômage	0,18	0,38
Trois fois ou plus au chôm.	0,68	0,01
Recherche 1er emploi	0,46	0,41
Agriculteur	-0,62	0,01
Artisan-Comm-Chef d'entr.	-0,47	0,09
Cadre sup-Prof.libér.	-0,48	0,06
Prof.intermédiaire	0,11	0,55
Employé	0,27	0,11
Personnel de service	0,05	0,87
Ouvrier	0,41	0,02
Ouvrier agricole	-0,18	0,73
Retraité	-0,30	0,13
Moins de 24 ans	0,74	0,01
25-40 ans	0,65	0,01
40-60 ans	0,32	0,10
Aucun diplôme-CEP	0,47	0,01
Diplôme niveau BEPC	0,38	0,01
Bac ou équivalent	0,12	0,50
Constante	-0,67	0,01
Bien classés : 64,9% Mal classés : 32,9% Eliminés : 2,2%		

Les modalités dont les coefficients ont les plus grandes valeurs absolues correspondent à l'âge et à la fréquence des situations de chômage passées. On voit donc que les résultats ne correspondent pas exactement à ceux de l'analyse factorielle discriminante.

Ceux-ci sont d'autre part relativement plus réguliers et conformes aux tableaux croisés initiaux : les personnes ayant été au chômage trois fois ou plus, ou ceux qui recherchent un premier emploi ont plus tendance à citer les entreprises comme responsables que ceux qui n'ont été au chômage qu'une fois ou deux.

Les effets de l'âge et du diplôme sont réguliers : plus on est jeune et moins on est diplômé, plus on attribue la responsabilité des difficultés d'embauche aux entreprises. L'hypothèse initiale de la neutralisation d'effets d'âge et de diplôme opposés semble confirmée par cette méthode, alors qu'elle était infirmée par l'analyse discriminante[11].

3.4 Quand on regroupe les modalités, les résultats obtenus par les deux méthodes sont plus comparables

L'adoption du même recodage que celui effectué dans le cas de l'analyse factorielle discriminante nous permet de comparer les deux méthodes de discrimination.

Pour chaque variable explicative, et comme nous l'avons déjà signalé, la régression logistique de SAS ne fournit pas, par convention, des coefficients pour chaque modalité, contrairement à l'analyse factorielle discriminante. Nous avons éliminé les modalités de référence pour la régression logistique de la fonction discriminante de Fisher en leur substituant leur combinaison linéaire fonction des autres modalités.

[11] Certaines colinéarités entre modalités (plus de 60 ans et retraités par exemple) peuvent induire des résultats instables, c'est une des raisons pour lesquelles nous allons effectuer des regroupements.

Modalité	Coefficient de l'analyse factorielle discriminante = cd	Coefficient de la régression logistique = cr	cd/cr
A déjà été au chômage	1,37	0,37	3,7
PCS1	-1,62	-0,79	2,0
PCS2	-0,58	-0,22	2,6
Age < 40	1,87	0,40	4,7
Diplôme inférieur au Bac	1,60	0,34	4,7
Constante	-1,68	-0,05	-

Rappel :
PCS1 = agriculteurs, artisans-commerçants-chefs d'entreprise, cadres supérieurs et professions libérales, retraités;
PCS2 = professions intermédiaires et autres inactifs.

Le rapport entre le coefficient d'une modalité dans la fonction discriminante calculée par l'analyse factorielle discriminante et le coefficient de cette même modalité dans la fonction de score calculée par la régression logistique est approximativement constant, et les signes des coefficients sont identiques : **les résultats obtenus sont proches**.

Si les résultats globaux sont donc cohérents entre les deux méthodes, il n'en reste pas moins que pour certaines modalités de variables (personnes non diplômées ou munies du seul CEP par exemple), les effets sont divergents selon les méthodes utilisées. Nous avons donc effectué une dernière série d'investigations sur plusieurs sous-populations, de façon à contourner l'hypothèse d'additivité, peut-être trop forte.

5. Analyses discriminantes filtrées sur différentes catégories socio-démographiques

Les résultats des deux méthodes, analyse discriminante et régression logistique, sont donc globalement identiques, mais différents dans le détail des effets de chaque modalité de chaque variable. Pour mieux comprendre les résultats peu réguliers de l'analyse discriminante et améliorer éventuellement le pourcentage d'individus bien classés, des analyses discriminantes ont été appliquées à des sous-populations. La segmentation avait en effet montré que les effets des variables socio-démographiques n'étaient pas parfaitement additifs et qu'il peut être utile de distinguer la sous-population ayant vécu une période de chômage (chômeurs et anciens chômeurs) du reste de la population.

On a donc réalisé deux analyses discriminantes : une sur la sous-population n'ayant jamais été au chômage, l'autre sur la sous-population ayant déjà connu le chômage. En effet, il était apparu lors de la segmentation que, dans le groupe des personnes "à risque vis-à-vis du chômage", le vécu face au chômage était la variable la plus explicative.

L'amélioration recherchée n'a pas été obtenue : la probabilité de bien classer les individus, pour l'une comme pour l'autre des fonctions discriminantes, avoisine 0,6, comme dans le cas de la fonction discriminante calculée sur l'ensemble de la population.

Par ailleurs, les coefficients des modalités explicatives sont proches de ceux qui avaient été précédemment calculés sur l'ensemble de la population.

Les fonctions discriminantes obtenues sont les suivantes :

- population n'ayant jamais été au chômage :

Pourcentages de classement dans chacun des groupes et au total

	Bien classés	Mal classés	Total
1er groupe	54,3	45,7	100,0
2ème groupe	61,2	38,8	100,0
Ensemble	**57,8**	**42,2**	100,0

Fonction linéaire de Fisher reconstituée à partir des variables d'origine

Variable	Coefficient
PCS1	-0,33
PCS2	-0,03
PCS3	0,28
Moins de 24 ans	0,09
25-40 ans	0,15
40-60 ans	0,01
Plus de 60 ans	-0,29
Aucun diplôme-CEP	-0,27
Diplôme < Bac	0,30
Bac ou équivalent	0,07
Diplôme supérieur	-0,10
Constante	-0,00

- population ayant déjà été au chômage au cours des dix années précédant l'enquête :

Pourcentages de classement dans chacun des groupes et au total

	Bien classés	Mal classés	Total
1er groupe	71,5	28,5	100,0
2ème groupe	41,0	59,0	100,0
Ensemble	**61,8**	**38,2**	100,0

Fonction linéaire de Fisher reconstituée à partir des variables d'origine

Variable	Coefficient
Une fois au chômage	-0,30
Deux fois au chômage	0,07
Trois fois ou plus au chom.	0,09
Recherche 1er emploi	-0,01
PCS1	-0,34
PCS2	-0,13
PCS3	0,98
Moins de 24 ans	0,15
25-40 ans	0,04
40-60 ans	-0,01
Plus de 60 ans	-0,14
Aucun diplôme-CEP	0,04
Diplôme < Bac	0,43
Bac ou équivalent	-0,05
Diplôme supérieur	-0,24
Constante	0,03

D'autres tentatives, à partir du recodage de la PCS en trois groupes, et d'analyses discriminantes distinctes au sein de chacun de ces groupes, n'ont pas donné de meilleurs résultats.

L'examen détaillé des probabilités d'affectation, pour chaque individu, révèle que, dans tous les cas cités ci-dessus, une forte proportion d'entre elles est voisine de 0,5. Ceci signifie, par exemple, que nombre d'individus ont été affectés au groupe 1, qui était considéré comme le plus probable pour eux, avec une probabilité à peine supérieure à 0,5 : cette affectation serait donc instable, surtout dans le cas de variables qualitatives où, les individus étant regroupés sur un petit nombre de modalités, une légère variation de la probabilité de classement a priori peut transformer notablement le résultat final. En résumé, le groupe de départ, composé de 2010 individus, est peut-être trop homogène par rapport à nos variables explicatives. Une solution consisterait alors à extraire de la population de départ un sous-groupe plus typé, afin de construire sur ce sous-groupe une fonction discriminante qu'on appliquerait par la suite à l'ensemble de la population.

Cette méthode d'amélioration de la discrimination par le biais d'un sous-groupe typé a fait l'objet de la dernière investigation décrite en annexe de cette note. Mais elle n'a pas permis d'améliorer de façon notable la discrimination, ni d'obtenir des coefficients de la fonction de Fisher relativement réguliers.

CONCLUSION

Des différentes investigations décrites dans cette note, trois enseignements principaux peuvent être tirés :

- Quand on s'est restreint à des modalités de classe suffisamment importantes en effectif et suffisamment homogènes (une segmentation permet de définir ces modalités regroupées), l'analyse discriminante sur coordonnées factorielles a donné les mêmes résultats que la régression logistique.

- Dans le cas étudié, où les effets des variables socio-démographiques sur la variable d'opinion étudiée ne sont pas purement additifs et où ces variables explicatives ne sont pas très nombreuses, la régression logistique a donné des résultats plutôt plus réguliers et un pourcentage d'individus bien classés, légèrement supérieur.

- Dans une méthode comme dans l'autre, il est imprudent d'interpréter le signe ou la valeur d'un coefficient de régression obtenu pour une modalité isolée.

Cette étude comparative a été réalisée assez rapidement. Il aurait été intéressant de faire intervenir un plus grand nombre de variables explicatives pour améliorer la normalité des composantes factorielles utilisées par l'analyse discriminante. Il serait aussi intéressant d'observer la variation des résultats obtenus en diminuant ou augmentant la taille de l'échantillon. L'application des mêmes méthodes à d'autres variables d'opinion pour lesquels des effets additifs des caractéristiques socio-démographiques sont plus nets serait à envisager.

Bibliographie

. **AGRESTI (A)**, *"Categorial data analysis"*, Florida Wily and Sons, 1989.

. **BOUROCHE (J.M) et SAPORTA (G)**, *"L'analyse des données"*, collection Que sais-je ?

. **COX (D.R)**, *"Analyse des données binaires"*, Dunod, Paris, 1972.

. **DEPARDIEU (D), et LOLLIVIER (S)**, *"Les facteurs de l'absentéisme"*. Encadré : *"Le modèle logit et les comparaisons de taux de pratique entre deux sous-populations"*, Economie et Statistiques, N° 176, avril 1985, pp. 17 et 20.

. **EFRON (B)**, *"The efficiency of logistic regression compared to normal discriminant analysis"*, JASA 70, pp 892-898, 1975.

. **GOURIEROUX (C)**, *"Econométrie des variables qualitatives"*, Ed. Economica.

. **INED**, *"L'utilisation de la régression logistique dans les enquêtes"*. Compte-rendu de la séance du séminaire de méthode d'enquêtes de l'INED du 16 avril 1991.

. **LEBART (L)**, *"Sept ans de perceptions - Evolution et structure des opinions en France de 1978 à 1984"*, CREDOC, 1986.

. **LION (S)**, *"Classification dichotomique descendante"*, CREDOC, Cahier de Recherche, N° 16, 1991.

. **PRESS (J.J) and WILSON (S)**, *"Choosing between logistic regression and discriminant analysis"*, JASA 73, pp 699-705, 1978.

. **SAS, version 6**, *"Manuel de référence"*, procédure LOGISTIC.

. **SAPORTA (G)**, *"Probabilités, analyse des données et statistiques"*, Ed. Technip, 1990.

. **SPAD.N**, *"Manuel de référence"*, procédure DIS 2 G.

. **VERGER (D)**, *"L'achat d'un logement ne va pas sans achat d'équipement"*. Annexe méthodologique : *"Le modèle logit."*, Economie et Statistiques, N° 161, décembre 1983.

. **VOLLE (M)**, *"L'analyse des données"*, Ed. Economica.

Annexe

Une amélioration de la discrimination par un meilleur choix de la population de départ

On réalise en premier lieu cette tentative d'amélioration sans tenir compte de la segmentation. Rappelons que les deux groupes de départ avaient été obtenus par le recodage de la variable suivante :

"Les entreprises déclarent rencontrer de plus en plus de difficultés pour recruter les personnes dont elles ont besoin, notamment pour des emplois qualifiés. Selon vous, quelle en est la raison principale ?" :

1. Les entreprises sont trop exigeantes
2. Il n'y a pas assez de personnes qualifiées ou compétentes
3. Les salaires proposés sont insuffisants
4. Les conditions de travail offertes sont pénibles
5. Les entreprises ne veulent pas payer la formation nécessaire
6. Les demandeurs d'emploi sont trop exigeants
7. Les emplois proposés sont souvent trop éloignés du domicile.

Le recodage avait été effectué comme suit : les modalités faisant appel à la responsabilité des entreprises (modalités 1, 3, 4 et 5) ensemble, celles faisant appel à la responsabilité des demandeurs d'emploi (modalités 2 et 6) constituant l'autre groupe ; la modalité 7 restante avait été affectée au groupe faisant appel à la responsabilité des chômeurs.

La sélection d'un sous-groupe plus typé s'est faite en choisissant moins de modalités dans le recodage de départ. Dans le groupe 1 (responsabilité entreprises), nous avons retenu les modalités 1 et 5 ; dans le groupe 2 (responsabilité chômeurs), les modalités 2 et 6. On fait l'hypothèse que ce choix conduira à une population mieux scindée entre groupe 1 et groupe 2 que la population de départ. On se retrouve ainsi avec 1582 individus, au lieu de 2010 initialement. C'est sur cette population de 1582 personnes qu'on va élaborer une fonction discriminante.

1. Une fonction discriminante sur une sous-population typée.

Dans cette étape, on ne fait pas appel à la segmentation. On réalise donc une analyse factorielle discriminante à partir de nos 1582 individus qu'on espère mieux typés. On introduit au préalable la probabilité a priori d'appartenance au groupe 1, qui est, sur cette sous-population, de 45% (ceci signifie simplement que 45% de notre population se trouvent initialement dans le groupe 1, et qu'en dehors de toute considération sur les variables explicatives, un individu a a priori 45 chances sur 100 de se retrouver dans le groupe 1).

Les classements obtenus sont alors les suivants :

Pourcentages de classement dans chacun des groupes et au total

	Bien classés	Mal classés	Total
1er groupe	56,0	44,0	100,0
2ème groupe	69,0	31,0	100,0
Total	**63,1**	**36,9**	100,0

La proportion totale d'individus bien classés n'est que légèrement améliorée par rapport aux analyses initiales, mais c'est la première fois que les pourcentages de bon classement sont supérieurs à 50% dans les deux groupes : ceci permet de supposer que l'analyse ainsi réalisée est d'une qualité légèrement meilleure.

On précise ce jugement en observant les probabilités d'affectation de chaque individu, et l'histogramme de répartition des individus dans les groupes 1 et 2 sur l'axe discriminant.

Les probabilités d'affectation sont plus grandes que dans les cas précédents (globalement comprises entre 0,6 et 0,8) : la sous-population choisie est donc plus typée que la population totale, comme on l'avait désiré.

L'histogramme de répartition sur l'axe discriminant (cf ci-contre) permet de calculer, pour chaque positionnement d'individus sur l'axe, le pourcentage d'individus dans le groupe 1 et le pourcentage d'individus dans le groupe 2, en fonction de leur coordonnée sur l'axe discriminant :

Coordonnée	% de population dans le groupe 1	% de population dans le groupe 2
-1,3	17	83
-0,9	33	67
-0,4	44	56
0	50	50
0,5	58	42
0,8	70	30

Les pourcentages sont "bien classés", c'est-à-dire que les individus sont de moins en moins nombreux dans le groupe 1 à mesure que leur coordonnée s'éloigne du domaine négatif. La fonction discriminante calculée donne donc une répartition convenable. L'amélioration obtenue par rapport aux premières fonctions discriminantes est faible, mais sensible.

2. Utilisation de la sous-population typée et de la segmentation.

On sait maintenant que la sous-population de 1582 individus, sur laquelle on a travaillé ci-dessus, est relativement bien typée par rapport à la population totale. On va donc à nouveau l'utiliser, en tenant compte de la segmentation réalisée sur la population totale (on fait l'hypothèse que la segmentation fournirait les mêmes résultats sur cette sous-population, à la première étape au moins). Il s'agit d'utiliser la première étape de la segmentation, qui montre que la population se partitionne en trois classes en fonction de la PCS. On réalise ensuite trois analyses discriminantes, une par classe de PCS.

Ainsi, chaque sous-population sur laquelle nous travaillons ici n'a pas les mêmes probabilités d'affectation a priori que la population totale.

	PCS1	PCS2	PCS3	Total
Groupe 1	31%	46%	61%	45%
Groupe 2	69%	54%	39%	55%

Pour la classe correspondant à la PCS1 (agriculteurs, artisans-commerçants-chefs d'entreprise, cadres supérieurs et professions intellectuelles supérieures, retraités), la probabilité, a priori d'appartenance au groupe 1, est donc de 0,31 ; pour la PCS2 (professions intermédiaires et autres inactifs), elle est de 0,46 ; et pour la PCS3 (employés, ouvriers, ouvriers agricoles, personnels de service), elle est de 0,61.

2.1. La PCS1

Pour la PCS1, on a vu que la probabilité a priori est de 0,31. L'analyse discriminante fournit les résultats suivants :

Variable	Coefficient
Déjà été au chômage	0,11
Jamais été au chômage	-1,01
Moins de 40 ans	0,12
Plus de 40 ans	-0,74
Diplôme < Bac	1,59
Bac et plus	-0,53
Constante	-0,02

Les pourcentages de bon et mauvais classement sont à peu près les mêmes que dans l'ensemble de la population (65% de bons classements).

Dans le cas des PCS agriculteurs, artisans-commerçants-chefs d'entreprise, cadres supérieurs et professions intellectuelles supérieures, retraités, le modèle explicatif fait intervenir en premier lieu le fait d'avoir un diplôme inférieur au bac et de n'avoir jamais été au chômage.

2.2. La PCS2

Pour la PCS2, on a vu que la probabilité a priori est de 0,46. L'analyse discriminante fournit les résultats suivants :

Variable	Coefficient
Déjà été au chômage	0,08
Jamais été au chômage	-0,26
Moins de 40 ans	1,10
Plus de 40 ans	-0,80
Diplôme < Bac	0,68
Bac et plus	-0,56
Constante	-0,01

Ici encore, la proportion d'individus bien classés est d'environ 60%. Les modalités les plus explicatives sont "Moins de 40 ans" et "Plus de 40 ans", c'est donc l'âge qui intervient le plus dans les catégories professions intermédiaires-autres inactifs.

2.3. La PCS3

Pour la PCS3, on a vu que la probabilité a priori est de 0,61. L'analyse discriminante fournit les résultats suivants :

Variable	Coefficient
Déjà été au chômage	0,75
Jamais été au chômage	-0,76
Moins de 40 ans	0,49
Plus de 40 ans	-0,20
Diplôme < Bac	0,73
Bac et plus	-0,14
Constante	0,01

Les modalités les plus explicatives concernent le vécu du chômage au cours des dix années précédant l'enquête. Il s'agit ici des PCS ouvriers-employés-ouvriers agricoles.

Si cette dernière investigation produit des résultats assez proches de la segmentation initiale, les pourcentages de bons classements restent aux alentours de 60%. Les analyses factorielles discriminantes séparées ne donnent pas de meilleur résultat que l'analyse globale.

CAHIER DE RECHERCHE

Récemment parus :

Entre école et emploi : les transitions incertaines, par Denise Bauer, Patrick Dubéchot, Michel Legros, N° 19, Septembre 1991.

Price expectations of french households : A test on INSEE panel data (1972 - 1988), par François Gardes, Jean-Loup Madre, N° 20, Octobre 1991.

Chômeurs au fil du temps, par Isa Aldeghi, N° 21, Novembre 1991.

Deux analyses lexicales : Les améliorations à apporter au fonctionnement de la société - L'image du milieu professionnel, Enquête "Conditions de vie et Aspirations des Français", par Laurent Clerc, Ariane Dufour, N° 22, Janvier 1992.

La codification des objets complexes : réflexions théoriques et application à un corpus de 8 000 produits alimentaires, par Saadi Lahlou, Joelle Maffre, Valérie Beaudouin, N°23 Décembre 1991.

Nature et traitement statistique des données textuelles : réflexions méthodologiques, par Anne-Lise Aucouturier, Valérie Beaudouin, Isabelle Blot, Didier Faivre, Saadi Lahlou, Julie Micheau. N° 24, Décembre 1991.

Président : Bernard SCHAEFER Directeur : Robert ROCHEFORT
142, rue du Chevaleret, 75013 PARIS - Tél. : (1) 40.77.85.00

CRéDOC

Centre de recherche pour l'Étude et l'Observation des Conditions de Vie