La « complexité » du social. Quelques réflexions sur l'usage de l'analyse des données symboliques en sociologie.

Frédéric Lebaron
Laboratoire Professions-Institutions-Temporalités (UMR 8085)
Université de Versailles-Saint-Quentin-en-Yvelines / Université Paris-Saclay
France
E-mail: frederic.lebaron@uvsq.fr

1. Introduction

Le thème de la « complexité » du monde social est présent depuis longtemps dans la réflexion méthodologique et épistémologique sur les spécificités des sciences sociales (par exemple Simiand, 1922)[1]. Passage obligé des manuels ou « pont-aux-ânes » d'un discours essayiste plus ou moins vague, on peut aussi le considérer comme un enjeu réel et important si l'on souhaite, comme c'est à l'ordre du jour, faire progresser le dialogue entre les sciences de la nature et les sciences sociales.

Ce thème met en jeu la notion-même de *modélisation* et ses limites lorsqu'elle est transférée de façon plus ou moins mécanique de la physique, la chimie ou la biologie vers le « social », au nom de l'efficacité éprouvée de la mathématisation de ces disciplines.

Dans le présent article, nous commençons par évoquer diverses stratégies développées dans les recherches en sciences sociales, en particulier depuis l'avènement de la « statistique des chercheurs » (Rouanet et al., 2008), pour traiter de façon rigoureuse cette « complexité », qui peut être définie de diverses manières (section 2). Cela nous permet d'aborder, dans une troisième section, quelques-unes des caractéristiques de l'analyse des données symboliques (ADS), qui en font un domaine de recherches stimulant dans la perspective de l'étude de phénomènes complexes en sciences sociales. Une étude de cas exploratoire, fondée sur l'exploitation des données européennes EU-SILC 2013 « en coupe », est présentée et discutée, à cette fin. Elle nous conduit à évoquer en conclusion un programme de recherche à la fois méthodologique et empirique auquel le recours à l'analyse des données symboliques peut contribuer.

2. La sociologie face à la « complexité »

Après avoir évoqué la notion fondatrice de « fait social total » pour partir d'une définition minimale préalable, on présente ensuite deux familles de réponses méthodologiques à l'enjeu de la complexité dans les sciences sociales, puis leurs évolutions récentes et leurs manifestations dans la méthodologie en sciences sociales, avant d'évoquer plus particulièrement la question des « échelles » dans l'analyse des faits sociaux.

2.1. Le fait social total, paradigme de la « complexité »

Par « complexité » du réel, on entend ici une réalité à première vue triviale : les faits sociaux sont toujours constitués d'un ensemble composite de réalités diverses simultanément présentes, qui sont situées sur différents « plans » qu'il est difficile, voire artificiel, de séparer par la pensée les uns des autres. Cette idée peut être référée à celle de « fait social total », développée par le sociologue et

[1] François Simiand écrit : « pour mériter ce nom de fait scientifique, pour entrer dans la science, il faut que toute cette abstraction, tout en se distinguant de la *complexité* concrète, se modèle cependant suffisamment sur elle, respecte, comme l'a dit un philosophe contemporain, les articulations de la réalité » (Simiand, 1922, p. 29).

anthropologue Marcel Mauss (Mauss, 1923) : toute réalité sociale est à la fois psychique et institutionnelle, individuelle et collective, économique, politique et juridique, matérielle et idéelle, locale et globale, etc. Face à cette caractéristique, les pensées du « social » n'ont cessé, depuis les origines, de découper les faits sociaux selon divers principes, et d'opposer telles ou telles « dimensions » (« secteurs », « domaines », « plans », etc.) entre elles. Les débats sans fin suscités par ces « découpages » ont nourri la pensée du social et ses controverses.

Avec l'avènement de la statistique sociale, au XVIIIème et surtout au XIXème siècle, émerge un nouveau domaine scientifique, au double sens d'ensemble de données empiriques et de méthodes, qui modifie sensiblement l'appréhension de ce problème. On doit ainsi à cette évolution une reformulation implicite du thème de la « complexité » en sciences sociales, et l'apparition de « solutions » particulières de natures assez diverses[2].

2.2. Deux familles de réponses

Deux grandes lignes de « réponse » émergent et se développent avec l'avènement de la « statistique des chercheurs » en sciences sociales : l'une peut être qualifiée de « réductionniste » et l'autre d' « englobante ». L'approche « réductionniste », qui a connu l'écho le plus fort en économie, est souvent associée, encore aujourd'hui, à l'idée courante de « modélisation » (sous-entendu : *a priori* et aussi, le plus souvent, *formelle*). Il s'agit de représenter le phénomène étudié en le réduisant au jeu, déterminé par le modèle déduit de la théorie et de l'état des connaissances, de quelques forces ou mécanismes générateurs bien identifiés (voir Bressoux, 2008). L'approche « englobante » étudie plutôt de la façon la plus « exhaustive » possible les systèmes de relation entre phénomènes tels qu'ils sont observables dans la réalité, sans préjuger du jeu de certaines forces plutôt que d'autres, avec l'idée fondamentale que c'est par l'interprétation analytique des données dans toute leur étendue que se dégagera progressivement un modèle, en quelque sorte *a posteriori* (pour l'économie dans la tradition de Benzécri, voir par exemple : Desbois, 2009). Dans ce deuxième cas, cependant, un « modèle-cadre » préside toujours au recueil initial de l'information, sans lequel aucune structure ne peut émerger du réel (voir Le Roux, Lebaron, 2015).

2.3. L'approfondissement contemporain des deux démarches, « réductionniste » et « englobante »

Ces deux approches, très fécondes, n'ont cessé de se développer et de se consolider depuis lors. Du côté de l'approche modélisatrice, la science économique, mais également la démographie, la science politique ou la sociologie –en tout cas dans leurs variantes « individualiste méthodologique » ou « analytique » - sont allés très loin dans la tentative de représenter de façon simplifiée les principaux mécanismes de la réalité étudiée, afin de tester l'existence des mécanismes identifiés. Les méthodes de régression, dans le cadre du modèle linéaire général, ont permis de mener à bien la réalisation de ce programme sur le plan plus proprement statistique. Mais, en cherchant à intégrer de plus en plus de variables, les usages de certaines méthodes associées à la démarche « réductionniste », au départ conçues selon des normes de parcimonie, en sont venus de plus en plus à se rapprocher de la visée « exhaustive » associée aux méthodes plus « englobantes ». Cela au point que, par exemple, certains spécialistes de l'économétrie des séries temporelles défendent aujourd'hui explicitement une approche précisément dite elle-aussi « englobante » (voir Meuriot, 2015, suivant sur ce point l'économiste britannique David Hendry).

La démarche « modélisante » continue aussi de se diversifier et de s'affiner pour répondre aux limites des modèles antérieurs, par exemple avec la sophistication des modèles dans l'analyse des

[2] Parmi celles-ci, la perspective durkheimienne, qui émerge dans la dernière décennie du dix-neuvième siècle, se caractérise par la mise en avant d'une conception très générale (« englobante ») du « fait social », qui intègre par exemple les faits économiques, juridiques, politiques comme autant de manifestations d'une réalité fondamentale, la tâche du sociologue étant d'étudier les relations entre ces différents phénomènes mais non de les séparer par une décision théorique arbitraire (Durkheim, 1895).

chroniques et les voies développées en finance pour sortir de l'hégémonie de représentations inadéquates (Lévy-Vehel et Walter, 2002, Walter, 2013).

La voie « réductionniste » stricte continue aussi bien sûr d'être explorée et appliquée de façon plus classique, en économie mais aussi en sociologie : les méthodes de simulation « multi-agents » prolongent par exemple, dans la sociologie contemporaine, la perspective plus clairement « analytique » ou « déductive » illustrée par les recherches de R.Boudon (1975).

De l'autre côté, celui d'approches couramment qualifiées d' « inductives », on observe également une forme d'approfondissement et de diversification. La naissance de l'analyse géométrique des données autour de Jean-Paul Benzécri a contribué au succès large des méthodes dites « exploratoires » et « multidimensionnelles », très souvent combinées avec les méthodes de classification (Benzécri, 1973). Dans cette tradition aussi, les innovations ont été nombreuses, et elles ont contribué à enrichir l'approche « englobante », sans exclure des tentatives pour se rapprocher de l'approche plus « parcimonieuse » associée à l'usage quasi-exclusif des méthodes de régression dans une grande partie de la littérature des sciences sociales. On peut ici parler de modélisation géométrique, dont l'idée est de rester au plus près de la complexité des phénomènes sociaux (voir Lebaron, Le Roux, 2015). De plus en plus, cette démarche est appliquée à des données elles-mêmes complexes, prenant en compte le temps et l'espace géographique, présentant des formats divers, situées sur des échelles différentes. La construction d'espaces multidimensionnels permettant la représentation géométrique de la réalité en est à la fois l'instrument et l'objectif, et constitue un objet-carrefour pour la méthodologie en sciences sociales.

2.4. Questions d'échelles et d'hétérogénéités

L'un des enjeux contemporains en méthodologie statistique des sciences sociales est la prise en compte des différentes échelles et, plus largement, des hétérogénéités de toutes sortes dans l'analyse des faits sociaux. Du côté des méthodes de modélisation *a priori*, cela conduit en particulier à l'affirmation des méthodes « multi-niveaux », plus aptes à prendre en compte l'emboîtement des individus dans des structures (unités géographiques en particulier) elles-mêmes emboîtées les unes dans les autres, jusqu'à un niveau très global (Courgeau, 2004).

Du côté des méthodes d'analyse « englobante », l'analyse des données symboliques (cf. Bock et Diday 2000, Billard et Diday 2006, Diday et Noirhomme 2008) est l'une des innovations visant à analyser simultanément des données de niveaux différents, hétérogènes par nature (puisque les objets analysés, appelés « concepts », peuvent être eux-mêmes composés d'histogrammes, d'échelles, etc.), en procédant à un codage approprié de cette information hétérogène. On peut également évoquer l'analyse des interactions spatiales qui enrichit la perspective de « modélisation » de données complexes (Pumain, Saint-Julien, 2010).

3. L'analyse des données symboliques en sciences sociales : un exemple

Après avoir rappelé quelle est la perspective de l'analyse des données symboliques, on présente brièvement une étude de cas menée sur des données européennes. Les premiers résultats obtenus font ensuite l'objet d'un commentaire sociologique.

3.1. L'analyse des données symboliques

L'analyse des données symboliques est, ainsi que le rappellent Verde et Diday (2014), « une branche de l'analyse des données qui développe des techniques exploratoires afin traiter de données « symboliques », c'est-à-dire ici de variables prenant leurs valeurs sous la forme d'intervalles, de multi-catégories, ou d'ensemble de catégories auxquelles est associé un « mode », qui peut être une probabilité, une fréquence, un poids ».

La démarche de ce type d'analyse s'apparente évidemment à celle des méthodes d'analyse géométrique de données (Le Roux, Lebaron, 2015), les originalités de l'ADS en tant que méthode étant le codage de variables initialement hétérogènes, et le caractère « multi-échelle » explicitement

revendiqué, qui conduisent à représenter l'hétérogénéité de façon géométrique et sous la forme d'un résumé dans les axes principaux.

Du point de vue méthodologique général, on peut dire que le codage de variables hétérogènes relève de la construction d'un tableau et donc d'une métrique pertinents, alors que le caractère « multi-échelle » de l'ADS la rapproche de la démarche de l'analyse des données structurées.

3.2. Une étude de cas exploratoire sur données européennes

Nous suivrons assez étroitement la démarche proposée par Alonso et Iaaksonen (2015) à propos des données de l'*European Social Survey*. Dans notre cas, les données retenues sont des variables catégorisées, représentées en ADS sous la forme d'*histogrammes*. Les variables numériques présentes dans notre base de données ont été, provisoirement, laissées de côté.

3.2.1. Les données

L'enquête *European-Union Statistics on Income and Living Conditions* (EU-SILC)[3] est une enquête européenne annuelle commanditée par Eurostat aux instituts statistiques nationaux depuis 2004[4][5]. Elle porte sur les conditions de vie aux niveaux des ménages et des individus, et comporte des variables socio-démographiques (âge, sexe, région...), des variables sur le logement (type de logement, situation, statut d'habitation...), sur la situation professionnelle au dernier emploi (revenus, type de contrat...), sur la consommation (possession de certains biens tels qu'un ordinateur, une voiture...), les revenus.

La base sur laquelle nous travaillons, celle de l'enquête EU-SILC 2013 en coupe, comporte 362 301 individus issus de 32 pays (dont les 28 pays de l'UE), après sélection des seuls individus déclarant une profession au sens de l'ISCO/CITP 2008 [6] et après élimination des individus présentant une non-réponse au moins à l'une des dix variables d'intérêt retenues, parmi lesquelles seront choisies les variables actives de l'analyse géométrique (voir plus loin la liste des variables étudiées).

3.2.2. Agrégation des micro-données et construction des concepts

Pour construire les concepts de l'analyse des données symboliques, on a choisi de croiser deux facteurs, le pays (n=32) et le groupe social défini par l'ISCO/CITP au niveau 1 (9 modalités).

ISCO 08 Code	Titre FR
1	Directeurs, cadres de direction et gérants
2	Professions intellectuelles et scientifiques
3	Professions intermédiaires
4	Employés de type administratif
5	Personnel des services directs aux particuliers, commerçants et vendeurs
6	Agriculteurs et ouvriers qualifiés de l'agriculture, de la sylviculture et de la pêche
7	Métiers qualifiés de l'industrie et de l'artisanat
8	Conducteurs d'installations et de machines, et ouvriers de l'assemblage
9	Professions élémentaires

[3] Alain Desrosières, Laurent Thévenot. *Les catégories socio-professionnelles*, Paris, La Découverte, 2000 ; sur le dispositif EU-SILC, *Economie et Statistique*, 469-470, juillet 2014.

[4] Les données EU-SILC ont été obtenues auprès d'Eurostat dans le cadre du laboratoire Printemps.

[5] Une rupture de série a donc lieu entre 2001 et 2004, avec un changement d'enquête et l'absence de données en 2002 et 2003.

[6] La population correspond donc aux actifs et à ceux des inactifs ayant déclaré une profession. On a, par ailleurs choisi de faire une analyse équi-pondérée sur les individus.

Les pays sont l'Allemagne (DE), l'Autriche (AT), la Belgique (BE), la Bulgarie (BG), Chypre (CY), la Croatie (HR), le Danemark (DK), l'Espagne (ES), l'Estonie (EE), la Grèce (EL), la Hongrie (HU), l'Irlande (IE), l'Islande (IS), l'Italie (IT), la Finlande (FI), la France (FR), la Lettonie (LV), la Lituanie (LT), le Luxembourg (LU), Malte (MT), la Norvège (NO), les Pays-Bas (NL), la Pologne (PL), le Portugal (PT), la république tchèque (CZ), le Royaume-Uni (UK), la Roumanie (RO), la Serbie (RS), Slovaquie (SK), la Slovénie (SI), la Suède (SE), la Suisse (CH).

3.2.3. Etude descriptive élémentaire des variables et visualisation sous forme histogrammes

Reprenant le schéma d'une précédente étude (Lebaron, 2015), on a ensuite sélectionné des variables en retenant principalement quatre thèmes d'intérêt:

- conditions économiques et exclusion sociale ;
- logement ;
- environnement matériel et sécurité physique ;
- santé.

Les quatre rubriques retenues ont été choisies pour rendre compte de façon synthétique du caractère multidimensionnel des conditions de vie concrètes. Elles expriment autant de type de capitaux: les ressources économiques, celles liées au logement, qui renvoient à un aspect concret du capital économique, l'environnement social, avec des questions sur le cadre de vie et la sécurité physique, et qui mesurent aussi des ressources liées au contexte immédiat, et, enfin, la santé, soit un capital physique, en partie « biologique ». On a donc ici une synthèse simplifiée des principales dimensions considérées comme fondamentales dans le rapport Stiglitz-Sen-Fitoussi (2009). Les données sont toutefois très fortement centrées sur l'exclusion sociale et les privations matérielles les plus fortes.

Variables socio-démographiques générales de l'individu

Sexe (RB090) : homme (1) / femme (2)

Statut marital (PB190) : jamais marié (1) / marié (2) / séparé (3) / veuf (4) / divorcé (5)

Union consensuelle (PB200) : oui sur une base légale (1) / oui sans base légale (2) / non (3)

Pays de naissance (PB210) : code spécifique

Citoyenneté (PB220A) : code spécifique

NACE (PL111) : code de la branche

Catégorie CITP (ISCO : voir supra).

Type de contrat (PL140) : emploi permanent ou CDI (1) / emploi temporaire ou CDD (2)

Position managériale (PL150) : supervision (1) / pas de supervision (2)

Degré d'urbanisation (DB100) : région densément peuplée (1) / région intermédiaire (2) / région faiblement peuplée (3)

Plus haut niveau de diplôme ISCED atteint (PE040) : éducation pré-primaire (1) / éducation secondaire basse (2) / éducation secondaire haute (3) / éducation post-secondaire non-tertiaire (4) / premier niveau de l'éducation tertiaire – ne conduisant pas à une qualification avancée en recherche (5) / deuxième niveau de l'éducation tertiaire – conduisant à une qualification avancée en recherche (6)

Conditions économiques, précarité et exclusion des ménages

Retards de paiement sur des factures de services dans les douze derniers mois (HS021) : oui une fois (1) / oui deux fois ou plus (2) / non (3)

Retards de paiement sur des factures d'installations en location (hire purchase) ou autres prêts (HS031) : oui une fois (1) / oui deux fois ou plus (2) / non (3)

Capacité à s'offrir une semaine annuelle de vacances loin de chez soi (HS040) : oui (1) / non (2). Cette question est de même nature que la précédente, avec une dimension sans doute encore plus dépendante de la perception de ce qu'est partir « loin de chez soi », dont la signification n'est pas simple

Capacité à s'offrir un repas avec de la viande, du poulet, du poisson (ou l'équivalent végétarien) tous les deux jours (HS050) : oui (1) / non(2). Il s'agit d'une question en partie « subjective » posée à la personne répondant pour le ménage. Elle fournit un bon indicateur de pauvreté, dans la mesure où elle renvoie bien à une « capabilité » ou encore à une possibilité matérielle, en partie subjective.

Capacité à faire face à des dépenses inattendues (HS060) : oui (1) / non (2)

Possession d'un téléphone, y compris mobile (HS070) : oui (1) / non (2)

Possession d'une TV couleur (HS080) : oui (1) / non (2)

Possession d'un ordinateur (HS090) : oui (1) / non ne peut pas se permettre (2) / non autre raison (3). Les modalités de réponse sont au nombre de trois, permettant de distinguer deux motifs (subjectifs) de la non-possession d'ordinateur. Possession d'une machine à laver (HS100) : oui (1) / non (2)

Possession d'une voiture (HS110) : oui (1) / non ne peut pas se permettre (2) / non autre raison (3)

Capacité à joindre les deux bouts (HS120) : très difficilement (1) / difficilement (2) / avec quelque difficulté (3) / relativement facilement (4) / facilement (5) / très facilement (6).

Charge financière du remboursement d'achats à crédit ou de dettes (HS140) : un fardeau lourd (1) / un fardeau léger (2) / pas de fardeau du tout (3).

Conditions de logement

Il s'agit ici de questions portant sur la situation résidentielle des ménages. On a retenu une question sur la charge financière du coût total du logement (qui dépend de la perception de ce qu'est un fardeau « lourd » ou « léger ») et une autre sur le type de logement. La première renvoie au poids de l'accès au logement dans le niveau de vie, la seconde décrit plutôt les conditions concrètes de logement et son environnement.

Problèmes avec le logement : trop obscur, pas assez de lumière (HS160) : oui (1) / non (2)

Charge financière du coût total du logement budget (HS140) : un fardeau lourd (1) / un fardeau léger (2) / pas de fardeau du tout (3). C'est une question à trois modalités de même nature que la précédente : un fardeau lourd, un fardeau léger, pas de fardeau du tout.

Type de logement (HH010) : maison séparée 1), semi-séparée (2), appartement dans un immeuble comptant moins de dix appartements (3), appartement dans un immeuble comptant plus de dix appartements (4), autre (5).

Statut de propriété (HH021) : propriétaire de plein droit (1) / accédant à la propriété (2) / locataire ou sous-locataire au prix de marché (3) / logé à prix réduit (4) / logé gratuitement (5)

Toit qui fuit, ou autres dégradations de l'appartement (HH040) : oui (1) / non (2)

Capacité à maintenir la chaleur dans l'appartement (HH050) : oui (1) / non (2)

Environnement résidentiel

Les trois questions dichotomiques (oui / non) posées impliquent une évaluation personnelle, les notions de « bruit », de « violence criminelle et vandalisme », ou encore de « pollution », etc., étant laissées à l'appréciation de l'enquêté.

Bruit des voisins ou de la rue (HS170) : oui (1) / non (2) ;

Pollution, saleté ou autre problème environnemental (HS180) oui (1) / non (2)

Violence criminelle ou vandalisme dans le quartier (HS190) : oui (1) / non (2)

Santé

A nouveau, on utilise deux questions en partie dépendantes d'une appréciation subjective (avec la notion de « limitation » et celle de « besoin non satisfait »).

Santé générale (PH010) : très bonne (1) / bonne (2) / correcte (3) / mauvaise (4) / très mauvaise (5)

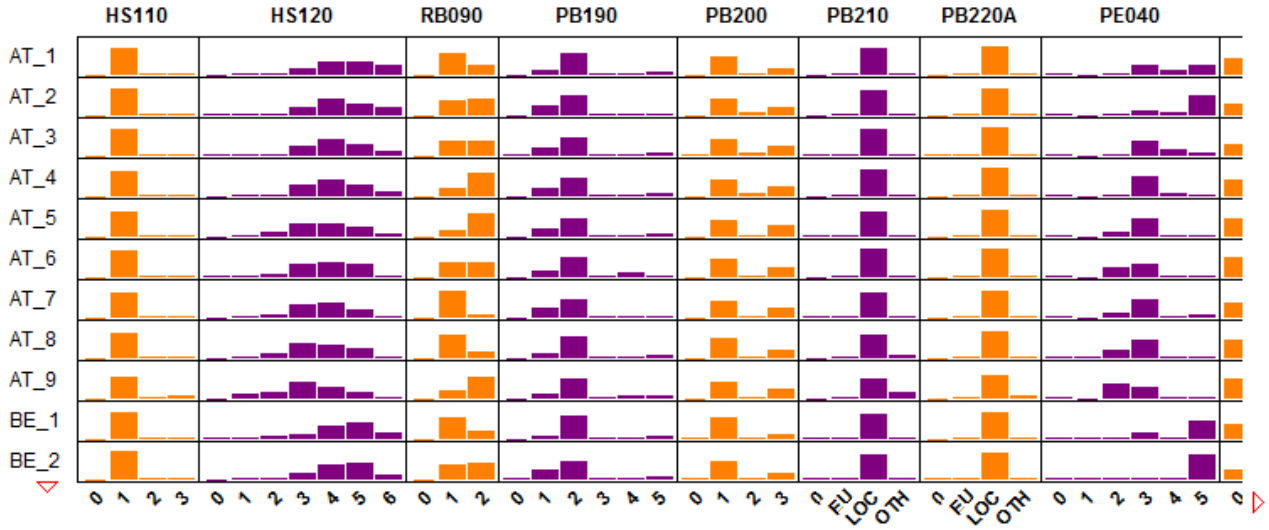
Souffre d'une maladie chronique (PH020) : oui (1) /non (2)

Activités limitées à cause de problèmes de santé (PH030) : oui (1) / oui en partie (2) / non (3) ;

Besoin non satisfait d'examen ou de traitement médical dans les douze derniers mois (PH040) : oui à au moins une occasion (1) /non à aucune occasion (2).

La base de données analysée, créée à l'aide du logiciel Syr, compte ainsi 288 concepts et 38 variables. Les données peuvent être présentées sous la forme d'un tableau tel que le tableau 1 (extrait d'un tableau plus large).

Tableau 1 : visualisation des données sous forme d'histogramme.



La visualisation sous forme d'histogrammes permet de faire apparaître et d'étudier de façon simple les *écarts entre distributions de fréquences* à l'intérieur des pays et entre les pays, variable par variable. Dans l'exemple retenu, on voit ainsi que les cadres dirigeants et les professions intellectuelles et scientifiques belges (BE_1 et BE_2) sont beaucoup moins nombreux que les « professions élémentaires » (AT_9) autrichiennes à rencontrer des difficultés pour « joindre les deux bouts » (variable HS120, modalités ordonnées de 1 « très difficile » à 6 « très facile »). Les écarts entre pays sont également notables.

On note aussi, avec la variable PE040 (ordonnée du plus faible au plus haut niveau de qualification) que les catégories sociales se différencient fortement en termes de niveaux de diplôme, ce qui correspond bien sûr au fait que le critère du niveau de qualification est utilisé pour les définir : les concepts choisis correspondent bien à des unités socio-démographiques spécifiques.

3.2.4. Analyse en composantes principales

On a procédé à l'ACP de 8 variables actives suivantes :

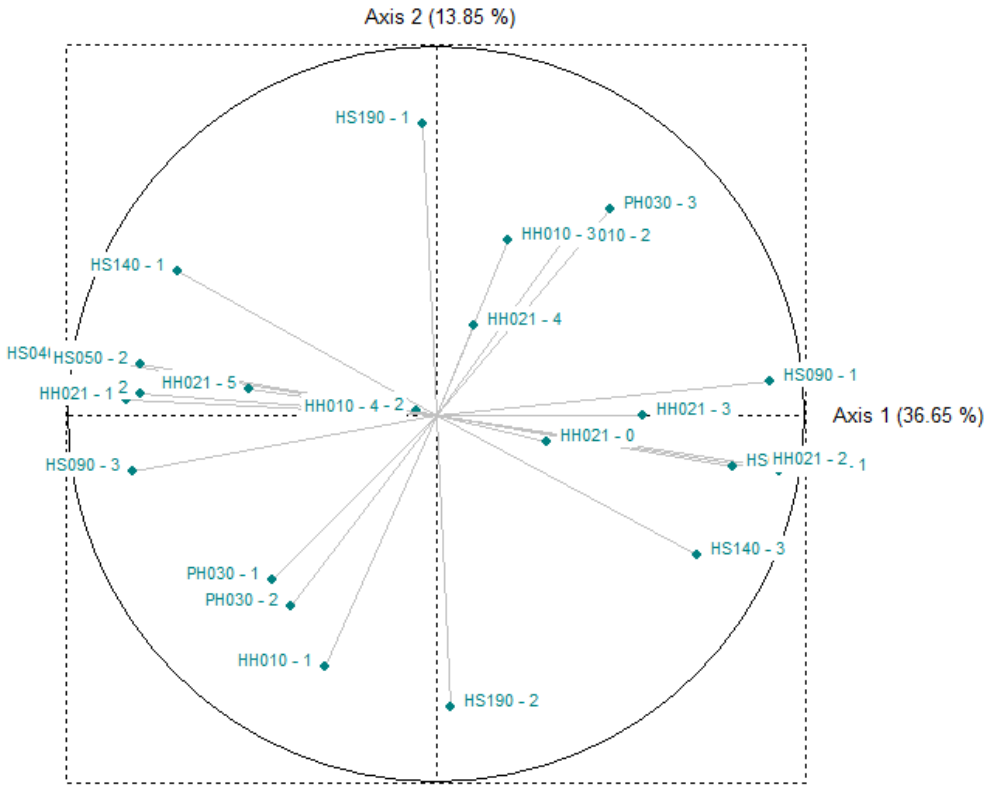
Conditions économiques : Possession d'un ordinateur (HS090) : oui (1) / non ne peut pas se permettre (2) / non autre raison (3). Capacité à s'offrir une semaine annuelle de vacances loin de chez soi (HS040) : oui (1) / non (2) ; Capacité à s'offrir un repas avec de la viande, du poulet, du poisson (ou l'équivalent végétarien) tous les deux jours (HS050) : oui (1) / non(2). ; Charge financière du remboursement d'achats à crédit ou de dettes (HS140) : un fardeau lourd (1) / un fardeau léger (2) / pas de fardeau du tout (3).

Environnement : Violence criminelle ou vandalisme dans le quartier (HS190) : oui (1) / non (2) ;

Santé : Activités limitées à cause de problèmes de santé (PH030) : oui (1) / oui en partie (2) / non (3).

Logement : Type de logement (HH010) : maison séparée (1), semi-séparée (2), appartement dans un immeuble comptant moins de dix appartements (3), appartement dans un immeuble comptant plus de dix appartements (4), autre (5) ; Statut de propriété (HH021) : propriétaire de plein droit (1) / accédant à la propriété (2) / locataire ou sous-locataire au prix de marché (3) / logé à prix réduit (4) / logé gratuitement (5).

Graphique 1 : variables actives de l'ACP dans le plan 1-2.



Le premier axe de l'ACP s'interprète comme un axe de niveau de confort économique et social général. Il est corrélé positivement avec la possession d'un ordinateur, l'accession à la propriété, la capacité à s'offrir une semaine de vacances loin de chez soi, la capacité à s'offrir un repas à base de viande, poisson ou équivalent végétarien tous les deux jours au moins.

Le deuxième axe de l'ACP est fortement corrélé avec une variable, le taux de délinquance dans le quartier, et dans une moindre mesure avec le mode de résidence et l'état de santé. Plus on descend, plus la situation est dégradée en matière de santé mais dans une résidence séparée et sans délinquance environnante. On retrouve ainsi un axe proche de ce que l'on observe sur les données individuelles.

L'axe 3 (cf. graphique 2) est lié au mode de résidence, à la santé, et à la situation de locataire. Il oppose des situations plus « précaires » en bas, à des situations plus « stables » en haut.

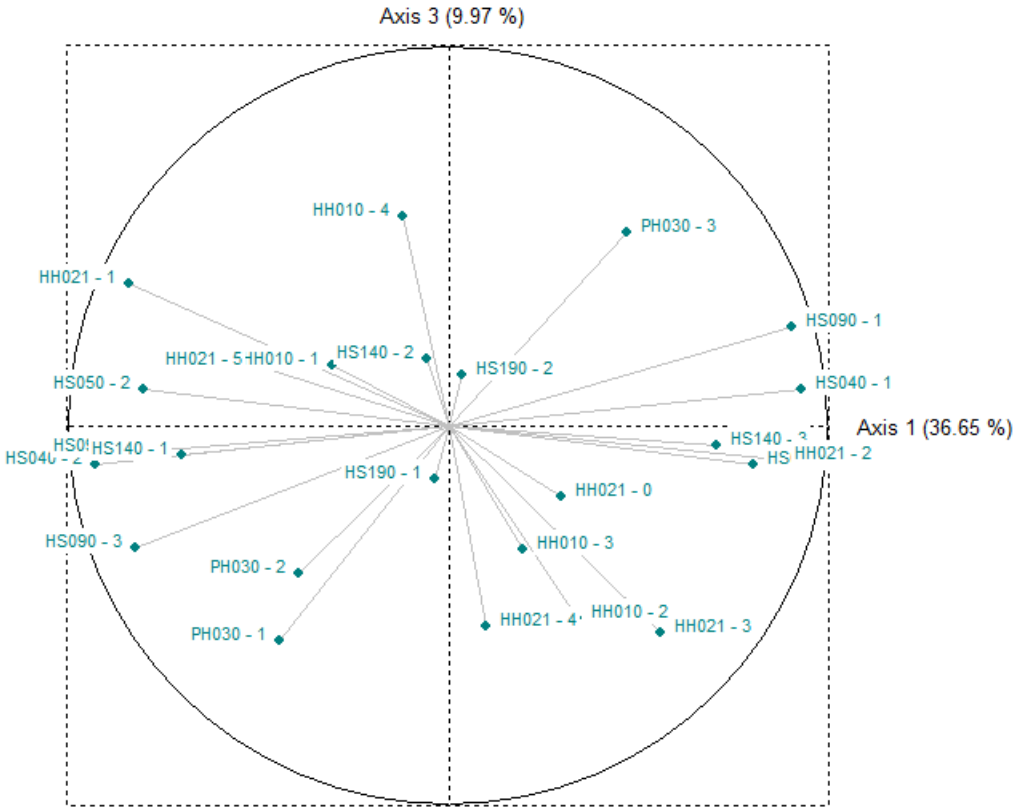
La projection des variables supplémentaires permet de compléter cette interprétation. L'axe 1 est un axe lié à la hiérarchie économique et sociale, alors que l'axe 2 est un axe « urbain / rural ».

Le nuage des concepts fait apparaître la forte dispersion des classes-pays en matière de conditions de vie. A gauche, on observe en particulier les catégories « professions élémentaires » et « agriculteurs » de Bulgarie et Roumanie, alors qu'à droite on voit se dégager les catégories supérieures et intermédiaires, voire même populaires, des pays d'Europe du Nord et de l'Ouest.

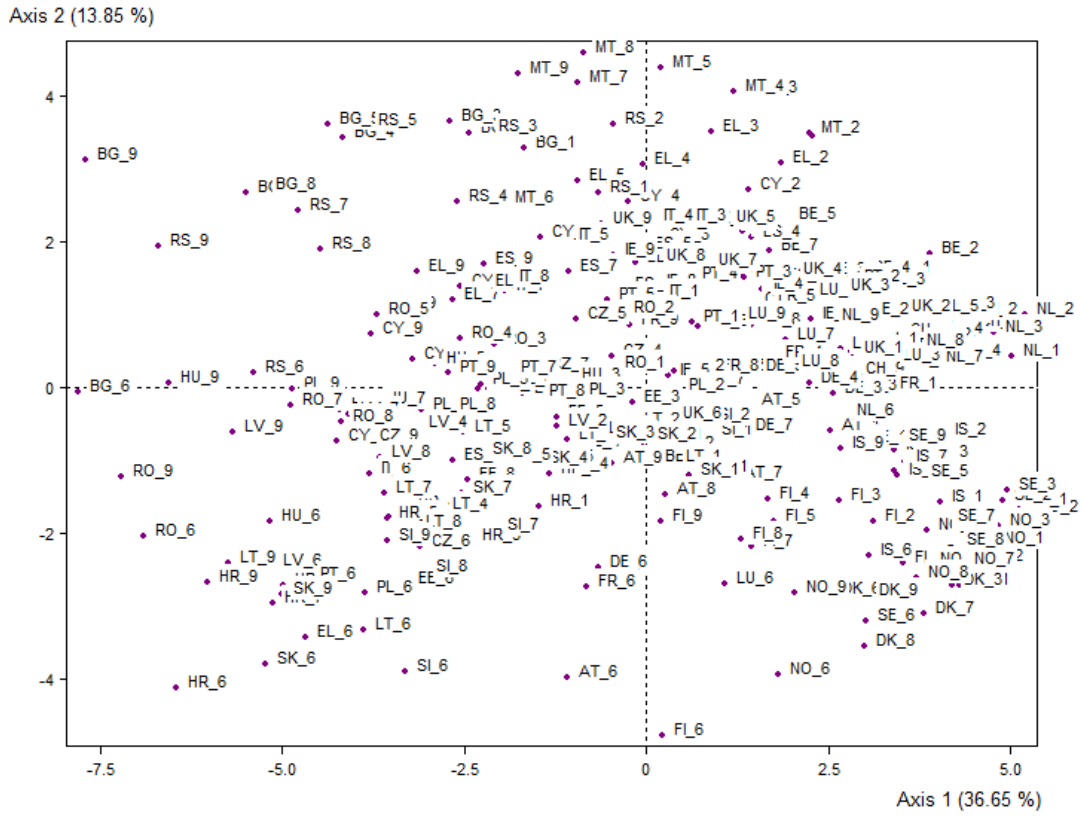
Sur l'axe 2, la situation des agriculteurs se démarque clairement vers le bas, par opposition à des groupes plus urbains, situés dans différents pays de niveaux de développement variables (Norvège, Finlande, Autrice, Slovénie, Slovaquie, Croatie…).

Enfin, l'axe 3 oppose des groupes populaires de certains pays comme le Royaume-Uni ou la Belgique, caractérisés par une forte précarité, de catégories supérieures dans d'autres pays, comme les pays d'Europe centrale et orientale.

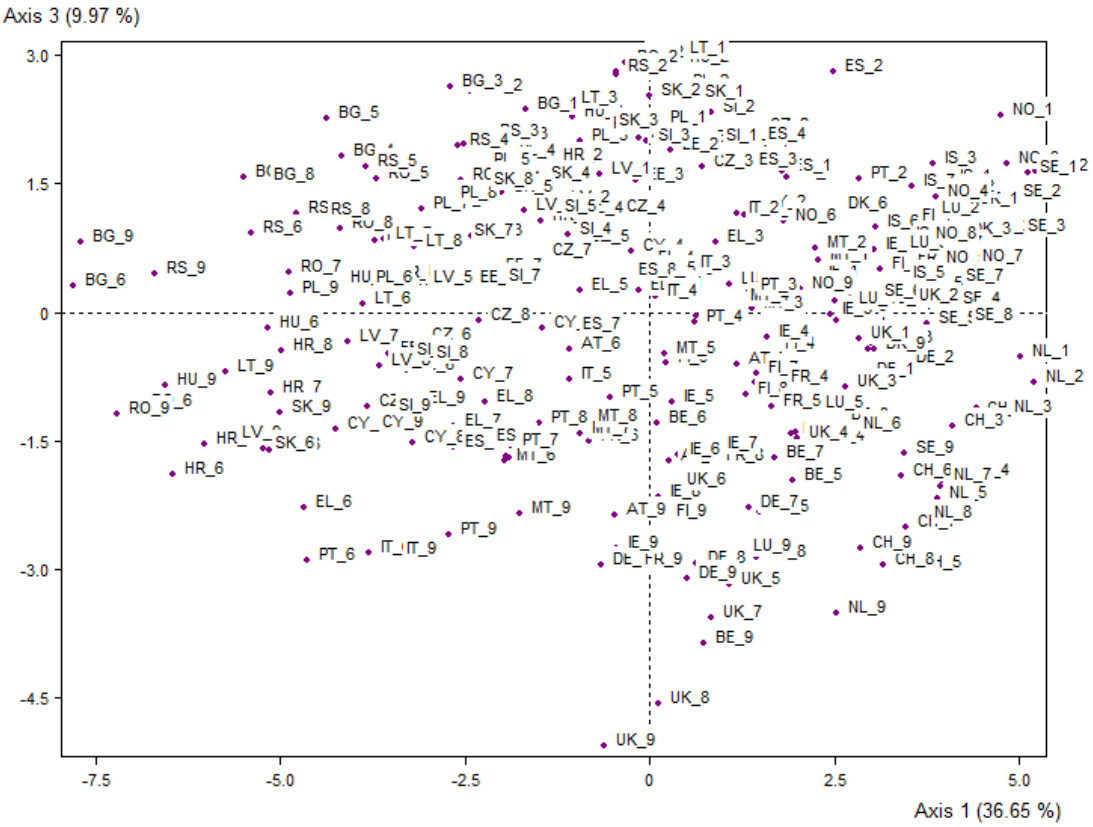
Graphique 2 : variables actives dans le plan 1-3 de l'ACP.



Graphique 3 : nuage des concepts (classes-pays) dans le plan 1-2 de l'ACP.

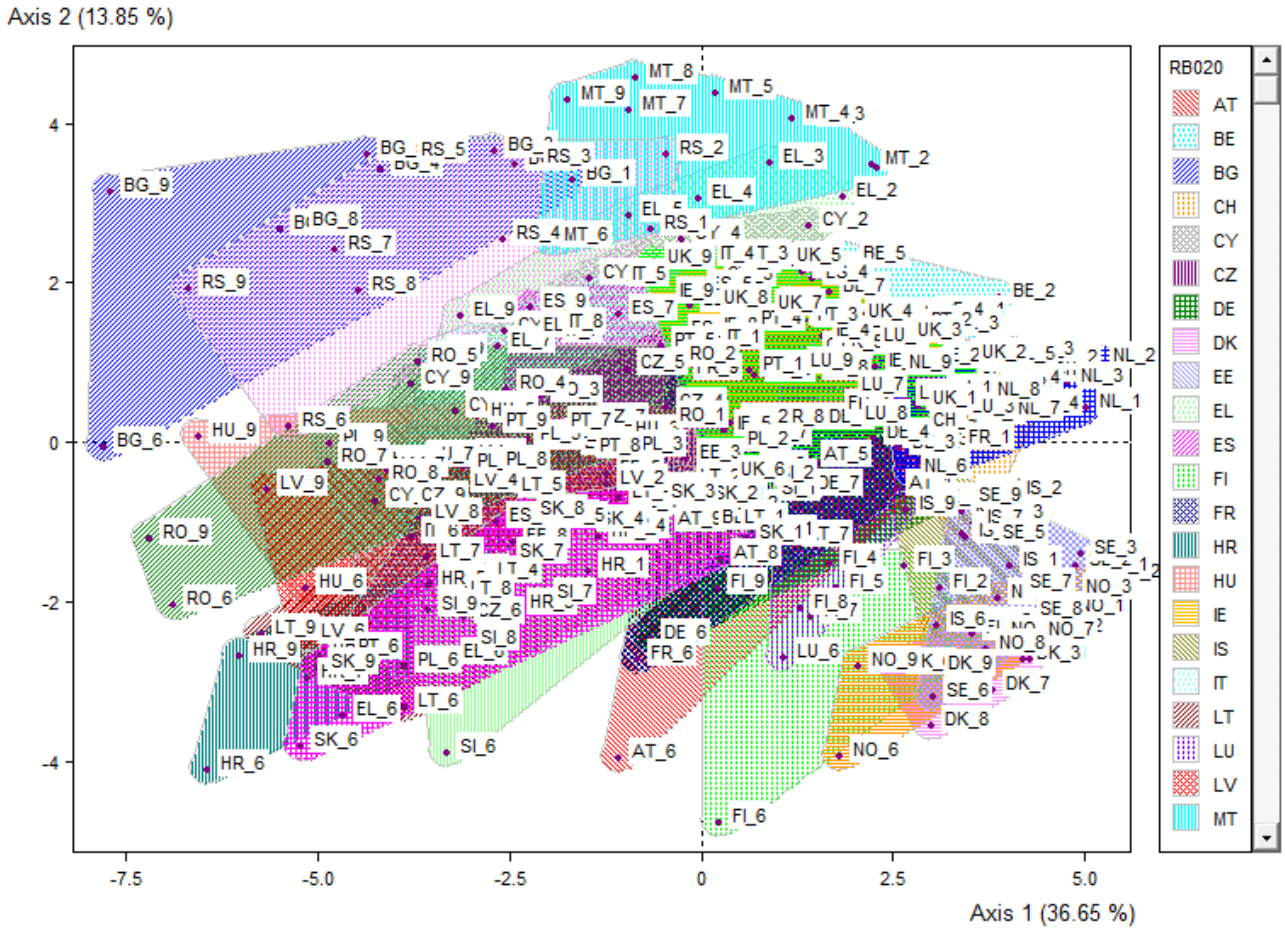


Graphique 4 : nuage des concepts dans le plan 1-3 de l'ACP.



On peut enfin représenter un « facteur structurant » dans l'espace des concepts. Le logiciel NetSyr utilise pour cela une représentation sous la forme de figures géométriques déterminées par les 9 points correspondant à l'autre variable. On voit sur le graphique suivant que les différents pays se superposent partiellement au centre du nuage mais que des spécificités les caractérisent quant à la position et la forme de la figure, reflétant une interaction importante entre les facteurs pays et classe sociale.

Graphique 5 : hypercubes des pays dans le plan 1-2 de l'ACP.



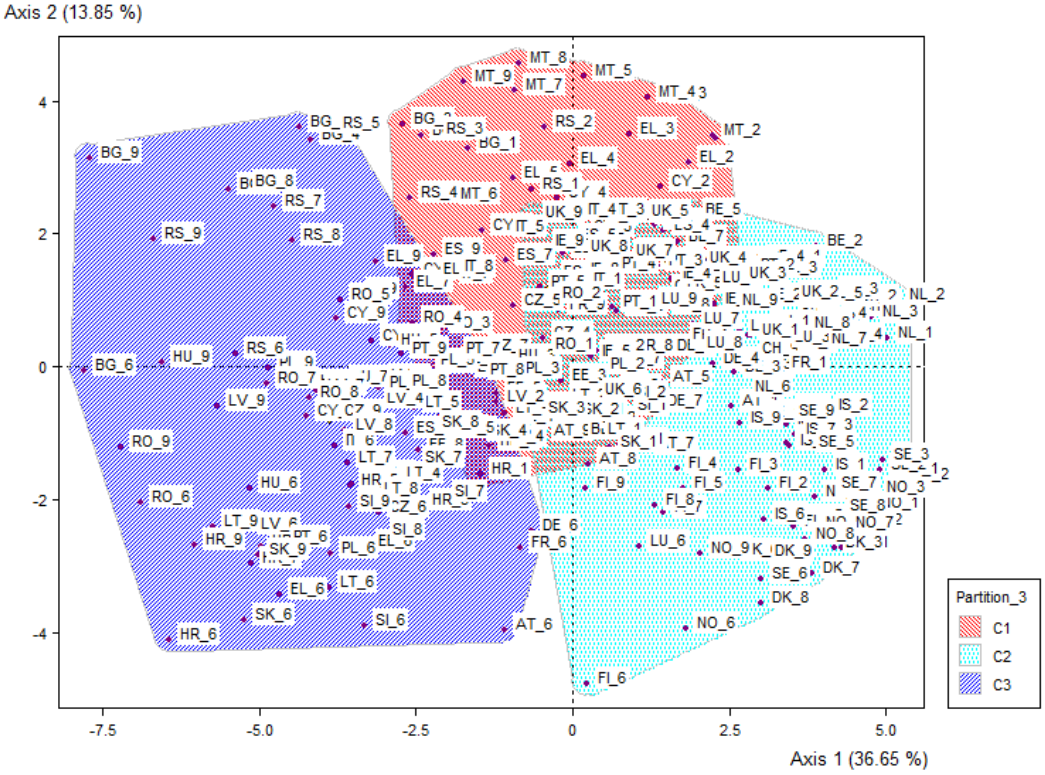
3.2.5. La visualisation des « hétérogénéités »

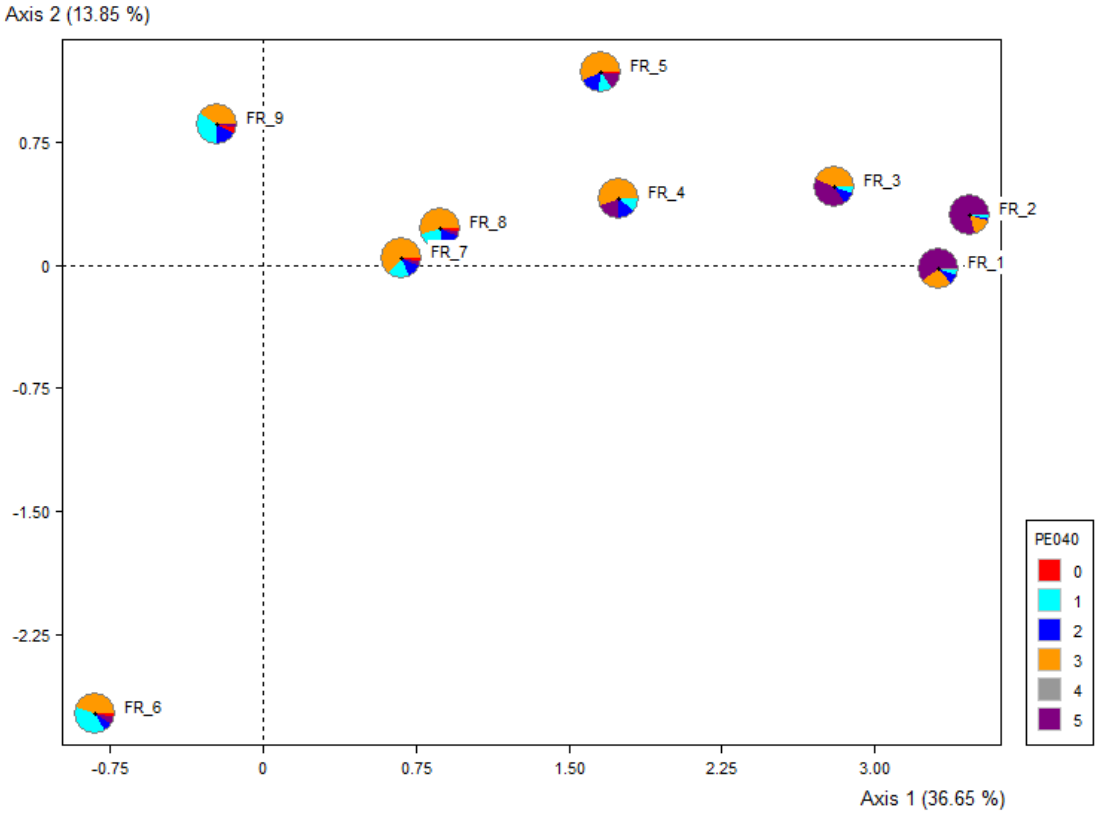
L'ADS (avec le programme NetSyr) permet aussi de représenter les hétérogénéités observées dans l'espace multidimensionnel construit à l'aide de l'ACP. À titre d'exemple, on représente ici les différentes configurations de diplômes parmi les 9 groupes ISCO français. On voit que, dans le plan 1-2 de l'ACP, seuls les groupes 6 et 9 se situent à gauche sur le premier axe et que le groupe 6 se distingue très nettement en bas sur le deuxième axe, les autres groupes étant clairement hiérarchisés (le groupe des professions intellectuelles et scientifiques un peu plus à droite) sur l'axe 1, conformément à l'interprétation socio-économique que nous en avons faite. Ces différents groupes se distinguent fortement du point de vue de la structure des diplômes en leur sein.

3.2.6. Classification

Le programme Netsyr procède par une classification par la méthode des *K-means*. Nous n'en présentons ici qu'un aperçu dans le plan 1-2, avec une classification en trois groupes de concepts. Ces trois classes peuvent ensuite être décrites par les propriétés des concepts qui les constituent.

Graphique 6 : nuage des groupes CITP de la France dans le plan 1-2 de l'ACP et représentation de la structure des diplômes en leur sein.





3.3. Une interprétation sociologique

On retrouve ici des résultats proches de ceux d'analyses menées au niveau des micro-données, mais situées ici d'emblée au niveau de concepts spécifiquement construits pour l'analyse, à savoir les classes-pays (pour une analyse très similaire du point de vue des unités d'analyse et de la démarche, cf. Hugrée *et al*., 2014).

Une forte opposition hiérarchique structure l'espace des pays et des groupes sociaux en Europe, et différencie nettement en leur sein les conditions de vie. Elle est redoublée par une différenciation liée à l'environnement résidentiel (urbain / rural) et une autre liée au degré de précarité. Chaque « classe-pays » constitue un point dans un espace ainsi structuré, mais, bien évidemment, les dispersions individuelles sous-jacentes sont importantes : on s'est situé ici d'abord au niveau des agrégats que constituent les groupes-pays.

La « carte géométrique » présentée plus haut ressemble à la « carte géographique » des pays, traduisant bien le fait que l'espace européen est aussi un espace social, et chaque zone étant spécifiquement différenciée. S'ajoute ici, résultat spécifique à l'ADS, la possibilité de visualiser à l'aide d'histogrammes dans l'espace construit les « hétérogénéités » relatives des différents classes-pays selon les critères concernés.

L'intérêt de ce type d'analyse est ainsi de rendre possible, dans un cadre global, la visualisation géométrique de données hétérogènes et « multi-niveaux », ce qui enrichit potentiellement l'appréhension de processus sociaux qui sont toujours situés sur plusieurs échelles (le système économique et social mondial, la zone régionale, le pays, et, au niveau le plus fin, l'individu voire la pratique ou l'activité) et présentent toujours des articulations subtiles de variations internes, qui sont situées à ces différentes échelles.

4. Conclusion

L'approche de la méthodologie statistique en sciences sociales évoquée rapidement ici peut donc être combinée à une perspective sociologique « globale », et contribuer ainsi à la description et à la formalisation progressive de structures sociales emboitées les unes dans les autres et caractérisées par des hétérogénéités fortes. Ce programme de recherche peut s'étendre à toutes sortes d'échelles d'observation, du très micro (l'activité individuelle ponctuelle) jusqu'au très macro (l'espace mondial).

À la différence de la démarche de l'analyse multi-niveau, cette perspective se situe dans la filiation d'une approche « englobante » de l'étude de faits sociaux complexes, qui vise à rendre compte de la multidimensionnalité du réel sans le réduire à quelques forces supposées *a priori* déterminantes. Cependant, elle participe de la même volonté intégratrice entre théories et méthodes des sciences sociales (Courgeau, 2004 ; Pumain, Saint-Julien, 2010).

L'intégration des deux traditions, aussi difficile soit-elle, du moins à première vue, apparaît dès lors comme un objectif important pour l'avenir des sciences sociales, en particulier si l'on songe à leurs interactions de plus en plus denses avec les sciences de la nature.

Les liens entre ces approches et des démarches plus « qualitatives » de la sociologie, comme l'observation ethnographique telle qu'elle s'est développée en sociologie et anthropologie (Weber, 2015), sont aussi parmi les enjeux futurs importants, auxquels une perspective plus attentive à la complexité et à sa formalisation rigoureuse peut également fournir un éclairage intéressant.

Bibliographie

Benzécri, J.-P. (1973). *Analyse des Données. Tome II: Analyse des Correspondances.* Paris: Dunod.

Billard L., Diday E. (2006) *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley.

Bock, H. H. and Diday, E. (Eds) (2000) *Analysis of Symbolic Data*, Springer.

Boudon R. (1975), L'inégalité des chances, Paris, Fayard/Pluriel [2011].

Bressoux P. (2008), Modélisation statistique appliquée aux sciences sociales, Bruxelles, de Boeck.

Hugrée C., Pénissat E., Spire A., Brousse C. (2014). « Capital culturel et pratiques culturelles. Les enjeux d'une comparaison européenne depuis l'enquête SILC-EU 2006 », communication au colloque Les classes sociales en Europe, AFS, décembre 2014.

Courgeau D. (2004), Du groupe à l'individu. Synthèse multiniveau, Paris, INED.

Desbois D. (2009), « La place de l'a priori dans l'analyse des données économiques ou le programme fort des méthodes inductives au service de l'hétérodoxie », *Revue Modulad*, 39, p. 176-181.

Diday E. and Noirhomme-Fraiture M., (eds.) (2008). *Symbolic Data Analysis and the SODAS Software*, Wiley.

Durkheim E. (1895), Les règles de la méthode sociologique, Paris, Flammarion [2010].

Greenacre MJ (1984). *Theory and application of Correspondence Analysis*. London: Academic Press, Inc.

Human Development Report 2010 (20th Anniversary Edition). The Real Wealth of Nations: Pathways to Human Development. Published for the United Nation Development Programme.

http://hdr.undp.org/en/reports/global/hdr2010/chapters/

Lebaron F. (2015), "L'espace des conditions de vie des actifs occupés en Europe", à paraître comme chapitre d'un document de travail de l'INSEE.

Lévy-Vehel J., Walter C. (2002), Les marchés fractals, Paris, PUF.

Le Roux (2014), Analyse géométrique des données multidimensionelles, Paris, Dunod.

Le Roux B., Lebaron F. (2015), « Les idées-clés de l'analyse géométrique des données », in F.Lebaron, B.Le Roux (dir.), La méthodologie de Pierre Bourdieu en action. Espace culturel, espace social et analyse des données, Paris, Dunod.

Mauss M. (1923), Essai sur le don, Paris, PUF [2012].

Meuriot V. (2015), Une étude critique et réflexive de l'économétrie des séries temporelles (1974-1982)", Habilitation à diriger des recherches, université de Versailles-Saint-Quentin-en-Yvelines / université Paris-Saclay.

Passeron J.-C. (1991), Le raisonnement sociologique. L'espace non-popperien du raisonnement naturel, Paris, Nathan.

Pumain D., Saint-Julien T. (2010), Analyse spatiale des interactions, Paris, Armand Colin.

Rouanet H, Bernard J. M., Lecoutre B. Lecoutre M. P. Le Roux B. (1998) New ways in statistical methodology : from significance tests to Bayesian inference (Foreword by P. Suppes). Berne, Peter Lang

Simiand, F. (1922), Statistique et expérience. Remarques de méthode. Paris. Marcel Rivière.

Verde R. Diday E. (2014), "Symbolic data analysis: a factorial approach based on fuzzy coded data", in Editors: Jorg Blasius, Michael Greenacre, Visualization and Verbalization of Data, Chapter: 16, CRC Press, p.255-270.

Walter C. (2013), Le modèle de marche aléatoire en finance, Paris, Economica.

Weber F. (2015), Brève histoire de l'anthropologie, Paris, Flammarion.

World Development Indicators 2010, Washington, D.C.: World Bank.