HARMONISATION DES MÉTHODES AU NIVEAU EUROPÉEN : UN PRÉALABLE POUR ASSURER LA COMPARABILITÉ OU UN MYTHE ?

Daniel Defays

Introduction

La qualité est devenue depuis quelques années un label qu'il importe d'afficher ; certains en font un argument de vente, d'autres un mode de gestion. Les services publics ne paraissent pas pouvoir échapper à cette tendance. Plus que d'une mode, il s'agit d'une préoccupation profonde liée à des exigences croissantes des consommateurs, un souci d'efficacité des gouvernants, une pression d'un environnement de plus en plus concurrentiel. Certains offices statistiques ont déjà réagi ; ils proposent à leurs clients, leurs fournisseurs et leurs partenaires des chartes qualité, introduisent des modes de gestion basés sur la qualité totale (TQM), envisagent de se faire certifier. Eurostat, l'office statistique des Communautés européennes, n'a pas échappé à ce mouvement. Il a défini sa mission en termes de fournitures de services statistiques de *qualité*. Cette ambition a replacé la comparabilité, qui au niveau européen est une composante essentielle de la qualité, au centre des préoccupations. Elle fait actuellement l'objet de nombreux débats : ne peut-on pas harmoniser les données a posteriori, faut-il tout harmoniser au même niveau, l'harmonisation complète a-t-elle un sens dans des contextes socio-économiques et surtout administratifs différents ? L'objet de cet article est de faire le point sur ces questions, en insistant particulièrement sur l'impact que différents choix méthodologiques peuvent avoir sur le niveau de comparabilité des données. Il débute par un rappel de ce qui justifie une plus grande harmonisation des données européennes et de ce qu'on appelle comparabilité, il examine ensuite à partir d'études récentes la possibilité réelle d'harmoniser *a posteriori* ou plutôt l'interaction entre mesures et méthodes de mesure. La possibilité de distinguer différents niveaux de comparabilité est ensuite évoquée avant d'aborder une question plus fondamentale, centrale pour la construction européenne : l'harmonisation est-elle possible et si la réponse est oui, que signifie exactement ce terme ?

Pourquoi harmoniser ?

Il est devenu commun de souligner les différences entre données, information, connaissance. Quel que soit le point de vue adopté, il apparaît clairement que la valeur informative d'une donnée est liée à sa capacité à renvoyer à d'autres informations, à ses connotations. Un nombre isolé ne signifie rien. Il faut lui attacher des éléments descriptifs, le situer dans le temps, le comparer à d'autres pour en faire naître une signification. La statistique n'échappe pas à cette logique. Un nombre d'entreprises innovantes dans un pays, un PNB, s'apprécient aussi par référence à des données similaires (comparables ?) collectées dans d'autres pays, à d'autres moments. Obtenir des données comparables apparaît donc dans un premier temps comme une nécessité pour donner de la substance, du contenu à nos informations statistiques (voir par exemple le débat sur ce thème en Intelligence Artificielle, Hofstadter, 1982).

La globalisation, si souvent évoquée, appelle également une internationalisation de nos données et par conséquent plus de comparabilité ; les chefs d'entreprises opèrent sur des marchés qui dépassent les frontières nationales, les pays échangent des biens et des services, les citoyens voyagent. Les pouvoirs publics doivent se préoccuper de ces flux, de ces échanges pour comprendre leur propre économie et décrire la société, les acteurs socio-économiques réclament plus d'informations sur ce qui se passe à l'étranger, un sentiment de citoyenneté supra nationale est en train d'émerger et suscite un intérêt croissant pour ce qui se passe au-delà des périmètres nationaux.

Faut-il mentionner ici la construction européenne, l'apparition d'une administration spécifique avec ses besoins propres en information sur l'Union ? Les politiques communes agricole, régionale, de recherche et développement, la mise en place d'un réel marché intérieur nécessitent à des fins de gestion des informations comparables et souvent qui puissent être agrégées au niveau européen. Ces données conditionnent la gestion de budgets impressionnants. L'utilisation, comme une des références pour définir les contributions nationales au budget communautaire, du PNB, concept central de la comptabilité nationale, elle-même principal élément d'articulation des systèmes statistiques, a également eu une influence déterminante sur l'harmonisation des statistiques.

Et le plaidoyer pourrait continuer. La comparabilité internationale est devenue une exigence incontournable en cette fin de vingtième siècle ; elle correspond à une demande forte des utilisateurs de statistique exprimée à de nombreuses occasions. Elle conditionne par conséquent la qualité de nos services, si cette qualité est définie, conformément à la norme ISO 8402, comme l'ensemble des propriétés et des caractéristiques d'une entité qui lui confèrent l'aptitude à satisfaire des besoins exprimés et implicites.

Si la reconnaissance de la nécessité de plus de comparabilité ne paraît pas poser de problème, la définition de ce qu'on entend par comparabilité est moins évidente.

Que signifie harmonisation ?.

Dans le début de cet article nous avons utilisé de manière quasi interchangeable les mots 'comparabilité' et 'harmonisation'. Leur signification est pourtant différente et mérite d'être précisée. L'harmonisation paraît un préalable à la comparabilité. Ceci est pourtant un peu court. Le recours aux dictionnaires peut aider à mieux comprendre les significations et rôles respectifs de ces deux concepts. Comparer, c'est 'examiner les rapports de ressemblance et de différence' ou c'est 'rapprocher en vue d'assimiler ; mettre en parallèle', nous dit 'Le petit Robert', alors qu'harmoniser, c'est 'mettre en harmonie, en accord ', l'harmonie étant elle-même définie comme les relations existant entre les diverses parties d'un tout qui font que ces parties concourent à un même effet d'ensemble. On peut ainsi comparer des statistiques relativement différentes comme le nombre de chômeurs dans une région donnée à une époque donnée avec la population totale correspondante ou des dépenses de fonctionnement moyennes dans une population d'entreprises avec un chiffre d'affaires moyen ; ces comparaisons ont un sens parce qu'elles permettent de mettre les données en parallèle, d'opérer des assimilations (de calculer des rapports par exemple), mais ces données ne doivent pas nécessairement être harmonisées. L'harmonisation est plus exigeante ; elle requiert l'existence d'un tout, une subdivision en parties, et l'existence d'un accord, d'une correspondance entre ces parties. Le fait d'utiliser de manière indifférenciée les mots comparabilité et harmonisation est donc un abus de langage. Dans ce qui suit, nous nous intéressons à la notion stricte d'harmonisation. Un des objectifs de cet article est du reste de lui donner un contenu plus précis.

La distinction étant établie, il importe cependant de répéter qu'une des justifications de la nécessité d'harmoniser est de pouvoir comparer des données d'origines différentes : statistiques relatives à différents secteurs, différentes périodes, différentes zones géographiques. La comparaison a pour objet de mettre en rapport des différences ou des ressemblances observées avec des effets liés aux secteurs, au temps, aux pays. Pour que ce rapprochement soit possible et surtout instructif, il importe que les effets soient aussi purs que possible. Des écarts qui sont attribuables en même temps à des différences méthodologiques (définitions de concepts différentes, méthodes de mesure non similaires ...) et à la variable explicative d'intérêt sont difficiles à interpréter. Si l'impact de la méthodologie peut être assimilé à un bruit, ceci n'est pas gênant pourvu qu'on puisse estimer les niveaux de précision atteints dans l'estimation des paramètres étudiés. N'est-il pas courant dans les sciences expérimentales de comparer des caractéristiques de population à partir d'estimateurs de variance différents ? Malheureusement la manière précise dont les différents choix méthodologiques opérés affectent les grandeurs estimées est généralement inconnue. En comparant deux pays, on risque donc d'interpréter des biais liés essentiellement à des méthodes différentes. Ceci semble plaider pour une harmonisation complète des méthodes de mesure. Cette position radicale rencontre comme on peut s'en douter de nombreuses résistances.

L'harmonisation *a posteriori*

La comparabilité, comme je l'ai signalé en début d'article, n'est qu'une des composantes de la qualité, à côté de la pertinence de la mesure ('relevance' en anglais), de sa précision, de sa fraîcheur, de sa cohérence avec les autres informations statistiques, par exemple (voir, par exemple, Depoutot, 1998). Dans certains cas, des conflits peuvent apparaître entre ces différentes exigences. Les enquêtes nationales visent généralement à apprécier des évolutions temporelles, des effets régionaux ou sectoriels ; la comparabilité internationale est alors perçue comme un bénéfice complémentaire mais qu'on n'est pas prêt à percevoir à n'importe quel prix. Les systèmes nationaux possèdent leur propre logique, leur propre cohérence, doivent répondre à des exigences de fraîcheur des données qui peuvent entrer en conflit avec les prescriptions internationales. Un concept harmonisé peut perdre sa pertinence dans certains pays. Ceci conduit à des arbitrages. Certains pays ont défendu le concept d'harmonisation a posteriori. Les méthodes de mesure restent de la compétence exclusive des pays et ceux-ci opèrent en fin de traitement les corrections nécessaires pour rendre les données comparables. Des données sur la population seront donc collectées par recensement dans certains pays, par exploitation de registres administratifs dans d'autres, par questionnaire ou par entretien, avec ou sans correction pour les non-réponses. L'idée sous-jacente est qu'il existe un concept, par exemple un paramètre de population précisément défini, qui peut être mesuré indifféremment de différentes manières. Les spécifications communautaires doivent porter sur la nature des résultats désirés et non sur les méthodes à utiliser pour les obtenir. Cette position n'est pas dénuée d'ambiguïté. Où finissent les résultats et où commencent les méthodes ? Admettons que l'on se soit mis d'accord sur ce qu'est l'innovation technologique, peut-on réduire les obligations communautaires au comptage, par exemple, des unités innovantes ? Non, bien sûr ; il faut encore spécifier le type d'unités, la population et l'époque de référence. Peut-on s'arrêter là en décrétant que la statistique qui nous intéresse est l'effectif de l'ensemble des entreprises innovantes du secteur manufacturier à la date du premier janvier 1993 ? Suivant que l'information est recueillie par téléphone ou par entretien individuel, par une interrogation du chef d'entreprise ou d'un comptable, par une question posée à la fin ou au début d'un long questionnaire, suivant que le taux de réponse est de 30, 60 ou 90 %, suivant qu'il y a eu correction ou non pour non-réponses, suivant que l'information est produite par exploitation d'un fichier administratif, par recensement ou sondage, obtiendra-t-on des résultats comparables ? La réponse est clairement non. Pour expliquer l'interaction entre méthode de mesure et phénomène mesuré Gribbin (Gribbin, 1984) reprend une anecdote racontée par le physicien quantique Wheeler. Celui-ci fut invité à jouer pendant un dîner au vieux jeu des devinettes. Il sortit de la pièce afin que les autres invités puissent décider quel était l'objet à découvrir mais fut exclu pendant un temps incroyablement long, situation qui attestait que ses partenaires choisissaient un mot singulièrement difficile ou s'apprêtaient à lui jouer un tour. De retour dans la pièce, il constata que les réponses à ses questions du type 'est-ce que ça vole ?", ou "est-ce un objet inanimé,?" étaient d'abord très rapides, mais au fur et à mesure que le jeu avançait, se faisaient de plus en plus lentes. Ceci lui paraissait étrange puisque le groupe était supposé s'être mis d'accord a priori sur un objet et qu'il suffisait de répondre par oui ou par non. Après un long interrogatoire de l'assemblée , Wheeler suggéra : "est-ce un nuage ?". La salle

répondit "oui" en choeur et dans un éclat de rire. En fait, ses amis s'étaient mis d'accord non sur l'objet qu'il convenait de deviner, mais sur le fait que chaque personne interrogée devait donner une réponse sincère concernant un objet réel auquel elle pensait et qui devait correspondre à toutes les réponses précédentes. Chaque joueur devait donc imaginer son propre objet et le modifier progressivement en fonction des réponses données aux questions de Wheeler. Chacun, à chaque réponse, révisait, le cas échéant, l'objet auquel il pensait, d'où la difficulté de la tâche non seulement pour Wheeler mais également pour ses amis. En opérant ainsi, l'objet qu'a découvert Wheeler est un pur produit du processus utilisé pour le trouver. Il est construit par les questions posées et les réponses données. Il est impossible dans certains cas de dissocier méthode de mesure et objet de mesure. Cet exemple emprunté à un exposé sur la mécanique quantique peut être transposé à la statistique ou du moins à certains domaines de la statistique. Comment définir une attitude par exemple sans faire référence au questionnement ?

L'impact des modes de collecte et de traitement de l'information sur les résultats obtenus est un sujet qui préoccupe Eurostat depuis longtemps. Dans les paragraphes qui suivent je présente brièvement quelques travaux récents sur ce sujet. Ils permettent de mieux apprécier les rapports complexes qui existent entre les instruments de mesure, les méthodes et les mesures réalisées.

L'étude de Statistique Pays-Bas sur l'interprétation des unités statistiques

Faire de la statistique, c'est estimer des paramètres de population. Les populations sont constituées d'unités sur lesquelles il importe de se mettre d'accord si l'on veut comparer des résultats. L'importance de ce problème est du reste reconnue par l'existence d'un règlement communautaire sur les unités statistiques. Les prescriptions internationales ne peuvent être formulées qu'en termes généraux compte tenu de l'hétérogénéité des situations nationales. Comment les pays interprètent-ils ces normes communautaires et quel est l'impact de ces différentes interprétations sur la comparabilité des résultats ? Ces questions ont fait l'objet d'une étude commanditée par Eurostat et exécutée par Statistique Pays-Bas. Pour ce faire, trois pays ont été invités à échanger des informations sur des unités statistiques et à comparer les délinéations obtenues pour ces unités en utilisant leurs propres critères. Plus précisément, chaque pays a identifié dans son répertoire national 16 ensembles d'unités (des secteurs manufacturier et des services) qu'il a proposés aux collègues des deux autres pays. Ceux-ci avaient pour tâche de définir dans ces ensembles ce qu'ils considéraient être des entreprises. Pour ce faire, ils étaient autorisés à poser des questions supplémentaires au pays donateur : existence de liens financiers, situation géographique précise, interdépendance des gestions etc. A partir de ces informations, chaque pays a proposé une subdivision en entreprises, au sens national du terme, des ensembles fournis par les deux autres et bien entendu de l'ensemble que lui-même avait proposé. Trois variables ont été attachées à ces unités : le secteur d'activité, l'emploi et le chiffre d'affaires. Trois délinéations différentes (une pour chaque pays) d'un même ensemble d'opérateurs économiques (environ 45 au total)

ont ainsi été définies. Elles correspondent à trois interprétations d'une même norme communautaire (en fait, 5 interprétations différentes ont été proposées deux pays ayant une interprétation théorique différente de celle qui était effectivement utilisée). Les résultats obtenus sont inquiétants, un pays a dans le secteur des services identifié 21 entreprises là où un autre n'en avait trouvées que 7. L'impact sur les totaux par secteur n'est pas non plus négligeable : des différences de l'ordre de 10% ont été observées dans certains cas. Les résultats de cette étude sont bien entendu difficiles à généraliser, les ensembles initiaux n'étant pas représentatifs des populations généralement étudiées.

Il invite cependant à la prudence et montre au minimum la nécessité de confronter les interprétations nationales pour mieux comprendre les disparités observées.

La population de référence

Un autre exemple intéressant de sensibilité des résultats à différentes interprétations données à un même concept est donné par Sirilli (Sirilli, 1997). Il s'agit cette fois-ci de la définition de la R&D. Dans le manuel de Frascati, le champ des enquêtes sur la recherche et le développement est défini à partir de la notion de recherche systématique ("travaux entrepris de façon systématique"), c'est-à-dire effectuée de manière continue. Cette restriction, comme le montre l'étude de cas qui suit, est fondamentale. Pour l'année 1992, l'Italie lors de deux enquêtes successives, une sur la R&D proprement dite et l'autre sur l'innovation, a utilisé des variantes. Dans l'enquête R&D, le champ couvre l'ensemble des entreprises qui effectuent de la R&D de manière stable et continue et qui ont une certaine taille ; dans l'enquête innovation, sont couvertes les firmes, quelle que soit leur taille, qui ont innové mais sans effectuer nécessairement elles-mêmes de la recherche. Une question commune permet de comparer le nombre de firmes qui font de la R&D et leurs dépenses. Il est normal de s'attendre à des différences puisque les populations de référence sont en fait différentes. Mais leur ampleur surprend : 748 enterprises pour l'enquête R&D et 4229 pour l'enquête innovation. Et des différences similaires sont observables en matière d'évolution de 1985 à 1992 (de 793 à 748 dans un cas, de 2557 à 4229 dans l'autre). Par contre les données de dépenses - comme on pouvait s'y attendre puisque les différences sont sûrement attribuables à l'inclusion des petites enterprises et de recherches occasionnelles dans une enquête - ne diffèrent que de 14%. Cet exemple bien entendu ne démontre rien du tout puisque les champs comme je viens de le préciser étaient différents. Il confirme la sensibilité des statistiques à la définition des populations de référence, ce qui est bien entendu une évidence, et invite à se méfier de l'impact que pourraient avoir différentes opérationalisations de la notion de recherche systématique sur les statistiques produites. De légères variantes risquent d'introduire des variations importantes et d'invalider des comparaisons internationales. Sans opérationalisation commune de la notion de recherche systématique les comparaisons pourraient être vaines....

L'importance du questionnaire

L'importance de la formulation et de la présentation des questions sur les réponses qui leur sont données est connue depuis longtemps. Des travaux récents de Piazza et Sniderman (Piazza, 1997) ont par exemple montré que des écarts considérables – passage d'une proportion d'approbation de 26 à 63 % - sont possibles suivant la manière dont une question est introduite et formulée. L'exemple donné porte sur la discrimination raciale. La question posée concerne la préférence qu'il faut donner en matière d'admission à l'université aux candidats présentant les qualifications nécessaires qui sont noirs de peau. Dans une première version du questionnaire, la question était précédée d'une note explicative qui soulignait l'existence dans le passé d'une discrimination qui avait profité aux blancs. Elle insistait sur la nécessité de compenser cet état de fait en donnant préférence aux noirs ; elle reconnaissait que ce point de vue n'était pas partagé par tout le monde, certains n'admettant pas l'utilisation de critère racial pour la sélection à l'université. L'autre version était assez semblable : une note explicative reconnaissait une discrimination dans le passé et de la nécessité de consentir un effort supplémentaire pour assurer que les noirs qui présentent les qualifications nécessaires soient considérés. Le point de vue opposé (pas de discrimination basée sur la race) était également présenté comme dans la première version. Sur les 889 personnes interrogées à partir de la version 1, 26% se sont déclarés en faveur d'une discrimination positive à l'égard des noirs. Ce pourcentage, calculé sur un échantillon de 911 personnes, est passé à 63% avec la version 2 où la nécessité d'un effort supplémentaire était mentionnée dans la question.

Ce qui étonne n'est pas l'existence d'un effet mais plutôt son ampleur. Et il n'est pas sûr que l'exemple donné soit caricatural. Les mots utilisés dans les questions sont quelquefois entourés de halo, de connotations qui ont souvent des origines culturelles et qui peuvent à eux seuls induire des différences non négligeables. Pensez simplement au concept d'innovation par exemple ou au mot développement qui admet différentes traductions dans certaines langues.

D'autres illustrations de l'influence de la manière dont sont formulées les questions existent. Van Bastelaer, par exemple, cite l'exemple de la statistique de l'emploi aux Pays-Bas où l'accroissement relatif du nombre de femmes qui ont un emploi rémunéré entre 85 et 88 est respectivement de 26 % ou 13 % suivant l'enquête de référence utilisée (enquête force de travail ou enquête auprès d'établissements). Selon l'auteur, cette différence est directement attribuable au mode de questionnement (Van Bastelaer, 1994). Il cite également l'exemple de l'Allemagne qui a modifié le questionnaire de son mini-census en 1990 en ajoutant une question qui faisait référence explicitement à l'exécution de travaux dits mineurs (moins de 15 jours par semaine, salaire inférieur à 450 DM et pas de contribution à la sécurité sociale), alors que ces travaux étaient déjà couverts dans les enquêtes précédentes (des instructions explicites à cet égard avaient été données aux enquêteurs). Suite à cette modification du questionnaire, la proportion de personnes exerçant un travail de cette nature est passée de 2 % en 1988 à 4 % en 1990.

La correction pour non-réponses

De plus en plus d'instituts nationaux de statistiques sont confrontés à des problèmes importants de non-réponses. Dans certains pays, il n'est pas exceptionnel pour des enquêtes non obligatoires d'observer des taux de réponse inférieurs à 50 %. La première enquête communautaire sur l'innovation est particulièrement illustrative à cet égard : 33 % en Irlande, 22 % en Allemagne, 50 % aux Pays-Bas etc. Dans des situations aussi extrêmes un examen attentif des non-répondants s'impose. Ceci peut se faire de différentes manières. Certains se sont contentés des résultats rassurants obtenus en Allemagne où aucun biais attribuable aux non-répondants n'a été mis en évidence et n'ont pas organisé d'enquête complémentaire ; ils ont simplement adapté leurs facteurs de pondération en conséquence. D'autres ont sondé les non-répondants et ont appliqué des corrections. A priori, les profils de non-répondants dans les différents pays, soumis à une même enquête harmonisée, ne devaient pas être substantiellement différents. Les corrections n'auraient pas dû changer les positions relatives des pays en matière de tendance à innover (chiffrée par le pourcentage d'entreprises innovantes). Les résultats ne confirmèrent absolument pas cette intuition. En Irlande par exemple, avant correction, le taux d'entreprises innovantes était de 71,2 % ; après correction il est devenu 33% alors qu'aux Pays-Bas, il est passé de 54,4 % à 58,4 % .

De nouveau ces résultats semblent confirmer des évidences. Ils montrent le soin qu'il convient d'apporter à des recommandations internationales si l'on veut minimiser les disparités liées aux méthodes de mesure et de traitement. Faut-il toujours imposer une enquête auprès des non-répondants ? Si non, à partir de quel seuil est-elle nécessaire ? Sous quelle forme (interrogation téléphonique, envoi d'un questionnaire simplifié, visite...) ? Il n'est pas certain que si ces choix sont laissés aux pays, ils ne soient pas source de biais comme dans l'exemple donné ci-dessus.

Les facteurs de pondération

Très souvent, l'objectif des enquêtes statistiques est d'estimer des totaux de variables sur des populations données (emploi total, somme des valeurs ajoutées, etc.). Pour ce faire, il est courant d'utiliser des estimateurs de type Horvitz-Thompson, où les observations de l'échantillon sont pondérées par les inverses des probabilités de tirage. Souvent, cependant, il est possible de prendre en compte des données auxiliaires (généralement les marges connues de certains tableaux à construire), et d'intégrer ces informations dans l'estimation du total. Cette pratique est courante dans les Etats membres et généralement non pilotée par des recommandations communautaires : quel type de données auxiliaires faut-il prendre en compte, comment "caler" les estimations à partir des données auxiliaires ? Ici encore, les différents choix possibles risquent d'induire des biais et par conséquent d'affecter, si ces biais ne sont pas mesurés ou corrigés, la comparabilité des résultats. Eurostat, dans le cadre du traitement d'une enquête sur l'impact du marché intérieur sur les entreprises, a réalisé certaines simulations qui peuvent aider à mieux comprendre l'ampleur de l'impact de ces différents choix méthodologiques. Un institut

allemand a, en effet, envoyé dans le cadre d'une enquête conduite en 1994 auprès de 1268 entreprises, des données individuelles, des facteurs de pondération et des données agrégées. Les poids fournis ont été recalculés suivant différentes méthodes :

- post-stratification à partir de données auxiliaires sur la distribution des entreprises allemandes par secteur et classe de taille (croisements) ;

- calage sur les marges du tableau précédent (c'est-à-dire totaux par secteurs et totaux par classe de taille) au moyen de différentes méthodes de minimisation de la distance entre les poids initiaux fournis par l'institut et de nouveaux poids (les distances sont calculées par des moindres carrés généralisés - MCG - par itération du quotient - IQ, par une méthode MCG restrictive, ou par une méthode IQ restrictive). Le lecteur intéressé trouvera dans (Bienvenue, 1998) plus de détails sur ces simulations.

Les résultats montrent que si le choix de la distance ne paraît pas affecter fortement les estimations des totaux (on passe, par exemple, de 41 % de "D'accord avec le fait que le programme du marché unique a été un succès pour mon entreprise " à 40,5 %), l'utilisation non seulement des marges, mais également des interactions entre secteur et taille pour post-stratifier les données a une influence considérable, certaines estimations passant de 48 % à 27 % par exemple.

Différents niveaux de comparabilité

Ces différents exemples semblent indiquer que la comparabilité est une exigence fort coûteuse. Soit les méthodes de mesure sont harmonisées, c'est-à-dire dans ce cas unifiées (et ce mot possède des connotations de pensée unique qui ne plaisent pas), et les résultats peuvent être mis en parallèle, les différences peuvent être interprétées, soit les pays sont libres à l'intérieur de recommandations communautaires d'effectuer les choix les plus indiqués pour leur système national, et la comparabilité des données devient faible. Ce dilemme a amené certains à préconiser une harmonisation à différentes vitesses ou à géométrie variable pour reprendre une expression du jargon communautaire (voir par exemple, l'évaluation de l'enquête communautaire sur l'innovation). Le niveau de comparabilité ne doit pas être identique pour toutes les statistiques. Elle a un prix que l'on n'est prêt à payer que dans des circonstances particulières, lorsque par exemple, une politique communautaire l'exige, comme pour la mesure du PNB, l'indice des prix à la consommation ou le chômage. D'un point de vue statistique, la notion de différents niveaux de comparabilité peut s'aborder autrement.

En fait la possibilité même de comparer des données renvoie à la théorie de la mesure ; elle établit quand il paraît justifié pour une mesure donnée (disons une mesure nominale comme le numéro d'une carte d'identité) de comparer, voire de combiner par des opérations arithmétiques données, des valeurs distinctes. On peut ainsi comparer des numéros d'identité et établir s'ils sont identiques ou distincts mais on ne peut pas les ordonner. Pourquoi ? Parce que mesurer c'est établir une correspondance entre un

système formel et un système empirique et que les seules opérations formelles permises sont celles qui correspondent à des opérations identifiées et sensées du système empirique : on peut comparer deux poids de manière empirique avec une balance par exemple mais on ne peut pas faire de même avec des identités : le poids est une mesure ordinale (au moins), l'identité est nominale. Malheureusement la théorie de la mesure se préoccupe assez peu à ma connaissance du problème qui nous intéresse ici, à savoir la possibilité de comparer des données issues de populations différentes.

Nous pouvons pourtant reprendre certains principes de cette théorie pour formaliser la notion de comparabilité et de niveaux de comparabilité.

Supposons par exemple que nous voulions étudier les dépenses moyennes consacrées à la recherche et au développement dans deux zones géographiques distinctes. Il apparaît raisonnable d'exiger pour que les deux moyennes puissent être comparées que

- les valeurs individuelles puissent être comparées ;

- les moyennes puissent être calculées.

En termes de théorie de la mesure, ceci signifie que les deux mesures sur les deux populations comparées doivent être des mesures d'intervalle, sinon le calcul d'une moyenne n'aurait pas de sens. Mais ceci n'est pas suffisant, pour comparer les moyennes ; il faut encore que, si a est un objet quelconque de la première population (une entreprise de la première zone géographique) et si b appartient à la seconde, l'inégalité $f(a) > g(b)$ ou $f(a) < g(b)$ - en notant respectivement f et g les mesures des dépenses de R&D dans les deux zones - ait un sens. En d'autres mots, il faut qu'il existe une relation W entre les deux populations qui soit telle que aWb si et seulement si $f(a) > g(b)$. Dans notre exemple, ceci signifie que l'on puisse comparer empiriquement les dépenses de deux entreprises de zones différentes. Remarquer que cette exigence est moins triviale qu'on ne pourrait croire car si les zones correspondent à des pays différents les dépenses sont en principe libellées en monnaies différentes et ne sont donc pas comparables a priori. De plus, si les niveaux de vie sont très différents, si les coûts d'équipement et les salaires sont beaucoup plus élevés dans une population que dans l'autre, la définition de la relation W doit être faite avec soin.

Une approche complémentaire du sens d'une comparaison de nos deux moyennes est également possible. Elle nécessite l'introduction de quelques notations. Soit A la population d'entreprises dans la première zone géographique et B la population correspondante de la deuxième zone. Les effectifs de ces populations sont respectivement N(A) et N(B). La question que nous nous posons est : quand l'affirmation $(S f(a)/ N(A)) > (S g(b) / N(B))$ a-t-elle un sens ? La théorie de la mesure ne répond pas directement à cette question. On peut cependant déduire facilement de ses principes mêmes que la comparaison a un sens seulement si (la condition est nécessaire mais sûrement pas suffisante) pour toutes transformations admissibles f de f et g de g, l'inégalité ci dessus est toujours respectée :

$(S (f \circ f)(a)/ N(A)) > (S (g \circ g)(b) / N(B))$

la classe des transformations admissibles étant définie par la nature des échelles f et g. Ceci signifie entre autres que si f et g sont libellés dans une monnaie commune (en écus par exemple), le passage à une autre monnaie commune ne devrait pas affecter le sens de la comparaison.

L'exigence de comparaison ordinale impose donc des contraintes aux mesures f et g. Mais d'autres exigences sont envisageables. Ainsi, il existe souvent un besoin fort de pouvoir constituer à partir de données relatives à des parties des totaux, des agrégations. Pour comparer les taux de chômage en Europe et aux États-Unis il faut combiner des données nationales pour pouvoir obtenir un total européen. La théorie de la mesure spécifie des conditions nécessaires et suffisantes pour que des mesures relatives à différents individus puissent être sommées, multipliées, combinées, à l'intérieur d'une même population, malheureusement, comme je l'ai déjà signalé, sans vraiment se préoccuper des problèmes que peuvent poser des mesures définies sur des populations différentes.

Pour calculer un total communautaire, il faut bien entendu pouvoir sommer des valeurs au sein d'un même pays, ce qui exige une mesure d'intervalle, c'est-à-dire, entre autres, une relation binaire o telle que $f(a \circ b) = f(a) + f(b)$ en reprenant les notations introduites précédemment. Mais ceci n'est pas suffisant. Il faut encore pouvoir additionner des mesures sur des populations différentes. Ceci nécessite d'étendre des opérations définies séparément sur des ensembles A et B à A U B. On devra ainsi s'interroger sur le sens qu'il y a à additionner des dépenses d'entreprises dans des pays différents, c'est-à-dire soumises à des fiscalités différentes et libellées initialement dans des monnaies différentes. De plus, il importe que l'opération qui associe des unités de A et de B soit de même nature que celle qui associe des unités de A ou des unités de B. Je m'explique. Si je compte par exemple des unités innovantes dans un pays A et que je les ajoute à celles d'un autre pays B, implicitement j'assume qu'il est neutre de supprimer une unité dans A et de l'ajouter dans B. Une entreprise innovante italienne est considérée en quelque sorte équivalente à une entreprise innovante belge. Nous retrouvons ici la notion d'équivalence qui soutend tous les propos de cet article, comme je l'explique dans le paragraphe suivant.

Ces considérations pourraient mener à définir différents niveaux de comparabilité basés sur les différents types de traitement que l'on veut faire subir aux données (simples comparaisons, agrégations). Cette approche fonderait théoriquement un concept dont on conçoit l'utilité, voire la pertinence, sans en apprécier les implications précises.

Si les considérations précédentes ont amené certains à proposer différents degrés d'harmonisation, elles invitent également à repenser le concept même d'harmonisation. La question n'est plus de savoir si le coût d'une méthode de mesure unique n'est pas excessif mais de s'interroger sur son sens.

L'harmonisation est-elle possible ?

J'ai abordé ce débat dans un article précédent et j'en reprends ici les éléments principaux (Defays, 1995). Une uniformisation totale des concepts et des méthodes présuppose des environnements géographiques, administratifs, légaux, économiques... identiques. Pour bien faire comprendre cet argument, utilisons des conditions extrêmes. Comparons par exemple la structure des entreprises dans un pays européen et dans un pays en voie de développement. D'un côté des unités sont répertoriées dès qu'elles exercent une activité économique, la main-d'oeuvre est enregistrée, les travailleurs ont des salaires, les systèmes postaux et téléphoniques fonctionnent plus ou moins et les unités sont habituées à remplir des formulaires, des questionnaires. De l'autre les unités sont plus instables, ne sont pas toutes déclarées, l'emploi est peut-être partiellement nomade, le troc existe toujours et les moyens de communication sont plus lents et en général moins efficaces ; l'interrogation systématique par questionnaire n'est pas possible. Que pourrait signifier dans un pareil contexte des méthodes identiques ? Ceci n'a pas de sens et on sent bien dans cet exemple que ce qui est recherché dans une harmonisation ce n'est pas une identité de concepts et de méthodes, mais une équivalence, une correspondance, un accord. Quel type de correspondance ? C'est ce que nous discutons dans le paragraphe suivant.

Quel accord entre les parties du tout pour qu'il y ait harmonisation ?

Le caractère exotique de l'exemple ci-dessus pourrait pour certains en diminuer la pertinence dans les problèmes que nous traitons dans cet article. Les pays de l'Union européenne ne sont pas si différents que cela, tout compte fait. Les mêmes concepts existent plus ou moins ou sont du moins facilement transposables. Toute personne qui fut un jour responsable de l'harmonisation d'un concept au niveau européen vous convaincra facilement du contraire. Prenons une notion aussi centrale et aussi simple a priori que la notion d'entreprise. Supposons que nous intéressions plus particulièrement aux naissances de nouvelles entreprises. Combien de créations en 1997 ? Si la définition d'une entreprise peut paraître difficile, particulièrement parce qu'il en existe de grandes avec des structures complexes, la question à laquelle nous cherchons à répondre ne concerne que les petites entreprises et paraît donc pouvoir s'aborder avec une définition simple et universelle de l'entreprise : une unité légale active. Pourtant un examen de ce concept dans les différents pays fait directement apparaître des différences notables : dans certains pays il n'est pas nécessaire d'être enregistré pour démarrer une activité commerciale : on peut en Angleterre, par exemple, sans aucune licence, ouvrir un petit magasin où l'on vend des fleurs ; de plus dans certains pays les procédures d'enregistrement sont plus longues, plus compliquées et beaucoup plus coûteuses que dans d'autres. La signification même de ce que représente une unité légale en terme de germe d'un futur opérateur économique risque d'en être affectée. Sans mentionner le fait que tous les statisticiens n'ont pas accès aux fichiers d'unités légales et risquent de devoir approcher ce concept au moyen de l'enregistrement d'unité administrative comme l'unité TVA, changeant implicitement la définition du concept. La vraie question à se poser

lorsqu'on veut comparer le nombre de créations d'unités TVA au Royaume-Uni avec ce qui se passe en France par exemple est : "quel est l'équivalent de l'unité TVA britannique en France ?". Le problème est donc un problème d'analogie : l'unité TVA est au Royaume-Uni ce que l'unité X est à la France. Le raisonnement par analogie a été largement étudié en psychologie et plus particulièrement en intelligence artificielle. La statistique pourrait, je crois, utilement s'inspirer de certains de ces travaux. Donnons un nouvel exemple pour mettre en évidence les particularités et les difficultés de ce type de raisonnement (voir D. Hofstdadter, 1995). La première dame des États-Unis est madame Clinton. Qui est la première dame de France ? Madame Chirac ? D'accord mais madame Jospin pourrait être un candidat plausible compte tenu du rôle important joué par un premier ministre en France. Et la première dame du Royaume-Uni ? Ici les choses se compliquent. Le prince Philippe est sûrement candidat mais madame Blair aussi. En effet, la reine d'Angleterre n'est pas un président et le rôle politique de M. Blair est sûrement plus important et plus proche de celui d'un président. Mais dans la notion de première dame, il y a les notions de "premier" et de "dame" et à cet égard la reine Élisabeth paraît un candidat plus approprié. On sent bien dans cet exemple que lorsque les deux structures à comparer ne sont pas identiques il importe de privilégier certaines caractéristiques de la situation initiale du type "être le conjoint du chef d'État ", ou "être très proche du pouvoir", au détriment d'autres aspects du type "être une femme" ou "être le conjoint du président" ; mais ce choix est partiellement arbitraire et dépend des inflexions que l'on veut donner au concept sous-jacent. Des problèmes similaires se posent en statistique. Comment harmoniser une date d'enquête par exemple ? Peut-on considérer que les dates de deux enquêtes - disons agricoles - organisées respectivement en Europe et en Australie sont harmonisées parce qu'elles ont lieu simultanément le 1er avril sur les deux continents ? En termes de saisons, de comportements des agriculteurs, de vacances, cette même date a des implications ou des significations différentes ; harmoniser dans ce cas consisterait sûrement à effectuer les enquêtes à des dates différentes dans le temps mais à une saison identique ou à un moment identique de l'année comptable. Comme dans l'exemple de la première dame des États-Unis, l'analogie nécessite de sacrifier certaines caractéristiques de la situation (sa date dans l'absolu) par rapport à d'autres (date relative par rapport aux saisons...). Harmoniser c'est établir des analogies. Un autre exemple emprunté cette fois à la statistique d'entreprises. Un pays X limite son enquête annuelle sur les entreprises à celles qui occupent 20 personnes ou plus. Ces entreprises représentent 90% de la valeur ajoutée du secteur manufacturier ; elles ont en moyenne un chiffre d'affaires supérieur à 2 millions de piastres (la monnaie nationale). Comment transposer ces concepts dans le pays Y dont les entreprises sont en moyenne beaucoup plus petites ? Une interprétation littérale amène à ne couvrir que les entreprises de 20 personnes ou plus. Malheureusement elles ne représentent que 80 % de la valeur ajoutée totale et leur chiffre d'affaires est en moyenne plus bas. Les deux populations sont-elles comparables ? Ne devrait-on pas prendre les entreprises qui couvrent également 90% de la valeur ajoutée totale en baissant le seuil de taille ? Nous sommes de nouveau confrontés aux choix difficiles du raisonnement par analogie : privilégier certaines caractéristiques au détriment des autres. Mais comment guider ce choix ? Sûrement en fonction des utilisations envisagées des statistiques nationales dans la mesure où elles peuvent être anticipées.

Conclusion

Nous sommes condamnés à comparer des mesures obtenues via des méthodes différentes sur des concepts au mieux équivalents. Mais la nécessité d'identifier et de quantifier des effets nationaux subsiste. Comment résoudre ce conflit ? Différentes voies de solution sont envisageables. Une meilleure compréhension des sources de différences paraît indispensable ; des études du type de celle réalisée par Statistique Pays-Bas devraient probablement être lancées de manière plus sytématique pour mieux contraster les différents choix nationaux et mesurer leur impact sur la comparabilité des résultats. Une théorie de l'erreur qui permette de chiffrer non seulement ce qui est lié aux aléas des échantillonnages mais également aux écarts observés par rapport à des prescriptions communes paraît également nécessaire. Enfin, l'utilisation de modèles dans l'analyse des résultats, comme proposé par exemple par Depoutot (Depoutot, 1997), pourrait permettre de purifier les mesures et d'obtenir des concepts, mesurés par des variables latentes, qui soient plus comparables.

Eurostat entend poursuivre et promouvoir des travaux dans ces différenes directions et invite tous ceux qui sont intéressés par le sujet à joindre leurs efforts au sien.