Dominique Joye, « Du jeu de données à la multiplication des sources. Remise en cause d'un paradigme de recherche ? »

Pour cette année 2014-2015, nous tenterons de poursuivre nos réflexions amorcées l'année précédente sur les bases de données numériques en l'étendant à la notion de corpus de recherche. En guise de point de départ de cette exploration qui se veut incrémentale, Dominique Joye (UNIL-SSP) nous fait part de quelques évolutions récentes qui, d'après lui, affectent l'analyse quantitative de données sociologiques, un domaine de recherche qu'il côtoie depuis le début de sa carrière. À travers l'exposition de « périodes » (volontairement schématiques), nous verrons que ce champ d'études a subi un certain nombre de transformations dans son rapport aux données ainsi qu'à leurs traitements.

Période 1970-1980 : le modèle du jeu de données

Les années 1970 correspondent en Suisse à la fondation de plusieurs grandes archives de données. Celles-ci étaient stockées sous forme de cartons remplis de cartes perforées et pouvaient être envoyés par la poste à des chercheurs ayant les moyens d'y mettre le prix. Selon cet idéal-type, le chercheur passait ensuite quelques années à exploiter son jeu de données pour produire des explications sociologiques de phénomènes inscrits sur ses cartes perforées.
Peu importe ici l'aspect quelque peu caricatural du tableau ; l'important est de saisir l'aspect *circonscrit*, *homogène* et *synchronique* du carton rempli de cartes perforées et capable de faire supporter aux données des modèles statistiques confirmatoires forts. Une certaine forme d'inférence statistique se met en place, des pratiques s'exercent et des habitudes d'analyse apparaissent.

Période 1980-1990 : apparition des panels

La grande nouveauté des panels qui se mettent en place dès le milieu des années 1980 est l'inclusion d'une dimension diachronique permettant de tisser des *trajectoires*. Avec ce type de jeu de données, le chercheur détient désormais des informations sur les *mêmes* individus (et leurs *mêmes* entourages) année après année, ce qui lui permet de traiter des évolutions d'agrégats. Le modèle reste – le plus souvent – celui du jeu de données homogènes et

UNIL | Université de Lausanne
LADHUL - Laboratoire de cultures et humanités digitales de l'UNIL

circonscrit, ce qui tend toujours à conférer aux données la capacité de supporter des modèles statistiques confirmatoires forts.

Période 1990-2000 : les modèles multi-niveaux

Historiquement, les modèles multi-niveaux sont issus d'études de pédagogie. Il s'agit de modèles assez puissants qui découpent la dichotomie « acteur/système » – fortement suggérée par les anciens jeux de données – en une multitude de couches qui se superposent les unes avec les autres. Soit par exemple un élève ; *en tant qu'élève*, il est doté d'un certain nombre d'attributs (e.g. genre, âge, situation familiale, lieux de vie) ; en tant qu'élève *d'une classe*, ses attributs s'insèrent dans les plis des attributs propres aux classes scolaires (e.g. nombre d'élèves, niveau scolaire) ; en tant qu'élève d'une classe *d'un établissement scolaire*, les attributs de l'élève et de sa classe s'insèrent eux-mêmes dans les plis des attributs propres aux établissements scolaires (e.g. budgets, situations géographiques, direction, etc.) ; en tant qu'élève d'une classe d'un établissement scolaire *d'un district*, les attributs de l'élève, de sa classe et de son établissement s'insèrent dans les plis des attributs des districts, etc. etc. jusqu'à – potentiellement – remonter jusqu'à l'univers tout entier.

En somme, la beauté du modèle multi-niveau réside dans le fait qu'une entité (élève, établissement, district) se singularise désormais par le déploiement de ses attributs, eux-mêmes pouvant se « brancher » sur d'autres entités dont les singularités passent également par le déploiement d'attributs, eux-mêmes pouvant se « brancher » sur d'autres entités, etc., etc.

Si les panels avaient rajouté une épaisseur diachronique aux jeux de données quantitatives, les modèles multi-niveaux leur ont rajouté une épaisseur *ontologique* : les classes, les districts ou les groupes sociaux obtiennent maintenant le droit d'être considérés comme des entités pour autant qu'ils soient spécifiés par une liste d'attributs.

Mais c'est également là que les choses commencent à se compliquer car si les différents niveaux obtiennent un poids d'existence qui permet la formulation de questions de recherche plus fines incluant davantage de variables, **le passage statistique d'un niveau à l'autre devient plus coûteux**. Soit par exemple le traitement statistique de deux niveaux en relation : il se peut tout à fait que le premier niveau soit issu d'un tirage aléatoire (par exemple un échantillon d'élèves) et le deuxième d'un tirage arbitraire (par exemple les districts lémaniques) ce qui empêcherait d'utiliser un modèle d'inférence statistique simple pour traiter de leurs relations. Une vigilance accrue au statut des couches de données utilisées devient dès lors une prérogative à un traitement quantitatif opérant sur le mode de la signification statistique.

2000-2014 : combinaison des sources d'information

Du jeu de données simple jusqu'aux modèles multi-niveaux, on peut déjà sentir un certain glissement qui engage dans le même mouvement les questions de recherche, les données d'enquête, leur saisie par des logiciels de gestion de bases de données et les outils de calculs statistiques. Car les cartes perforées laissent la place à des formats de stockage et de traitement moins volumineux, plus puissants et plus accessibles ; grâce au développement des panels, des trajectoires individuelles deviennent repérables ; grâce aux modèles multi-niveaux, des groupements d'attributs obtiennent un poids ontologique et rendent la frontière entre l'individu et son contexte de plus en plus difficile à cerner : ces trois évolutions contribuent à changer l'horizon des possibles et participent peut-être – subrepticement – à une certaine complexification des modèles d'analyse et des questions de recherche.

Ça ne sont là que des suggestions. Ce qui compte vraiment ici est l'impression générale de Dominique Joye selon laquelle, aujourd'hui, **de plus en plus de chercheurs en analyse quantitative de données sociologiques ne se limitent plus à un seul jeu de données mais « piochent » dans plusieurs pour se constituer leur propre base de données de recherche.** Et cette nouvelle forme de pratique n'apparaît pas sans conséquence, comme lorsque des chercheurs en sciences humaines moissonnent au sein d'une diversité d'archives dispersées.

Des conséquences se ressentent déjà pour les membres des institutions qui précédemment fournissaient les jeux de données qui circonscrivaient le corpus du chercheur. Aujourd'hui, les membres de ses institutions ne peuvent plus se contenter de fournir des jeux de données ; ils se doivent d'entamer **un travail d'aiguillage du chercheur vers des portions de jeux de données disparates qui n'ont pas encore été mis en relation**. A ce travail d'orientation s'ajoute également souvent un travail de formation visant à rendre ces mêmes chercheurs capables d'effectivement faire fusionner des portions de base de données différentes.

La deuxième conséquence – en lien avec la première – est une nouvelle prérogative pour ceux ou celles qui souhaitent proposer un modèle confirmatoire fort à partir d'un corpus fait de données disparates : **il leur faut d'avantage s'intéresser aux modalités de production des données**. Car dans le schéma du jeu de données unique, le chercheur était facilement capable de connaître les conditions de production de ses données et donc également de calculer les probabilités d'inclusion de chaque individu dans l'échantillon. Ceci lui permettait de faire de la statistique « classique » et proposer des explications plus ou moins significatives. Mais lorsque sont mélangées des données d'enquêtes réalisées dans des conditions différentes, **il devient très difficile d'obtenir une connaissance précise de la variance de son échantillon** et donc *in fine* de proposer des modèles confirmatoires forts. En fait, selon Dominique Joye, pour obtenir une connaissance précise de la variance de son échantillon – et donc être en droit de proposer des modèles explicatifs plus ou moins significatif –, il est nécessaire de se replonger dans les différentes façons dont ont été produites chaque jeu de données qui a été mélangé. En somme, **le prix de la signification statistique est une connaissance fine de ce que *peuvent* les données rassemblées en une même base**.

Une troisième conséquence – en lien très étroit avec la deuxième – est **une remise en cause lancinante du modèle de la signification statistique et des modèles confirmatoires forts**. Car est-il véritablement possible de payer le prix ? Est-il véritablement possible de remonter jusqu'aux modalités de production des données pour normaliser les différences de variance et poursuivre la quête de résultats significativement solides ? Le jeu en vaut-il vraiment la chandelle ? Rien n'est moins sûr puisqu'au-delà du travail de recherche nécessaire à une bonne connaissance de l'aspect qualitatif des données quantitatives, le travail proprement statistique de normalisation devient de plus en plus complexe. Ne serait-il pas plus sage – et plus *intéressant* – de privilégier d'autres analyses statistiques plus adaptées à ces nouveaux corpus de recherche constitués de données issues de différentes sources ? On pourrait par exemple penser à des analyses géométriques, mieux adaptées aux nouvelles possibilités de croisement des jeux de

UNIL | Université de Lausanne
LADHUL - Laboratoire de cultures et humanités digitales de l'UNIL

données.

En somme, rétroactivement, peut-être est-il aujourd'hui possible de faire cette observation : **il se peut tout à fait que les habitudes statistiques développées dans le champ de l'analyse quantitative de données sociologiques aient été fonction de corpus délimités pour lesquels les capabilités des données recueillies étaient facilement reconnaissables**. La toute-puissance de la signification statistique deviendrait alors contingente et fonction d'un certain type de corpus aujourd'hui de moins en moins en vogue. Si l'on accepte cette proposition, il apparait que la puissance de la signification statistique n'est pas absolue mais bien relative à l'homogénéité et la connaissance qualitative des données rassemblées dans le corpus sur lequel elle opère. À partir de là, si les corpus deviennent hétérogènes et la connaissance des données fragile, l'exploration d'autres modèles et opérations statistiques semble pour le moins judicieuse.

Cette problématique, manifeste dans le cas des sciences sociales quantitatives, fait peut-être également écho à des évolutions récentes dans d'autres domaines des sciences humaines et sociales. On peut par exemple penser à la tendance à l'agrégation des données issues d'enquêtes qualitatives ainsi qu'au moissonnage de bases de données et d'archives qui conduisent des chercheurs – notamment des historiens – à engager des travaux plus comparatifs sur des thématiques plus transversales.

Florian Jaton