Le couplage des registres de population belges, XIX^{e} -XX^{e} siècles

Michel ORIS

Introduction

Il est quelque peu irritant de produire une tantième description d'une source comme le registre de population belge, qui a déjà suscité plusieurs présentations critiques excellentes. Toutefois, le système introduit précocement en Belgique, dès 1846 grâce au génie d'Adolphe Quetelet, a connu un tel succès que son usage comme outil de gestion administrative s'est généralisé dans de nombreux pays occidentaux. C'est une tendance relativement récente, directement liée à la mise en place de systèmes informatiques performants[1]. Parallèlement, ces nouvelles potentialités techniques de traitement des données ont poussé plusieurs équipes d'historiens démographes à tenter de reconstituer des registres de populations anciennes en couplant des sources diverses, essentiellement les actes d'état civil et les listes nominatives[2]. La précision des termes n'est donc qu'apparente lorsque l'on aborde la problématique du couplage automatique des registres de population, et il importe de lever les ambiguïtés.

Après une genèse intellectuelle d'au moins un demi-siècle, l'usage du registre de population se généralise en Belgique à partir de 1846. A cette date, seuls les pays scandinaves disposent d'outils administratifs analogues. Ils s'articulent sur la transcription dans d'épais volumes des bulletins de ménages remplis lors du recensement du 15 octobre 1846, qui reprennent pour chaque individu les éléments nominatifs, le sexe, l'état civil, l'âge, le lieu de naissance et la profession. Hormis

[1] R. VERHOEF et D.J. VAN DER KAA, "Population registers and population statistics", in *Population Index*, t. 53, 1987, n° 4, pp. 633-642.

[2] Quelque nombreuses qu'elles soient, les références ci-dessous ne donne qu'un aperçu de leurs travaux.

le souci de correction scientifique qui a présidé à l'opération initiale de recension, il n'y a là rien de très original; de tels travaux ont été et continueront à être réalisés dans l'ensemble de l'Europe. Ce qui fait des registres de population des documents uniques, c'est qu'ils ne se contentent pas de ce qui pour eux n'est qu'un socle. Ils y joignent l'enregistrement de la dynamique de la population.

Qu'un individu vienne au monde, il est inscrit au bas du ménage de ses parents, qu'il décède, le fait est transcrit dans une colonne à côté de son nom, qu'il émigre vers une autre localité, celle-ci est de même indiquée ainsi que la date du mouvement. La même catégorie de renseignements est mentionnée pour les nouveaux venus, immigrants ou ménages de jeunes mariés qui s'établissent, dont l'existence est signalée dans des registres dits "de supplétifs".

Ainsi conçu, le système aurait dû permettre une détermination exacte de la population de n'importe quelle localité à n'importe quel moment. Concrètement, il présentait un certain nombre de failles, en particulier le sous-enregistrement des mouvements migratoires, surtout des sorties. Il est donc rapidement apparu qu'une remise à jour périodique était nécessaire. La loi du 2 juin 1856 consacre le principe de recensements tous les dix ans, à charge des registres fondés à partir de cette opération de fournir des données suffisamment fiables durant les périodes intercensitaires[3].

Notre but tout à fait concret à été de constituer un registre de population à partir des registres de population. En une énonciation moins paradoxale, disons qu'il s'agit de rassembler, de "coupler" toutes les informations relatives à un individu dans les cinq séries décennales qui reprennent les habitants de la ville de Huy de 1846 à 1900. Il n'est pas question de détailler ici les enjeux analytiques d'un tel travail; ce serait en soi la matière d'un gros article. A défaut, dans une introduction décidément non exempte de sujets d'irritation, tenons-nous-en à une déclaration d'intentions qui, quant à son fond, est plus classique encore que la description du registre de population.

Le couplage de la documentation nominative ne permet pas seulement l'étude déjà complexe des biographies individuelles. L'agrégation des données décrit aussi les milieux immergeants dans les limites imposées par les sources compilées. Le

[3] Pour plus de détails sur les registres de population, voir notamment : E. VAN DE WALLE et O. BLANC, "Registres de population et démographie : La Hulpe (1846–1880)", in *Population et Famille*, 1975, n° 36, pp. 113–128; M. POULAIN, "Du registre de population aux statistiques de migration interne en Belgique : critique des sources et correction des données", *Ibidem*, t. 45, 1978, n° 3, pp. 1–45; M.P. GUTMANN et E. VAN DE WALLE, "New sources for social and demographic history : the Belgian population registers", in *Social Science History*, t. 2, 1978, n° 2, pp. 121-143; R. OBOTELA, *Les recensements belges au 19e siècle (...)*, thèse inédite de l'U.Lg., 1983; N. MALPAS, *Inventaire des registres de population conservés dans la province de Liège*, Bruxelles, 1986; M. ORIS, *Economie et démographie de Huy au 19e siècle (...)*, thèse inédite de l'U.Lg., 1990–91, pp. 1080–1120.

sommet ici, c'est une société, urbaine en l'occurrence. Dans le décor de la petite ville de Huy sur Meuse, elle est plus que la troupe des figurants, elle est la pièce que composent ses multiples acteurs. Le registre de population à vocation universelle dresse la typologie de leur recrutement et de son évolution, la formation et le renouvellement d'une population et de ses diverses composantes, la fixation, la hiérarchie et le réagencement des rôles. Au-delà de la description, les mécanismes émergent et se dévoile une dynamique de la stabilité et du changement.

On se rapproche alors du vieux rêve de l'histoire totale, qu'il faut pourchasser sans illusion excessive mais avec obstination. Il n'existe aucune solution idéale pour ce faire, et encore moins de solutions simples. Traiter les registres de population, en maximiser les potentialités, est seulement une des moins mauvaises, des plus riches, des plus enthousiasmantes esquisses[4].

D'évidence, le recours à l'informatique s'imposait pour traiter une masse de quelque 100.000 notices individuelles. Mais au-delà de l'expérience hutoise, la richesse de la source, les enjeux historiographiques qui dépendent de son exploitation, sont tels qu'il était impensable de ne pas en tirer un système général, apte à traiter aussi bien les populations réduites issues d'un sondage que les bases les plus massives, telle celle de Seraing avec son quart de million de notices. Ces exigences ont orienté une réflexion qui a pu s'appuyer sur des travaux nombreux, importants, et ... intransposables au cas belge. L'inspiration qu'ils offrent est particulièrement féconde lorsqu'il s'agit de définir un système probabiliste nécessairement imparfait[5].

A l'origine, fin des années 1960, début des années 1970, le thème du couplage était familier à divers chercheurs, en particulier à des médecins[6]. C'est d'ailleurs à un statisticien et épidémiologiste de l'agence atomique du Canada, H.B. Newcombe, que l'on doit l'introduction de la problématique et de ses techniques de base en

[4] Opinions convergentes de F. NAULT et B. DESJARDINS, "Recent advances in Computerized population registers", in *Historical Methods*, t. 21, 1988, n° 1, p. 31; ou de J.D. WILLIGAN et K.A. LYNCH, *Sources and methods of historical demography*, New-York, 1982; G. BOUCHARD, Ch. POUYEZ, R. ROY, "L'avenir des fichiers de population dans les sciences humaines : le projet de fichier-réseau de la population saguenayenne", in *Etudes Canadiennes*, décembre 1980, n° 9, pp. 31–45.

[5] Pour un historique du couplage et une présentation des "parrains" de cette étude, on peut se reporter à G. BOUCHARD, R. ROY, B. CASGRAIN, *Reconstitution automatique des familles. Le système* SOREP, t. 1, Chicoutimi, 1985, pp. 5–15, et M. ORIS, *Economie et démographie... op. cit.*, pp. 1143 sv. Sur l'ensemble du processus probabiliste, nous donnons beaucoup plus de détails dans le même travail, pp. 1146–1207.

[6] I. WINCHESTER, "A brief survey of the algorithmic, mathematical and philosophical literature relevant to historical record linkage", in E.A. WRIGLEY (éd.), *Identifying People in the past*, Londres, 1973, pp. 128–154.

démographie historique[7]. En particulier, il a diffusé la méthode probabiliste, qui consiste à accorder à chaque élément pris en compte une valeur inverse à sa fréquence, une donnée rare étant plus discriminante qu'une occurrence courante[8]. On le voit, le postulat probabiliste est élémentaire, ce qui le rend dans une large mesure inattaquable. Pourtant, une objection de taille naît immédiatement de son application à des données réelles. Celles-ci peuvent diverger radicalement (DUPONT-DEMARET) ou de manière mineure (DUPONT-DUPOND).

Comment définir la gravité d'une variation et l'intégrer dans le processus logique de prise de décision? Depuis ses premiers théoriciens, toute la problématique du couplage est enfermée dans cette interrogation. En première approximation, l'ordinateur, apte à traiter de grandes masses de données, est une bête simpliste qui tranche une alternative élémentaire, égalité ou dissemblance, sans tenir compte des nuances qu'introduit le raisonnement humain. Celui-ci n'est pas pour autant la référence, comme on le lit encore trop souvent. Ce qui en lui ne peut être formalisé, détaillé, structuré, bref programmé, est sa subjectivité qui, jusqu'à preuve du contraire, n'est pas une qualité. Si la programmation semble moins performante, c'est d'une part en raison de contraintes techniques, obstacle que le développement de l'intelligence artificielle lève de plus en plus, d'autre part parce qu'elle avoue franchement ses failles en explicitant ses règles, ce que ne fait pas ou pas assez la démarche humaine, dite manuelle. Il faut donc affirmer avec force que l'approche informatique est la plus saine et aborder sans complexe un problème complexe[9].

---

[7] H.B. NEWCOMBE, "Products from the early stages in the development of a system of linked records", in E.D. ACHESON (éd.), *Record Linkage in Medecine*, Edimbourg et Londres, 1968, pp. 295–303; ID., "Couplage de données pour les études démographiques", in *Population*, t. 24, 1969, n° 4, pp. 653–684; ID. et M. SMITH, "Changing patterns of family growth. The value of linked vital records as a source of data", in *Population Studies*, t. 4, 1970, pp. 193–203.

[8] G. BOUCHARD, R. ROY, B. CASGRAIN, *Op. cit.*, pp. 6–8, détaillent davantage les apports de Newcombe.

[9] Une évolution du scepticisme ou de l'inquiétude à l'affirmation est évidente au travers des travaux de I. WINCHESTER, "The linkage of historical records by man and computer: techniques and problems", in *The Journal of Interdisciplinary History*, t. 1, 1970, n° 1, pp. 107–124; L. HENRY, "Simulation d'une reconstitution de familles par ordinateur", in *Annales de Démographie Historique*, 1972, pp. 245–250; A. CHAMOUX, "La reconstitution des familles : espoirs et réalités", in *Annales E.S.C.*, t. 27, 1972, n° 4–5, pp. 1083–1090; G.A. CONDRAN, "Linkage of the 1880–81 Philadelphia death register to the 1880 manuscript census : a comparison of hand- and machine-record linkage technique", in *Historical Methods*, t. 14, 1981, n° 2, pp. 73–84; G. BOUCHARD, R. ROY, B. CASGRAIN, ... *Le système* SOREP... *op. cit.*, pp. 485–487. Evidemment, comme le notent les derniers cités, il faut tenir compte du contexte technique dans lequel se sont faits les premiers essais.

I. Les critères et leur préparation

Le premier mode de gestion des variations est leur résorption préalable. Cette opération implique nécessairement une perte d'information; aussi ne faut-il prendre que des décisions aussi neutres que possible. A ce stade, les lieux de naissance sont remplacés par des codes communaux. La logique et les principes de ce travail sont assez connus pour qu'il ne soit pas nécessaire d'y revenir[10]. Le cas des prénoms est un peu différent mais peut se traiter de manière analogue. Dans l'un et l'autre cas, le résultat est la réduction des variantes mineures d'une même valeur, soit un objectif modeste mais qui, par sa modestie même, exclut le subjectivisme qui irait de pair avec toute tentative d'associations de formes plus poussée[11]. Cet obstacle devient apparent dès qu'il s'agit de gérer le stock des patronymes.

De nombreux chercheurs ne disposent dans leur corpus documentaire que de cette variable, ainsi que des prénoms qui sont beaucoup moins discriminants. Aussi de nombreuses méthodes de traitement ont-elles été proposées, que l'on peut ramener à quatre types.

- 1. Les trames et lemmatisations, qui tentent de dégager du mot son squelette et qui, malgré leur parti pris de généralité, se sont révélées peu ou pas du tout applicables à des ensembles linguistiques hétérogènes, telle la Belgique du 19e siècle, à la population brassée par les intenses mouvements migratoires induits par la révolution industrielle[12].

---

[10] C. DESAMA, "Informatique et histoire. Le codage automatique des professions et des provenances", in *Revue. Informatique et Statistique dans les Sciences humaines*, t. 22, 1986, n° 1–4, pp. 67–82; M. ORIS, "Les relations entre bases de données nominatives et agrégées comme outil pour une analyse différentielle des migrations (...)", communication au *III^{e} Colloque international des Démographes de Langue française.* Montréal 7–10 juin 1988.

[11] Les résultats sont d'ailleurs déjà fort appréciables puisque parmi les prénoms, 2.324 mentions de base ont été réduites à 876 codes, soit 37,7 % du total initial.

[12] C. DESAMA, "Le couplage automatique des données en démographie historique", in *Annales de Démographie historique*, 1972, pp. 297–301; ID., *Les structures de la population verviétoise pendant la révolution industrielle (1799–1850). Essai d'une méthodologie de traitement par ordinateur de la documentation historique*, thèse de l'U.Lg., 1978, pp. 240–242. Sur les rapports entre linguistique, peuplement et composition du stock patronymique, E. HÉLIN, "La dénomination des personnes dans quelques régions de la Belgique francophone", in L. HENRY (éd.), *Noms et prénoms: aperçu historique sur la dénomination des personnes en divers pays*, Dolhain, 1974, pp. 21–32. Sur les trames les plus anciennes et les plus connues : L. HENRY, "Variations des noms de famille et changements de prénom. Problèmes qui en résultent pour le couplage automatique des données", in *Annales de Démographie historique*, 1972, pp. 245–250; H. CHARBONNEAU, Y. LAVOIE, J. LEGARE, "Etude des caractéristiques nominatives dans l'état civil et les recensements canadiens du 17^{e} siècle. Prélude à l'utilisation des ordinateurs pour le jumelage des données", *Ibidem*, pp. 269–295, qui présentent en annexe les règles du "Russel Soundex Code" (p. 292) et de la méthode Henry (pp. 293–295); C. STEPHENSON, "The methodology of historical census record linkage : a user's guide to the Soundex", in *Journal of Family History*, t. 5, 1980, n° 1, pp. 112–115.

- 2. Les codes phonétiques, particulièrement nombreux, qui suscitent une critique semblable[13]. On peut en outre leur reprocher qu'une transduction phonétique ne traite qu'un type d'erreur et peut être une source de confusion supplémentaire appliquée à des écarts d'une autre nature[14].
- 3. Mêlant les deux idées, Chiaramella a commencé à la fin des années 1970 à développer un transducteur phonétique, capable de produire plusieurs niveaux de simplification, du toilettage élémentaire à la trame en passant par la phonétisation. Cette solution a été reprise par l'équipe de Salt Lake City avec le souci d'en accroître la "portabilité", l'applicabilité, puisqu'il s'agit de traiter l'immense base de données des Mormons, laquelle rassemble des individus de toute origine[15].
- 4. Au Québec, l'équipe de SOREP a dressé une typologie des variantes de manière à leur appliquer un traitement propre. Cette méthodologie prévoit un code phonétique mais aussi une "table universelle des équivalences", sorte de dictionnaire des formes et de leurs associations[16]. Le P.R.D.H. de Montréal

---

[13] Il en existe pour la France avec L. HENRY, "Projet de transcription phonétique des noms de famille", *in Annales de Démographie historique*, 1976, pp. 201–214; pour une partie du Québec avec G. BOUCHARD, P. BRARD, Y. LAVOIE, "FONEM : un code de transcription phonétique pour la reconstitution automatique des familles saguenayenne", in *Population*, 1981, n° 6, pp. 1085–1103; pour les Pays-Bas avec M. BOOT, P. LOURENS, J. LUCASSEN, "A linguistic pre-processor for record linkage in socio-economic historical research", in *Computers and the Humanities*, t. 17, 1983, pp. 45–64; pour la Belgique avec J.P. BOUGARD, "La reconstitution semi-automatique des familles par la standardisation des patronymes. Le cas de Wasmes et Warquignies, 1803-1850", in *Actes du Congrès de Comines du 28–31 août 1980*, Comines, 1983, pp. 103–136; pour l'Espagne avec D.S. REHER, "Introduction à l'étude de l'information nominative à partir de la mise sur ordinateur des archives paroissiales espagnoles", in *Annales de Démographie historique*, 1984, pp. 137–145. Et, plus que vraisemblablement, la liste n'est pas exhaustive...

[14] Formalisation très intéressante de Y. CHIARAMELLA en terme de modèles de variations indépendants ou complémentaires : *Traitement de données ambiguës dans un système de base de données. Application aux bases de données démographiques*, thèse de l'Université scientifique de Grenoble et de l'Institut national polytechnique de Grenoble, 1981, pp. 52–62.

[15] Voir M. CHIARAMELLA avec M. BORNAREL, "Un programme interactif de reconstitution de population : 'Mercure'. Application à la population de la paroisse Saint-Laurent au 18e siècle", in M. SKOLNICK *et al*, "Colloque sur les méthodes de reconstitution automatique des familles, in *D.H.*, 1978, pp. 24–26, puis avec M. SKOLNICK, "Couplages d'informations : quelles perspectives nouvelles pour la reconstitution automatique des populations?", in *Informatique et Sciences humaines*, n° 56–57, 1983, pp. 57–74, ainsi que divers travaux cités *infra* ou *supra*, en particulier sa thèse de doctorat, qui est un des deux ou trois travaux qui ont renouvelé la problématique du couplage au cours de la dernière décennie.

[16] La typologie est la suivante : variations graphiques ou orthographiques superficielles, variations phonétiques mineures et majeures (diminutifs), dissociations d'éléments composés, substitutions. Voir G. BOUCHARD et C. POUYEZ, "Name variations and computerized record linkage", in *Historical Methods*, t. 13, 1980, n° 1, pp. 119–125; G. BOUCHARD, R. ROY, "Les variations nominatives dans

a récemment développé le caractère dynamique de cette dernière technique en lui associant une dimension généalogique, dans un système de registre de population évolutif qui postule la préexistence de fiches reconstituées[17].

Les solutions de type 3 et 4 constituent le sommet actuel de la recherche en matière de traitement qualitatif des éléments nominatifs. Toutefois, indépendamment de l'impossibilité concrète d'appliquer immédiatement les méthodes du type 4, il n'est pas évident qu'il soit nécessaire de pointer sur les données belges une artillerie aussi lourde. Le choix d'une démarche doit avant tout être fonction des difficultés rencontrées. Or, de premières études par sondage dans trois communes au milieu du 19^{e} siècle ont donné des résultats très encourageants.

Les divergences entre deux patronymes ne se présentent que dans des proportions de 7 % à Amay, entité mi-rurale, mi-industrielle, de 16 % à Tilleur et 17 % à Seraing, en plein bassin d'industrie lourde. Par rapport aux expériences menées à Verviers sur des données de la première moitié du siècle, l'amélioration est notable[18]. Divers auteurs ont eux aussi mis en évidence une régression des problèmes depuis le 17^{e} jusqu'au 20^{e} siècle[19]. Cette évolution renvoie à une causalité complexe qui n'est pas assez documentée à ce jour (alphabétisation, unification linguistique, civilisation de l'écrit dont la machine administrative, en ce compris dans ses pratiques de recensement et de tenue d'un état civil, est un levier puissant, etc).

Quoi qu'il en soit, profitant de ces conditions favorables, nous nous sommes contenté de recourir à un logiciel liégeois qui rassemble vingt et une règles de standardisation, à mi-chemin de la trame et de la phonétisation[20]. Comme nous

---

les registres paroissiaux du Saguenay", in *Annales de Démographie Historique*, 1982, pp. 354–368; ID., "La reconstitution automatique des familles et le problème des substitutions nominatives : mesures de rendement", in *Informatique et Sciences Humaines*, t. 14, 1982, n° 56–57, pp. 75–94; G. BOUCHARD, R. ROY, B. CASGRAIN ... *Le système SOREP... op. cit.*, pp. 111–139.

[17] B. DESJARDINS et F. NAULT, "Recent advances... *op. cit.*, p. 32. L'idée de tels dictionnaires semble remonter aux études sur la vallée de Parme de M. SKOLNICK, L.L. CAVALLI-SFORZA, A. MORONI, E. SIRI et L. SOLIANI, "A reconstruction of historical persons from the parish registers of Parma Valley, Italy", in *Genus*, t. 29, 1973, n° 3–4, pp. 103–155.

[18] M. ORIS, "Le couplage automatique des données dans les registres de population liégeois", in *DH Bulletin d'Information*, n° 46, 1986, pp. 22–25; C. DESAMA, "Le couplage automatique des données nominatives : nouvelles perspectives", in *Historiens et Populations. Liber Amicorum Etienne Hélin*, Louvain-la-Neuve, 1991, p. 123.

[19] R.S. SCHOFIELD, "The standardization of names and the automatic linking of historical records", in *Annales de Démographie Historique*, 1972, pp. 360–361; J.P. BARDET, "Les variantes des noms de personnes à Rouen aux 17^{e} et 18^{e} siècles", in *Idem*, p. 255; Y. BLAYO, "Name variations in a village in Brie, 1750–1860", in E.A. WRIGLEY (éd.), *Identifying... op. cit.*, pp. 57–63; G. BOUCHARD, R. ROY, B. CASGRAIN, ... *Le système SOREP... op. cit.*, pp. 258–259.

[20] S. PASLEAU, *Une population dans le développement économique (...). Seraing 1846–1914*, thèse inédite de l'U.Lg., 1990, pp. 301–308, 314–315. Cette solution est proche de celle adoptée par

le verrons, cette opération a peu modifié le tableau des fréquences d'apparition des patronymes. Cela s'interprète moins en terme de carences du programme, même s'il n'est pas parfait, qu'en raison du peu de variations mineures entre des formes proches. C'est la confirmation indirecte de la stabilisation des mentions patronymiques évoquée dans le paragraphe précédent.

Parmi les variables susceptibles de suppléer aux carences résiduelles dans l'enregistrement des noms de famille, la date de naissance est sans conteste la meilleure. Dans leur projet de reconstitution de la population insulaire, les chercheurs islandais lui ont attribué un rôle privilégié; dans la Belgique contemporaine, elle constitue la première clé d'accès au registre national informatisé[21]. C'est toujours la date qui a été utilisée ici, quitte à calculer automatiquement une année quand un âge seul était connu.

Lors de la minutieuse préparation du recensement de la population de 1846, les commissions provinciales de statistique d'Anvers, du Brabant et de Liège s'accordent à récuser la demande de la date de naissance, car "il importe de ne demander que des renseignements que chacun puisse facilement donner (...) bon nombre de personnes, même de la classe élevée, ignorent cette date"[22]. Pour les mêmes raisons, l'idée de demander d'exprimer l'âge non seulement en années mais aussi de le préciser en mois est rejetée comme aberrante.

Il est rapidement apparu que l'âge s'intégrait mal dans un document dynamique comme le registre de population. Il en résultait toute une série d'ambiguïtés, en particulier pour l'inscription des nouveau-nés, qui par définition ont tous 0 an, ou pour ceux qui s'établissent après le recensement alors qu'ils sont issus de la localité, donc dépourvus de date d'entrée en observation pour peu que celle de mariage n'en tienne pas lieu. En 1846, 5,2 % des années de naissance n'ont pu être calculées à partir de l'âge, faute d'une référence à partir de laquelle la soustraction puisse s'opérer.

Cette proportion aurait été plus importante si, spontanément, plusieurs administrations communales ne s'étaient mises à signaler la date de naissance[23]. A Huy, l'âge n'est mentionné systématiquement que dans le registre de 1846 pour les individus recensés au 15 octobre. Pour ceux qui sont arrivés après, c'est la date qui est indiquée

---

J.P. BOUGARD, *op. cit.*, et du même : *Les familles de Wasmes et de Warquignies aux 18e et 19e siècles. Essai de reconstitution automatique*, thèse inédite de l'U.C.L., 1978.

[21] S. FRIDRIKSSON, "The Icelandic demographic records and their linking", in *Méthodes de reconstitution automatique des familles. Colloque de l'I.U.S.S.P.*, Florence, 1977.

[22] *Bulletin de la Commission centrale de Statistique*, t. 2, 1845, p. 65.

[23] C'est le cas à Amay et à Tilleur. M. ORIS, "Le couplage..." *op. cit.*, pp. 23-24.

dans un peu moins de 75 % des cas. Elle est alors pratiquement toujours complète, reprenant non seulement l'année mais aussi le jour et le mois (94,5 % des dates).

Dès 1856, dix ans avant que cela ne devienne une obligation, ce renseignement précis devient la mention normale. Un calcul à partir des déclarations d'âge n'est plus nécessaire que dans 8 % des cas. A partir de 1866, il n'y est plus procédé qu'accidentellement. Toutefois, un autre problème se présente. Quelque 14 % des dates de naissance en 1866, 9 % en 1890 et surtout 32 % en 1880 sont incomplètes, limitées à l'année. Ce phénomène est largement dû à une transcription partielle des bulletins de recensement dans les registres de population. En effet, il n'y a guère que des individus présents lors de leur mise en service qui soient affectés.

II. L'apport des critères

L'apport des critères de couplage est mesuré par le tableau n° 1, qui reprend deux indices proposés par Chiaramella[24]. L'entropie désigne la quantité moyenne d'information apportée par chaque type de variables. Elle s'obtient par la formule $H = -\sum_{i=1}^{n} p_i \log_2 (p_i)$, où p_i est la probabilité d'apparition d'une valeur i, qui correspond à l'inverse de sa fréquence.

Tableau 1

Entropie et pouvoir discriminant des variables retenues pour le couplage

	Entropie	pouvoir discriminant
Patronyme	10,94	9,70
Patronyme normalisé	10,57	9,49
Prénom	6,36	4,72
Date de naissance	14,29	13,98
Année de naissance	6,67	6,50
Année + mois de naissance	10,07	9,87
Lieu de naissance	4,34	1,54

[24] V.H. WESLEY, S.M. DINTELMAN, A.T. MANESS, Y. CHIARAMELLA, G.P. MINEAU, L.L. BEAN, M. SKOLNICK, "Couplage d'informations dans un système de gestion de bases de données généalogiques", in *Revue, Informatique et Statistique dans les Sciences humaines*, t. 23, 1987, pp. 195.

Tableau 2
Distribution des quantités d'informations par type de variables

	Chiffres bruts					chiffres relatifs				
QUANTITE D'INFORMATIONS	PATRO-NYMES	PATRO-NYMES NORMA-LISES	PRE-NOMS CODES	DATE DE NAIS-SANCE	LIEUX DE NAIS-SANCE CODES	PATRO-NYMES	PATRO-NYMES NORMA-LISES	PRE-NOMS CODES	DATE DE NAIS-SANCE	LIEUX DE NAIS-SANCE CODES
0 à 0,99	0	0	0	0	58249	0,00	0,00	0,00	0,00	58,44
1 à 1,99	0	0	0	0	0	0,00	0,00	0,00	0,00	0,00
2 à 2,99	0	0	25180	0	0	0,00	0,00	12,95	0,00	0,00
3 à 3,99	0	0	22787	0	0	0,00	0,00	11,72	0,00	0,00
4 à 4,99	0	0	7652	0	0	0,00	0,00	3,93	0,00	0,00
5 à 5,99	0	0	12749	0	8348	0,00	0,00	6,56	0,00	8,38
6 à 6,99	0	0	53918	0	3164	0,00	0,00	27,73	0,00	3,17
7 à 7,99	6706	7538	31109	7507	4151	6,60	7,41	16,00	7,46	4,16
8 à 8,99	14407	18845	18216	10189	4333	14,17	18,53	9,37	10,12	4,35
9 à 9,99	22718	23422	9211	5498	4936	22,34	23,03	4,74	5,46	4,95
10 à 10,99	16413	16925	6008	912	5076	16,14	16,65	3,09	0,91	5,09
11 à 11,99	12721	12255	3264	378	3238	12,51	12,05	1,68	0,38	3,25
12 à 12,99	8819	7954	1965	1669	2500	8,67	7,82	1,01	1,66	2,51
13 à 13,99	6559	5145	999	16277	2036	6,45	5,06	0,51	16,17	2,04
14 à 14,99	4537	3394	570	26765	1307	4,46	3,34	0,29	26,59	1,31
15 à 15,99	4467	3426	451	23174	1303	4,39	3,37	0,23	23,02	1,31
16 à 16,99	4334	2777	194	8294	1027	4,26	2,73	0,10	8,24	1,03
17 à 17,99	0	0	196	0	0	0,00	0,00	0,10	0,00	0,00
TOTAL	101681	101681	194469	100663	99668	100	100	100	100	100
Moyenne	10,94	10,57	6,36	13,19	4,34	Mode	16,63	16,63	2,95	15,03	0,77
						1è quartile	8,97	8,86	7,92	8,34	9,23
Ecart type	2,34	2,19	2.52	2,80	4,65	Médiane (Me)	9,69	9,66	9,75	9,16	10,35
						3è quartile	10,69	10,67	11,45	10,98	11,38
Coeff. de variat°	0,21	0,21	0,40	0,21	1,07						
						Asymétrie (skewness) :					
Quantité d'information :											
minimale	7,06	7,06	2,95	7,58	0,77	Coefficient de Yule = (Q1 + Q3 - 2 Me) / (Q3 - Q1)					
maximale	16,63	16,63	17,57	16,62	16,60						
							0,16	0,12	-0,04	0,38	-0,04

Les chiffres confirment le constat empirique tiré d'une première enquête. La date de naissance est la source d'information la plus riche tandis que l'apport du lieu est faible. Les éléments nominatifs sont en position intermédiaire, les patronymes se révélant comme partout plus instructifs que les prénoms. Avec une entropie de 10,94, ceux qui sont présents dans la base de données hutoises sont un peu moins intéressants que ceux de la généalogie de l'Utah, qui arrivent à 11,91[25]. Le pouvoir discriminant est évalué de la manière suivante : $D = -log_2[1/\sum_{i=1}^{n}(p_i^2)]$. La hiérarchie qui se dégage de son examen est logiquement semblable.

Ces indicateurs n'ont toutefois qu'une valeur de résumé, de moyenne, dont ils ont les défauts. Or toutes les séries concernées se caractérisent par une hétérogénéité plus ou moins prononcée. Il est possible, à partir des fréquences, de calculer les quantités d'information propres à chaque valeur, à chaque forme distincte[26]. Le tableau n° 2 résume ces opérations tout en indiquant les principales caractéristiques de chaque distribution, pour lesquelles le graphique n° 1 fournit une visualisation.

Une première famille est constituée par les variables fortes, le patronyme et la date de naissance. Pour le premier, la série n'est certes pas une classique courbe de Gauss, mais c'est celle qui s'en rapproche le plus, qui a l'écart-type et le coefficient de variation le plus faible. Le peu d'effet de la normalisation des patronymes est évident puisque l'opération ne réduit l'entropie, soit l'information moyenne, que d'à peine 3,38 %, alors que l'écart-type de la série ne régresse que de 2,34 à 2,19.

Les fréquences des dates de naissance présentent par contre une structure à deux composantes. Un groupe dominant, qui conditionne d'ailleurs la moyenne, rassemble les quantités d'information entre 13 et moins de 17, particulièrement de 14 à 15,99 (près de 50 % des observations). Un groupe secondaire se situe entre 7 et 9 (23 %). En dehors de ces deux agglomérats, il n'y a pratiquement rien. Ces traits indiquent que deux populations distinctes sont arbitrairement mêlées : les dates calculées et incomplètes, limitées à l'année de naissance, et les dates complètes, reprenant alors le mois et le jour. Un retour au tableau n° 1 montre que les premières ont une entropie faible, de l'ordre de celle des prénoms, alors que l'association mois-année, quand elle se présente, atteint déjà le niveau des patronymes.

Ces derniers et les dates présentent des distributions asymétriques positives, ce qui signifie que les valeurs les plus proches de la quantité d'information maximum sont mieux représentées que dans une courbe gaussienne classique. A l'inverse, les prénoms et les lieux de naissance se distinguent par une asymétrie négative. Les prénoms constituent une variable un peu particulière car chaque individu peut en

[25] *Idem*, p. 194.

[26] Il s'agit simplement du logarithme en base 2 de l'inverse de sa fréquence relative d'apparition.

avoir plusieurs, chacun représentant une quantité d'information supplémentaire[27]. Pour le couplage, seules les quatre premières mentions ont été utilisées, et encore aurait-on pu sans dommage se limiter aux trois premières. Près de 68 % des personnes reprises dans la base ont au moins deux prénoms, un peu plus de 20 % au moins trois mais ils ne sont plus que 2,7 % à en avoir quatre. Ces valeurs globales camouflent une évolution et une différenciation interne constante. Entre les registres de 1846 et 1890, la proportion des prénoms de rang 2, exprimée en pour-cent de ceux de rang 1, passe de 56 à 81 %, le rang 3 grimpe de 10 à 33 %.

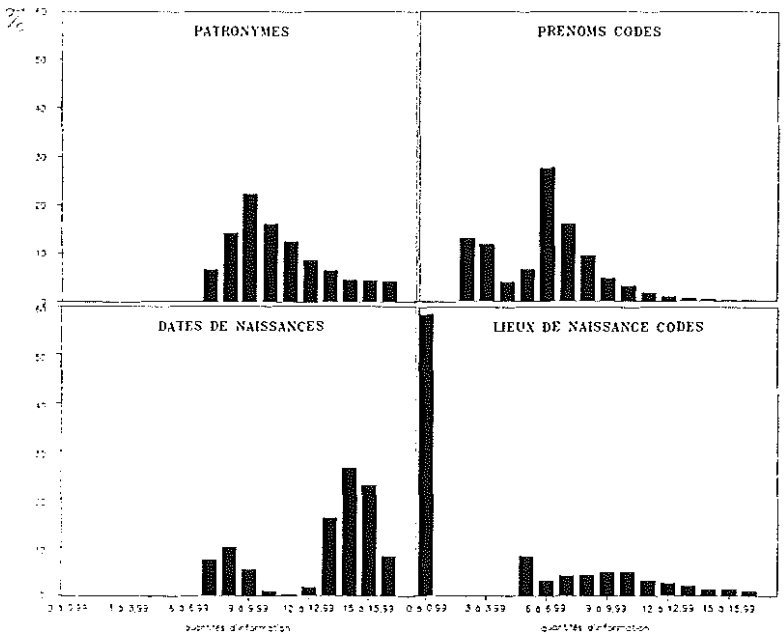
Les prénoms sont beaucoup moins discriminants que les patronymes. Le nom qui apporte la quantité d'information la plus faible, soit 7,09, a une fréquence absolue de 761 occurrences présentes dans la base. Pas moins de 62 prénoms font mieux et rassemblent 63 % des mentions analysées. Ce score est notamment dû à l'inclusion dans ce groupe de la tranche modale, de 6 à moins de 7 (155 à 178 occurrences), qui correspond à lui seul à 27 % (Graphique n° 1). En tête, Joseph et Marie se distinguent nettement puisqu'on les retrouve à peu près une fois sur quatre. Ils représentent une information à peu près nulle.

Les lieux de naissance forment la série la plus hétérogène. En fait, comme le montre éloquemment le graphique n° 1, il faut distinguer les natifs de Huy, qui sont si nombreux (58,44 %) que cette indication n'a aucune valeur, de tous les autres. Pour eux, le lieu constitue une information d'un bon niveau. Dans plus de deux cas sur cinq, la quantité excède 10.

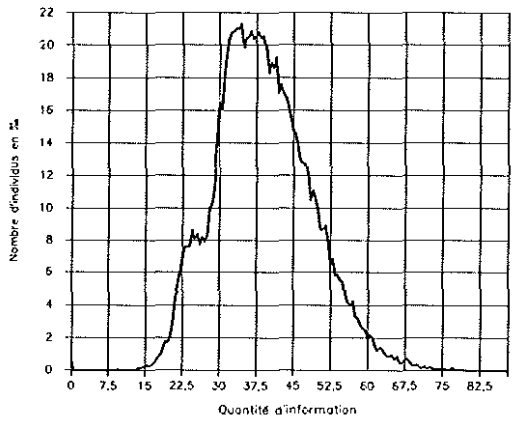
Au-delà des interrogations sur l'apport des critères isolés, il convient encore de s'interroger sur la manière dont ils s'agrègent au sein d'une notice individuelle. Le graphique n° 2 représente une série pratiquement gaussienne avec pour moyenne pondérée une valeur de 39,14, soit plus du double de l'élément isolé le plus discriminant. L'écart-type est faible et la coupole des quantités d'informations les plus fréquentes, entre 19,7 (moyenne–2 écarts-types) et 58,55 (moyenne+2 écarts-types), rassemble plus de 96 % des observations. Les notices sur- et sous-documentées vont nécessairement mal s'intégrer dans un processus probabiliste;

[27] Dès le stade du dépouillement, il a été décidé que l'on ne se hasarderait pas sur le terrain des distinctions contestables entre prénoms simples, uniques, composés, doubles, multiples. Toutes les formes distinctes, même celles explicitement associées par un tiret, ont été séparées et considérées comme une occurrence particulière. Sur les typologies et sur les considérations qui fondent notre choix, voir J. DUPAQUIER, "Introduction", in *Le prénom, mode et histoire. Les entretiens de Mahler 1980*, Paris, 1984, et J.C. SANGOI, "La prénomination en Bas-Quercy, 18^{e} –19^{e} siècles", in *Annales de Démographie historique DH Bulletin d'Information*, n° 51, 1988, pp. 22–23.

Graphique 1
Répartition des critères selon leur quantité d'information



Graphique 2
Répartition des notices individuelles selon leur quantité d'information



nous en ferons la démonstration plus loin. Toutefois, elles forment un résidu trop modeste pour remettre l'option méthodologique en cause.

III. La détermination d'une population cible

Pour décrire le produit développé, nous recourrons à un système de notation mathématique classique. Il impose au lecteur un certain effort de mémorisation, mais nous croyons que le gain en clarté le vaut. Chaque individu X_i est individualisé par ses attributs a. Par commodité le patronyme sera a_1, les quatre premiers prénoms a_2 à a_5, la date de naissance a_6, le lieu de naissance a_7. En outre, on dispose de a_{1*}, soit le patronyme normalisé, de a_{6*}, soit l'année et le mois de naissance, et a_{6**}, soit l'année seule. Leur valeur ou quantité d'information est désignée par la lettre v (de v_1 à v_7). La quantité d'information totale d'une notice individuelle est V_i. La population cible d'un individu donné est $P(X_i)$.

Définir cette dernière est classiquement réservé à la phase d'appariement, qui consiste à former des sous-ensembles de données de taille raisonnable. Cette réunion ne préjuge pas de la décision d'accepter un lien, prise dans la phase dite de jumelage[28]. L'équipe de SOREP a, à juste titre, noté que l'empirisme est souvent la règle lors de l'appariement, ce qui se traduit par des exposés laconiques sur la question, alors que la prise de décision a suscité de nombreux débats et des réflexions théoriques approfondies[29].

Dans l'expérience rapportée ici, nous refusons cette scission des opérations à la fois pour des raisons techniques et logiques. On peut postuler que la population cible complète d'un individu est composée de l'ensemble des personnes qui ont au moins un paramètre en commun avec lui, sous réserve que tous les attributs sans exception ne soient l'objet de variations. D'évidence, $P(X_i)$ atteindra régulièrement une taille telle que le coût du jumelage deviendra prohibitif[30]. Procéder ainsi serait d'autant plus absurde que la dimension de l'ensemble sera d'autant plus grande que l'élément pris en compte est peu discriminant, donc que la probabilité qu'il puisse fonder une décision d'acceptation est faible.

Il faut donc adopter une démarche inverse et définir une population cible à partir d'attributs qui représentent une information réelle. Soit $P1(X_i)$ est l'union des populations dotées d'un attribut $k(P_{ak})$. Les éléments a sont considérés du plus au moins discriminant, jusqu'à ce que la taille maximale (n=150) soit dépassée. Il

---

[28] M. SKOLNICK *et al.*, "Colloque sur les méthodes... *op. cit.*, pp. 31–32. Ce sont les Canadiens du P.R.D.H. qui ont popularisé ce vocabulaire et cette scission des opérations.

[29] G. BOUCHARD, R. ROY, B. CASGRAIN, ... *Le système SOREP... op. cit.*, pp. 74–76.

[30] Sur la croissance exponentielle des coûts à ce stade, voir R.S. SCHOFFIELD, *The standardization... op. cit.*, pp. 359–360.

est entendu que plusieurs critères envisagés successivement peuvent introduire dans l'ensemble une même notice individuelle, qui ne compte que pour une unité.

Le pari que nous faisons est d'affirmer que $P1(X_i)$ contient l'ensemble des X susceptibles d'être couplés avec X_i. En effet, d'après la logique de décision probabiliste retenue pour le jumelage, les individus inclus dans la population cible sur base de critères peu discriminants, dans le cadre d'une approche laxiste, n'atteindront le seuil d'acceptation que s'ils présentent une égalité sur d'autres critères discriminants, lesquels conduiraient de toute façon à les intégrer à $P1(X_i)$ dans l'approche par sélection. Ce postulat est soumis à d'autant plus de restrictions que grande est la variabilité des données discrimantes et que leur nombre est petit, donc que la probabilité d'une alternative est réduite[31].

Or il ne s'agit pas là d'une hypothèse gratuite puisque, comme dit plus haut, 58 % des individus repris dans la base ont pour lieu de naissance Huy, qu'un sur quatre se prénomme Joseph ou Marie, que les années de naissance dont la fréquence excède la limite de 150 sont largement majoritaires et que les patronymes qui sont dans le même cas sont portés par un bon tiers de la population étudiée. Un exemple montre ce qu'il en résulte. Voici un X_i recensé en 1846 avec ses attributs et, entre parenthèses, la fréquence d'apparition de ceux-ci : GILLARD (256) [GILAR (258)], Jean (4.871), Marie (22.781), Joseph (25.180), né dans un village proche de Huy, Wanze (556), en 1803 (320). Il est recensé comme X_j en 1856 avec exactement les mêmes indications, sauf la date de naissance qui devient le 27-11-1802 (10).

Dans ce cas, tout à fait réel, que l'on parte de 1846 ou de 1856, la notice individuelle de X_j ne sera pas incluse dans la population-cible de X_i et inversement. Pour qu'elle soit intégrée, il faudrait faire monter la limite admise de 150 à 300. Une telle décision aurait pour conséquence immédiate de doubler les coûts de jumelage et pour conséquence secondaire de produire une masse de liens de valeur faible, sujets à caution. En outre, même à un niveau de 300, des risques de pertes subsistent. La solution adoptée consiste à maintenir le système exposé en prévoyant cependant une exception. Si quand la limite de 150 est dépassée, le patronyme normalisé n'a pas déjà été pris en compte, tous les individus qui en sont porteurs et dont l'année de naissance est égale à celle de X_i dans une fourchette de quatre ans sont inclus dans $P1(X_i)$.

Comme on le voit, le souci de théoriser l'opération d'appariement n'a pas suffi à nous empêcher de devoir recourir à un complément empirique, dont la valeur ne

[31] Point évoqué plus haut et bien connu des praticiens de la reconstitution des familles : les mentions de couples se relient aisément car elles donnent quatre éléments, tandis que les mentions individuelles forment la masse des cas d'ambiguïtés.

tient qu'à l'efficacité concrète. Elle règle parfaitement le cas GILLARD et tous ceux du même genre, sans générer des populations-cibles de taille prohibitive. Qui plus est, ces dernières sont réutilisables.

Après la sélection des liens, explicitée ci-dessous, on envisage en priorité les individus avec lesquels une relation a été provisoirement acceptée, pour leur établir à leur tour une population cible. Ce type de récursivité arborescente a été choisi pour des raisons techniques de limitation du coût, qui s'expliquent par des raisons logiques. En effet, deux possibilités se présentent. Soit il y a parfaite égalité entre X_i et un X_j quelconque, et en ce cas la transitivité est complète, tous les liens retenus pour X_i valent aussi pour X_j, tout le travail est déjà fait. Soit l'égalité est imparfaite et il faut redéfinir une population cible pour X_j à partir de la hiérarchie de ses critères discriminants. Toutefois, ils sont nécessairement plus ou moins, et plutôt plus que moins, les mêmes que ceux de X_i, puisque les deux individus ont assez en commun pour avoir été liés. Il suffit donc d'adapter et non de recommencer ce qui a déjà été fait pour X_i, tant que de notice individuelle en notice individuelle, on reste dans l'arbre qu'il a fondé.

IV. La sélection des liens

Une fois la population cible définie, la notice de X_i est comparée à celle de tous les X_j composant $P1(X_i)$. Tout d'abord, les valeurs v de chaque attribut de X_i sont calculées. La formule

$$v_k = -log_2\,(prob(a_k))$$

fournit la quantité d'information pour un critère k donné.

$$V_i = \sum_{k=1}^{n} v_k$$

permet d'obtenir la quantité d'information associée à l'individu i. Ces paramètres servent à une comparaison qui adopte systématiquement le schéma suivant. Pour l'individu X_j considéré appartenant à $P1(X_i)$, sont mesurées les valeurs de ses attributs et sa valeur globale. On est alors en mesure de définir une quantité idéale d'information, qui est simplement

$$V_{max}(X_i, X_j) = max(V_i, V_j).$$

Elle se distingue de la quantité moyenne, au calcul aussi élémentaire :

$$V_{moy}(X_i, X_j) = (V_i + V_j)/2,$$

ainsi que d'une notion plus complexe, la quantité réelle d'information, qui envisage en tant que tel le lien $X_i - X_j$:

$$V_r(X_i, X_j) = \sum_{k=1}^{n} \alpha_k \times V_{moy}(v_{ik}, v_{jk}),$$

où αk varie entre 0, quand a_{ik} diffère de a_{jk}, et 1 quand ils sont égaux. En cas de divergence estimée partielle, suivant des règles définies plus loin, ce système permet de paramétrer la sanction de l'écart.

Enfin, le taux d'identité est le rapport entre la quantité réelle et la quantité maximale d'information :

$$I(X_i, X_j) = V_r(X_i, X_j)/V_{max}(X_i, X_j).$$

Cet indice varie entre 0 et 1. Essentiellement, il faut noter que le recours aux quantités moyennes est indispensable. Dans une autre version de cet algorithme, la référence était X_i seul, ce qui implique qu'une comparaison $X_i - X_j$ ou $X_j - X_i$ ne donnait pas nécessairement le même résultat. En matière de couplage, le principe de symétrie est un absolu et, par là-même, un test sans appel de la logique de la méthode.

En outre à la base, il faut souligner que le calcul de V_i et V_j ne peut se faire que sur les attributs communs à X_i et X_j. En effet, admettons que le premier ait une date de naissance limitée à l'année et le second une date complète. Si tous les critères sont considérés, V_{max} sera normalement égal à V_j. Or, lors du calcul de $I(X_i, X_j)$, il va sanctionner $V_r(X_i, X_j)$ qui lui n'a pris en compte que la seule égalité ou inégalité analysable, soit celle sur l'année. Le taux d'identité s'en trouverait artificiellement et lourdement affaibli.

V. La sanction des écarts

Ce paragraphe se limite à définir les modes de calcul de α, qui varient selon l'attribut considéré. Pour deux d'entre eux, la règle est simple. S'il y a divergence sur le lieu de naissance codé, α est mis à 0, sinon à 1. C'est l'alternative binaire classique. La variable étant codée, il n'y a pas lieu de compliquer les choses. Certes, on sait qu'à Huy, il arrive que des immigrés, après quelque temps, se "transforment" en natifs du lieu, mais une étude antérieure de reconstitution manuelle des familles a montré que le cas était rare[32]. De plus, Huy étant une occurrence extrêmement fréquente, la sanction de l'écart sera faible. On procède de même pour les prénoms, qui sont eux aussi codés. Dans leur cas, une règle postule qu'un changement de

[32] M. ORIS, *Pauvreté et transition démographique à Huy au 19^{e} siècle (...)*, mémoire inédit de l'U.Lg., 1984, p. 14.

rang n'a aucune importance. Une égalité entre un premier et un troisième prénom débouchera sur un α égal à 1.

*

* *

Des problèmes plus importants sont posés par la date de naissance. Il faut tenir compte du caractère composite de cette donnée, mais aussi de son remarquable pouvoir discriminant quand elle est complète. Donc,

$$V_r(a_{i6}, a_{j6}) = max(\alpha \times V_{moy}(v_{i6}; v_{j6}), \alpha \times V_{moy}(v_{i6*}; v_{j6*}), \alpha \times V_{moy}(v_{i6**}; v_{j6**}))$$

où $\alpha = 1$ en cas d'égalité;

$\alpha = 0,5$ si a_{i6} diffère de a_{j6} par un seul chiffre;

$\alpha = 1 - (abs(a_{i6**} - a_{j6**})/10)$, pour autant que la différence absolue soit inférieure à 10;

$\alpha = 0$ dans tout autre cas.

Cette manière de procéder peut surprendre mais se justifie avec une relative simplicité. Toutes autres choses restant égales, la quantité d'information sera d'autant plus grande que les renseignements seront complets. Or, la date peut être en quatre positions (année), six (année + mois) ou huit (année + mois + jour). Concrètement, le cas intermédiaire est très rare. Pour effectuer la comparaison, il faut utiliser le plus petit commun dénominateur. Admettons qu'il corresponde au maximum. Dans ce cas, si les deux dates sont identiques à un et un seul chiffre près, la probabilité que la divergence soit réelle est faible; il est vraisemblable qu'il s'agit d'une faute de lecture ou d'encodage résiduelle. Comme ce raisonnement est subjectif, α est mis à 0,5, ce qui équivaut à diminuer de moitié la quantité moyenne d'information.

Cette sanction est si lourde que le procédé ne sert réellement que lorsque l'écart implique un chiffre du mois ou de l'année. Si c'est le jour qui est concerné, une deuxième solution donnera pratiquement tout le temps un meilleur résultat et sera donc retenue. Elle consiste à utiliser simplement le bloc mois-année, qui est nécessairement inclus dans toute date complète et doté d'une moindre quantité d'information (Voir ci-dessus les entropies et pouvoirs discriminants). C'est cette perte qui tient lieu de sanction "naturelle", directement fonction de la composition du corpus étudié, à une inégalité sur le jour seul. La contribution à la quantité réelle, donc la probabilité que le taux d'identité atteigne le seuil d'acceptation du lien, s'en trouve réduite d'autant, mais l'information que représente une égalité sur les six positions mois-année est prise en compte.

On peut encore descendre d'un niveau, suivant la même logique, en ne tenant plus compte que de l'année. A ce stade, une divergence n'implique toujours pas l'annulation pure et simple d'une quantité d'information encore amoindrie. Ce refus persistant d'une alternative binaire classique peut surprendre, mais plusieurs études ont montré que d'un document à l'autre, un même individu pouvait fournir des renseignements très divergents. Il est vrai que ce sont des déclarations d'âge qui sont alors le plus souvent en cause, quoique R. Obotela ait montré à partir des registres de quatre communes que des problèmes réels subsistaient après que les dates de naissance les ont eu remplacées. Comme Tillot en Angleterre, il a toutefois constaté que la grande majorité des écarts se situaient dans une fourchette réduite de $\pm$ 2, au pire $\pm$ 4 ans[33].

Ce type de constat a poussé certains chercheurs à considérer qu'une variation dans ces limites pouvait être ignorée ou tolérée, ce qui revient au même[34]. Le système développé pour le traitement des registres de population autorise une solution alternative. Le paramètre α sera calculé comme l'écart absolu entre les deux années de naissance divisé par dix, étant entendu qu'il est égal à zéro dès que la divergence atteint ou dépasse dix ans. Si par exemple elle est de huit ans, la quantité réelle d'information n'est plus que de 20 % de la quantité moyenne, alors que pour une variation d'un an, elle grimpe à 90 %. Cette méthode peut être qualifiée de semi-linéaire, puisque la sanction est proportionnelle à l'écart à un seuil près.

Au terme de ce processus, nous croyons n'avoir rien perdu de l'information apportée par la variable la plus discriminante, ni en avoir usé abusivement.

*

* *

Pour traiter le patronyme, la philosophie est assez semblable mais les outils diffèrent. Une version normalisée est disponible, ce qui permet de procéder comme pour les dates mais avec seulement deux possibilités. Une troisième est introduite par le biais du calcul d'un indice de similitude entre deux patronymes. Ce type de mesure sert parfois de substitut aux méthodes de réduction phonétique ou de standardisation, parfois comme complément de celles-ci. La mesure la plus connue est celle de Guth, surtout utilisée pour le couplage des listes nominatives nord-américaines, mais les

---

[33] P.M. TILLOT, "Sources of inaccuracy in the 1851 and 1861 censuses", in E.A. WRIGLEY (éd.), *Nineteenth-century society (...)*, Cambridge, 1972, pp. 107–108; R. OBOTELA, *op. cit.*, pp. 210–213.

[34] Par exemple M.E. MINTOU, *Méthodes statistique et lexicographique optimisées d'appariement automatique de listes nominatives évolutives et périodiques*, mémoire inédit (ingénieur-Paris), 1987; S. PASLEAU, *Une population... op. cit.*, pp. 309–310, 326–327.

équipes du P.R.D.H., de Laredo (Houston) ou du projet d'étude des populations industrielles au Creusot (Patrice Bourdelais) ont elles aussi proposé un outil[35].

A l'exemple de SOREP, nous avons repris l'indice de similitude de Guth (ISG). A la différence de la distance de Damereau, il peut porter sur des formes non normalisées et il ne se contente pas d'un simple comptage des lettres communes, ce qui peut être source de graves mécomptes. Une égalité n'est acceptée qu'avec un caractère qui est à la même position, ou une en amont ou deux en aval, la somme des égalités ainsi définies étant rapportées au nombre total de lettres. L'outil peut directement définir la valeur du paramètre α puisque le résultat varie entre 0 et 1.

La critique la plus fondamentale adressée aux mesures de ce type est leur absence de symétrie. Autrement dit, la comparaison d'un patronyme A à un patronyme B ne donne pas forcément le même résultat que la confrontation de B vers A. Pour que l'ISG évite ce reproche, il faut tout d'abord toujours diviser le nombre de lettres égales par celui de la plus grande des deux chaînes de caractères comparées. Ensuite, il ne faut tolérer qu'un seul déplacement latéral, aussi bien à gauche qu'à droite, étant entendu qu'une fois qu'une lettre a été engagée dans une égalité, elle ne peut l'être dans une seconde. C'est une option nettement restrictive, comme le montre l'exemple suivant :

SCHMIT — SMIT dont la comparaison donne un ISG de $1/6^e$ = 16,7 % pour nous, $4/6^e$ = 66,7 % dans la version initiale. Au vu de ce cas, notre choix semble contestable puisqu'une évolution plausible est sévèrement sanctionnée, mais un autre exemple peut être :

SAIMIT — SMIT, avec des ISG respectifs de 33,4 % et 66,7 %, soit une variante peu plausible en première approximation qui est plutôt bien jugée par le test dans sa formulation usuelle.

Il n'y a pas de solution miracle en ce domaine mais l'option restrictive a l'avantage d'assurer la symétrie. En outre, pour le couplage des registres de population, nous utilisons l'indice en concurrence avec le patronyme normalisé. Il appartient à chacun d'effectuer sa part du travail, soit de gérer un type de divergence, sur quelques caractères pour le premier, sur les variations phonétiques pour le second. Le meilleur résultat prime et définit la contribution à la quantité réelle d'information.

---

[35] G.J.A. GUTH, "Surname spellings and computerized record linkage", in *Historical Methods*, t. 10, 1976, n° 1, pp. 10–16; D. DEBROU, M. OLSEN, "The Guth algorithm and the nominal record linkage of multi-ethnic populations", *Ibidem*, t. 19, 1986, n° 1, pp. 20–24; G. BOUCHARD, R. ROY, B. CASGRAIN, ... *Le système SOREP... op. cit.*, pp. 147–163; P. BEAUCHAMP, R. ROY, J. LEGARE, "Reconstruction automatique des familles par le programme "Hochelaga II"", in *Population et Famille*, t. 33, 1974, n° 3, pp. 1–40; M.E. MINTOU, *op. cit.*.

Pour cet usage et dans ce contexte, une mesure plutôt trop sévère que trop peu s'impose.

VI. Les résultats de la démarche probabiliste

La démarche probabiliste produit 134.043 liens qui ont atteint le seuil minimal de 0,6 ou 60 %. Le tableau n° 3 répartit cet ensemble selon le taux d'identité et la quantité réelle d'information, croisement dont le graphique n° 3 propose une représentation visuelle. La quantité réelle forme un série très groupée. Près de 90 % des cas se concentrent entre les niveaux 20 et 39, bien centrés sur une moyenne de 27,9. La distribution des taux est plus surprenante. Divisée par classes de cinq pourcent, elle accueille dans la première, juste au-delà du seuil, plus d'un cinquième des liens. Leur proportion se réduit rapidement à 5-6 % avant de remonter brutalement à 27 % au niveau 100, qui équivaut à une identité parfaite entre deux notices.

Les pourcentages en lignes et en colonnes mettent en évidence les intersections dominantes. En ligne, il apparaît que plus la quantité réelle croît, en particulier au-delà de la moyenne de 35,4, plus la part des égalités parfaites se développe. Dès la classe Vr=30-39, les liens entre deux notices exactement semblables frôlent la majorité absolue. Ils la dépassent largement au-delà. L'inverse est aussi vrai. Dans le groupe des Vr de 10 à 19, soit parmi les relations les plus faiblement documentées, les taux d'identité qui sont juste supérieurs au seuil minimal, entre 60 et 69 %, représentent 86 % de l'effectif de la ligne. L'analyse en colonne donne des résultats cohérents.

Bref, une double intersection partage le tableau. La transition se situe au niveau de la moyenne pour les quantités réelles. Par contre pour les taux d'identité (moyenne de 0,811 ou 81,1 %), une vaste plage sépare deux concentrations situées aux extrêmes de l'échelle de vraisemblance. Globalement, un coin supérieur gauche de liens sujets à caution s'oppose nettement à un coin inférieur droit de liens sûrs. Ceci suggère que le seuil minimal de 60 % était trop faible, à tout le moins qu'une modulation selon la quantité réelle aurait évité de charrier de nombreux liens faux. Il est toutefois illusoire de croire possible la définition d'une solution idéale en ce domaine.

Tableau 3

Ventilation des liens produits au terme de la démarche probabiliste selon la quantité réelle d'information et le taux d'identité

Chiffres bruts

Taux : / Quantité	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100	TOTAL
10-19	19266	9714	3425	599	177	81	112	193	214	33781
20-29	9274	6870	6517	5503	3236	3954	1855	3178	7857	48244
30-39	825	1207	1659	2708	3446	3092	3031	1933	17272	35173
40-49	16	96	268	402	821	1005	1398	754	8842	13602
50-59	1	9	17	33	79	148	235	189	2158	2869
60 & +	0	0	3	5	15	9	21	34	287	374
TOTAL	29382	17896	11889	9250	7774	8289	6652	6281	36630	134043

Chiffres relatifs (% en ligne)

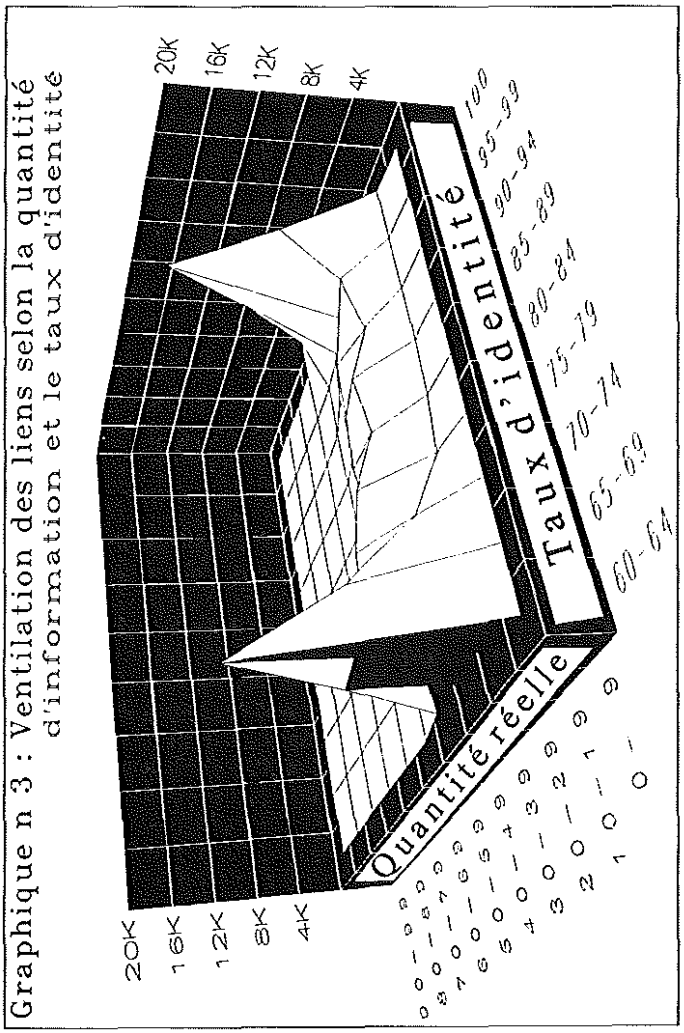
Taux : / Quantité	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100	TOTAL
10-19	57,0	28,8	10,1	1,8	0,5	0,2	0,3	0,6	0,6	100
20-29	19,2	14,2	13,5	11,4	6,7	8,2	3,8	6,6	16,3	100
30-39	2,3	3,4	4,7	7,7	9,8	8,8	8,6	5,5	49,1	100
40-49	0,1	0,7	2,0	3,0	6,0	7,4	10,3	5,5	65,0	100
50-59	0,0	0,3	0,6	1,2	2,8	5,2	8,2	6,6	75,2	100
60 & +	0,0	0,0	0,8	1,3	4,0	2,4	5,6	9,1	76,7	100
TOTAL	21,9	13,4	8,9	6,9	5,8	6,2	5,0	4,7	27,3	100

Chiffres relatifs (% en colonne)

Taux : / Quantité	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100	TOTAL
10-19	65,6	54,3	28,8	6,5	2,3	1,0	1,7	3,1	0,6	25,2
20-29	31,6	38,4	54,8	59,5	41,6	47,7	27,9	50,6	21,4	36,0
30-39	2,8	6,7	14,0	29,3	44,3	37,3	45,6	30,8	47,2	26,2
40-49	0,1	0,5	2,3	4,3	10,6	12,1	21,0	12,0	24,1	10,1
50-59	0,0	0,1	0,1	0,4	1,0	1,8	3,5	3,0	5,9	2,1
60 & +	0,0	0,0	0,0	0,1	0,2	0,1	0,3	0,5	0,8	0,3
TOTAL	100	100	100	100	100	100	100	100	100	100

Graphique 3

Ventilation des liens selon la quantité d'information et le taux d'identité



VII. Les imperfections de la démarche probabiliste et l'approche critique de la source

Certes, les modes de calcul de α, soit les procédés utilisés pour transformer des écarts "discrets" entre deux variables en écarts "continus", sont inévitablement discutables, quels que soient les efforts et les précautions. Pourtant, si d'un projet à l'autre, les méthodes diffèrent, ce n'est pas en termes de procédures que se pose pour le couplage un problème de fond rencontré par toutes les équipes. La remise en cause est venue de deux directions. Pauvres en état civil, les Nord-Américains ont beaucoup travaillé sur l'étude du renouvellement des populations par confrontation de listes nominatives périodiques[36]. Différents auteurs ont dénoncé avec force l'incapacité à dépasser la subjectivité au cas par cas érigée au rang de méthode. Il ont affirmé l'impossibilité de fixer une procédure équilibrée qui ne pêche ni par laxisme, ni par excès de rigorisme, bref qui ne fournisse des résultats sur- ou sous-estimés qui conditionnent l'objectif même des études de listes nominatives, la détermination des mouvements migratoires et du renouvellement de la population[37].

Une critique analogue a porté sur les registres de population formés de familles reconstituées automatiquement. Elle souligne ce que l'équipe de SOREP appelle la relation contradictoire du couplage automatique, soit l'impossibilité de concilier la fiabilité des liens, qui conduit à un sous-jumelage, et leur exhaustivité qui implique à peu près nécessairement un sur-jumelage[38]. Chiaramella est allé plus

---

[36] L'étude modèle est sans doute celle de M.B. KATZ, *The people of Hamilton, Canada West. Family and Class in a mid-nineteenth century City*, Harvard, 1975.

[37] W.R. PREST, "Stability and Change in Old and New England : Clayworth and Dedham", in *Journal of Interdisciplinary History*, t. 6, 1976, pp. 359–374; R. ROY, Ch. POUYEZ, Fr. MARTIN, "Le jumelage des données nominatives dans les recensements : problèmes et méthodes", in *Histoire Sociale/Social History*, t. 18, 1980, n° 25, pp. 173–193; ID., "The linkage of census name data : problems and procedures", in *Journal of Interdisciplinary History*, t. 14, 1983, n° 1, pp. 129–152; D.H. PARKERSON, "How mobile are the nineteenth-century American?", in *Historical Methods*, t. 15, 1982, n° 3, pp. 99–109. Le problème a aussi été posé en France par Y. BLAYO, "Couplage des données d'état civil et des listes nominatives du recensement", in *Annales de Démographie historique*, 1972, pp. 265–268.

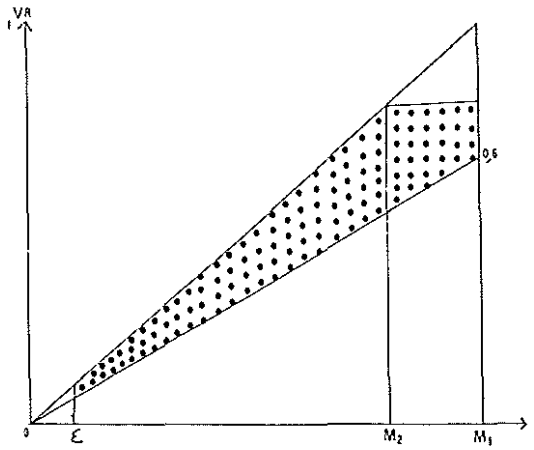
[38] G. BOUCHARD, R. ROY, B. CASGRAIN, ... *Le système SOREP... op. cit.*, pp. 74–79; G. BOUCHARD, "The processing of ambiguous links in computerized family reconstruction", in *Historical Methods*, t. 19, 1986, n° 1, pp. 9–19. Le problème n'a jamais été nié. Voir E.A. WRIGLEY et R.S. SCHOFIELD, "Nominal record linkage by computer and the logic of family reconstitution", in E.A. WRIGLEY (éd.), *Identifying... op. cit.*, pp. 64–101; M. SKOLNICK, A. MORONI, L.L. CAVALLI-SFORZA, E. SIRI, L. SOLIANI, K. de NEVERS, *Automatic reconstruction of families from Parish registers : Parma valley*, Colloque de l'I.U.S.S.P. à Florence en 1977; Th. HERSHBERG, A.N. BURSTEIN, R. DOCKHORN, "Record linkage", in *Historical Methods*, t. 9, 1976, n° 2–3, pp. 137–163; M.P. GUTMANN, "The future of record linkage in history", in *Journal of Family History*, t. 2, 1977, n° 2, pp. 151–158; J. LEGARE, *Le registre de population : laboratoire contesté du démographe*, 1978, 12 p.; R.J. SCHWARTZ, K.M. WEISS, A.V. BUCHANAN, "Looking into the black box : practical approaches to record linkage", in *Annales de Démographie historique*, 1984, pp. 119–128; H. WESLEY *et al.*, "Record linking using a Genealogical

loin. A une longue expérience dans le domaine du couplage, il a joint un doctorat en mathématiques où, usant d'éléments divers de la théorie de l'information, il démontre l'impossibilité de fixer, au sein d'un processus de choix probabiliste, un seuil clair, un niveau au-delà duquel le lien est confirmé, en-deça duquel il est infirmé de manière sûre. *"De nombreuses mesures ont été proposées en classification automatique notamment, ainsi que dans le domaine de la reconstruction automatique des familles; aucun de ces critères ne répond clairement à la question suivante : un degré de ressemblance élevé correspond-il à une forte probabilité pour que deux enregistrements soient homologues?"*[39].

Un tel énoncé paraît paradoxal. Une réflexion théorique sur le couplage permet de mieux le comprendre.

Graphique 4

L'approche de la quantité maximale d'information dans un processus de décision probabiliste



Database System", communication à la section *Information Systems for Differential Demographic Analysis, I.U.S.S.P. Conférence générale, Florence, 5–12 juin 1985*, pp. 2–3.

[39] Y. CHIARAMELLA, *Traitement de données... op. cit.*, pp. 125–126. Des éléments de cette démonstration se retrouvent dans les travaux cités à la note précédente.

Sur le graphique n° 4, l'axe des X est dévolu à la quantité maximale d'information des liens $[V_{max}(X_i, X_j)]$, l'axe des Y à leur quantité réelle $[V_r(X_i, X_j)]$. Par définition, V_r est $\leq$ à V_{max}, donc l'aire des résultats observables est limitée vers le haut par la droite $V_r = V_{max}$. En outre, $_{V max}$ n'est jamais nulle, ce qui équivaudrait à une absence complète d'information. La borne inférieure correspond au score le plus bas possible, celui d'un individu qui serait doté de l'ensemble des attributs les moins discriminants. L'hypothèse inverse renvoie à une notice dont tous les éléments ont une fréquence de 1. La valeur globale correspondante définit une droite $M1$ qui borne l'aire à droite. Toutefois, cette limite est sans signification puisque d'évidence, une somme d'attributs uniques ne trouvera aucune correspondance dans la base, donc aucun lien ne sera établi et de V_{max} ou V_r calculées. Il existe une deuxième borne $M2$, quand toutes les fréquences sont à deux. Pour passer outre, il faudrait que de 1 à n-1 attributs ai(ent) une fréquence de 1 et, en tout état de cause, V_r ne sera jamais égal à V_{max} au-delà de $M2$.

Ces différentes contraintes sont reprises sur le graphique n° 4. De son examen sortent deux constats :

- plus V_{max} est faible, moins la distance est grande entre les deux obliques;
- plus V_{max} est grand, plus cette distance augmente (jusqu'à $M2$).

Il en découle que plus V_{max} est faible, plus grand est le risque de commettre un faux jumelage, car le taux d'identité pourrait être supérieur au seuil pour des individus présentant de notables différences. Inversement, plus V_{max} est élevé, plus considérable est le risque de refuser des liens, car la plage des valeurs qu'est susceptible de prendre la quantité réelle s'accroît (absence d'un prénom rare par exemple). Ceci est une des manières de formuler la "relation contradictoire" du couplage automatique.

Chiaramella souligne cependant que si le risque est irréductible, son intensité est fonction de la qualité des données analysées[40]. A cet égard, ce qu'il appelle les "règles de compatibilité" jouent un rôle essentiel. Il s'agit en fait des tests qui veillent, par exemple, à ce qu'un individu décédé ne soit jumelé après sa mort. Grâce à l'enregistrement des migrations et du mouvement naturel de la population, le registre belge offre normalement une gamme absolument complète. La densité, l'exhaustivité théorique des règles de compatibilité, est partie intégrante de l'originalité du document belge. Elle représente un atout décisif qu'il ne faut toutefois pas surestimer.

En effet, derrière la théorie, il y a le fonctionnement concret du système, où plusieurs failles se repèrent. Ce n'est pas le lieu d'entamer une analyse détaillée de

[40] Y. CHIARAMELLA, *Traitement... op. cit.*, pp. 145-148.

la source et de la validité des indications qu'elle renferme. Elle a été faite ailleurs et souligne que les leçons du registre ne peuvent être considérées comme une référence absolue[41]. Nous travaillons sur une source imparfaite, susceptible de reprendre pour un même individu des dates de naissance divergentes, des patronymes phonétisés ou mal orthographiés, des prénoms atrophiés, etc., sans parler du sous-enregistrement des migrations et d'opérations de révision un peu alambiquées menées en 1867, 1881 et 1891 par les fonctionnaires communaux. Dès lors, en matière de critères de décision, l'abondance de biens n'est pas un vice.

Il faut au contraire jongler avec l'ensemble des variables pour en tirer une évaluation critique, tant des documents de base que du travail de couplage. L'approche probabiliste et l'approche critique ne peuvent être opposées; elles doivent s'épauler l'une l'autre pour fonder des principes et des échelles de décision. C'est ce que documente le point VIII.

VIII. Orientation et tests des liens. Vers une échelle de décision

Il s'agit tout d'abord de déterminer l'orientation des liens. Explicitement, pour un lien $X_i - X_j$ donné, il faut savoir si X_i précède X_j ou inversement. En effet, de la manière dont fonctionnent les programmes, il est parfaitement possible de coupler une notice individuelle de 1880 avec une autre de 1890, mais aussi de 1866, 1856, 1846. Rétablir un ordre chronologique n'est pas toujours aisé dans un document dynamique, complexe et imparfait par endroit comme l'est le registre de population. Il est indispensable de bien gérer les multiples critères de décision en les agençant et en fixant des ordres de priorité dans leur consultation. Cette opération est par là même le niveau privilégié pour développer une série de tests. Quelque 152.911 ont été effectués.

Au terme de ces vérifications, 14.542 liens, soit 10,2 %, sont exclus de la suite des opérations dans les conditions suivantes :

1. une orientation n'a pu être fixée (5.944 liens);
2. – un conflit d'orientation mineur survient et un test quelconque a rendu un verdict négatif;
   – au moins deux tests ont rendu un verdict négatif (4.240 rejets);
3. X est décédé avant la date d'entrée en observation de Y (4.358 liens refusés).

Bref, une redondance de problème ou une invraisemblance grave sont rédhibitoires. Ces décisions sont confortées par la confrontation des démarches critique et probabiliste. Dans le cadre de cette dernière, le tableau n° 3 a présenté la statistique

[41] M. ORIS, *Economie et démographie... op. cit.*, pp. 1080-1120.

Tableau 4

Localisation selon la quantité réelle d'information et le taux d'identité des liens rejetés, conservés malgré un problème et confirmés par le fichier mariage (en % de l'ensemble des liens)

Liens rejetés

Quantité \ Taux :	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100	TOTAL
10-19	25,7	24,8	21,7	15,7	13,6	12,3	13,4	12,4	0,9	24,5
20-29	21,5	16,4	12,1	7,3	3,5	2,9	3,0	1,8	1,6	9,9
30-39	13,5	8,5	4,4	3,6	2,5	1,5	1,5	1,5	2,1	2,7
40-49	-	4,2	5,6	4,7	3,3	3,9	2,6	2,1	3,1	3,2
50-59	-	-	-	-	3,8	4,1	1,7	1,6	3,9	3,5
60 & +	-	-	-	-	-	-	-	-	4,2	3,5
TOTAL	24,0	20,3	13,6	6,6	3,3	2,6	2,3	2,1	2,4	10,8

Liens conservés malgré un problème

Quantité \ Taux :	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100	TOTAL
10-19	37,6	33,0	29,6	24,2	18,6	9,9	11,6	9,3	6,5	34,6
20-29	23,3	20,8	15,3	11,0	8,0	6,4	6,6	4,9	5,0	13,2
30-39	21,1	16,9	15,7	11,5	12,1	10,7	9,5	8,4	9,1	10,6
40-49	-	25,0	17,2	18,2	17,7	18,4	17,0	12,6	16,3	16,5
50-59	-	-	-	-	24,1	28,4	20,9	20,6	24,2	24,0
60 & +	-	-	-	-	-	-	-	-	31,4	29,9
TOTAL	32,6	27,2	19,5	12,4	11,3	9,9	10,8	7,6	11,0	18,5

Liens confirmés par le fichier mariage

Quantité \ Taux :	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100	TOTAL
10-19	0,0	0,0	0,1	0,7	1,1	2,5	0,9	0,5	1,4	0,1
20-29	0,5	1,0	1,6	1,9	3,0	2,4	1,3	1,4	2,0	1,5
30-39	1,7	2,8	2,7	3,1	3,4	4,1	3,8	2,7	5,1	4,2
40-49	-	5,2	3,4	4,0	5,7	5,4	4,3	3,8	5,2	5,0
50-59	-	-	-	-	6,3	6,8	7,7	2,6	5,5	5,6
60 & +	-	-	-	-	-	-	-	-	5,6	5,6
TOTAL	0,2	0,6	1,3	2,3	3,5	3,5	3,3	2,2	4,5	2,3

n.b. : Pourcentages obtenus en rapportant les valeurs de l'annexe n° 1 à celles du tableau n° 3.

des 134.043 liens de base. Son analyse a mis en évidence une scission nette entre un groupe de relations fortes, avec simultanément un taux d'identité et une quantité réelle d'information élevés, et un groupe de relations faibles. Il est du plus haut intérêt d'appliquer une grille semblable à des sous-populations "sensibles". Le tableau n° 4 présente le relevé des liens rejetés en pour-cent du chiffre contenu dans la même cellule du tableau n° 3, pour autant que l'effectif de référence excède 50.

Il en ressort qu'une relation absolue n'existe pas, puisque même des liens très forts sont rejetés, mais que le rejet n'est pas pour autant un phénomène indépendant des deux critères de décision d'ordre quantitatif. Il est doublement déterminé par la faiblesse du taux d'identité (I) et celle de la quantité réelle d'information (Vr). Quatre groupes peuvent être dégagés empiriquement en utilisant comme bornes les moyennes arrondies. Dans l'extrême supérieur gauche du tableau, les liens les plus faibles, de $Vr < 30$ et de $I < 80$, représentent 45,6 % de l'ensemble de ceux produits par la première phase du couplage, et un cinquième d'entre eux sont rejetés, plus d'un quart même pour les valeurs les plus faibles.

Par contre, pour les relations qui ont simultanément un $Vr \geq 30$ et un $I \geq 80$, les pourcentages sont de 33,4 pour à peine 2,4 % de rejets. Les deux coins restants du tableau rassemblent des valeurs intermédiaires. Un peu moins de 16 % des liens, avec une quantité d'information inférieure ou égale à 30, ont un taux d'identité supérieur à 80. La proportion de rejets en leur sein atteint 5,85 %. Quant au dernier groupe ($Vr > 30, I < 80$), il est sans grande importance (7.249 liens, 5,4 % de l'ensemble, 2,6 % de rejets).

Ce qui ressort de ces chiffres, c'est que le choix d'un seuil laxiste de 60 % a limité les risques de sous-jumelage, puis que la démarche critique repère une quantité respectable de sur-jumelages dans la zone à risque des liens d'information et d'identité également faibles.

A l'opposé de cette distribution, le tableau n° 4 localise également 3.097 liens qui ont pu être confirmés grâce à l'existence d'un fichier préexistant, celui des mariages. Il n'y en a pratiquement pas parmi les liens faibles les plus sujets à caution. Globalement, quand le niveau du taux et surtout de la quantité réelle croît, les pourcentages de confirmés font de même. Dans leur cas comme dans celui des rejets, démarche critique et démarche probabiliste se révèlent globalement cohérentes.

C'est dès lors ce degré de cohérence qui permet de déterminer la gravité des autres anomalies repérées, elles aussi localisées globalement par le tableau n° 4. La méthode étant explicitée, il n'est pas nécessaire d'infliger au lecteur tout le détail de son cheminement statistique[42]. Il en ressort que les liens acceptés malgré un conflit

[42] *Ibidem*, pp. 1198-1202.

mineur d'orientation (5.562, soit 4,15 %), et plus encore ceux marqués du fait d'une incohérence dans l'enregistrement des migrations (7.755, soit 5,79 %), sont soit indifférents au niveau du taux d'identité et de la quantité d'information, soit voient même leur fréquence s'accroître parallèlement. Là, c'est la source qui apparaît en faute et le problème ne peut être sanctionné que faiblement ou pas du tout. Par contre, pour les liens caractérisés par un écart sur le sexe (11.537, soit 8,61 %), la sensibilité aux paramètres probabilistes est évidente. La distribution est semblable à celle des rejets avec une concentration au niveau des liens déjà statistiquement faibles. Une telle anomalie critique apparaît de ce fait comme grave.

IX. Les principes de décision

Sur base de l'évaluation des critères et de leur cohérence, une table et une méthode de prise de décision peuvent être construites. Premier croisement à réaliser : le taux d'identité est d'autant plus significatif que la quantité réelle d'information est plus élevée. En ce qui concerne cette dernière, l'analyse des différents tableaux permet de dégager aisément trois groupes : moins de 20, de 20 à la moyenne arrondie à 30, de 30 inclus et au-delà. Par contre pour le taux d'identité, si les extrêmes se distinguent sans problème, la plage qui les sépare est trop vaste pour ne former qu'un groupe. A défaut de critères évidents, des classes de dix pour-cent sont découpées arbitrairement.

Ces groupes sont désignés par un numéro, placés en ligne et colonne pour former des marges dont le produit donne un premier score. Ce processus est illustré ci-dessous :

	I	60-64	65-69	70-79	80-89	90-99	100
V_r		1	2	3	4	5	6
- 30	2	2	4	6	8	10	12
30-39	4	4	8	12	16	20	24
40 et +	6	6	12	18	24	30	36

La critique interne des liens se traduit plus simplement par une échelle de sanctions positives et négatives :

Confirmation par mariage :	+ 8
Un test ou plus passé(s) avec succès :	+ 4
Infirmation (migrations) :	0
Infirmation (orientation) :	- 4
Infirmation (sexe) :	- 8

Nous disposons ainsi de deux échelles qui peuvent être additionnées[43]. Au terme de cette opération, les problèmes critiques seront plus sanctionnés quand ils surviennent au sein de liens statistiquement faibles, ce qui applique bien la leçon tirée de l'analyse des critères de décision. Chaque lien se voit affecter d'un score variant entre -6 et 48. Il est aisé de convertir les données pour les faire varier entre des bornes plus conventionnelles, comme 0 et 10, mais cela n'a pas grande importance.

A l'échelle d'évaluation se joint un principe général fondé sur l'heuristique de la source. Tout individu dont la date d'entrée en observation se situe entre le 1-1-1847, date d'ouverture du premier registre, et le 31-12-1857, date de clôture de la période intercensitaire qu'il couvre, doit être lié à un individu présent au 1-1-1857 (codés A des registres de 1857-66), sauf s'il est mort ou a émigré sans mention d'un retour. L'avantage pratique du principe est sa réversibilité. Toute personne recensée au 1-1-1857 doit avoir une antériorité dans le registre de 1847-56. Il est aisé de déterminer les deux sous-populations, 1847-56 et 1-1-1857, qui ne satisfont pas à cette double règle.

L'arsenal des tests et des contraintes heuristiques s'applique aux maillons, mais aussi à l'entièreté de chaque chaîne, ce que l'on peut synthétiser grâce aux concepts d'ouverture et de fermeture. Le second en particulier, ne doit pas être confondu avec la notion proposée par M. Skolnick, même si l'objectif est commun : la levée des ambiguïtés[44]. Dans notre cas, il s'agit de notions au contenu très concret. Une ouverture signifie qu'une biographie individuelle doit débuter par une présence au premier recensement (15 octobre 1846, le registre étant ouvert au 31 décembre), ou par une naissance à Huy ou par une immigration, l'une ou l'autre après le 31-12-1846. Une fermeture se traduit par un décès, ou par une émigration sans retour, ou par une

[43] Notons que rétrospectivement, on peut considérer qu'il aurait été plus cohérent de ne pas rejeter d'emblée les liens marqués par le test sur la date de décès, mais de les sanctionner lourdement à ce stade.

[44] M.H. SKOLNICK, "The resolution of ambiguities in record linkage", in E.A. WRIGLEY (Ed.), *Identifying people... op. cit.*, pp. 102–127.

présence au recensement de 1900, qui permet de "refermer le couvercle" sur la base de données[45].

Du principe général et de son application découle la définition de deux autres sous-populations, de liens cette fois. En se plaçant toujours par convention entre 1847–56 (ensemble des X) et le 1-1-1857 (ensemble des Y), deux genres de relations peuvent se présenter : les élémentaires, du type $X_i - Y_i$, et les intermédiaires, du type $X_i - X_j - ... - X_n - Y_i$.

Il importe de traiter d'abord l'ensemble des liens dits directs. Préalablement, le fichier des individus est examiné de manière à repérer les sous-populations "à problème", soit tous ceux qui, normalement, devraient faire partie d'un lien, devraient disposer d'une antériorité ou d'une postériorité dans le registre, sans qu'au moins une solution vraisemblable n'ait été produite par la première phase du couplage.

La constitution des biographies individuelles peut dès lors commencer. Les concurrences éventuelles sont levées par comparaison des scores et, en cas d'égalité, des taux d'identité. Le pari a été ici que les relations les plus faibles, les plus problématiques, apparaissaient dans des situations de concurrence avec un lien plus fort. Par exemple, un X-Y quelconque d'une part, et d'autre part le frère de X avec le même Y. Le premier permet de rejeter le second. S'il existe une solution pour le frère de X, c-à-d. s'il est couplé à un autre Y, il n'y a pas lieu d'incrémenter une des sous-populations à problème. A défaut, c'est toutefois ce qu'il convient de faire.

Le tableau n° 5 montre que le pari était raisonnable. Les liens directs retenus ont des scores globalement bien plus élevés que ceux de l'ensemble des liens produits par la démarche probabiliste et non rejetés par la démarche critique. Les moyennes sont respectivement de 25,9 et de 15,7. Cet écart s'explique en particulier parce qu'aucune des relations retenues ne se situe en-dessous de 0 sur l'échelle de décision. Au niveau juste au-dessus, entre 0 et 4, le rapport des liens directs sélectionnés et des liens produits indique que les premiers ne représentent qu'à peine 5 % des seconds, 17 % pour les scores de 5 à 9, déjà 30 % pour ceux de 10 à 14 et entre 37 et 62 % avec une moyenne de 43,05 % au-delà de 14.

[45] Pour des raisons matérielles, nous n'avons pu élaborer un fichier informatique des codés A au 1er janvier 1901, ce qui aurait été l'idéal. Il a fallu procéder par recours manuel aux index.

Tableau 5
Répartition des liens selon leur type et leur score

Scores	Liens non rejetés	Liens directs	Liens inter-médiaires	Liens non rejetés	Liens directs	Liens inter-médiaires	Liens directs	Liens inter-médiaires
	n	n	n	%	%	%	en % des non rejetés	
- 0	10690	0	0	8,9	0,0	0,0	0,0	0,0
0-4	18593	1024	19	15,6	2,8	2,6	5,5	0,1
5-9	9301	1585	23	7,8	4,4	3,2	17,0	0,2
10-14	11158	3387	79	9,3	9,4	10,8	30,4	0,7
15-19	11016	4152	61	9,2	11,5	8,4	37,7	0,6
20-24	20123	8374	127	16,8	23,2	17,4	41,6	0,6
25-29	3277	2007	33	2,7	5,6	4,5	61,2	1,0
30-34	7976	3442	95	6,7	9,6	13,0	43,2	1,2
35-39	20131	7683	142	16,8	21,3	19,5	38,2	0,7
40-44	6977	4320	138	5,8	12,0	18,9	61,9	2,0
48	259	55	13	0,2	0,2	1,8	21,2	5,0
TOTAL	119501	36029	730	100	100	100	30,1	0,6

Une fois l'analyse des liens directs effectuée, il reste à traiter un petit stock de relations dites intermédiaires, c-à-d. celles pour lesquelles le respect du principe général dépourvu de sens, mais pour lesquelles, néanmoins, le respect d'une cohérence invite à sélectionner une option. C'est en fait le cas des individus dont la date d'entrée en observation est celle de leur mariage. Posons un lien entre un jeune célibataire recensé dans le ménage de ses parents en 1847 (X_i), lié à un marié recensé en 1857 (Y_i). Une relation intermédiaire avec un X_j marié en 1854 s'intègre tout naturellement dès lors que le système probabiliste a produit deux liens $X_i - X_j$ et $X_j - Y_i$. En cas de concurrence, c'est la moyenne de leur score et taux d'identité qui lève l'ambiguïté. Cela fait, X_j est assimilé à un élément de relation directe, de telle sorte qu'un maillon supplémentaire puisse à son tour s'intégrer. C'est cependant très rare au sein d'une population peu mobile comme celle de Huy.

Il ne reste alors plus qu'un résidu de relations indirectes, en fait concrètement les cas de migrations-retours, de personnes qui s'en vont et reviennent à Huy 10, 20

ou 30 ans plus tard. Dans ce cas, il est impossible de lever la relation contradictoire du jumelage automatique, autrement dit d'atteindre une solution idéale. Par contre, en analysant la statistique des scores des liens directs, on peut définir un optimum, soit le seuil dont le choix provoque le minimum de décisions erronées. Celles-ci sont irritantes mais inévitables. De toute façon, la migration-retour est certes une thématique intéressante, mais n'est tout de même qu'une parmi les innombrables problématiques sur lesquelles le registre de population permet de jeter un regard neuf.

Conclusions

Dans le cadre de cet article, nous avons voulu présenter une méthodologie de couplage hybride. Elle tente de dégager, de confronter et d'associer deux maxima de vraisemblance, probabiliste et critique, pour annuler les carences d'une technique et d'une source. Il en résulte 36.759 liens directs ou intermédiaires qui ont bien sûr fait l'objet d'un examen critique, de même que les populations dites à problème. Ce travail nous amène à balayer devant notre porte, à repérer plusieurs failles. Revues systématiquement, les relations de score inférieur à 5 s'avèrent manifestement fausses dans 15 % des cas. Cette proportion chute à 4 % entre 5 et 9 et devient nulle au-delà[46]. Le logiciel n'a donc produit que 261 erreurs, soit même pas un pour-cent. Elles sont dues pour l'essentiel au rapprochement de deux notices pourvues d'une date de naissance complète et identique, partagée par deux individus dotés de patronymes tout à fait différents, mais peu discriminants car fréquents l'un et l'autre.

Lors de l'identification des liens erronés, un phénomène surprenant est apparu. Plusieurs relations sont faibles en raison de la forte sanction qui accompagne un écart sur le sexe. Or celui-ci n'est pas indiqué par la source mais déduit à partir des prénoms au moment du dépouillement. Lors de la codification de ces derniers, le premier chiffre a servi à désigner le sexe et les trois suivants à identifier le prénom. Ainsi Clément et Clémentine ou Jean et Jeanne ont le même code au premier chiffre près. Ce système était conçu pour servir dans le cadre du couplage, puis a été considéré comme une sophistication excessive et négligé. Ce fut une erreur car près de 3.000 relations n'ont pu être établies entre des notices où seul le code sexe des prénoms variait. Le cas est si fréquent qu'il a fait l'objet d'un traitement semi-automatique.

A ce jour, l'examen du reste des populations à problème n'est pas clôturé. En première approximation, il semble que le sous-enregistrement migratoire ait été plus important que prévu, de l'ordre de 8 % avec une concentration dans

[46] Les liens de score égal ou supérieur à 10 n'ont été examinés que par le biais d'un sondage au 1/10e.

les premières années de mise en service des registres. Ceux-ci contiennent de nombreuses remarques (plus de 3.400) du type "parti sans laisser d'adresse vers telle date", ou "a déclaré à ses voisins partir pour Liège". Encodées dans une zone "observations", elles ont semblé trop variables pour pouvoir être traitées. Ce fut une seconde erreur, car l'attribution d'un simple code de départ ou d'arrivée aurait permis de régler automatiquement près de 40 % des cas qui ont dû être examinés manuellement.

C'est une leçon pour un avenir qui se présente plutôt bien. Le système mérite certes d'être amélioré, mais tel qu'il a été conçu, il ne produit pratiquement pas d'erreurs, par ailleurs bien localisées dans les liens de score faible. Il a en outre réussi à surmonter nombre de variations, puisqu'entre les notices couplées 19,6 % des dates de naissance diffèrent, ainsi que 11,8 % des patronymes. De manière un peu surprenante, c'est le lieu de naissance qui s'est révélé la donnée la plus stable (4,2 %). Les prénoms (8,4 %) et les patronymes normalisés (7 %) se situent en position intermédiaire.

Au terme de l'expérience, nous nous retrouvons face à un document qui nous est désormais beaucoup mieux connu. La moindre fonction du couplage n'est pas la densité critique qu'il impose puis permet[47]. C'est d'ailleurs dans ce domaine qu'il convient de prolonger la démarche en analysant non plus les maillons mais les chaînes qui correspondent aux biographies individuelles, puis sur cette base les cycles de vie des familles et des ménages et leur cohérence interne.

Nous nous retrouvons aussi face à un fichier couplé. On ne peut prétendre que le traitement qui l'a généré lève absolument la relation contradictoire du couplage, mais à tout le moins, l'objectif est approché au plus près. En outre, comme plusieurs auteurs l'ont proposé, la case et sous-case du tableau de décision sert de marque aux liens. Elle tient lieu de résumé de l'histoire du couplage réalisé, d'estimation de son degré de sûreté, et peut servir de modulateur dans les traitements socio-démographiques[48].

A nouveau en termes de perspectives, les techniques de couplage peuvent largement profiter de l'expérience poursuivie ici. Récemment, deux auteurs ont mis en évidence l'apport de la variable "ménage" lors d'expérimentations sur des listes nominatives et des registres de population. Il s'agit de joindre au raisonnement

[47] La démarche des Québécois est à cet égard exemplaire (J. LEGARE, "A population register for Canada under the French regime : context, scope, content and applications", in *Canadian Studies in Population*, t. 15, 1988, pp. 1-16; G. BOUCHARD, R. ROY, B. CASGRAIN, ... *Le système SOREP... op. cit.*), quoique cette idée soit aussi exprimée ou sous-jacente dans tous les projets de couplage.

[48] E.A. WRIGLEY et R.S. SCHOFIELD, *op. cit.*; Th. HERSHBERG, A.N. BURSTEIN, R. DOCKHORN, *op. cit.*; G. BOUCHARD, R. ROY, B. CASGRAIN, *op. cit.*, pp. 404-405.

horizontal qu'est la comparaison des données individuelles, un raisonnement vertical sur la position de l'individu, d'user du contexte familial sur base du principe de la préférence familiale[49]. Or ce dernier n'est qu'un postulat, certes raisonnable mais pas absolu.

C'est pourquoi, travaillant sur des données globalement de bonne qualité dans un contexte heuristique remarquablement favorable, qui permet de lever à peu près complètement la relation contradictoire du couplage, nous avons voulu développer une démarche aussi "pure" que faire se peut, précisément pour offrir une opportunité de tester les limites de cette approche, dont nous ne contestons nullement l'intelligence mais qui assurément doit donner à l'occasion des résultats erronés. Il faut apprécier quand, dans quelles circonstances et surtout, comment on peut limiter au maximum les risques de dérapage.

La réflexion en ce domaine est susceptible de s'orienter dans de nombreuses autres directions encore. Citons, lors du couplage de deux listes nominatives, l'appréciation statistique de la prise en compte systématique du mouvement naturel, à partir de l'état civil. Ce travail très lourd est-il justifié par une réduction notable de la relation contradictoire du jumelage? Dans quel ordre de grandeur celle-ci se situe-t-elle? Quelle en est la part irréductible en l'absence d'un enregistrement des migrations? En étendant l'expérience hutoise à la grande cité industrielle de Seraing, puisqu'aussi bien le fichier existe, il sera possible d'estimer la réponse à chacune de ces questions. On le voit, le registre de population, malgré ses carences, s'avère bien un laboratoire méthodologique exceptionnel.

[49] Mélange des travaux de S. PASLEAU, *Une population... op. cit.*, pp. 323-325; M. DEMONET, *Appréciation des résultats obtenus grâce au logiciel SAAP*, document dactylographié, *s.l., s.d.*.

Annexe 1

Ventilation selon la quantité réelle d'information et le taux d'identité des liens rejetés, conservés malgré un problème et confirmés par le fichier mariage

Liens rejetés

Taux : / Quantité	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100	TOTAL
10-19	4944	2409	743	94	24	10	15	24	2	8265
20-29	1995	1124	786	400	112	114	55	56	128	4770
30-39	111	103	73	98	87	47	45	29	368	961
40-49	0	4	15	19	27	39	36	16	276	432
50-59	0	1	0	0	3	6	4	3	84	101
60 & +	0	0	0	0	0	0	0	1	12	13
TOTAL	7050	3641	1617	611	253	216	155	129	870	14542

Liens conservés malgré un problème

Taux : / Quantité	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100	TOTAL
10-19	7247	3209	1015	145	33	8	13	18	14	11702
20-29	2165	1431	994	608	259	254	122	156	392	6381
30-39	174	204	261	311	418	331	288	162	1574	3723
40-49	3	24	46	73	145	185	237	95	1440	2248
50-59	1	1	3	11	19	42	49	39	523	688
60 & +	0	0	1	0	4	1	7	9	90	112
TOTAL	9590	4869	2320	1148	878	821	716	479	4033	24854

Liens confirmés par le fichier mariage

Taux : / Quantité	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100	TOTAL
10-19	8	4	2	4	2	2	1	1	3	27
20-29	46	69	103	103	96	93	24	46	160	740
30-39	14	34	45	85	118	127	116	53	878	1470
40-49	0	5	9	16	47	54	60	29	459	679
50-59	0	1	1	1	5	10	18	5	119	160
60 & +	0	0	0	0	1	0	2	2	16	21
TOTAL	68	113	160	209	269	286	221	136	1635	3097