Nº 11-522-x au catalogue
ISSN : 1709822-X

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Données administratives pour l'estimation des chiffres de la population : apprentissage statistique à partir des premières vagues du recensement permanent de la population italienne

par Angela Chieppa, Nicoletta Cibella, Antonella Bernardini, Giampaolo de Matteis et Silvia Farano

Date de diffusion : le 22 octobre 2021

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Données administratives pour l'estimation des chiffres de la population : apprentissage statistique à partir des premières vagues du recensement permanent de la population italienne

Angela Chieppa, Nicoletta Cibella, Antonella Bernardini, Giampaolo de Matteis et Silvia Farano[1]

Résumé

Le Recensement de la population et des logements permanent est la nouvelle stratégie de recensement adoptée en Italie en 2018; il est fondé sur des registres statistiques combinés à des données recueillies au moyen d'enquêtes spécifiquement conçues pour améliorer la qualité des registres et garantir les produits du recensement. Le registre au cœur du recensement permanent est le registre de base de la population (RBP ou RBI en italien, pour *Registro Base degli Individui*), dont les principales sources administratives sont les registres locaux de la population.

Les chiffres de la population sont déterminés par correction des données du RBI au moyen de coefficients basés sur les erreurs de couverture estimées à l'aide des données d'enquête, mais le besoin de sources administratives supplémentaires est clairement ressorti pendant le traitement des données recueillies pendant le premier cycle du recensement permanent. La suspension des enquêtes en raison de l'urgence de la pandémie, combinée à une réduction importante du budget du recensement pour les prochaines années, rend plus impératif encore de modifier le processus d'estimation afin d'utiliser les données administratives comme source principale.

Un registre thématique a été mis en place afin d'exploiter toutes les sources administratives supplémentaires; la découverte de connaissances à partir de cette base de données est essentielle pour mettre en évidence les tendances pertinentes et créer de nouvelles dimensions, appelées « signes de vie », utiles pour l'estimation de la population.

La disponibilité des données recueillies lors des deux premières vagues du recensement offre un ensemble unique et précieux aux fins d'apprentissage statistique; l'association entre les résultats d'enquête et les « signes de vie » pourrait servir à créer un modèle de classification permettant de prédire des erreurs de couverture dans le registre de base de la population (RBI). Le présent article présente les résultats du processus visant à produire des « signes de vie » qui se sont révélés importants dans l'estimation de la population.

Mots clés : données administratives; recensement de la population; registres statistiques; découverte de connaissances à partir de bases de données.

1. Introduction

1.1 Recensement permanent de la population italienne et chiffres de population

Le *Recensement permanent de la population et du logement* est la stratégie de recensement adoptée en 2018 en Italie, fondée sur des registres statistiques continuellement mis à jour au moyen de données administratives et d'enquêtes. Deux enquêtes-échantillons sont menées chaque année afin d'améliorer la qualité des registres et de les compléter avec l'information qu'ils ne fournissent pas, de sorte que les produits du recensement aient le même niveau de diffusion que les éditions classiques antérieures.

Les chiffres de population du recensement, c.-à-d. les chiffres totaux de la population qui réside habituellement dans chaque municipalité, constituent le principal produit du recensement : dans le nouveau cadre d'une production statistique officielle fondée sur des registres, l'estimation des chiffres de population se fonde sur le *registre de base*

[1] Institut national italien de statistique (ISTAT), Département du recensement de la population (chieppa@istat.it , cibella@istat.it).

de la population (*RBI*) corrigé au moyen des résultats d'enquête et, éventuellement, par une intégration avec d'autres données administratives. Les principales sources administratives du RBI sont les registres de population municipale locale.

Le RBI peut être sujet à des erreurs de surdénombrement, c.-à-d. l'inclusion dans le registre des personnes qui ne sont plus sur le territoire, ainsi qu'à des erreurs de sous-dénombrement, c.-à-d. la non-inclusion dans le registre de personnes qui sont sur le territoire.

Pour le cycle de démarrage du recensement permanent, le plan méthodologique de calcul des facteurs de correction du RBI comprend deux étapes.

- La première étape est une estimation directe pour les municipalités dans l'échantillon de l'enquête annuelle : on adopte le modèle de capture-recapture pour estimer les erreurs de couverture du registre de base de la population; la « première capture » est la présence dans le registre, tandis que la « deuxième capture » (la mesure indépendante nécessaire de la même population cible) correspond aux données collectées dans le cadre des enquêtes du recensement.

- La deuxième étape est une estimation indirecte pour les municipalités non échantillonnées, au moyen de modèles sur petits domaines.

1.2 « *Signes de vie* » administratifs

Une base de données thématique (registre), appelée *AIDA-Integrated Archive of Usual Resident Population*, a été établie pour exploiter toutes les sources administratives utiles à l'estimation de la population.

On a élaboré des flux de travail particuliers pour gérer les données provenant de sources administratives, notamment sur les sujets suivants : Travail et études, Déclarations de revenus, Gains, Retraite et prestations non liées à la pension, Permis de séjour, Location et vente d'appartements. Les signaux administratifs de chaque individu, à savoir sa présence dans une ou plusieurs sources, pourraient être considérés comme *des signes de vie (SdV)* d'une personne en Italie, un indicateur possible de résidence habituelle. Ces signes de vie sont classés en fonction de profils de durée, du type et de la fiabilité de la source spécifique, par rapport à d'autres signaux individuels (p. ex. les relations du ménage).

Les signes de vie et les données d'enquête de l'AIDA sont des sources précieuses pour l'estimation du surdénombrement et du sous-dénombrement des données du RBI. Pour les premières vagues du recensement permanent, on a utilisé les SdV afin de gérer le sous-dénombrement de l'enquête : toutes les unités d'échantillon tirées du RBI qui sont des non-répondantes de l'enquête mais associées à des SdV élevés dans l'AIDA ont été récupérées et considérées comme « habituellement résidentes ».

1.3 Facteurs incitant à améliorer le plan du recensement permanent et domaines d'étude actuels

Le premier cycle du recensement permanent a atteint ses principaux objectifs : les deux vagues d'enquêtes annuelles du recensement ont enregistré des taux de réponse très élevés et le recensement permanent permet la diffusion des résultats du recensement dans le respect des contraintes de coût et d'actualité.

Il faut toutefois le repenser en raison de certaines questions critiques concernant le plan d'estimation. La réflexion générale doit porter sur les éléments suivants.

- Questions d'échantillonnage :

    - les variations même minimes des taux de réponse et des résultats d'enquête produisent des effets très importants sur les chiffres de population, tant en termes absolus que relatifs;

    - compte tenu de la taille des échantillons, les correcteurs sont significatifs au niveau municipal*citoyenneté, mais il faut des domaines plus détaillés;

    - il faut gérer le biais municipal.

- Coût des enquêtes, fardeau des répondants, qualité de l'enquête :

    - il faut tirer parti des dossiers administratifs pour réduire les coûts;

    - non-réponse d'enquête non ignorable;

    - complexité et fortes dépendances du réseau d'enquête.

- Nécessité d'un processus d'estimation commun et responsable de dénombrement de la population :

    - les résultats du recensement de la population ont de fortes répercussions administratives et politiques, car ils déterminent les niveaux de financement, la disponibilité des services, etc.; le processus d'estimation doit être communiqué à des intervenants publics (processus d'estimation compréhensible);

    - les facteurs de correction décimale pour les groupes ou les enregistrements ne sont pas faciles à diffuser quand les estimations municipales sont contestées par les autorités locales; des listes d'enregistrements individuels pour réviser les données municipales sont souvent nécessaires (résultats utilisables du processus d'estimation).

De plus, la pandémie de COVID-19 a été un facteur qui a fortement incité à l'utilisation de données administratives en l'absence d'enquêtes, étant donné que les enquêtes du recensement avaient été suspendues en 2020.

Deux domaines principaux font actuellement l'objet de recherches : (a) un nouveau plan pour les enquêtes du recensement du cycle post-2021, qui doit assurer un compromis optimal entre les coûts et la qualité des estimations; (b) la conception et la mise à l'essai d'autres solutions méthodologiques aux fins d'estimation des chiffres de population, utilisant le plus efficacement possible les données administratives intégrées dans la base de données AIDA. Ce dernier domaine comprend l'expérimentation décrite dans les paragraphes qui suivent.

2. Concevoir une nouvelle solution par une méthode fondée sur la science des données

Le fait de disposer des deux premières vagues d'enquêtes du recensement permanent, d'une part, et de sources administratives, d'autre part, couvrant la même période de référence permet de déterminer des tendances dans les données pertinentes pour l'estimation des chiffres de population et d'élaborer une solution fondée sur les données aux fins de prédiction de la résidence habituelle.

Compte tenu de la complexité du phénomène et du fait que les données administratives ne sont pas spécifiquement conçues à des fins statistiques, il est extrêmement important d'extraire des connaissances de ces bases de données afin d'en améliorer la qualité et de choisir ou de construire des dimensions appropriées pour l'estimation des résultats du recensement. Il faut une équipe interdisciplinaire pour élaborer le processus de découverte des connaissances, car les questions méthodologiques sont aussi cruciales que les compétences en gestion de bases de données ou en automatisation des processus et que les connaissances d'experts sur les questions thématiques (comme la résidence habituelle, la qualité des sources administratives et la qualité des données d'enquête).

Une équipe interdisciplinaire, l'apprentissage statistique à partir de données pour calculer un modèle prédictif de la résidence habituelle, la gestion des bases de données pour améliorer les flux de travail des données administratives sont des composantes essentielles de ce qu'on appelle une *méthode fondée sur la science des données.*

Une expérimentation de ce type de méthode, utilisant toutes les données disponibles, est en cours à l'ISTAT. Elle vise à améliorer la production des résultats du recensement dans les prochaines années par l'exploitation de données administratives.

Pour ce qui est du problème d'apprentissage statistique, l'objectif final de cette expérience est de trouver le meilleur « apprenant » pour prédire la résidence habituelle en Italie au niveau individuel (cible/variable de réponse) en classant les signes administratifs de vie et de présence dans le RBI (covariables/variables auxiliaires/variables indépendantes).

De plus, les résultats intermédiaires de ce processus d'apprentissage sont essentiels aux fins d'amélioration du plan de recensement permanent : (1) améliorer les flux de travail qui utilisent les données administratives (flux de travail normalisé pour gérer différentes sources et gérer les variables de chaque source, dont le nombre peut varier d'une année à l'autre); (2) sélection des caractéristiques aux fins d'estimation de la résidence habituelle (détermination d'un sous-ensemble de variables pertinentes); (3) mise au point de la base de données de SdV AIDA (normalisation des variables calculées à partir de sources administratives qui pourraient varier d'une année à l'autre); (4) détection des tendances critiques associées à des prédictions très incertaines.

Les étapes de l'expérimentation sont les suivantes.

- *Mise en place de bases de données expérimentales* : prétraitement et chargement des données tirées d'enquête, du registre de base de la population, de sources administratives; exécution du couplage au niveau individuel entre différentes sources.

- *Exploration des données* : étude des associations entre les signaux administratifs et l'information du registre de base de la population, par l'utilisation des données d'enquête comme mesure des résultats de la résidence habituelle, qui permet de déterminer les tendances de sous-dénombrement et de surdénombrement dans le registre.

- *Ingénierie des caractéristiques* :

    - formalisation des connaissances d'experts;

    - introduction/création de nouvelles variables, fondées sur les résultats de l'exploration des données et les règles expertes.

- *Modélisation prédictive* :

    - entraînement/test de modèles de classification (modèles latents, arbre de décision et autres);

    - règles déterministes définies par des experts.

- *Évaluation statistique et thématique des résultats* :

    - correction du registre de base de la population en fonction de la classification individuelle pour tenir compte du surdénombrement et du sous-dénombrement;

    - analyse des résultats; comparaison entre les résultats des différents modèles et avec les estimations de 2018-2019.

Les étapes suivent l'ordre de la séquence, mais elles s'alimentent aussi continuellement les unes les autres.

Les résultats partagés et interprétés de chaque étape, y compris la préparation des données, sont des objectifs intermédiaires extrêmement importants pour l'élaboration d'une nouvelle solution réalisable (remaniement de la collecte des données et du plan d'échantillonnage des enquêtes; nouveau flux de travail aux fins de gestion des sources administratives, etc.).

3. Premières étapes de l'expérimentation

3.1 Analyse exploratoire des données sur la base de données expérimentale intégrée

Des données provenant de différentes sources ont été intégrées dans une base de données, qui est utile aux objectifs d'apprentissage, avec couplage au niveau individuel, et qui assure la cohérence de la référence de temps/période entre les sources.

- On a chargé les données d'enquête en tenant compte des règles expertes de validation de la collecte de données; en plus des microdonnées recueillies, les indicateurs de qualité tirés des opérations sur le terrain sont chargés.

- Pour ce qui est des sources administratives (SdV), on a normalisé les flux de travail de chargement de plus de 50 archives différentes pour produire des « signaux » comparables, différenciés selon la durée et la qualité de la source. La référence temporelle est gérée au moyen d'une fenêtre temporelle appropriée sur la durée du signal, centrée sur le 31 décembre de chaque année. Différentes périodes sont testées, d'une fenêtre longitudinale de plusieurs années à une période d'au moins 12 mois. Les profils de continuité quant à la durée des signes administratifs reflètent ce qu'on appelle la « force du signal ».

- À partir du registre (RBI), les variables démographiques de base et le lieu habituel de résidence sont chargés; la référence temporelle est le 31 décembre de chaque année et le jour de l'enquête du recensement, afin d'assurer la même référence/période de données d'enquête.

Cette base de données intégrée expérimentale est continuellement modifiée au moyen des résultats du processus d'acquisition de connaissances. C'est pourquoi il y a différentes versions de la base de données, chacune ayant une dimension nouvelle ou transformée.

Au point de départ du processus d'apprentissage, la distribution conjointe de la « présence dans le registre de base de la population » et de la « présence dans les sources administratives » montre différentes sous-populations qui doivent faire l'objet d'études plus approfondies. Sur plus de 60 millions d'enregistrements individuels dans l'ensemble de personnes ayant une présence dans le RBI au cours des années 2018-2019 et/ou ayant des signaux administratifs pendant cette période, on constate que :

- 2,5 % des personnes présentes dans le RBI n'ont pas de signe administratif : elles doivent faire l'objet d'une enquête qui vérifie quelle partie est réellement due à un surdénombrement du registre;

- environ 1,8 million de personnes ne sont pas présentes dans le RBI, mais ont des signes administratifs au cours de la même période : 28,7 % d'entre elles ont des signes forts, c'est-à-dire une présence constante/continue provenant de sources qualifiées; correspondent-elles toutes vraiment à un sous-dénombrement du registre de base de la population?

- environ 97 % des personnes présentes dans le RBI ont des signes dans des sources administratives, mais

- 5 % des personnes ont des signes « épisodiques » (la durée du signe ou la source n'est pas significative selon la définition de la résidence habituelle) et pourraient être associées à un surdénombrement.

- Parmi ces personnes, 4 millions ont des signes à l'extérieur de la province du RBI : quelle partie correspond à une erreur de lieu habituel de résidence dans le RBI (composante locale de l'erreur de couverture)?

Dans ces études, la réponse à l'enquête pourrait aider à différencier la sous-population et à détecter les enregistrements individuels admissibles aux erreurs de couverture dans le RBI.

Les dimensions analytiques sont individuelles, comme l'âge, le sexe, les signes administratifs/le profil (quand il y a des signes), l'information sur le lieu de résidence (quand la personne est enregistrée dans le registre) et le lieu du signal administratif, mais aussi les attributs municipaux et régionaux. Enfin, pour le sous-ensemble de personnes qui ont participé aux échantillons des enquêtes du recensement, les réponses aux enquêtes et les indicateurs de qualité du travail sur le terrain sont inclus. L'analyse des correspondances multiples et les arbres de classification constituent de très bons outils pour détecter les tendances dans les données et pour partager les résultats avec les experts en thématiques.

Certaines tendances ressortent clairement, ainsi que des « caractéristiques » et variables pertinentes :

- *La taille de la population municipale est clairement une dimension discriminante* : les petites villes sont associées à un taux de réponse très élevé aux enquêtes et à une bonne qualité dans le RBI, tandis que dans les grandes villes, il y a des tendances essentielles pour lesquelles il est plus difficile d'évaluer la qualité du registre.

- *Les personnes âgées/retraitées* ainsi que les *relations familiales dérivées des signaux fiscaux* sont des profils de personnes qui permettent de prédire une bonne qualité dans le registre.

- D'autres données sont nécessaires pour mieux comprendre et réduire l'incertitude sur les *tendances essentielles*, comme *les personnes sans signes, les ménages constitués d'une personne, les étrangers dans les grandes villes,* parce que la qualité des prédictions, qui s'appuie sur les variables disponibles à partir des signes administratifs et des enquêtes actuelles, est trop faible.

- Dans les *villes de taille moyenne,* il y a un groupe significatif de personnes dont les signes administratifs se situent à l'extérieur de la province du RBI. Dans ces cas, il faut également effectuer une analyse plus approfondie pour mesurer correctement la classification erronée réelle du lieu de résidence du RBI.

- Pour les *étrangers*, une dimension pertinente de prédiction de la résidence habituelle est la zone ou le pays de citoyenneté.

- Les *SdV longitudinaux*, sur plus d'un an, saisissent mieux la *composante étrangère du sous-dénombrement du registre de base de la population*.

Les résultats de l'analyse et la comparaison entre les données d'enquête et les données administratives ont également servi à améliorer le flux de travail de chargement administratif et le couplage au microniveau (pour résoudre certains faux signaux).

3.2 Utiliser des connaissances d'experts pour calculer de nouvelles variables

Les résultats de l'analyse montrent la nécessité d'un plus grand nombre de renseignements, qui pourraient être partiellement calculés à partir de variables existantes ou qui pourraient entraîner certains changements dans les

processus de chargement à partir des sources originales. Les connaissances d'experts, concernant les caractéristiques sociales et les sources administratives, sont essentielles dans cette phase connue dite *étape d'ingénierie des caractéristiques*. L'ingénierie des caractéristiques a un effet important sur la performance des modèles de prédiction finaux.

Dans l'expérimentation actuelle, la formalisation des « règles expertes » pour détecter certaines sous-populations particulières au moyen des variables existantes a permis de convertir les connaissances d'experts en nouvelles variables/données. À titre d'exemple, des règles expertes ont servi à détecter, parmi toutes les personnes ayant des signes de travail ou d'études situés à l'extérieur de la province de résidence habituelle enregistrée dans le RBI, celles qui étaient des navetteurs, de façon à séparer les cas de classification erronée ou d'erreur de couverture. Après la formalisation, on a ajouté de nouvelles variables pertinentes à la base de données expérimentale pour mettre en œuvre ces règles expertes, comme les « liens familiaux », les « itinéraires de voyage », la « typologie des municipalités fondée sur les habitudes de déplacement ».

De plus, on a demandé à des experts de définir des règles pour détecter dans quels cas (quel type de SdV et de caractéristique démographique dans la registre de base de la population) un enregistrement individuel pourrait être considéré comme une erreur de couverture du registre de base. Cette classification déterministe est utile pour l'entraînement et la mise à l'essai du modèle statistique : d'une part, elle peut servir de référence et, d'autre part, cette formalisation des connaissances d'experts peut grandement contribuer à l'interprétation des résultats des modèles prédictifs statistiques.

L'utilisation de règles déterministes fondées sur des connaissances d'experts ne suffit pas toujours pour classer certaines données individuelles, qui résistent au classement. Des difficultés essentielles demeurent. Un modèle statistique fondé sur les données réduira probablement cet agrégat, mais la qualité de la prédiction de la résidence habituelle concernant certaines tendances essentielles impliquerait probablement d'établir de nouvelles cibles précises pour le plan des enquêtes.

4. Prochaines étapes et conclusions

Après la mise en place de l'expérimentation et après les premières étapes de l'analyse exploratoire ainsi que de la sélection et de l'ingénierie des caractéristiques, le processus d'apprentissage en est actuellement aux étapes décisives finales, à savoir :

- la gestion des nouvelles sources (liées à la COVID-19, particularités de l'année 2020) dans le registre de l'AIDA;
- l'application de règles améliorées et de nouvelles variables dans la base de données;
- la sélection d'un modèle prédictif statistique;
- l'évaluation du modèle choisi et la comparaison avec les résultats des règles déterministes d'experts.

La stratégie du recensement permanent de la population a permis de réaliser un changement radical dans la production statistique des résultats de recensement.

À l'heure actuelle, il faut franchir une autre étape vers une utilisation massive de données administratives pour améliorer la qualité des données du registre de base de la population et des chiffres de population du recensement.

La meilleure solution implique non seulement des aspects méthodologiques et des hypothèses de modèle, mais aussi des coûts d'enquête et des problèmes de faisabilité, ainsi qu'une expertise aux fins de la conception et de la gestion des flux de travail des bases de données. C'est pourquoi une démarche interdisciplinaire est nécessaire.

Les données des deux premières vagues d'enquêtes du recensement constituent une source précieuse pour apprendre à partir des données comment intégrer les signes administratifs de vie dans le processus d'estimation.

Les premières étapes de ce type d'étude donnent :

- une base de données intégrée expérimentale utile pour l'entraînement, la mise à l'essai et l'évaluation de modèles d'estimation de rechange;
- la sélection des dimensions pertinentes, y compris des nouvelles variables calculées à partir de données existantes et de règles expertes;
- la détection des tendances essentielles : le nouveau plan des enquêtes pourrait servir à collecter des données afin de résoudre cette incertitude.

On évalue actuellement de nouveaux modèles d'estimation fondés sur un modèle prédictif statistique dérivé du processus d'apprentissage. On insistera sur l'utilité de solutions mixtes fondées sur des classificateurs qui tiennent également compte des critères entrés par les experts en thématiques.

Bibliographie

Bernardini, A., A. Chieppa, N. Cibella, F. Solari (2021), « Administrative data for population counts estimations in Italian Population Census », dans Perna, C., et coll. (éd.) *Book of short papers – SIS 2021*, Pearson, p. 274-278.

Bernardini A., N. Cibella, G. Gallo, et coll. (2019) , « Empirical evidence for population counting : the combined use of administrative sources and survey data », présentation à l'atelier de l'ESS sur l'utilisation des données administratives et les statistiques sociales, Valence, Espagne.

Casari, A, A. Zheng (2018), *Feature Engineering for Machine Learning*, Boston, O'Reilly Inc.

Chieppa, A., G. Gallo, V. Tomeo, et coll. (2018), « Knowledge discovery for inferring the usually resident population from administrative registers », *Mathematical Population Studies. International Journal of Mathematical Demography*, 26:2, p. 92-106.

Hastie, T., R. Tibshirani, et J. H. Friedman (2009), *The elements of statistical learning: data mining, inference, and prediction*, New York, Springer

Kim, J.K., J.N.K. Rao (2012), Combining data from two independent surveys : a model assisted approach, *Biometrika*, 99(1), p. 85-100.

Pfeffermann, D, J. Eltinge, et L. Brown (2015), « Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture », *Journal of Survey Statistics and Methodology*, 3, p. 425–483

CEE-ONU (2018), *Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses*.

CEE-ONU (2020), *New frontiers for censuses beyond 2020.*

Zhang, Li-Chun (2019), « A note on dual system population size estimator », *Journal of Official Statistics*, 35(1), p. 279-283.