Mettre les technologies cloud au service de la production statistique

Romain Avouac
Insee
romain.avouac@insee.fr

Thomas Faria
Insee
thomas.faria@insee.fr

Frédéric Comte
Insee
frederic.comte@insee.fr

2024-12-06

Résumé French

0.1.1 Introduction

L'exploitation de sources de données non traditionnelles afin d'améliorer le processus de production statistique est une orientation majeure du Système Statistique Européen (SSE), résumée à travers le concept de *Trusted Smart Statistics* (F. Ricciato, A. Wirthmann, K. Giannakouris, M. Skaliotis, et others [1]). Cette dynamique s'accompagne d'innovations dans les processus statistiques, permettant de tirer parti du potentiel de ces sources — plus grande disponibilité, résolution spatio-temporelle accrue, etc. — tout en faisant face à leur complexité et à leurs limites. Parmi ces innovations figurent les méthodes d'apprentissage automatique et leurs applications prometteuses dans les domaines du codage et de la classification, des redressements et de l'imputation (T. Gjaltema [2]). Les multiples défis auxquels font face les instituts statistiques dans ce contexte d'évolution sont abordés dans le *Mémorandum de Bucarest sur les statistiques officielles dans une société numérisée*, qui prévoit que « la variété des nouvelles sources de données, paradigmes computationnels et outils nécessitera des adaptations de l'architecture métier statistique, des processus, des modèles de production, des infrastructures informatiques, des cadres méthodologiques et de qualité, ainsi que des structures de gouvernance correspondantes », et invite en conséquence le SSE à évaluer les adaptations requises et à les prioriser (DGINS [3]). L'orientation B (« innover et être en première ligne sur les sources de données ») de la stratégie Insee Horizon 2025 traduit l'opérationnalisation de cette orientation dans le cadre du service statistique public (INSEE [4]).

Dans l'optique de ces transformations, de nombreux travaux ont été menés dans le cadre de projets successifs à l'échelle européenne pour opérationnaliser l'utilisation de sources de données non-traditionnelles dans la production de statistiques officielles. Dans le cadre du projet ESSnet Big Data II (2018-2020), les instituts statistiques nationaux (INS) ont travaillé sur une large gamme de thématiques (offres d'emploi en ligne, transactions financières, traces GPS, etc.) afin de constituer les briques nécessaires pour intégrer ces sources dans les processus de production et identifier leurs limites (EUROSTAT [5]). En France, les travaux sur l'exploitation des données mobiles (B.

Sakarovitch, M.-P. d. Bellefon, P. Givord, et M. Vanhoof [6]) ou des données de caisse (M. Leclair, I. Léonard, G. Rateau, P. Sillard, G. Varlet, et P. Vernédal [7]) ont permis d'illustrer le potentiel de ces sources pour construire de nouveaux indicateurs ou raffiner des indicateurs existants. Néanmoins, si un travail considérable a été consacré au développement de cadres méthodologiques (P. Descy, V. Kvetan, A. Wirthmann, et F. Reis [8], D. Salgado, L. Sanguiao-Sande, S. Barragán, B. Oancea, et M. Suarez-Castillo [9]), de lignes directrices sur la qualité (A. Kowarik et M. Six [10]), ainsi qu'à la conception de processus sécurisant l'acquisition de données auprès de tiers (F. Ricciato, F. De Meersman, A. Wirthmann, G. Seynaeve, et M. Skaliotis [11]), les infrastructures informatiques et les compétences nécessaires pour gérer ces nouveaux objets sont restées peu abordées dans la littérature.

Les caractéristiques de ces nouvelles sources rendent leur traitement particulièrement complexe. On qualifie souvent de *big data* ces données qui se distinguepar leur volume (souvent de l'ordre de plusieurs centaines de Go voire du To), leur vélocité (vitesse de génération, proche du temps réel) ou de leur variété (données structurées mais aussi non structurées, telles que les textes et les images). Pourtant, les « compétences pour automatiser, analyser et optimiser ces systèmes complexes ne font souvent pas partie des compétences traditionnelles de la plupart des instituts statistiques nationaux » (A. Ashofteh et J. M. Bravo [12]). Au cours des dernières années, un nombre croissant de statisticiens publics sont formés aux méthodes de *data science* permettant d'envisager l'intégration de ces sources dans des processus de production statistique. Dans ses multiples acceptions, le terme « data scientist » reflète en effet l'implication croissante des statisticiens dans le développement informatique et l'orchestration de leurs opérations de traitement des données, au-delà des seules phases de conception ou de validation (T. H. Davenport et D. Patil [13]). Toutefois, on observe en pratique, à l'Insee et dans d'autres organisations, que la capacité de ces profiles à tirer parti des sources *big data* et des méthodes d'apprentissage automatique est limitée par plusieurs défis.

Un premier défi réside dans l'absence d'infrastructures informatiques adaptées aux nouvelles sources de données auxquelles les INS ont désormais accès, ainsi qu'au besoin croissant de nouvelles méthodes statistiques. Par exemple, les sources *big data* nécessitent d'énormes capacités de stockage et s'appuient souvent sur des infrastructures et des méthodes de calcul distribué pour être traitées (L. Liu [14]). De même, l'adoption de nouvelles méthodes statistiques basées sur des algorithmes d'apprentissage automatique requiert des capacités informatiques — en particulier des GPU (unités de traitement graphique) dans le cadre du traitement du texte ou de l'image — pour paralléliser massivement les calculs (A. Saiyeda et M. A. Mir [15]). De telles ressources sont rarement disponibles dans les infrastructures informatiques traditionnelles. Lorsque des infrastructures de calcul adaptées sont disponibles, comme les supercalculateurs (HPC) utilisés dans certains domaines de recherche, elles nécessitent des compétences spécifiques — notamment pour leur mise en place et leur maintenance — qui sont rarement disponibles au sein des INS.

Un autre défi majeur est d'équiper les statisticiens avec des environnements de développement leur permettant d'expérimenter plus librement. L'essence de l'innovation dans les travaux statistiques réside dans la capacité à intégrer rapidement de nouveaux outils et méthodologies. Cette agilité est limitée lorsque les statisticiens dépendent excessivement des départements informa-

tiques pour provisionner des ressources ou installer de nouveaux logiciels. Dans les configurations traditionnelles — ordinateurs personnels ou bureaux virtuels sur des architectures centralisées[1] — les départements informatiques privilégient généralement la sécurité et la stabilité du système au détriment de la fourniture de nouveaux services, ce qui limite le potentiel d'innovation. De plus, ces environnements rigides rendent difficile la mise en œuvre de bonnes pratiques de développement, telles que le travail collaboratif — nécessitant des environnements permettant de partager facilement des expérimentations avec ses pairs — et la reproductibilité.

Un troisième défi concerne la difficulté de passer des expérimentations innovantes à des solutions en production. Même lorsque les statisticiens ont accès à des environnements leur permettant d'expérimenter aisément, la transition vers le déploiement d'une application ou d'un modèle reste généralement difficile. Les environnements de production diffèrent souvent des environnements de développement, ce qui entraîne des coûts de développement supplémentaires importants pour passer d'une preuve de concept à une solution industrialisée qui rend du service dans la durée. Par exemple, dans le cas des projets d'apprentissage automatique, les modèles déployés nécessitent un suivi rigoureux pour s'assurer qu'ils conservent leur précision et leur utilité au fil du temps, et requièrent généralement des améliorations périodiques ou continues. Ces besoins plaident pour des environnements plus flexibles permettant aux statisticiens de gérer de manière autonome le cycle de vie complet de leurs projets de *data science*.

Ces différents défis ont un thème sous-jacent commun : le besoin d'une plus grande autonomie. La capacité des méthodes de *data science* à améliorer et potentiellement transformer la production des statistiques officielles dépend crucialement de la capacité des statisticiens à mener des expérimentations innovantes plus librement. Pour ce faire, ils doivent avoir accès à des ressources informatiques substantielles et diversifiées leur permettant de gérer le volume et la diversité des sources *big data* et d'exploiter les méthodes d'apprentissage automatique. Ces projets expérimentaux nécessitent à leur tour des environnements de développement flexibles favorisant le travail collaboratif pour tirer parti de la diversité des profils et compétences des équipes de projet. Enfin, pour tirer pleinement parti de ces expérimentations, les statisticiens ont besoin d'outils pour déployer des applications sous forme de preuves de concept et orchestrer leurs opérations statistiques en toute autonomie.

Dans ce contexte, l'Insee a développé `Onyxia` : un projet open source permettant aux organisations de déployer des plateformes de *data science* favorisant l'innovation en offrant aux statisticiens une plus grande autonomie[2]. Cet article vise à décrire le processus de réflexion ayant conduit à ce projet et à illustrer comment il autonomise les statisticiens à l'Insee, devenant ainsi un pilier de notre stratégie d'innovation. La section 2 offre une analyse approfondie des derniers développements de l'écosystème de la donnée, mettant en lumière les choix technologiques qui ont façonné le développement d'un environnement moderne de *data science*, adapté aux besoins spécifiques des statisticiens. En particulier, nous montrons comment les technologies *cloud* — en particulier

[1] `AUSv3` est un exemple d'une telle infrastructure. Le statisticien utilise son poste de travail comme point d'accès à un bureau virtuel qui « reproduit » l'expérience habituelle du poste de travail. Néanmoins, les calculs qui sont lancés — via `R` ou `Python` par exemple — sont effectués sur des machines virtuelles (VM) de calcul dédiées, et non sur le poste de travail.

[2] https://github.com/InseeFrLab/onyxia

la conteneurisation et le stockage objet — sont essentielles pour créer des environnements évolutifs et flexibles qui favorisent l'autonomie tout en promouvant la reproductibilité des projets statistiques. Toutefois, malgré leurs atouts pour les applications modernes de *data science*, la complexité de configuration et d'utilisation des technologies *cloud* pose souvent des obstacles à leur adoption. Dans la section 3, nous détaillons le projet `Onyxia` vise précisément à rendre les technologies cloud accessibles aux statisticiens grâce à une interface conviviale et un catalogue étendu d'environnements de *data science* prêts à l'emploi. Enfin, à travers l'étude de cas de la classification des activités des entreprises françaises (APE), la section 4 illustre comment l'utilisation de ces technologies a considérablement facilité la mise en production de modèles d'apprentissage automatique à l'Insee en permettant d'appliquer les meilleures pratiques issues du *MLOps*.

TOTO

TOTO

TOTO

TOTO

Bibliographie

[1] F. Ricciato, A. Wirthmann, K. Giannakouris, M. Skaliotis, et others, « Trusted smart statistics: Motivations and principles », *Statistical Journal of the IAOS*, vol. 35, nº 4, p. 589-603, 2019.

[2] T. Gjaltema, « High-Level Group for the Modernisation of Official Statistics (HLG-MOS) of the United Nations Economic Commission for Europe », *Statistical Journal of the IAOS*, vol. 38, nº 3, p. 917-922, 2022.

[3] DGINS, « Bucharest Memorandum on Official Statistics in a Datafied Society ». 2018.

[4] INSEE, « Horizon 2025 ». 2016.

[5] EUROSTAT, « ESSnet Big Data 2 - Final Technical Report ». 2021.

[6] B. Sakarovitch, M.-P. d. Bellefon, P. Givord, et M. Vanhoof, « Estimating the residential population from mobile phone data, an initial exploration », *Economie et Statistique*, vol. 505, nº 1, p. 109-132, 2018.

[7] M. Leclair, I. Léonard, G. Rateau, P. Sillard, G. Varlet, et P. Vernédal, « Scanner data: advances in methodology and new challenges for computing consumer price indices », *Economie et Statistique*, vol. 509, nº 1, p. 13-29, 2019.

[8] P. Descy, V. Kvetan, A. Wirthmann, et F. Reis, « Towards a shared infrastructure for online job advertisement data », *Statistical Journal of the IAOS*, vol. 35, nº 4, p. 669-675, 2019.

[9] D. Salgado, L. Sanguiao-Sande, S. Barragán, B. Oancea, et M. Suarez-Castillo, « A proposed production framework with mobile network data », in *ESSnet Big Data II - Workpackage I - Mobile Network Data*, 2020.

[10] A. Kowarik et M. Six, « Quality Guidelines for the Acquisition and Usage of Big Data with additional Insights on Web Data », in *4th International Conference on Advanced Research Methods and Analytics (CARMA 2022)*, 2022, p. 269-270.

[11] F. Ricciato, F. De Meersman, A. Wirthmann, G. Seynaeve, et M. Skaliotis, « Processing of mobile network operator data for official statistics: the case for public-private partnerships », in *104th DGINS conference*, 2018.

[12] A. Ashofteh et J. M. Bravo, « Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical systems », *Statistical Journal of the IAOS*, vol. 37, nº 3, p. 771-789, 2021.

[13] T. H. Davenport et D. Patil, « Data scientist », *Harvard business review*, vol. 90, nº 5, p. 70-76, 2012.

[14] L. Liu, « Computing infrastructure for big data processing », *Frontiers of Computer Science*, vol. 7, p. 165-170, 2013.

[15] A. Saiyeda et M. A. Mir, « Cloud computing for deep learning analytics: A survey of current trends and challenges. », *International Journal of Advanced Research in Computer Science*, vol. 8, nº 2, 2017.