Nº 11-522-X au catalogue
ISSN 1709-822X

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Boutique multiservice d'intelligence artificielle et d'apprentissage automatique pour les statistiques officielles

par Francesca Kay

Date de diffusion : le 8 septembre 2025

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2024 de Statistique Canada
Le futur des statistiques officielles

Boutique multiservice d'intelligence artificielle et d'apprentissage automatique pour les statistiques officielles

Francesca Kay [1]

Résumé

Dans le présent article, plusieurs défis rencontrés lors de l'établissement des estimations de la population américaine d'après les dossiers administratifs (DA) en 2020 sont identifiés. Parmi ces enjeux, mentionnons l'exactitude au niveau de l'emplacement, la couverture des personnes et la cohérence dans le temps, le filtrage des non-résidents et des personnes décédées à la date de référence, le dévoilement des couplages manquants entre les enregistrements de personnes et d'adresses et la prévision des caractéristiques démographiques. Plusieurs façons d'aborder ces questions sont examinées. Les résultats des régressions illustrent la façon dont les défis et les solutions ont une incidence sur les estimations de la population des comtés basées sur les DA.

Mots-clés : Statistiques officielles; qualité; intelligence artificielle; apprentissage automatique.

1. Introduction

L'utilisation de l'intelligence artificielle et de l'apprentissage automatique (IA/AA) pour produire des statistiques officielles est l'un des domaines stratégiques qui doivent être développés davantage et pour lesquels il est avantageux d'agir de manière coordonnée et systémique. La Boutique multiservice d'intelligence artificielle et d'apprentissage automatique pour les statistiques officielles (IAAASO) jouera un rôle important dans l'élaboration de solutions innovantes en ce qui concerne les produits et les processus statistiques, ce qui permettra de produire plus rapidement des statistiques officielles et de mieux répondre aux besoins des utilisateurs. L'IAAASO rassemble un consortium de 14 pays pour développer des connaissances et des cas d'utilisation soutenant l'utilisation de solutions d'IA et d'AA pour produire des statistiques officielles.

Objectifs du projet d'IAAASO

Le projet a deux objectifs principaux :

- fournir une valeur ajoutée en s'efforçant d'offrir des conseils et une assistance sur mesure au système statistique européen (SSE) et au-delà;
- continuer à acquérir des connaissances et à élaborer des cas d'utilisation qui appuient l'utilisation de solutions d'IA et d'AA pour produire des statistiques officielles.

1.1 Objectifs spécifiques

Les objectifs spécifiques du projet d'IAAASO consistent à fournir un point d'entrée unique au personnel du SSE prenant part à l'innovation afin de cibler des occasions concrètes. Le projet fournira des conseils et un soutien pour le déploiement de solutions d'IA et d'AA dans des cadres de qualité, de méthodologie et de mise en œuvre, afin de faciliter la création de nouveaux produits et processus statistiques tirant parti de la recherche et du développement de pointe dans les domaines de l'IA et de l'AA.

Les principales activités sous-jacentes du projet d'IAAASO sont les suivantes :

- développer, maintenir et faire évoluer un ensemble cohérent de capacités pertinentes (y compris des méthodologies, des lignes directrices, des bacs à sable, des données étiquetées, des processus et des cadres méthodologiques, de mise en œuvre et de qualité) pour déployer des solutions d'IA et d'AA dans le domaine des statistiques officielles

[1] Francesca Kay , Central Statistical Office, Irlande, Francesca.kay@cso.ie. L'auteur tient à remercier les membres du personnel d'Eurostat qui ont travaillé sur le projet, soit Albrecht Wirthmann, Konstantinos Giannakouris et Jean-Marc Museux.

dans l'ensemble du SSE;
- mettre en place une plateforme ou un carrefour offrant un point d'entrée unique au personnel du SSE pour accéder aux capacités pertinentes;
- fournir un soutien et des conseils pour l'intégration et la maintenance des solutions pertinentes d'IA et d'AA dans les organisations du SSE en offrant de la formation et un soutien actif et efficace;
- créer des communautés autour des solutions à source ouverte développées et maintenues par les membres du SSE;
- partager des idées, des expériences, des réussites et des enseignements afin de stimuler l'innovation basée sur l'utilisation de l'IA et l'AA;
- favoriser et faciliter le passage de la mise au point et de l'expérimentation de solutions d'IA et d'AA à l'industrialisation et à la production.

1.2 Résultats du projet d'IAAASO

Les principaux résultats attendus du projet sont les suivants :
- créer un cadre d'élaboration de solutions d'IA et d'AA à utiliser dans le contexte de la production de statistiques officielles et européennes;
- permettre au personnel et aux partenaires du SSE d'accéder à des solutions et à des ressources d'IA et d'AA établies et éprouvées qu'ils peuvent utiliser dans le contexte de la production de statistiques officielles;
- inciter les organisations du SSE à innover et éventuellement tirer parti de l'utilisation d'approches d'IA et d'AA à l'égard des produits statistiques;
- réaliser des économies d'échelle et de ressources grâce à la coopération au sein et en dehors du SSE et accélérer le passage des idées à la production réelle.

2. Boutique multiservice d'IA et d'AA pour les statistiques officielles

2.1 Présentation du projet d'IAAASO

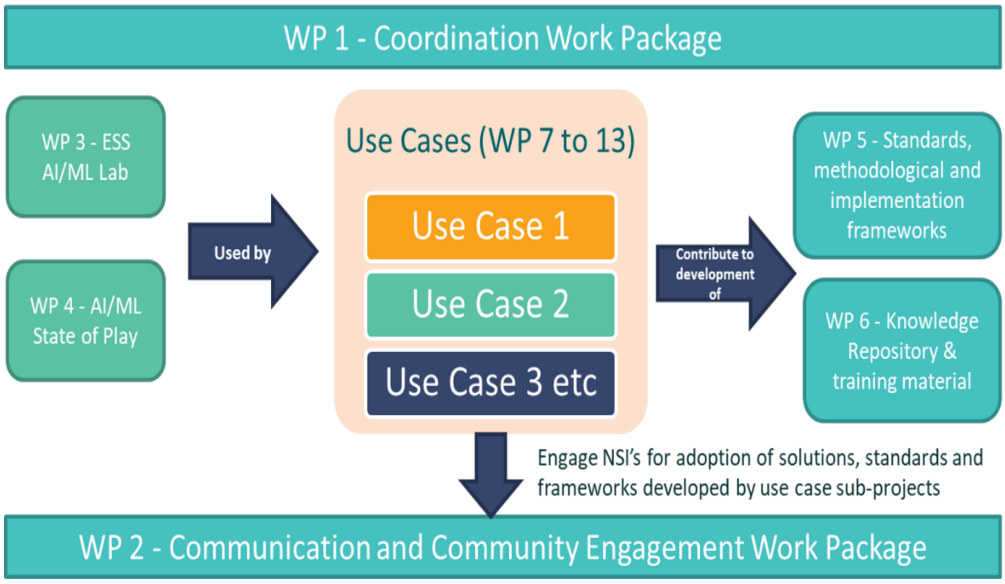
Le programme d'innovation approuvé par le Comité du système statistique européen (CSSE) en février 2023 reconnaît l'IA et l'AA comme des outils clés pour le développement transversal, le traitement de nouvelles sources de données et la prise en charge de processus complexes. L'utilisation de l'IA et de l'AA permettra, par exemple, d'utiliser des données provenant de capteurs, d'images satellites et du moissonnage du Web, ce qui permettra d'obtenir plus fréquemment de nouveaux résultats statistiques plus actuels et plus granulaires, ainsi que d'améliorer les processus statistiques.

On s'emploie encore à dégager les possibilités liées à l'IA et à l'AA. Il est donc opportun d'entreprendre dès maintenant une mesure coordonnée pour favoriser l'apprentissage systématique, le partage des expériences, la détermination des bonnes pratiques et la réutilisation des solutions. Ainsi, chaque institut national de statistique (INS) réalisera des économies d'échelle et verra ses coûts diminuer. Afin d'accélérer l'adoption de l'IA et de l'AA par les INS dans la production de statistiques officielles, Eurostat a lancé un appel à des subventions pour créer une Boutique multiservice d'IA et d'AA pour les statistiques officielles.

Le Central Statistics Office d'Irlande a réuni un consortium regroupant 14 pays (voir en annexe A les pays participants) pour mener à bien le projet. Ce consortium facilitera la mise à l'échelle ou la réutilisation de solutions existantes et nouvelles, la normalisation de la méthodologie, la compréhension de l'écosystème de l'IA et de l'AA dans le domaine des statistiques, l'élaboration de pratiques exemplaires appropriées et de lignes directrices sur la qualité, ainsi que l'accès à des ressources partagées telles qu'un environnement de bac à sable et de la formation.

Le projet a démarré en avril 2024 et durera quatre ans. Dans ses sections subséquentes, le présent article décrit le projet et la manière dont celui-ci entend placer la qualité au cœur de ses travaux.

Figure 3-1
Structure du projet d'IAAASO



3. Cadre théorique

La présente section décrit le cadre général du projet et la manière dont celui-ci sera mis en œuvre. La conception de la structure d'exécution du projet repose sur des approches courantes de gestion de projet.

La structure du projet s'articule autour de deux grands concepts (figure 3-1) :

- Des **lots de travaux sur les cas d'utilisation** qui se concentrent sur l'examen d'un problème spécifique qui pourrait être résolu en utilisant l'IA et l'AA ou d'un domaine de l'IA et l'AA qui pourrait être utilisé plus largement aux fins des statistiques officielles. Ces cas d'utilisation fourniront des normes, des méthodes, des codes et de la formation qui seront intégrés dans les lots de travaux transversaux afin de pouvoir les rassembler de manière centralisée et de fournir certaines des ressources de base mises à disposition par le projet.
- Des **lots de travaux transversaux** qui fournissent des services au nom de l'ensemble du projet aux membres du consortium, mais aussi à la communauté plus large de l'IA et de l'AA que le projet cherche à bâtir et à mobiliser.

La principale méthode de réalisation du projet consiste à mettre en place des lots de travaux individuels, comme suit :

Lots de travaux transversaux

- LT1 : Gestion et coordination de projet
- LT2 : Communication et mobilisation de la communauté
- LT3 : Laboratoire d'IA et d'AA du SSE : Infrastructure technique et configuration organisationnelle
- LT4 : État d'avancement de l'IA et de l'AA et surveillance de l'écosystème
- LT5 : Normes et cadres méthodologiques et de mise en œuvre
- LT6 : Répertoire de connaissances et matériel de formation

Lots de travaux sur les cas d'utilisation

- LT7 : Cas d'utilisation : IA et AA pour les données d'observation de la Terre et d'imagerie satellitaire
- LT8 : Cas d'utilisation : Accent sur la vérification – Vérification et imputation statistiquement valides et efficaces dans le domaine des statistiques officielles par IA et AA – avec un accent particulier sur la vérification
- LT9 : Cas d'utilisation : Accent sur l'imputation – Vérification et imputation statistiquement valides et efficaces dans le domaine des statistiques officielles par IA et AA – avec un accent particulier sur l'imputation
- LT10 : Cas d'utilisation : Du texte au code – Expériences et potentiel d'utilisation de l'IA et l'AA aux fins de classification et de codage

- LT11 : Cas d'utilisation : Appliquer l'AA pour estimer les réseaux de chaînes d'approvisionnement au niveau des entreprises
- LT12 : Cas d'utilisation : Grands modèles de langage
- LT13 : Cas d'utilisation : Génération de données synthétiques dans le domaine des statistiques officielles : techniques et applications

(Voir l'annexe B pour obtenir plus de détails sur chaque lot de travaux).

La qualité sera prise en compte principalement de deux façons :

Assurance de la qualité des lots de travaux
On s'emploie à mettre sur pied un comité de révision pour offrir une assurance de la qualité de chaque cas d'utilisation. Un ou deux examinateurs experts dans le domaine particulier du cas d'utilisation seront affectés au cas d'utilisation pour effectuer un examen par les pairs au début, au milieu et à la fin de chaque cas d'utilisation. Les examinateurs peuvent également fournir des conseils et des orientations spéciales pendant la réalisation du cas d'utilisation.

Élaboration de normes, de pratiques exemplaires, de cadres et d'orientations
L'un des principaux résultats du projet est de fournir :

- des lignes directrices méthodologiques et de mise en œuvre pour appliquer l'IA et l'AA dans le domaine des statistiques officielles;
- des connaissances généralisées et des normes et pratiques exemplaires définies pour la mise au point et l'exploitation de solutions d'IA et d'AA dans le domaine des statistiques officielles;
- des normes pour appliquer l'IA et l'AA dans le domaine des statistiques officielles. Cela permettra et facilitera le passage de la mise au point et de l'expérimentation de solutions d'IA et d'AA à l'industrialisation et à la production.

Pour ce faire, les cas d'utilisation individuels chercheront à élaborer ces éléments dans le cadre de leurs lots de travaux individuels et les verseront dans un dépôt central afin de déterminer et de fournir des produits et des conseils globaux pour améliorer la qualité globale de l'utilisation de l'IA et l'AA dans le domaine des statistiques officielles, tel qu'il est indiqué ci-dessous :

Normes

Alors que de nombreux bureaux de statistique ont mis à l'essai des méthodes d'intelligence artificielle et d'apprentissage automatique au cours des dix dernières années, la nécessité d'établir des normes est devenue évidente, d'une part en raison de la maturité croissante des applications et d'autre part de la fréquence de leur utilisation.

Il ne s'agit pas d'une solution unique pour toutes les solutions d'IA et d'AA. Il s'avère toutefois évident que les normes sont utiles jusqu'à un certain niveau. Souvent, ces normes n'ont pas besoin d'être élaborées de toutes pièces, mais peuvent être dérivées des pratiques exemplaires des bureaux de statistique ou de travaux antérieurs (par exemple, par la CEE-ONU). Le projet permettra d'aborder des questions conceptuelles sur les aspects qualitatifs (y compris la clarté, l'interprétabilité, la transparence, la qualité des données, l'auditabilité et l'éthique).

Un manuel de normes pour les statistiques officielles sera élaboré, peaufiné et mis à jour tout au long du projet, contenant des sujets tels que les méthodes normalisées, les aspects de la qualité et l'interprétabilité. Ce manuel est utile pour les statisticiens qui travaillent dans le domaine des statistiques officielles et qui utilisent les nouvelles méthodes d'IA et d'AA. Les bases des matrices de confusion et les graphiques de dépendance partielle en sont des exemples.

Cadres méthodologiques

D'énormes progrès ont été réalisés ces dernières années dans le domaine de l'IA et de l'AA. Par conséquent, un nombre croissant d'applications sont apparues dans le domaine des statistiques officielles, lesquelles se concentrent sur la mise au point de nouveaux produits et services, l'amélioration de la prise de décision et l'augmentation de l'utilisation de nouvelles sources de données en tant qu'indicateurs rapides.

L'application aveugle des techniques d'IA et d'AA ne suffit toutefois pas pour produire un résultat de haute qualité. Il est clair que des méthodes et des instructions (une méthodologie) doivent être élaborées pour permettre d'appliquer l'IA et l'AA d'une manière fiable et digne de confiance. D'un point de vue mathématique, la méthodologie vise à produire des résultats

statistiques impartiaux, reproductibles et valides pour la population cible dans le cadre d'études qui commencent généralement par des données dites « trouvées ». Un cadre méthodologique sera élaboré, perfectionné et mis à jour au cours du projet.

Cadre de mise en œuvre

Les normes et les cadres méthodologiques disponibles peuvent accroître considérablement la durabilité des applications d'IA et d'AA dans le domaine des statistiques officielles. Pour que celles-ci aient une incidence, il faut absolument disposer de moyens efficaces et pratiques pour passer de la théorie à la mise en œuvre pratique.

Le projet cherchera à établir des mécanismes efficaces pour soutenir la transition de l'expérimentation à la production. La réussite de cette transition repose sur la rationalisation et l'automatisation du cycle de vie de l'apprentissage automatique, une pratique définie comme les opérations d'apprentissage automatique (OAA). Pour ce faire, nous déterminerons les unités élémentaires standard pour soutenir les OAA et fournirons spécifiquement des conseils et des modèles pour leur mise en place dans les laboratoires d'IA et d'AA. Nous décrirons également les caractéristiques que les laboratoires d'IA et d'AA devraient avoir afin d'utiliser efficacement les pratiques exemplaires et les unités élémentaires standard.

Des cas d'utilisation individuels serviront à vérifier que les normes, les méthodes et les cadres d'OAA sont pratiques et bénéfiques et qu'ils facilitent la transition de l'expérimentation à la production, et ce, de façon durable, tout en produisant des statistiques de qualité appropriée. Un cadre de mise en œuvre global sera élaboré, perfectionné et mis à jour au cours du projet.

4. Résultats

Le projet vise à présenter les résultats de son travail lors d'une conférence annuelle pour le projet d'IAAASO :

- NTTS 2025
- Conférence 2026 (possiblement une conférence sur la Qualité)
- NTTS 2027
- Conférence de clôture – Dublin 2028

Ces travaux seront exécutés par la production de documents, de guides et de cadres, de présentations et d'articles lors de conférences, de webinaires et de formations. Le matériel sera disponible sur la plateforme CROS d'Eurostat qui est en cours de développement.

5. Conclusion

Grâce à la réalisation des treize lots de travaux distincts ciblés par le consortium, les principaux résultats escomptés sont les suivants :

- créer un cadre d'élaboration de solutions d'IA et d'AA à utiliser dans le contexte de la production de statistiques officielles et européennes;
- permettre au personnel et aux partenaires du SSE d'accéder à des solutions et à des ressources d'IA et d'AA établies et éprouvées qu'ils peuvent utiliser dans le contexte de la production de statistiques officielles;
- inciter les organisations du SSE à s'intéresser à l'IA et l'AA pour innover et faciliter leur compréhension et leur exploitation des avantages de l'IA et l'AA;
- réaliser des économies d'échelle et de ressources grâce à la coopération au sein et en dehors du SSE et accélérer le passage des idées d'IA et d'AA à la production réelle.

Si ces résultats sont atteints, le projet devrait fournir un ensemble significatif de ressources qui présenteront des conseils, des méthodes et des normes pour la production de produits statistiques de haute qualité grâce à l'application de l'IA et de l'AA.

Annexe A. Pays du consortium

- Statistics Austria – Autriche
- Statistical Service of Cyprus – Chypre
- Statistics Denmark – Danemark
- Institut national de la statistique et des études économiques (INSEE) – France
- Statistisches Bundesamt (DESTATIS) – Allemagne
- Central Statistics Office (CSO) – Irlande
- Istituto Nazionale di Statistica (Istat) – Italie
- Statistiques Luxembourg
- Statistics Netherlands (CBS) – Pays-Bas
- Statistics Norway – Norvège
- Statistics Poland – Pologne
- Instituto Nacional De Estatistica – Portugal
- Statistical Office of the Republic of Slovenia – Slovénie
- Instituto Nacional de Estadistica (INE) – Espagne
- Statistics Sweden (SCB) – Suède

Annexe B. Description des lots de travaux

Tableau A.1
Description des lots de travaux

Nom du lot de travaux	Pays responsable du lot de travaux	Description du lot de travaux
Coordination	CSO	Soutenir, coordonner et faciliter les activités du consortium dans le cadre du projet, tout en prenant des mesures pour garantir des niveaux de qualité adéquats des résultats qui peuvent être réalisés et rester durables dans l'ensemble du SSE.
Communication	Istat	Créer des communautés autour des solutions à source ouverte développées et maintenues par les membres du SSE. Partager des idées, des expériences, des réussites et des enseignements afin de stimuler Partager des idées, des expériences, des réussites et des enseignements afin de stimuler l'innovation basée sur l'utilisation de l'IA et l'AA. Communiquer à l'externe et diffuser les résultats du projet, en particulier au sein du SSE.
Plateforme et infrastructure technique	INSEE	Fournir un environnement de test de l'IA et l'AA (bac à sable) qui peut être utilisé par les groupes de travail sur les cas d'utilisation (LT 7 à LT 13) à des fins d'expérimentation et de développement. Fournir la documentation et l'assistance technique nécessaires à l'intégration des cas d'utilisation. Fournir de la documentation et des conseils pour permettre aux INS de créer leur propre bac à sable d'IA et d'AA (« plateforme en tant que lot »).
État d'avancement et surveillance de l'écosystème	DESTATIS	Obtenir des preuves de l'état d'avancement de l'utilisation de l'IA et de l'AA dans le SSE et au-delà, c'est-à-dire dresser et tenir à jour un tableau complet des développements et des activités en matière d'IA et d'AA dans le domaine des statistiques. Cerner les besoins des INS en matière d'IA et d'AA.

Nom du lot de travaux	Pays responsable du lot de travaux	Description du lot de travaux
Normes et méthodologies	DESTATIS et CBS	Élaborer des lignes directrices méthodologiques et de mise en oeuvre pour appliquer l'IA et l'AA dans le domaine des statistiques officielles. Développer des connaissances généralisées et des normes et pratiques exemplaires définies pour la mise au point et l'exploitation de solutions d'IA et d'AA dans le domaine des statistiques officielles. Parvenir à uniformiser l'application de l'IA et l'AA dans le domaine des statistiques officielles. Cela permettra et facilitera le passage de la mise au point et de l'expérimentation de solutions d'IA et d'AA à l'industrialisation et à la production.
Partage de connaissances	Statistics Poland	Fournir un soutien et des conseils pour l'intégration et la maintenance des solutions pertinentes d'IA et d'AA dans les organisations du SSE en offrant de la formation et un soutien actif et efficace. Fournir des conseils pour la sélection et le développement de modèles d'IA et d'AA pour des produits statistiques spécifiques.
lA et AA pour les données d'observation de la Terre et d'imagerie satellitaire	CBS	Les modèles d'IA et d'AA existants utilisant l'observation de la Terre peuvent-ils être généralisés dans l'espace (pays) et dans le temps (périodes) et dans quelles conditions? Si cela est possible, des lignes directrices méthodologiques pour la mise en oeuvre seront élaborées.
Accent sur la vérification – Vérification statistiquement valide et efficace dans les statistiques officielles par IA et AA	DESTATIS	Découvrir les possibilités d'utilisation de l'IA et l'AA en matière d'automatisation, d'efficacité et d'amélioration de la qualité du processus de vérification et d'imputation (en mettant l'accent sur la vérification). Explorer ces possibilités à l'aide d'exemples pratiques, en mettant l'accent sur la vérification, sans oublier les normes de qualité des statistiques officielles. Étudier les effets sur les charges de travail correspondantes pour la vérification dans les bureaux de statistique.
Accent sur l'imputation – Imputation statistiquement valide et efficace dans les statistiques officielles par IA et AA	INE	Englober les différentes situations de production statistique dans lesquelles les techniques d'IA et d'AA peuvent être utilisées pour l'imputation. Imputation anticipée/actualité – Imputation de valeurs qui seront connues et mesurées après l'achèvement de certaines étapes du processus (par exemple, après la collecte ou après la vérification pendant la collecte). Imputation postérieure à la collecte/exactitude – Prise en compte des situations typiques de non-réponse à la fois aux questions et aux unités dans des conditions de production habituelles, en abordant l'imputation de valeurs qui ne seront jamais connues, mais qui se rapportent à des unités qui font partie de l'échantillon. Imputation au-delà de l'échantillon/granularité – Examiner les situations dans lesquelles nous devons imputer des bases de données de population ou des ensembles de données (par exemple, des sources administratives) qui ne font pas partie d'un échantillon en reconstruisant des informations (au moins partiellement) pour une population entière à l'aide de données provenant d'un ensemble de données beaucoup plus petit.
Du texte au code – Expériences et potentiel de l'utilisation de l'IA et l'AA aux fins de classification et de codage	DESTATIS	L'utilisation de l'IA et de l'AA pour assigner des descriptions textuelles provenant de sources potentiellement différentes à des systèmes de classification internationalement reconnus (par exemple la NACE, la Classification de la consommation individuelle par objet, la Classification internationale type des professions) de statistiques officielles est étudiée, y compris le défi que représentent les textes rédigés dans des langues différentes en tant qu'obstacle à l'échange de données. Une recommandation pour un pipeline Text2code (y compris, par exemple, le prétraitement). Premières solutions à certains défis particuliers tels que les différentes langues.

Nom du lot de travaux	Pays responsable du lot de travaux	Description du lot de travaux
Appliquer l'AA pour estimer les réseaux de chaînes d'approvisionnement au niveau des entreprises	CBS	Développer des modèles d'apprentissage automatique qui peuvent être utilisés pour estimer des ensembles de données de réseaux de chaînes d'approvisionnement au niveau des entreprises à l'échelle de la population en tirant seulement parti des données sur les caractéristiques et les variables au niveau des entreprises. Entraîner les modèles d'apprentissage automatique à partir d'ensembles de données de réseaux à l'échelle de la population basés sur des données administratives détaillées sur les transactions entre utilisateurs et fournisseurs. Élaborer une méthodologie pour estimer les poids des couplages, en utilisant les tableaux d'entrées et de sorties des comptes nationaux comme référence.
Utiliser de grands modèles de langage génératifs dans le domaine des statistiques	SCB	Étudier la capacité des futurs grands modèles de langage (GML) (2024-2027) à s'intégrer dans la production et le soutien des statistiques officielles. Étudier les avantages d'un réglage de précision du GML existant sur les cas d'utilisation. Pour soutenir les objectifs ci-dessus, définir les facilitateurs et les contraintes architecturales en relation avec l'utilisation du GML.
Génération de données synthétiques dans le domaine des statistiques officielles : techniques et applications	Istat et Statistics Austria	Étudier différents algorithmes d'IA et d'AA pour générer des données synthétiques dans les domaines des statistiques officielles en équilibrant l'utilité et la protection de la vie privée. Réaliser une preuve de concept pour générer des données synthétiques dans le domaine des statistiques officielles, en appliquant des techniques d'IA et d'AA.