

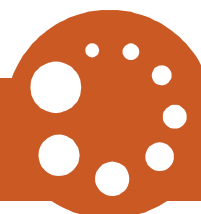


ML_SoSi – Apprentissage automatique Sécurité sociale

Projet-pilote dans le cadre de la stratégie d'innovation
sur les données de l'OFS

Sommaire de gestion

EXPERIMENTAL STATISTICS



Neuchâtel, 2023

Éditeur:	Office fédéral de la statistique (OFS)	Concept de mise en page:	Section PUB
Rédaction:	Luzius von Gunten, Frank Schubert, Brandon Qorri Gonzalez, Athanassia Chalimourda	Téléchargement:	www.statistique.ch
Domaine:	00 Bases statistiques des généralités	Copyright:	OFS, Neuchâtel 2023 La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée
Langue du texte original:	Allemand		
Traduction:	Services linguistiques de l'OFS		

Sommaire de gestion

Le chômage peut se dérouler de manière très différente selon les personnes touchées. Les parcours de vie sont ainsi marqués par l'octroi (répété) de prestations du système de sécurité sociale (assurance-chômage, assurance-invalidité, aide sociale), par des reprises de l'activité rémunérée ou encore par le retrait de la vie active et la migration. Dans le cadre du projet «ML_SoSi», les informations relatives aux parcours individuels sont analysées à l'aide de méthodes statistiques inductives afin de repérer des parcours types. Au-delà des résultats ainsi obtenus, le projet vise à développer une approche axée sur l'analyse des parcours individuels dans les données longitudinales pour les besoins de la statistique publique.

Les données de base utilisées sont constituées d'un jeu de données appariées et anonymisées contenant des informations mensuelles sur les prestations sociales individuelles de l'aide sociale (AS), de l'assurance-invalidité (AI) et de l'assurance-chômage (AC) ainsi que sur le compte individuel (CI) de l'activité professionnelle. Dans le présent rapport, ce jeu de données est désigné par l'abréviation «SHIVALV+IK»¹. L'univers de base comprend des personnes de 18 à 65 ans qui ont commencé à toucher des indemnités journalières de l'assurance-chômage (AC) durant la période 2010 à 2015. L'analyse est effectuée sur la base de cohortes annuelles. Elle inclut les informations relatives à la perception de prestations des assurances sociales et de l'aide sociale ainsi qu'à l'activité professionnelle durant les 48 mois ou 4 ans suivants.

Pour ce qui est de la mise en œuvre méthodologique, la cohorte 2010 a été soumise, dans un premier temps, à une procédure en deux phases de regroupement des séquences (apprentissage automatique non supervisé) pour identifier des parcours types, qui ont ensuite fait l'objet d'une analyse et d'une interprétation du contenu à l'aide d'indicateurs longitudinaux et de représentations graphiques («State Distribution Plots»). Dans un deuxième temps, ce partitionnement initial en parcours types (modèle) a été appliqué aux cohortes des années suivantes (2011–2015) à l'aide de l'apprentissage automatique supervisé (prévision). La validité du modèle a ainsi été vérifiée pour chaque cohorte, en se fondant sur l'évaluation de différents critères.

Cette démarche met l'accent sur l'identification et l'analyse des parcours types agrégés et l'application du modèle de référence à d'autres cohortes. L'utilisation de prédictions individuelles à quelque fin que ce soit est exclue.

Dix groupes de parcours types ont été déterminés pour les personnes ayant perçu des indemnités journalières de l'assurance-chômage pour la première fois en 2010 (tableau 1), groupes qui sont en majorité restés stables dans les cohortes de 2010 à 2015 (8 des 10 parcours types). Sur le plan du contenu, plusieurs groupes émergent dans lesquels les personnes réintègrent le marché du travail après une phase de perception d'indemnités de chômage. Les groupes se différencient par la durée de perception des indemnités journalières (groupes 1 et 2), par l'existence d'une phase de gain intermédiaire (groupe 3) ainsi que par des périodes répétées de perception de l'AC interrompues par des phases d'activité rémunérée (groupe 4). En outre, des groupes se dessinent qui tendent clairement à une perception durable de rentes AI ou de prestations de l'aide sociale (tableau 2; groupes 5, 6, 7, 8 et 9). Parmi ceux-ci, on distingue deux groupes de personnes ayant touché ces prestations pour la première fois (groupes 5, 9), deux autres ayant eu des phases prolongées de revenu professionnel complémentaire (groupes 6, 7) et un groupe de personnes qui, avant de toucher des indemnités journalières de l'assurance-chômage, dépendaient déjà de l'aide sociale de façon permanente ou répétée. Le dernier groupe (10) enfin réunit les personnes qui, pendant la période d'observation, n'étaient plus saisies durablement dans les systèmes considérés (AS, AI, AC,

¹ SH de l'allemand «Sozialhilfe», aide sociale

IV de l'allemand «Invalidenversicherung», assurance-invalidité

ALV de l'allemand «Arbeitslosenversicherung», assurance-chômage

IK de l'allemand «Individuelle Konten des Beitrags zur Altersvorsorge», compte individuel de l'activité professionnelle

CI/activité professionnelle). Ce projet a montré que la méthode de regroupement choisie, le «sequence clustering», est un algorithme prometteur pour produire des résultats pertinents du point de vue du contenu comme de l'analyse. Il permet de réduire considérablement la complexité des données sur les parcours et élargit ainsi les possibilités d'analyse grâce à l'identification de schémas qu'il n'est pas possible de déceler par déduction de manière anticipée.

Les données de séries temporelles revêtent une grande importance pour rendre ces informations encore plus pertinentes pour les destinataires de la statistique publique, notamment pour la gestion politique. Le modèle de regroupement initial ne peut cependant pas être appliqué sans autre à une nouvelle cohorte. Dans le présent projet, on a remédié à cette difficulté en appliquant ce modèle à de nouvelles cohortes par le biais d'un algorithme de prédiction. Ce procédé fonctionne bien et les critères utilisés pour déterminer à partir de quel moment il n'était plus possible d'appliquer ce modèle ont fait leur preuve dans le cas présent. Les constats tirés du projet présentent une valeur ajoutée concrète pour la production statistique standard, aussi bien pour les indicateurs longitudinaux nouvellement créés et leur visualisation que pour la définition de profils de parcours descriptifs et quantitatifs².

Les connaissances acquises, les possibilités et les limitations de l'application de méthodes d'apprentissage automatique dans la statistique publique font l'objet d'une analyse approfondie dans le rapport. En partant des principaux enseignements, des recommandations sont formulées pour des projets similaires de l'OFS. Les conclusions débouchent sur l'application d'une approche analytique générique et inductive pour les données relatives aux parcours individuels dans la production statistique de l'OFS.

² Voir la publication «Parcours dans le système de sécurité sociale en 2021»: <https://www.bfs.admin.ch/bfs/fr/home/statistiques/securite-sociale/analyses-parcours-systeme/analyses.assetdetail.25385873.html>