Prévisions du nombre de chômeurs et de demandeurs d'emploi dans le canton de Vaud

Rapport méthodologique

Statistique Vaud, juin 2026

Auteurs[1] : Cécile Hediger, Maxime Carron, Marie Dupraz

STATISTIQUE VAUD

Résumé

Le taux de chômage est un indicateur essentiel de la situation économique d'une région. Le suivi de son évolution au cours des mois à venir est une préoccupation importante. Dans ce but, Statistique Vaud a développé un modèle de prévision du chômage pour le canton, avec un horizon de 24 mois. Ce modèle combine des méthodes traditionnelles d'analyse de séries temporelles et l'apprentissage automatique (*machine learning*). Cette combinaison permet de renforcer la précision des prévisions et la robustesse du modèle. Les prévisions sont réalisées chaque mois et publiées sur le site web de Statistique Vaud.

[1] Ce travail a été réalisé avec l'aide du Swiss Data Science Center (SDSC), en particulier d'Aitor Ganuza Izagirre dans le cadre d'un travail de semestre et d'Alessandro Nesti. Nous remercions également Mohaddeseh Jandaghian et Teodoro Bevilacqua pour leur travail sur ce projet à Statistique Vaud.

1. Introduction

Le taux de chômage est un élément clé pour comprendre la situation économique d'une région ou d'un pays. Prévoir le taux de chômage est utile à l'analyse du contexte économique et au pilotage des offices régionaux de placement. A cette fin, Statistique Vaud a développé un modèle de prévision du taux de chômage cantonal pour les 24 prochains mois. Ce modèle combine une approche classique d'analyse de séries temporelles avec une approche basée sur l'apprentissage automatique. Chacune de ces approches apporte ses propres avantages, et leur combinaison permet d'améliorer la précision des prévisions ainsi que la robustesse du modèle.

Ce modèle de prévision, en développement depuis 2022 sous le sigle de la « statistique expérimentale », a été testé et validé rétroactivement sur plusieurs années. Au vu des résultats satisfaisants sur une période étendue, il est dorénavant utilisé de manière standard. Les prévisions sont communiquées chaque mois sur le site internet de Statistique Vaud[2].

C'est quoi la statistique expérimentale ?

La statistique expérimentale complète les produits de statistique classique en proposant des projets d'innovation qui font intervenir des nouvelles méthodes, de nouvelles sources de données ou qui présentent des résultats novateurs. Elle offre un espace d'expérimentation qui permet également de faire preuve de réactivité et de souplesse dans le processus de diffusion ainsi que de solliciter les retours du public.

Contrairement aux produits de statistique classique, les produits issus de la statistique expérimentale présentent un potentiel de maturité et peuvent être amenés à évoluer dans le temps. Pour faciliter l'interprétation, chaque publication s'accompagne d'une description des limites et incertitudes.

Le projet de refonte du modèle de prévision chômage

La révision du modèle de prévision du chômage s'inscrit dans une démarche de statistique expérimentale. L'utilisation de l'apprentissage automatique (*machine learning*) constituait en effet une expérimentation dont les résultats n'étaient pas garantis : rien n'assurait que cette nouvelle approche résoudrait les problèmes rencontrés avec le modèle traditionnel, ni qu'elle produirait des résultats d'une qualité jugée suffisante.

L'usage du *machine learning* dans le cadre de la statistique publique soulève par ailleurs plusieurs enjeux. La confiance des utilisateurs et utilisatrices, pilier fondamental de la statistique publique, repose sur la transparence, l'explicabilité et la reproductibilité des méthodes mises en œuvre par l'institution. Ces principes permettent aux utilisateurs et utilisatrices de comprendre la manière dont les résultats sont produits et, idéalement, de les reproduire à partir des mêmes données.

[2] https://www.vd.ch/etat-droit-finances/statistique/tableaux-de-synthese/economie « Chômage en chiffres et en graphiques »

Or, les modèles de *machine learning*, souvent perçus comme des « boîtes noires », sont plus difficiles à expliquer que les modèles usuels de statistique classique. Certains de leurs choix ou optimisations ne sont pas toujours directement interprétables, tandis que leurs performances dépendent fortement des données d'entraînement utilisées. Une attention particulière doit donc être portée à la qualité et à la représentativité de ces données, ainsi qu'aux risques de biais qu'elles peuvent introduire.

Par ailleurs, certaines exigences méthodologiques propres à la statistique publique – comme la mesure des erreurs, la quantification de l'incertitude ou la reproductibilité des résultats dans le temps – deviennent plus complexes à satisfaire avec des modèles de *machine learning*.

2. Chronologie

Depuis 2014, Statistique Vaud prévoit le nombre mensuel de chômeurs et de demandeurs d'emploi dans le canton pour les 24 prochains mois. Le modèle initial est basé sur des méthodes traditionnelles d'analyse de séries temporelles. Diverses variables liées à la conjoncture économique sont employées dans le modèle.

Ce modèle traditionnel offre des résultats globalement de bonne qualité, mais souffre de quelques défauts, notamment de produire des résultats non plausibles en cas de forte hausse du chômage, comme ce fut le cas au début de la période Covid. Aujourd'hui, les méthodes d'apprentissage automatique (*machine learning*) sont largement utilisées pour prévoir des indicateurs économiques. Après douze années d'utilisation du modèle traditionnel, une mise à jour était pertinente.
Cette mise à jour comprend :

- une revue des variables utilisées dans le modèle traditionnel et l'ajout de nouvelles variables ;
- le test de plusieurs modèles d'apprentissage automatique appliqués aux séries temporelles du nombre de chômeurs et le choix du meilleur modèle ;
- la comparaison des résultats du nouveau modèle basés sur l'apprentissage automatique et du modèle traditionnel.

Le nouveau modèle de prévision utilisé dès 2026 est une pondération des résultats du modèle traditionnel et d'un modèle basé sur l'apprentissage automatique (*machine learning*). Pour cette partie, un modèle *XGBoost* a été choisi. L'utilisation simultanée de ces deux modèles permet de corriger les biais de chacun d'eux, notamment en cas de brusque hausse du chômage. L'emploi du *machine learning* pour produire des résultats constitue une première expérience à Statistique Vaud. C'est pourquoi il est particulièrement intéressant de combiner une approche traditionnelle à de l'apprentissage automatique, dans le but de mettre en parallèle les deux approches sur une longue durée.

3. Définition des chômeurs et des demandeurs d'emploi

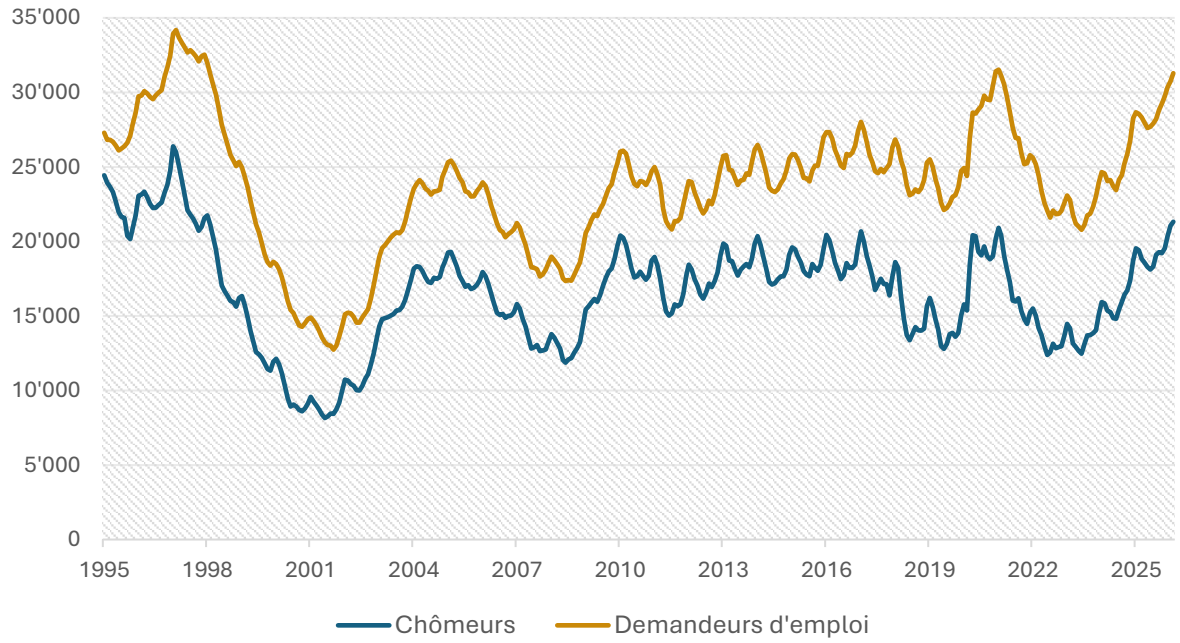
Les séries historiques du nombre de chômeurs et de demandeurs d'emploi dans le canton de Vaud provient du Secrétariat d'Etat à l'économie (SECO). Tous les mois, les offices régionaux de placement (ORP) du pays fournissent des données sur leurs bénéficiaires au SECO. Le SECO met

ensuite ces données à disposition du public et des cantons, par exemple dans son bulletin mensuel[3]. Selon la définition du SECO, est considéré comme **demandeuse d'emploi** toute personne inscrite dans un ORP, alors qu'est considéré comme **chômeuse** toute personne inscrite dans un ORP, qui ne travaille pas et qui est disponible de suite en vue d'un placement. La définition de chômeur est donc plus restrictive que celle de demandeur d'emploi, car elle comprend la notion de disponibilité immédiate. Tous les chômeurs sont aussi des demandeurs d'emploi, mais l'inverse n'est pas vrai. Par exemple, si la personne est inscrite dans un ORP mais touche un gain intermédiaire ou participe à une mesure relative au marché du travail, elle sera considérée comme demandeuse d'emploi, mais pas chômeuse[4]. Le fait de toucher ou non une indemnité chômage ne fait pas partie de la définition de chômeur.

Le **taux de chômage** est défini comme le nombre de chômeurs divisé par la population active. La population active est calculée au niveau du canton par le SECO, sur la base du Relevé structurel. Elle est mise à jour tous les trois ans. En 2026, la population active du canton de Vaud se monte à 427'805 personnes.

Une autre mesure du nombre de chômeurs existe : le chômage selon la définition du Bureau International du Travail (BIT). Ce taux est mesuré tous les trimestres par l'OFS via l'Enquête suisse sur la population active. Dans cette définition, les personnes au chômage n'ont pas besoin d'être inscrites dans un ORP pour être qualifiées comme telles. Les conditions sont qu'elles n'exercent pas d'activité rémunérée, soient activement à la recherche d'un emploi et immédiatement disponibles. Les résultats sont publiés par l'OFS par grande région, mais pas au niveau des cantons.

Figure 1 : Nombre de chômeurs et de demandeurs d'emploi dans le canton de Vaud, 1995-2025



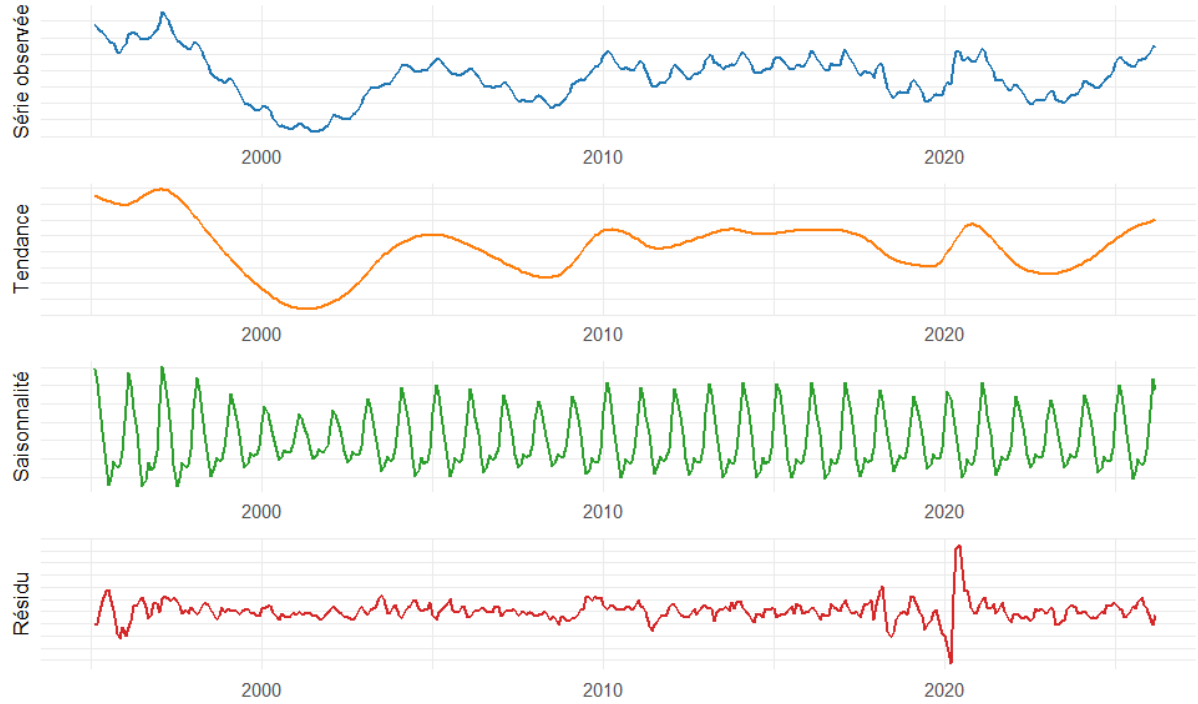
Source : SECO

[3] https://www.amstat.ch/v2/amstat_fr.html « Dates de publication »
[4] Plus de détails sur la différence de définition entre chômeurs et demandeurs d'emploi : https://dievolkswirtschaft.ch/fr/2024/07/chomage-des-taux-differents-selon-les-statistiques/

4. Qu'est-ce qui influence le chômage ?

Pour prévoir les fluctuations futures d'une série temporelle, il est nécessaire de comprendre sa structure. Dans le cas du nombre de chômeurs ou de demandeurs d'emploi, la série peut être décomposée en trois parties (voir figure 2) : la tendance, la saisonnalité et le terme d'erreurs (ou résidu).

Figure 2 : Décomposition de la série temporelle des chômeurs, Vaud, 1995-2025



La tendance est la direction générale, à la hausse, à la baisse ou à la stabilité. Cette composante est la plus difficile à prévoir car elle dépend de la conjoncture économique du canton et du pays.

La saisonnalité est le mouvement de la série observé chaque mois durant une année et qui se répète régulièrement d'une année à l'autre. Par exemple, pour les chômeurs et les demandeurs d'emploi, il existe une saisonnalité hivernale forte, avec un pic entre décembre et février. Ceci est expliqué par toutes les activités qui emploient moins de personnes durant l'hiver, principalement dans les secteurs de la construction et de l'agriculture. Un autre pic de moindre ampleur survient en août-septembre dans le canton de Vaud ; il s'agit des étudiants et étudiantes qui terminent leur diplôme et s'inscrivent au chômage juste après. La saisonnalité est facilement mesurée et peut être éliminée des données observées, appelée ensuite « séries désaisonnalisées ».

Pour prévoir la tendance, qui dépend de la conjoncture économique, des variables captant cette conjoncture sont employées. Nous utilisons trois indicateurs provenant des tests conjoncturels de l'institut KOF pour le canton de Vaud (marche des affaires, carnet de commandes et production). Au niveau suisse, nous employons l'indice de climat conjoncturel, l'indice du climat de consommation, le PIB et l'évolution des salaires. Selon la théorie économique, le chômage est aussi influencé par le niveau d'inflation[5] et les taux d'intérêts, nous employons donc aussi

[5] Voir la théorie de la courbe de Phillips par exemple.

l'indice des prix à la consommation et un indicateur du taux directeur de la BNS. Les variables utilisées dans les modèles de prévision sont présentées dans la partie suivante.

5. Variables utilisées

Après examen des données disponibles et revue d'articles sur le sujet, une base de données a été conçue comprenant des données économiques en lien avec le chômage. Les données récoltées sont de temporalités mensuelles ou trimestrielles. Toutes les séries sont disponibles depuis janvier 1995.

Le tableau 1 présente les variables employées pour effectuer les prévisions. Le modèle d'apprentissage automatique utilise toutes les variables de la liste, tandis que le modèle traditionnel en utilise dix.

A l'exception des variables provenant du KOF et de l'emploi vaudois, toutes les données sont accessibles librement sur internet. Les liens vers les pages web spécifiques sont ajoutés au tableau.

Tableau 1 : Liste des variables utilisées

Variables	Temporalité	Source	Utilisées dans le modèle traditionnel
Chômeurs VD	Mensuelle	SECO	X
Demandeurs d'emploi VD	Mensuelle	SECO	X
Chômeurs CH	Mensuelle	SECO	X
Places vacantes VD	Mensuelle	SECO	X
Places vacantes CH	Mensuelle	SECO	X
Indice des prix à la consommation CH	Mensuelle	OFS	
Taux d'intérêts court terme BNS (CH)	Mensuelle	BNS	
Tests conjoncturels : Marche des affaires VD	Mensuelle	KOF	
Tests conjoncturels : Carnet Commande VD	Mensuelle	KOF	
Tests conjoncturels : Production VD	Mensuelle	KOF	X
Climat conjoncturel suisse (CH)	Mensuelle	SECO	
Chômeurs de longue durée VD	Mensuelle	SECO	
Climat de consommation (CH)	Trimestrielle*	SECO	X
PIB (CH)	Trimestrielle	OFS	X
Salaires (CH)	Trimestrielle	OFS	X
Emploi VD	Trimestrielle	Statem, StatVD	
Emploi région lémanique	Trimestrielle	OFS	X
Chômeurs BIT région lémanique	Trimestrielle	OFS	

*Est passé en mensuel depuis 2024, mais est gardé en valeurs trimestrielles dans le modèle pour pouvoir utiliser la série depuis 1995. Le résultat du premier mois du trimestre est utilisé pour le trimestre en entier.

6. Modèles de prévision testés

Une série temporelle est une série d'observations prises successivement à des intervalles réguliers. La prévision de séries temporelles se base sur l'historique de ces observations, avec l'hypothèse qu'il est possible d'identifier des structures qui se répètent dans les valeurs historiques et de les utiliser pour prédire les valeurs futures (Petropoulos, 2022). Lorsque l'objectif est de prévoir uniquement l'intervalle suivant, on l'appelle la prévision en une étape (*one-step forecasting*). Si l'objectif est de prévoir plusieurs intervalles suivants, 24 mois dans notre cas, cela se nomme la prévision multi-étapes (*mutli-step forecasting*).

Il existe deux approches pour réaliser une prévision multi-étape : l'approche directe ou l'approche itérative. L'approche directe consiste à construire un modèle qui prévoit directement les résultats des différents intervalles. L'approche itérative consiste à prévoir l'intervalle numéro un, puis appliquer le modèle à cette itération pour prévoir l'intervalle numéro deux, etc. jusqu'à atteindre l'intervalle maximum souhaité. Nous privilégions l'approche directe pour ne pas propager une erreur initiale à tous les horizons suivants.

Pour évaluer les résultats d'un modèle, celui-ci est exécuté sur des données antérieures afin de prévoir les prochaines observations. Puis la différence entre chaque observation prédite et la valeur réellement observée est calculée. La métrique la plus souvent utilisée pour ceci est la racine carrée de l'erreur quadratique moyenne (*Root Mean Square Error – RMSE*), définie ainsi :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2}$$

Où $i = (1,2,3, \dots ,24)$ sont les périodes à prévoir, P_i les prévisions et O_i les observations. Plus le RMSE est bas, plus le modèle est performant.

A) Modèle traditionnel

Depuis 2014, un modèle basé sur des méthodes traditionnelles de prévision de séries temporelles est utilisé à Statistique Vaud pour prévoir le taux de chômage pour les 24 prochains mois. Il prévoit le nombre de chômeurs et demandeurs d'emploi en utilisant la méthode VEC (Vector Error Correction). Des modèles ARIMA (Autoregressive Integrated Moving Average) et VAR (Vector Auto Regression) avaient aussi été testés, mais non retenus en raison d'une performance moins élevée[6].

Ce modèle traditionnel combine les prévisions de plusieurs modèles VEC (voir annexe i pour des détails sur les modèles VEC). Chaque modèle teste une combinaison possible des huit variables explicatives et en garde entre trois et cinq. Sont retenus uniquement les modèles qui ont le mieux performé durant les trois dernières années, soit là où les erreurs de prévision sont minimales. La prévision finale est la moyenne des prévisions des cinq meilleurs modèles.

[6] Un modèle BVAR était aussi utilisé de 2014 à 2021 en combinaison du modèle VEC, mais a dû être abandonné à cause de problèmes de fonctionnement sur le logiciel R.

La performance globale du modèle traditionnel est satisfaisante, mais deux défauts sont apparus avec l'expérience de plusieurs années d'utilisation :

1) Le modèle détecte la tendance du nombre de chômeurs ou de demandeurs d'emploi – à la hausse ou à la baisse – et va le continuer avec la même pente sans infléchissement jusqu'au prochain changement de tendance. De ce fait, des valeurs non-plausibles (trop hautes ou trop basses) du nombre de chômeurs ou de demandeurs d'emploi peuvent être régulièrement prévues par le modèle.
2) Durant la pandémie, à cause de la brusque hausse du nombre de chômeurs, le modèle prévoyait une explosion de leurs effectifs, atteignant des niveaux irréalistes. Il a mis huit mois à se réajuster.

L'objectif principal de l'utilisation de nouveaux modèles basés sur l'apprentissage automatique est de remédier à ces deux problèmes.

B) Modèle basé sur l'apprentissage automatique

Les modèles de prévision basés sur l'apprentissage automatique ne prescrivent pas de relations spécifiques entre les données, au contraire des modèles traditionnels. Il n'y a pas d'ensemble d'équations qui structure le système, ce qui permet aux relations entre données d'être apprises automatiquement. Les méthodes qui s'appuient sur des algorithmes de régression non-linéaire et non-structuré, par exemple les réseaux de neurones ou les arbres de décisions, sont considérés comme relevant de l'apprentissage automatique (Petropoulos, 2022).

Les méthodes d'apprentissage automatique ont l'avantage de pouvoir absorber une grande quantité de données. Elles sont aussi plus génériques que les méthodes traditionnelles et peuvent être adaptées rapidement à des séries d'observations avec différentes caractéristiques. En revanche, elles demandent de plus grandes bases de données pour fonctionner correctement et sont moins explicables que les méthodes traditionnelles (phénomène de la boîte noire). Concernant les séries temporelles, une difficulté réside dans le fait que ces méthodes ne sont pas performantes pour gérer la temporalité et les saisonnalités (Makridakis, 2018). Une solution est d'ajouter des éléments temporels aux données, d'utiliser les séries décalées (*lags*) et de désaisonnaliser les séries.

Deux modèles d'apprentissage automatique ont été testés : XGBoost (eXtreme Gradient Boosting) et Ridge Regression. Le modèle XGBoost était plus performant, c'est donc celui-ci que nous avons sélectionné. Des méthodes de *deep learning* avaient aussi été envisagées, mais le nombre d'observations est trop limité dans notre cas pour ce genre de modèle.

XGBoost est un algorithme construisant des arbres de décisions de manière séquentielle, chaque nouvel arbre visant à corriger les erreurs commises par son prédécesseur. Cette méthode permet à XGBoost de capturer efficacement les relations non linéaires dans les données (Chen, 2017). Pour l'adapter à la prévision de séries temporelles, nous désaisonnalisons préalablement les séries des chômeurs et des demandeurs d'emploi, puis les resaisonnalisons en fin de processus. Des prédicteurs supplémentaires sont construits sur la base des variables de base pour inclure de la temporalité : lags, moyennes mobiles, différentiels, etc. Afin de prévenir des risques de surajustement, un certain nombre de modèles sont lancés en faisant varier le processus d'expansion de variables, puis la moyenne de ces modèles est utilisée comme prévision. La description de la préparation des données et du processus d'expansion de variables se trouve dans l'annexe ii.

Nous utilisons l'approche directe et non l'approche itérative, c'est-à-dire qu'il y a 24 modèles différents construits, un pour chaque horizon de prévision. Chaque modèle est formé indépendamment des autres, en utilisant les informations disponibles jusqu'au temps t − 1. Le premier modèle prédit le nombre de chômeur au mois t, le second au mois t + 1, et ainsi de suite. Cela permet d'éviter de propager une erreur initiale à tous les horizons suivants, risque inhérent à l'approche itérative.

Comme XGBoost est capable de prendre en charge de large base de données, nous avons ajouté trois variables dans le modèle par rapport à l'approche traditionnelle : le climat conjoncturel suisse, fourni mensuellement par le SECO, le nombre de chômeurs de longue durée dans le canton de Vaud (>1 an en tant que chômeur) et le nombre de chômeurs selon la définition du BIT dans la région lémanique, mesuré par l'OFS. La liste complète se trouve dans le tableau 1. Toutes les variables sont disponibles depuis janvier 1995.

7. Résultats

Pour évaluer la performance des différents modèles, la période allant de juin 2022 à décembre 2025 a été privilégiée afin de minimiser l'impact de la pandémie de Covid sur les résultats[7]. Les modèles sont exécutés rétroactivement pour chaque mois, produisant des prévisions pour les 24 mois à partir du mois sélectionné. Ces prévisions sont ensuite confrontées aux données réelles afin de mesurer la performance, par modèle et par horizon.

Les trois modèles à évaluer sont les suivants : le modèle traditionnel (qui est une combinaison de modèles VEC), le modèle *machine learning* (combinaison de modèles XGBoost), et un modèle hybride pondérant les deux premiers. Ce dernier modèle calcule une pondération entre les deux approches pour chaque horizon de prévision de sorte à minimiser l'erreur quadratique des prévisions antérieures.

Globalement, le modèle *machine learning* (ML) reste moins performant que le modèle traditionnel, quel que soit l'horizon de prévision (figure 3). Toutefois, son comportement permet de combler certaines lacunes du modèle traditionnel, en particulier pour les prévisions à long terme. À certains égards, les deux modèles sont complémentaires, dans la mesure où le modèle traditionnel tend à reproduire la tendance en cours pour prédire sa série, là où le modèle ML est plus prudent et ne prolonge pas systématiquement la tendance dans ses prévisions. De ce fait, le modèle hybride fondé sur la pondération de ces deux approches donne de meilleurs résultats que les modèles considérés séparément. Ceci se confirme pour l'ensemble des horizons ainsi qu'en période d'instabilité, comme lors de la pandémie de Covid. Par ailleurs, bien que le modèle pondéré incorpore davantage le modèle traditionnel, lui accordant une pondération moyenne de deux tiers sur l'ensemble des horizons, la contribution du modèle ML est croissante pour les horizons longs, traduisant sa capacité à compenser les erreurs de long terme propres au modèle économétrique traditionnel (figure 4).

[7] Il est également difficile de remonter avant la période Covid, car une partie des variables utilisées sont périodiquement rétro-corrigées. L'historique de ces variables sans rétro-correction a été conservé uniquement depuis 2022.

Les résultats présentés ci-après concernent les chômeurs, cependant les conclusions restent valables pour les demandeurs d'emploi. Pour ces derniers, des résultats peuvent être fournis sur demande.

Figure 3 : Erreur moyenne des modèles en fonction de l'horizon de prévision

Période de test : juin 2022 à décembre 2025

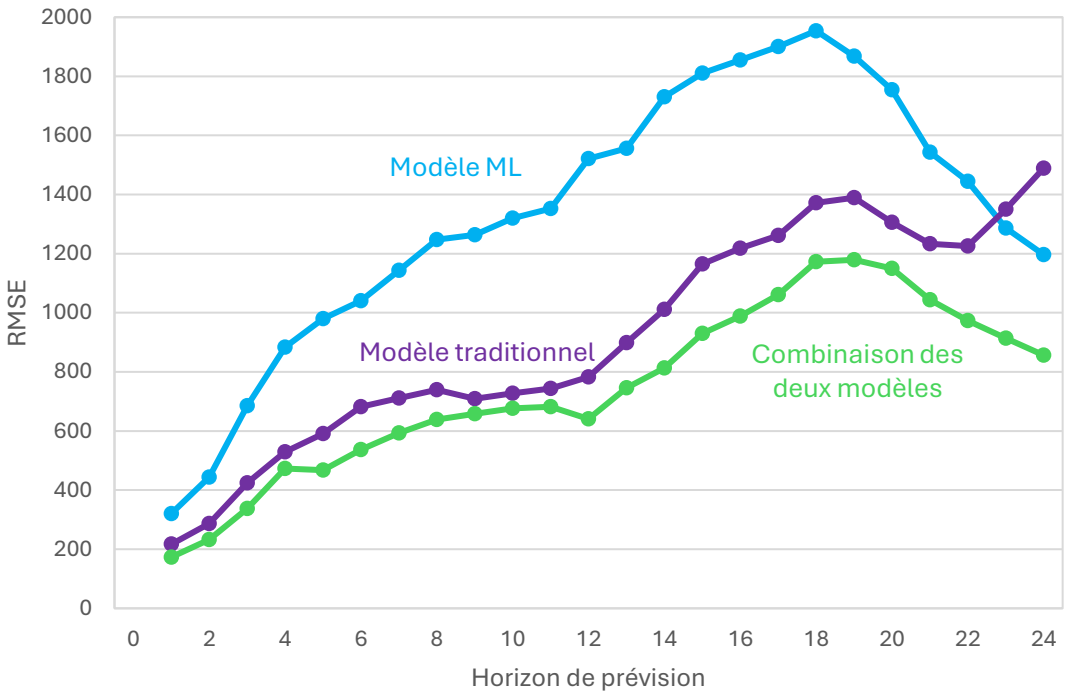
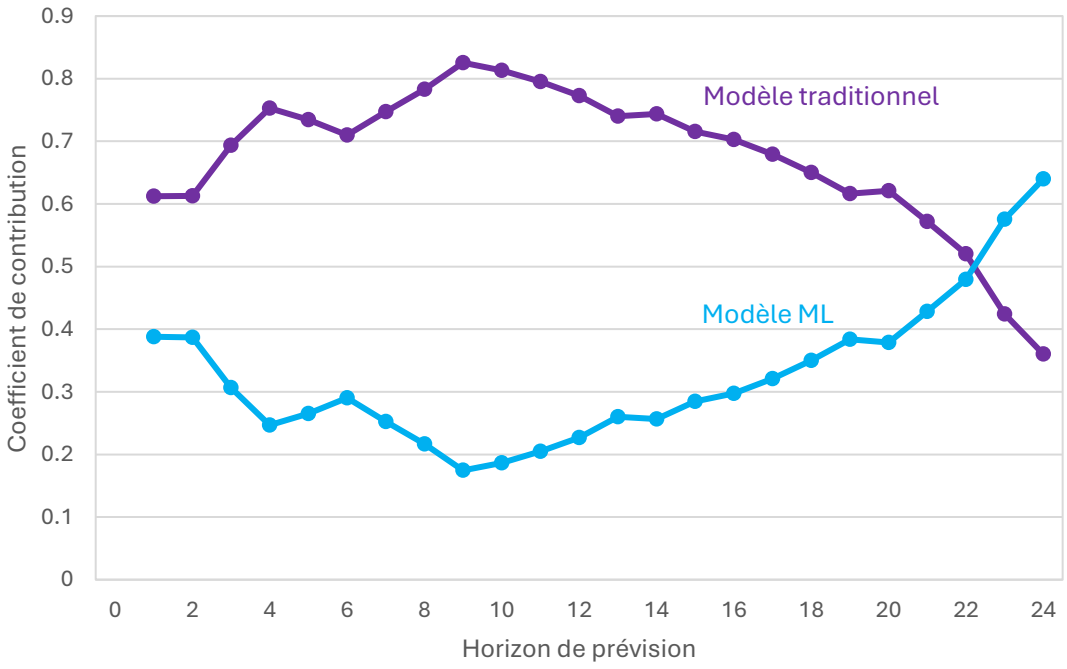


Figure 4 : Coefficients de contribution des modèles en fonction de l'horizon de prévision

Période de test : juin 2022 à décembre 2025

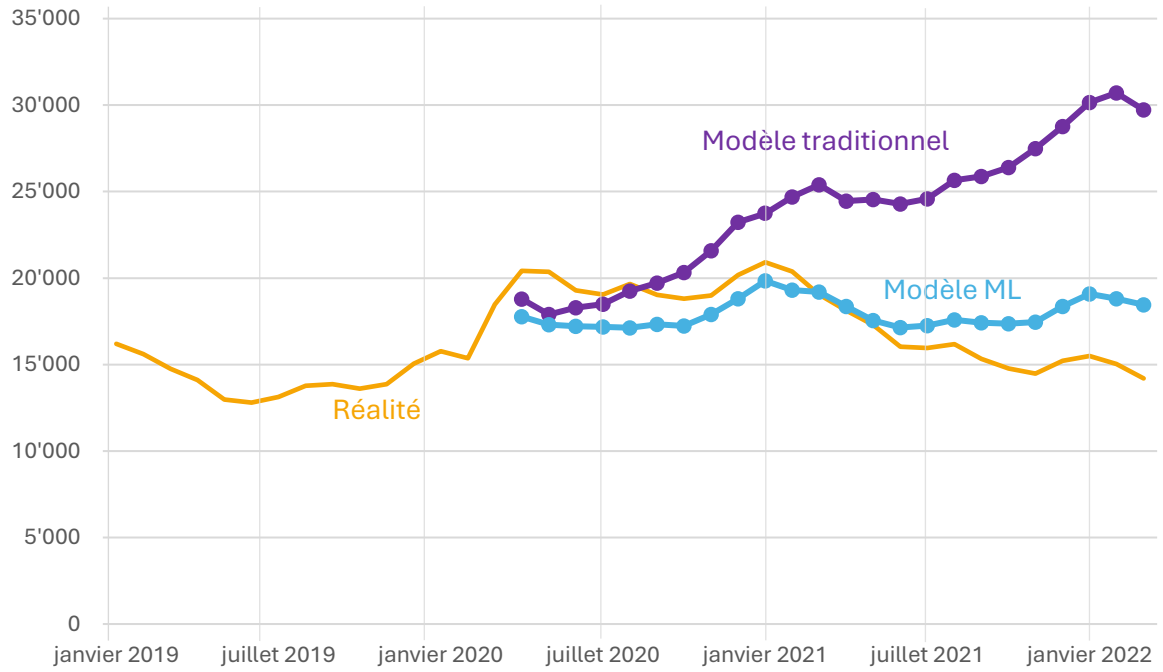


8. Performance durant la pandémie

Les mauvaises performances du modèle traditionnel durant la pandémie étaient l'une des raisons de la mise à jour du modèle. Le modèle traditionnel a en effet donné des résultats non-plausibles dès avril 2020 et pendant les huit mois suivants, prévoyant parfois jusqu'à 70'000 chômeurs dans le canton. Finalement, le maximum réellement atteint a été de 21'000 chômeurs en janvier 2021.

Le modèle ML n'est pas sensible à la brusque hausse du chômage débutée en mars 2020 (figure 5). Avec le modèle ML, les prévisions en avril 2020 pour les 24 prochains mois sont proches de la réalité observée. Si une nouvelle crise survient provoquant une brusque montée du chômage, nous savons que le modèle ML y sera plus robuste que le modèle traditionnel. Il restera à déterminer la façon d'ajuster le poids à attribuer à chacun des deux modèles.

Figure 5 : Nombre de chômeurs vaudois, prévisions dès avril 2020



9. Conclusions

Les résultats ont pu mettre en évidence la pertinence de **combiner le modèle *machine learning* et le modèle traditionnel** en vue de produire des prévisions du chômage plus fiables et robustes. Les modèles basés sur l'apprentissage automatique démontrent à la fois leurs atouts et leurs faiblesses : ils permettent de stabiliser le modèle économétrique traditionnel, mais ils restent moins performants du fait de leurs difficultés inhérentes à capturer la dimension temporelle, et ne peuvent donc pas se substituer aux modèles traditionnels tels quels.

Plusieurs pistes ont été identifiées pour pouvoir améliorer le modèle *machine learning* dans le futur : réajustement des techniques d'expansion de variables selon l'horizon de prédiction, prévision des biais temporels du modèle pour pouvoir les corriger et diverses techniques pour minimiser le surajustement du modèle aux données d'entraînement. Ces développements possibles ne sont pas envisagés à l'heure actuelle, dans la mesure où le modèle *machine learning* remplit la fonction principale pour laquelle il a été conçu, à savoir combler les limites du modèle traditionnel.

L'ensemble du code R associé aux différents modèles est disponible sur demande.

Références

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Ganuza Izagirre, A. (2024). Forecasting the unemployment rate in the canton of Vaud with machine learning algorithms. *Semester Project Report, EPFL*.

Gogas, P., Papadimitriou, T., & Sofianos, E. (2022). Forecasting unemployment in the euro area with machine learning. *Journal of Forecasting*, *41*(3), 551-566.

Kreiner, A., & Duca, J. (2020). Can machine learning on economic data better forecast the unemployment rate? *Applied Economics Letters*, *27*(17), 1434-1437.

Lones, M. A. (2021). How to avoid machine learning pitfalls: a guide for academic researchers. *arXiv preprint arXiv:2108.02497*.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS one*, *13*(3), e0194889.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., ... & Ziel, F. (2022). Forecasting: theory and practice. *International Journal of forecasting*, *38*(3), 705-871.

Weber, B. (2024). Chômage : des taux différents selon les statistiques. *La Vie économique*, juillet 2024.

Annexes

i. Modèles VAR et VEC

Un modèle VAR (*Vector AutoRegressive*) est un modèle multivarié qui explique chaque variable par ses propres valeurs passées ainsi que par les valeurs passées des autres variables du système. Un modèle VEC (*Vector Error Correction Model*) est un modèle VAR auquel on a corrigé les erreurs en ajoutant des liens de long terme dans le modèle à l'aide de vecteurs de cointégration. Ces liens de cointégration sont détectables à l'aide des tests de Granger-Engel ou Johansen par exemple. Le test de Granger-Engel ne peut vérifier qu'une relation à la fois, alors que le test de Johansen est valable pour plusieurs relations de cointégration, pour autant que les séries soient intégrées d'ordre 1 et que l'échantillon soit suffisamment grand.

Sous l'hypothèse de l'existence d'au moins une relation de cointégration dans le panel de variables choisies et le chômage, une sélection automatique du nombre de relations avec une borne inférieure et supérieure peut être établie par le test de Johansen à l'aide d'un logiciel statistique.

Les modèles VAR nécessitent que les séries soient rendues stationnaires avant la modélisation alors que pour les modèles VEC, les séries en différence ou en niveau peuvent être utilisées sous certaines conditions. Selon la littérature sur le sujet et après divers tests empiriques, nous avons retenu des séries en niveau pour les modèles à correction d'erreurs.

ii. Préparation des données pour le modèle de *machine learning* (ML)

La première phase du traitement consiste à :

- Mensualiser l'ensemble des variables trimestrielles : les premières valeurs de chaque trimestre sont simplement projetées sur les mois concernés.
- Désaisonnaliser les séries de chômeurs et de demandeurs d'emploi. Ces séries sont resaisonnalisées à la fin du processus.
- Eliminer la période Covid (mars 2020 à mars 2022). Cette période était spéciale au niveau du chômage, avec notamment des changements dans la réglementation (prolongation du droit au chômage). Le modèle ML fonctionne mieux sans cette période.
- Rendre l'ensemble des séries stationnaires, de sorte à éviter un surapprentissage du modèle sur le niveau des séries temporelles.

Les variables de base ayant été préparées, elles sont soumises à plusieurs types d'expansions (*feature expansion*) afin d'alimenter le modèle XGBoost avec un set de prédicteurs diversifiés et permettant au modèle de mieux capturer la dimension temporelle.

Plusieurs formes d'expansion de variables ont été mobilisées :

- Lags de variables : des lags de variables ont été produits afin d'élargir le contexte temporel des modèles. Pour limiter les risques de surajustement, seuls trois lags par variables sont choisis : le lag qui est le mieux corrélé à la série à prédire, ainsi que deux lags

suffisamment espacés dans le temps (six mois) pour éviter la redondance d'information dans le set de prédicteurs.

- Différentiels de variables : sur le même modèle de sélection, des différentiels sont produits entre le mois le plus récent des données et trois différents mois antérieurs.

- Moyennes et écarts-type mobiles : des moyennes et écart-types mobiles sont calculés sur les données mensuelles antérieures pour différentes tailles de fenêtres (nombre de mois). Pour chaque taille de fenêtre, on choisit la fenêtre la plus récente, ainsi que celle qui lui est directement antérieure. Pour les moyennes uniquement, on calcule également le différentiel entre les deux moyennes mobiles de chaque taille de fenêtre. Les tailles de fenêtre choisies sont 6, 12 et 24 mois afin de capturer les effets de court, moyen et long terme.

Afin de limiter les risques de surajustement, plusieurs modèles sont produits en faisant varier le processus d'expansion. Pour chaque modèle, on fait le choix de prendre l'ensemble des lags et différentiels, toutes variables confondues, mais de réserver les autres expansions (moyennes mobiles, écart-type mobiles…) à seulement quatre variables. De cette façon, on garantit un socle commun de prédicteurs entre les modèles, tout en permettant à chaque modèle de représenter davantage certaines variables ciblées. Sont ainsi produits autant de modèles qu'il existe de combinaisons de quatre variables parmi une pré-sélection de variables, qui inclut nécessairement la série à prédire (chômeurs ou demandeurs). La pré-sélection de variables a été fixée expérimentalement, de sorte à exclure les variables dont la contribution était systématiquement faible, voire inexistante, et ainsi accélérer sensiblement l'exécution des modèles. L'exécution finale du modèle ML se fait en 20-30 minutes dans le logiciel R.

Contact
Statistique Vaud
Rue de la Paix 6 – 1014 Lausanne
021 316 29 99
info.stat@vd.ch
www.vd.ch/statvd