
Étude empirique de l'évolution de l'immigration en Europe par des méthodes statistiques et d'intelligence artificielle

Projet réseau neurone



Projet réalisé par :

Hugo SUSCOSSE

Oumou Jasmine NGWAYA KANDE

Enseignants responsables du cours :

Houtmaley

Buoiyour Jamal

Table des matières

Introduction.....	3
Revue de littérature.....	3
Présentation des données.....	5
Analyse descriptive des variables.....	5
Modélisation statistique.....	7
Modélisation par régression linéaire multiple.....	7
Régression quantile.....	8
Analyse temporelle.....	9
Décomposition de la série temporelle de l'immigration.....	10
Test de stationnarité.....	11
Modèles ARIMA/SARIMAX.....	11
Intelligence artificielle et prévision.....	12
Intelligence artificielle et modélisation des dynamiques migratoires.....	12
Mise en œuvre des modèles de prévision.....	13
Comparaison des performances des modèles.....	14
Conclusion et perspectives.....	14
Bibliographie.....	16
Annexe.....	17
Analyse descriptive).....	19

INTRODUCTION

La question de l'immigration et de ses dynamiques occupe une place centrale dans les débats contemporains, en particulier en France, où près d'un quart de la population a au moins un parent ou un grand-parent issu de l'immigration. Au cœur des tensions politiques et idéologiques, elle est souvent abordée sous un angle émotionnel, parfois détaché des faits objectifs.

Ce projet vise à étudier l'évolution de l'immigration en France et dans les pays de l'Union européenne, à l'aide d'outils de modélisation statistique et d'intelligence artificielle. L'objectif est double : d'une part, analyser les déterminants économiques et sociaux de l'immigration, et d'autre part, prévoir ses tendances futures à travers le temps, en exploitant des données chronologiques structurées.

Notre démarche s'inscrit dans un contexte sensible, marqué notamment par la polémique autour de la « théorie du grand remplacement ». Contrairement aux approches idéologiques, nous proposons ici une étude empirique, rigoureuse et scientifique, s'appuyant sur des données fiables pour éclairer les phénomènes migratoires. Par rapport aux travaux existants, notamment ceux du doctorant Yaya qui explore cette problématique en profondeur, notre étude se veut plus modérée dans le ton mais plus innovante dans la méthode, en s'appuyant sur une approche moderne utilisant les technologies d'analyse les plus avancées, tout en élargissant la focale à l'échelle de l'Union européenne.

Pour cela, nous avons reconstruit la base de données initiale (au format panel) en séries temporelles annuelles, permettant une analyse dynamique de l'immigration entre 2000 et 2021. Afin de répondre à la problématique de recherche suivante : Quels sont les principaux déterminants économiques, démographiques et sociaux de l'immigration ? Quelles sont les tendances temporelles observables en matière d'immigration en France et en Europe ? Peut-on prévoir l'évolution future du nombre d'immigrés dans les pays de l'Union européenne ? À partir de ces interrogations, le projet vise à construire des modèles prédictifs robustes, intégrant des variables explicatives économiques, sociales et démographiques, afin d'apporter une lecture objective et prospective des dynamiques migratoires en Europe.

REVUE DE LITTÉRATURE

De nombreuses recherches ont été menées ces dernières années sur les effets économiques de l'immigration, que ce soit aux États-Unis ou dans les pays de l'OCDE. Ces travaux permettent de mieux comprendre l'impact réel des mouvements migratoires sur les économies développées, en apportant des résultats empiriques éloignés des discours idéologiques.

Une première étude, réalisée aux États-Unis et publiée dans le PMC Journal, s'intéresse à l'effet de l'immigration sur la distribution des salaires. Les auteurs partent du constat que l'immigration est souvent perçue, dans le débat public, comme un facteur pouvant exercer

une pression à la baisse sur les salaires des travailleurs natifs les moins qualifiés. L'analyse, fondée sur des données micro-économiques couvrant plusieurs décennies, montre que l'impact global de l'immigration sur les salaires reste modeste. Les seules catégories légèrement affectées sont les travailleurs peu qualifiés, qui subissent une faible diminution de leur salaire relatif. À l'inverse, les travailleurs plus qualifiés, ou ceux dont les compétences sont complémentaires à celles des immigrés, peuvent même bénéficier d'une augmentation salariale. L'étude souligne enfin que la flexibilité du marché du travail américain facilite l'intégration des nouveaux arrivants, et que d'autres facteurs, comme les évolutions technologiques ou la mondialisation ont un poids bien plus important que l'immigration dans la structuration des salaires.

Un autre rapport, publié en 2024 par le Congressional Research Service, propose une vue d'ensemble plus large des effets économiques de l'immigration aux États-Unis. Il met en évidence des apports significatifs de l'immigration sur la croissance économique, l'innovation et l'entrepreneuriat. Les immigrés, notamment lorsqu'ils sont hautement qualifiés, contribuent fortement au dépôt de brevets, à la création d'entreprises et au développement de nouveaux secteurs. Le rapport constate que l'impact de l'immigration sur les salaires et l'emploi des travailleurs natifs est, en règle générale, faible ou non significatif. Il reconnaît cependant que certains effets négatifs localisés peuvent apparaître dans des secteurs ou chez des groupes vulnérables, ce qui justifie le besoin de politiques d'accompagnement. À moyen et long terme, l'immigration est perçue comme un levier favorable pour la soutenabilité des systèmes de protection sociale, en particulier celui des retraites.

Une troisième étude élargit le champ d'analyse en intégrant plusieurs pays de l'OCDE. L'objectif est d'évaluer les effets de l'immigration internationale sur l'emploi et les salaires des natifs. En s'appuyant sur des données panel de long terme et en utilisant des méthodes économétriques robustes (effets fixes, variables instrumentales), les auteurs concluent que les effets agrégés sont neutres voire légèrement positifs. Ils constatent également que l'impact de l'immigration varie selon les secteurs d'activité et les régions, mais que les économies s'ajustent rapidement. Un point important mis en avant concerne la complémentarité des compétences : les immigrés ayant souvent des profils différents de ceux des natifs, ils ne sont pas nécessairement en concurrence directe. Enfin, l'étude insiste sur le rôle de l'immigration dans le renouvellement démographique, notamment dans les pays confrontés au vieillissement de leur population.

À ces travaux économiques s'ajoute une étude sociopolitique réalisée dans le cadre d'un travail de recherche universitaire. Ce travail, conduit par un doctorant, porte sur la réception sociale et politique de la théorie du « grand remplacement », une idée selon laquelle les populations européennes seraient progressivement remplacées par des populations immigrées, en particulier originaires d'Afrique ou du Maghreb. Le contexte historique français, marqué par une longue tradition d'immigration notamment liée à la période coloniale, crée une tension entre sentiment d'inclusion et peur identitaire. Cette peur est exploitée par certains mouvements politiques, principalement à l'extrême droite, pour légitimer des discours hostiles à l'immigration. Des chercheurs tels qu'Hervé Le Bras ont pourtant largement réfuté ces thèses, rappelant l'incertitude des projections démographiques et la complexité des dynamiques migratoires.

Le travail du doctorant mobilise des outils d'analyse de sentiments sur un corpus d'articles de presse. En classifiant les discours médiatiques selon leur tonalité (positive ou négative), les résultats indiquent que 67 % des articles expriment un rejet ou un scepticisme envers la théorie du grand remplacement. Cette approche empirique montre que, malgré la diffusion

importante de cette idée dans les réseaux sociaux et certains médias, elle reste largement critiquée dans la presse généraliste. Néanmoins, sa réception dans l'opinion publique reste significative, notamment sous l'effet de la désinformation et des peurs collectives. Le travail souligne également que cette théorie, loin d'être une simple construction intellectuelle, a pu inspirer des actes violents, comme l'ont montré les attentats de Christchurch ou d'El Paso. En résumé, cette revue de littérature met en lumière deux points essentiels. D'une part, les effets économiques de l'immigration sont globalement faibles ou positifs, et ne justifient pas les inquiétudes alimentées par certains discours politiques. D'autre part, la théorie du grand remplacement, bien qu'idéologiquement puissante, repose sur des bases empiriques fragiles et comporte des risques réels pour la cohésion sociale. L'enjeu scientifique est donc de produire des analyses rigoureuses et fondées sur les données, afin de distinguer les faits des représentations.

Analyse descriptive

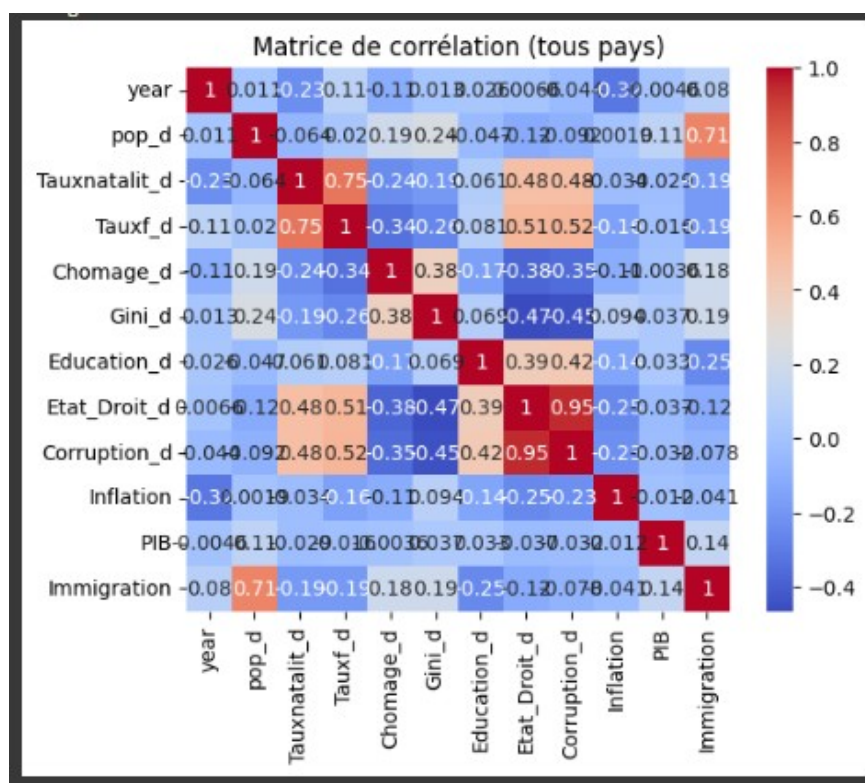
Présentation des données

Les données utilisées dans cette étude viennent d'un fichier appelé "Série temporelle 2000-2021.xlsx", contenant des informations pour plusieurs pays entre 2000 et 2021. Elles proviennent principalement de l'INSEE, de l'OCDE, et ont été complétées par la base de données construite par un doctorant, Yaya, qui travaille sur la question du « Grand Remplacement ».

Chaque ligne du fichier correspond à une année pour un pays donné, avec les informations suivantes : **pays** (le nom du pays), **annee** (l'année d'observation), **Immigration** (nombre d'immigrants enregistrés), **PIB** (en millions), **Chomage_d** (taux de chômage en %), **Tauxnatal_d** (taux de natalité), **Education_d** (indice ou taux d'éducation), **Gini_d** (coefficient de Gini (inégalités)), **Corruption_d** (indice de perception de la corruption), **Inflation** (taux d'inflation annuel en %).

Nous avons choisi ces données car elles permettent de comparer l'évolution de l'immigration avec différents facteurs économiques et sociaux. Les variables concernent uniquement les pays d'accueil, et non les pays d'origine, afin de savoir si les conditions économiques dans les pays de destination influencent le nombre d'immigrés.

Analyse descriptive des variables



L'analyse statistique préliminaire a permis de mieux appréhender la structure et la variabilité des données collectées. Le coefficient de variation (CV), qui mesure la dispersion relative des variables par rapport à leur moyenne, a été calculé pour les principales dimensions étudiées. Pour la variable immigration, le CV atteint 1,30, ce qui révèle une forte hétérogénéité des flux migratoires entre les pays et au fil des années. Une telle valeur suggère que les niveaux d'immigration varient considérablement d'un pays à l'autre, tant en volume absolu qu'en dynamique temporelle. La population présente également un coefficient de variation élevé, de l'ordre de 1,27. Cela indique que les pays considérés dans l'échantillon se distinguent par des tailles démographiques très contrastées, ce qui est cohérent avec la diversité géographique et économique de l'Union européenne. Les autres variables, telles que le taux de chômage, le taux de natalité ou le produit intérieur brut (PIB), présentent des coefficients de variation modérément élevés, ce qui traduit une certaine diversité structurelle mais dans des proportions plus contenues que celles observées pour l'immigration ou la population. Une analyse de la matrice de corrélation a ensuite été conduite afin d'examiner les liens entre les différentes variables explicatives. Cette matrice montre des corrélations relativement marquées entre certaines variables d'ordre sociodémographique, comme le taux de natalité et le taux de fécondité, ou encore entre le niveau d'éducation, l'état de droit et la perception de la corruption. Ces liaisons ne sont pas surprenantes dans la mesure où ces variables décrivent des dimensions proches du développement humain ou institutionnel. En revanche, la variable à expliquer, l'immigration, présente des corrélations faibles à modérées avec les autres variables du modèle, ne dépassant pas en valeur absolue le seuil de 0,25. Ce résultat confirme que l'immigration ne dépend pas d'un seul facteur, mais de plusieurs éléments combinés. Cela souligne le caractère multifactoriel du phénomène migratoire et limite la pertinence d'approches purement linéaires ou univariées. Par ailleurs, aucune corrélation supérieure à 0,8 n'a été observée entre les variables explicatives, ce qui

signifie qu'il n'existe pas de colinéarité excessive. Cela est un point positif pour la stabilité des modèles de régression multiple que nous appliquerons dans la suite de l'étude, bien que certaines variables fortement proches devront être surveillées.

La distribution statistique de la variable immigration a ensuite été examinée pour chaque pays, à l'aide du test de Shapiro-Wilk, afin d'évaluer la normalité des séries. Environ la moitié des pays étudiés présentent une distribution de l'immigration compatible avec la loi normale, c'est-à-dire que l'hypothèse nulle de normalité n'est pas rejetée au seuil de 5 %. C'est notamment le cas de l'Italie, du Luxembourg, des Pays-Bas, de la Belgique, de la Suède, de la Grèce, de l'Autriche, de l'Irlande, de la Tchéquie et de la Lettonie. À l'inverse, pour des pays comme la France, l'Espagne, le Portugal, la Hongrie ou la Lituanie, la distribution de l'immigration s'écarte significativement de la normale. Cela peut s'expliquer par des épisodes migratoires atypiques, des chocs économiques ou des politiques migratoires particulières sur la période étudiée.

Enfin, l'examen des statistiques de tendance centrale et de dispersion révèle une grande hétérogénéité des échelles entre les variables. Par exemple, la population moyenne par pays est de l'ordre de 18 millions d'habitants, tandis que le PIB moyen est exprimé en montants très élevés et que l'immigration moyenne annuelle tourne autour de 2,4 millions. Cette diversité d'échelle nécessitera une standardisation des variables dans certaines modélisations, notamment celles impliquant des réseaux de neurones ou des techniques de machine learning.

En conclusion, cette analyse descriptive met en évidence une forte variabilité des flux migratoires et une diversité significative entre les pays européens étudiés, tant sur le plan démographique qu'économique. Les corrélations observées confortent l'idée que l'immigration est un phénomène complexe, influencé par de multiples facteurs, et justifient pleinement le recours à des modèles multivariés dans la suite du travail.

MODÉLISATION STATISTIQUE

Modélisation par régression linéaire multiple

Une première approche économétrique a été menée à l'aide d'un modèle de régression linéaire multiple, appliqué aux données temporelles agrégées pour les pays de l'Union européenne. Ce modèle vise à expliquer le niveau annuel d'immigration à partir d'un ensemble de variables explicatives d'ordre économique, démographique et institutionnel.

Les résultats obtenus montrent que le modèle présente un coefficient de détermination (R^2) de 0,633, ce qui signifie qu'environ 63 % de la variance observée dans les niveaux d'immigration est expliquée par les variables introduites dans le modèle. Dans le contexte de données économiques et sociales, un tel niveau de R^2 est considéré comme satisfaisant, car il reflète une capacité explicative relativement forte malgré la complexité inhérente au

phénomène étudié. Le test global de la régression (F-statistique) est très significatif, avec une p-value proche de zéro, ce qui confirme que le modèle dans son ensemble permet de mieux expliquer l'immigration que le simple hasard.

Sur le plan des variables individuelles, plusieurs coefficients associés aux variables explicatives ressortent comme étant très significatifs. En particulier, le produit intérieur brut (PIB), la population, le taux de natalité, le taux de fécondité et le taux de chômage affichent tous des p-values inférieures à 0,001. Cela suggère qu'il existe une relation statistiquement robuste entre ces variables et le niveau d'immigration. Les effets fixes liés aux pays, modélisés à travers des coefficients spécifiques pour chaque pays (identifiants nationaux), sont eux aussi très significatifs. Cela indique l'existence d'effets propres à chaque pays qui ne sont pas entièrement capturés par les seules variables économiques ou sociales — ce qui est logique étant donné les différences historiques, politiques ou institutionnelles entre les États membres.

L'interprétation des signes des coefficients permet de mieux comprendre le sens des relations identifiées. Pour la majorité des variables significatives, les coefficients estimés sont positifs. Ainsi, toutes choses égales par ailleurs, une augmentation du PIB, de la population, du taux de natalité, du taux de fécondité ou du taux de chômage est associée à une hausse du niveau d'immigration. Cette tendance générale est également observée pour d'autres indicateurs comme le coefficient de Gini (mesure des inégalités), ou encore les indices de corruption ou d'état de droit, dont les effets sont eux aussi positifs et significatifs.

Cependant, une attention particulière doit être portée à la question de la multicolinéarité entre certaines variables explicatives. Plusieurs d'entre elles présentent des corrélations élevées, comme le taux de natalité et le taux de fécondité, ou encore le PIB et la population. Cette redondance peut affecter la stabilité des coefficients estimés et conduire à des interprétations erronées. Il conviendra donc, dans les développements ultérieurs, de tester et éventuellement réduire cette redondance en sélectionnant les variables les plus pertinentes.

En résumé, le modèle de régression linéaire multiple appliqué au panel de pays européens met en évidence un lien significatif entre l'immigration et un ensemble de variables économiques et sociales. Il confirme le caractère multifactoriel du phénomène migratoire et justifie le recours à des approches multivariées pour l'analyse et la prévision. Toutefois, les résultats doivent être interprétés avec prudence en raison de possibles interactions entre les variables explicatives.

Régression quantile

Afin de compléter l'analyse menée par régression linéaire classique, une régression quantile a été appliquée, centrée sur la médiane de la distribution de l'immigration. Cette méthode permet de réduire l'influence des valeurs extrêmes (outliers) et d'offrir une lecture plus robuste du comportement central du phénomène étudié. Contrairement à la régression ordinaire des moindres carrés (OLS) qui modélise la moyenne conditionnelle, la régression

quantile s'intéresse à une estimation plus résistante aux déformations provoquées par les observations atypiques.

Le coefficient de détermination obtenu (pseudo- R^2) est de 0,256, ce qui signifie que le modèle explique environ 25 % de la variance conditionnelle de la médiane de l'immigration. Ce niveau d'explication est nettement inférieur à celui observé dans la régression OLS (63 %), mais cela est attendu dans la mesure où la régression quantile ne cherche pas à maximiser la variance expliquée globale, mais à fournir une estimation plus stable et moins sensible aux variations extrêmes.

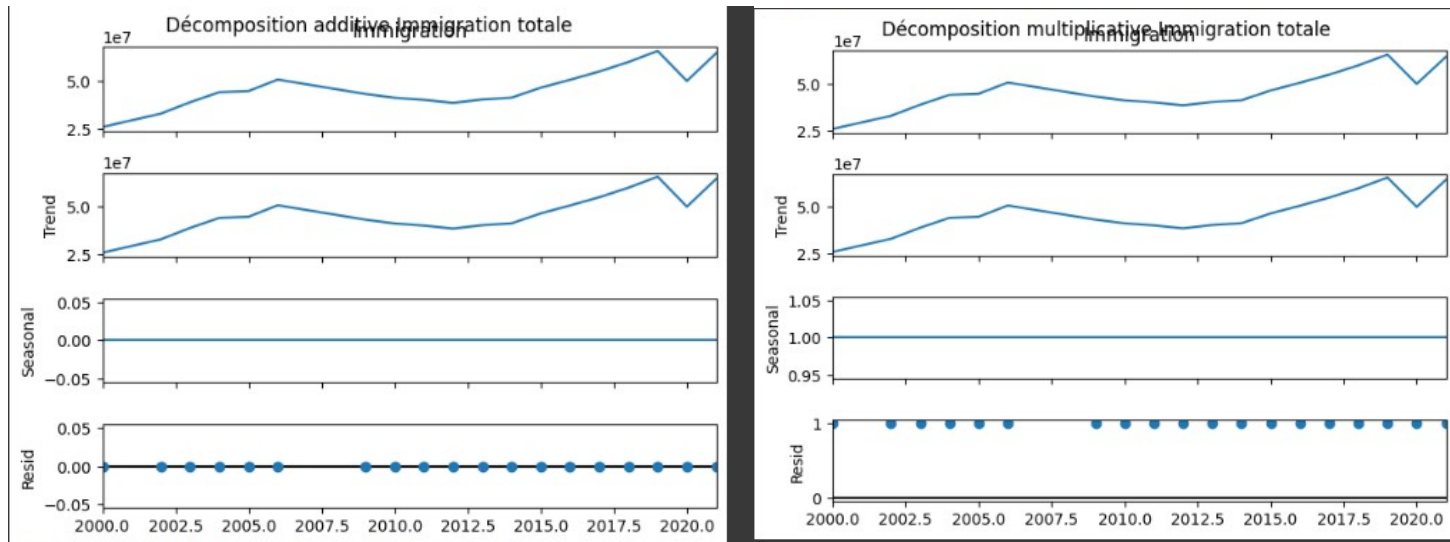
Concernant la significativité des variables explicatives, seuls quelques prédicteurs ressortent de manière statistiquement significative au seuil de 5 %. En particulier, le taux de natalité, le taux de fécondité et le taux de chômage affichent des coefficients positifs et significatifs, ce qui suggère qu'une augmentation de ces variables est associée à une hausse de la médiane du niveau d'immigration. En revanche, d'autres variables telles que le PIB, la population, le coefficient de Gini, le niveau d'éducation, l'indice de corruption, l'état de droit ou encore le taux d'inflation ne présentent pas d'effets significatifs sur la médiane. Cela contraste avec les résultats obtenus dans le modèle OLS, où plusieurs de ces variables apparaissaient comme fortement contributives.

Par ailleurs, les effets spécifiques par pays, modélisés à travers les identifiants nationaux, ne ressortent pas comme significatifs dans cette configuration. Plusieurs coefficients sont soit non significatifs, soit indéterminés (valeurs nulles ou absentes), ce qui pourrait s'expliquer par des limitations techniques liées à la structure du panel ou à la nature du modèle. Une autre explication plausible est que les différences médianes entre pays sont absorbées par les variables explicatives principales, ce qui rend les effets pays redondants dans cette spécification.

En conclusion, la régression quantile confirme certaines tendances mises en évidence par l'approche linéaire classique, mais avec une sensibilité plus marquée aux facteurs démographiques. Elle met en évidence le rôle du chômage, de la natalité et de la fécondité comme principaux déterminants de l'évolution centrale de l'immigration, tout en relativisant l'effet des autres variables économiques ou institutionnelles. Ce type de modèle, plus robuste mais moins explicatif en termes de variance globale, apparaît donc comme un complément utile à l'analyse classique. Pour aller plus loin dans la robustesse et la précision, d'autres approches plus avancées pourraient être envisagées, telles que les méthodes bayésiennes, les modèles non paramétriques, les régressions discontinues (RDD), ou encore les modèles de différence-en-différence (Diff-in-Diff).

ANALYSE TEMPORELLE

Décomposition de la série temporelle de l'immigration



Une analyse par décomposition de la série temporelle de l'immigration totale en Europe a été réalisée sur la période 2000–2021. Cette méthode permet de distinguer trois composantes fondamentales : la tendance de long terme, la saisonnalité éventuelle, et les résidus (ou bruits aléatoires).

La courbe observée montre une augmentation progressive du niveau d'immigration au cours de la période étudiée. Cette tendance est particulièrement marquée à partir de l'année 2016, ce qui pourrait refléter des dynamiques migratoires renforcées par des crises géopolitiques ou économiques dans certaines régions du monde. Une légère inflexion est observée en 2020, qui peut raisonnablement être attribuée aux restrictions de mobilité imposées par la crise sanitaire liée à la pandémie de Covid-19. La composante de tendance extraite par le modèle de décomposition confirme cette évolution haussière, captant de manière lissée la trajectoire générale de l'immigration au fil du temps.

Concernant la composante saisonnière, l'analyse révèle qu'elle est inexistante. Cela s'explique par la nature annuelle des données, qui empêche de capter des variations intra-annuelles. La ligne correspondant à la saisonnalité reste ainsi constante sur toute la période, ce qui confirme l'absence de fluctuations périodiques régulières au sein de chaque année. Ce résultat est cohérent avec les attentes et confirme la pertinence d'un traitement non saisonnier de la série.

L'examen des résidus montre qu'ils sont globalement faibles, proches de zéro, et ne présentent aucune structure particulière ni autocorrélation apparente. Cela signifie que la majeure partie de la dynamique de la série est capturée par la composante tendancielle, sans qu'il subsiste de bruit systématique ou de perturbation récurrente.

Enfin, une comparaison entre la décomposition additive et multiplicative a été menée. Les résultats sont très similaires dans les deux cas, ce qui s'explique par la faible variabilité relative de la série et l'absence de saisonnalité. En d'autres termes, l'évolution de

l'immigration ne dépend pas proportionnellement de son niveau passé, mais suit plutôt une dynamique linéaire en valeur absolue. Cela justifie le recours à un modèle additif, qui apparaît comme le plus adapté au comportement structural de la série étudiée. Toutefois, l'application d'un modèle multiplicatif ne présenterait pas d'inconvénient majeur dans ce cas précis, du fait de la stabilité de l'amplitude et de l'absence de composante saisonnière.

En conclusion, cette décomposition montre que la trajectoire de l'immigration européenne entre 2000 et 2021 est marquée par une croissance linéaire régulière, sans effets saisonniers ni changements d'amplitude significatifs, ce qui renforce la pertinence des modèles statistiques classiques pour modéliser cette tendance.

Test de stationnarité

L'analyse des séries temporelles nécessite, en amont de toute modélisation, la vérification de la stationnarité des données. Une série est dite stationnaire lorsque ses propriétés statistiques – notamment la moyenne et la variance – sont constantes dans le temps. Cette hypothèse est fondamentale pour garantir la validité des modèles autorégressifs classiques.

Afin de tester cette propriété, le test de Dickey-Fuller augmenté (ADF) a été appliqué à la série annuelle représentant le nombre total d'immigrés en Europe sur la période 2000–2021. Le résultat du test initial montre une p-value égale à 1, ce qui est bien supérieur au seuil usuel de 5 %. Cela indique que la série n'est pas stationnaire dans sa forme brute. Une première différenciation logarithmique a ensuite été réalisée, mais elle s'est révélée insuffisante : la p-value obtenue restait trop élevée (0,778), confirmant la persistance de la non-stationnarité.

Une deuxième différenciation a alors été effectuée sur la série log-transformée. Cette fois-ci, le test ADF donne une p-value de 0,012, ce qui permet de rejeter l'hypothèse nulle de non-stationnarité. La série devient donc stationnaire après deux différenciations, ce qui permet d'appliquer des modèles de type ARIMA (intégrant la différenciation d'ordre 2) ou ses extensions.

Modèles ARIMA/SARIMAX

Plusieurs modèles classiques de séries temporelles ont été considérés pour modéliser et prévoir l'évolution du nombre d'immigrés en Europe : AR, MA, ARIMA, et SARIMAX. Le choix du modèle le plus approprié dépend à la fois des propriétés de la série (stationnarité, présence de tendance) et de la possibilité d'introduire des variables explicatives exogènes (comme le PIB ou la population) afin de mieux expliquer les dynamiques migratoires.

Le modèle ARMA est adapté à des séries stationnaires sans tendance, tandis que le modèle ARIMA permet de traiter une série non stationnaire grâce à des différenciations successives. Le modèle SARIMAX, quant à lui, constitue une extension du modèle ARIMA en intégrant des variables exogènes, ce qui le rend particulièrement pertinent dans un contexte socio-

économique où des facteurs explicatifs sont disponibles.

Pour comparer ces modèles, plusieurs critères statistiques ont été mobilisés, notamment l'AIC (Akaike Information Criterion), le BIC (Bayesian Information Criterion) et le HQIC, dont les valeurs les plus faibles indiquent un meilleur ajustement. Par ailleurs, la significativité des coefficients ($p\text{-value} < 0,05$) a été examinée, tout comme le signe des effets estimés, afin d'en déduire des interprétations économiques cohérentes. Enfin, les résidus ont été soumis à des tests de qualité : test de normalité de Jarque-Bera, test d'hétéroscédasticité (test H), et test d'autocorrélation (Ljung-Box).

Sur l'ensemble des configurations testées, le modèle SARIMAX(1,2,1), intégrant le PIB et la population comme variables explicatives, s'est avéré être le plus performant, notamment pour les cas de l'Espagne, de la Hongrie et de la Lituanie, où les variables exogènes ressortaient comme significatives. Pour d'autres pays, le modèle ARIMA classique n'apportait pas d'amélioration substantielle. Dès lors, le modèle SARIMAX(1,2,1) a été retenu comme modèle de référence dans la suite de l'étude, en raison de sa capacité à intégrer des déterminants économiques explicites tout en tenant compte des propriétés temporelles de la série.

Enfin, une analyse de la multicolinéarité a été conduite via le Variance Inflation Factor (VIF). Deux variables ont été écartées du modèle final : l'état de droit et le taux de natalité, dont les VIF élevés traduisaient une forte redondance avec, respectivement, l'indice de corruption et le taux de fécondité. Cette sélection a permis d'améliorer la robustesse et la stabilité des estimations.

INTELLIGENCE ARTIFICIELLE ET PRÉVISION

Intelligence artificielle et modélisation des dynamiques migratoires

L'intelligence artificielle (IA) et les techniques de machine learning (ML) offrent aujourd'hui des outils performants pour analyser des phénomènes économiques complexes, en particulier lorsque les relations entre les variables sont non linéaires, multiples ou évolutives. Dans un contexte économique contemporain caractérisé par la disponibilité massive de données, qu'elles soient macroéconomiques, démographiques ou financières, les modèles classiques comme la régression linéaire ou les modèles autorégressifs (ARIMA) peuvent se révéler insuffisants pour saisir l'ensemble des dynamiques en jeu.

L'IA et le ML permettent ainsi d'exploiter pleinement le potentiel informatif de ces données. Ils sont capables de détecter des régularités cachées, d'identifier des structures complexes, de gérer des interactions non linéaires, et de modéliser des effets de seuil ou de rupture que les méthodes traditionnelles captent difficilement. Ces approches peuvent également automatiser la sélection des variables les plus pertinentes, ajuster dynamiquement les

paramètres d'un modèle, et fournir des prévisions plus robustes. Appliquées aux séries temporelles, elles permettent de dégager des tendances de long terme à partir de l'évolution historique des données, comme c'est le cas pour l'étude de l'immigration.

Principe de fonctionnement du Machine Learning
Le machine learning regroupe un ensemble d'algorithmes capables d'apprendre automatiquement à partir des données, sans qu'une règle explicite ne soit programmée pour chaque tâche. Ce processus passe par plusieurs étapes standards.

La première consiste à préparer les données : cela inclut le nettoyage, la normalisation, et la transformation des variables pour les rendre exploitables. Vient ensuite la séparation du jeu de données en deux parties : un jeu d'entraînement, utilisé pour ajuster les paramètres du modèle, et un jeu de test, sur lequel la performance du modèle est évaluée de manière indépendante.

Une fois le modèle choisi (régression, réseau de neurones, etc.), l'algorithme apprend à minimiser l'erreur de prédiction sur le jeu d'entraînement. Cela implique l'optimisation de paramètres internes comme les poids et les biais dans le cas des réseaux de neurones. Après cette phase d'apprentissage, le modèle est appliqué au jeu de test, ce qui permet de mesurer sa capacité de généralisation à l'aide d'indicateurs de performance tels que l'erreur quadratique moyenne (RMSE), l'erreur absolue moyenne (MAE) ou l'erreur quadratique moyenne (MSE).

Enfin, une phase d'optimisation est souvent nécessaire pour améliorer les performances du modèle. Elle peut consister à ajuster les hyperparamètres (comme le nombre de neurones ou le taux d'apprentissage), ou à comparer plusieurs architectures pour sélectionner la plus efficace.

Dans le cadre de cette étude, nous avons opté pour des réseaux de neurones récurrents (RNN), spécifiquement adaptés à la modélisation des séries temporelles. Les architectures LSTM (Long Short-Term Memory) et GRU (Gated Recurrent Unit) ont été privilégiées pour leur capacité à mémoriser des dépendances de long terme dans les données, ce qui est particulièrement pertinent pour prévoir l'évolution de l'immigration à partir des tendances historiques observées.

Mise en œuvre des modèles de prévision

Dans le cadre de cette étude, plusieurs modèles ont été utilisés pour prédire l'évolution de l'immigration dans les pays de l'Union européenne. L'objectif était d'évaluer leur capacité à fournir des prévisions fiables sur les cinq dernières années de la série (2017–2021), en comparant la qualité des résultats obtenus avec différentes approches statistiques et d'intelligence artificielle.

Le modèle SARIMAX, retenu précédemment pour son intégration de variables explicatives exogènes (PIB, population), a été appliqué en priorité. Toutefois, des contraintes importantes sont apparues lors de sa mise en œuvre. En effet, certaines séries nationales présentaient des valeurs manquantes ou des discontinuités dans les données, ce qui a limité la faisabilité

de la prévision sur plusieurs années. Dans de nombreux cas, la prédiction n'a pu être réalisée que pour l'année 2021, ce qui s'est avéré insuffisant pour évaluer correctement les performances du modèle dans le temps.

Pour surmonter ces limites, des modèles de réseaux de neurones ont été mis en œuvre. Trois architectures ont été testées : le Multi-Layer Perceptron (MLP), le Long Short-Term Memory (LSTM) et le Gated Recurrent Unit (GRU). Ces réseaux sont spécifiquement adaptés à la modélisation de séries temporelles grâce à leur capacité à capturer les dépendances temporelles et les non-linéarités. Les modèles ont été entraînés sur les données historiques, puis testés sur les années 2017 à 2021 afin de mesurer leur précision de prédiction.

Comparaison des performances des modèles

Les performances des différents modèles ont été évaluées à l'aide de deux métriques standards : la Root Mean Squared Error (RMSE) et la Mean Absolute Error (MAE). Ces deux indicateurs permettent de quantifier l'écart entre les prévisions produites et les valeurs réelles observées. Plus les valeurs de RMSE et MAE sont faibles, meilleure est la qualité de la prévision. Les résultats obtenus sont les suivants :

- GRU : RMSE = 1 254 144 ; MAE = 684 138
- LSTM : RMSE = 1 264 998 ; MAE = 697 035
- MLP : RMSE = 1 841 784 ; MAE = 827 710
- SARIMAX : RMSE = 1 971 438 ; MAE = 1 975 782

Ces résultats montrent clairement que les modèles de réseaux de neurones récurrents, en particulier GRU et LSTM, surpassent les autres approches en termes de précision. Le modèle GRU obtient les meilleures performances, avec le plus faible RMSE et MAE, suivi de très près par le LSTM. Le modèle MLP, de type perceptron multicouche, se situe en retrait. Enfin, le modèle SARIMAX, malgré sa pertinence théorique, présente les performances les plus faibles sur cet exercice de prévision, avec des erreurs environ 1,5 fois plus élevées que celles des modèles GRU et LSTM.

Ces résultats confirment l'intérêt des approches issues du machine learning, notamment les architectures séquentielles, pour modéliser des phénomènes dynamiques complexes comme l'évolution de l'immigration. Leur capacité à apprendre les régularités temporelles dans les données permet d'obtenir des prévisions plus précises que celles fournies par les méthodes économétriques classiques.

CONCLUSION ET PERSPECTIVES

Cette étude avait pour objectif de modéliser et de prévoir l'évolution de l'immigration dans les pays de l'Union européenne en mobilisant à la fois des méthodes classiques de séries temporelles (ARIMA, SARIMAX) et des approches issues de l'intelligence artificielle, en particulier le machine learning et les réseaux de neurones.

Les résultats obtenus mettent clairement en évidence la supériorité des modèles de réseaux de neurones récurrents, notamment les architectures LSTM (Long Short-Term Memory) et GRU (Gated Recurrent Unit). Le modèle LSTM se distingue par le plus faible RMSE (Root Mean Squared Error), tandis que le GRU présente la meilleure MAE (Mean Absolute Error). Ces deux modèles offrent une capacité prédictive nettement supérieure à celle des autres approches testées, en raison de leur aptitude à modéliser les dépendances temporelles complexes et les relations non linéaires entre les variables.

En comparaison, les performances du modèle MLP (perceptron multicouche) et du modèle SARIMAX se sont révélées sensiblement inférieures. Le MLP, bien qu'appartenant également à la famille des réseaux de neurones, n'intègre pas de mémoire temporelle, ce qui limite sa pertinence dans un contexte de séries chronologiques. Quant au SARIMAX, il reste performant dans certaines configurations, mais ne parvient pas à capter toute la complexité des dynamiques migratoires, en particulier les effets combinés et non linéaires entre les différentes variables explicatives.

Plus largement, cette recherche montre l'intérêt d'appliquer l'intelligence artificielle à des problématiques sociales complexes comme celle de l'immigration. Les modèles d'apprentissage automatique permettent non seulement d'améliorer la qualité des prévisions, mais également d'intégrer un grand nombre de facteurs explicatifs d'ordres économique, démographique ou institutionnel, tout en s'adaptant à l'évolution des données dans le temps.

Les perspectives futures de cette recherche sont multiples. Un premier axe d'approfondissement consisterait à intégrer des données sur les pays d'origine des migrants, en développant des modèles de migration bilatérale, permettant d'analyser les déterminants à la fois du pays de destination et du pays d'émigration. Cela nécessiterait l'introduction de variables supplémentaires, comme les inégalités, la croissance ou la stabilité politique dans les pays d'origine.

Un second axe pourrait être l'élargissement du périmètre géographique de l'étude. Plutôt que de se limiter à l'Union européenne, il serait pertinent d'inclure d'autres pays de l'OCDE, voire certains pays africains ou asiatiques fortement impliqués dans les dynamiques migratoires internationales. Une telle extension permettrait de construire un modèle global de l'immigration, capable de capter les grandes tendances mondiales sur le long terme.

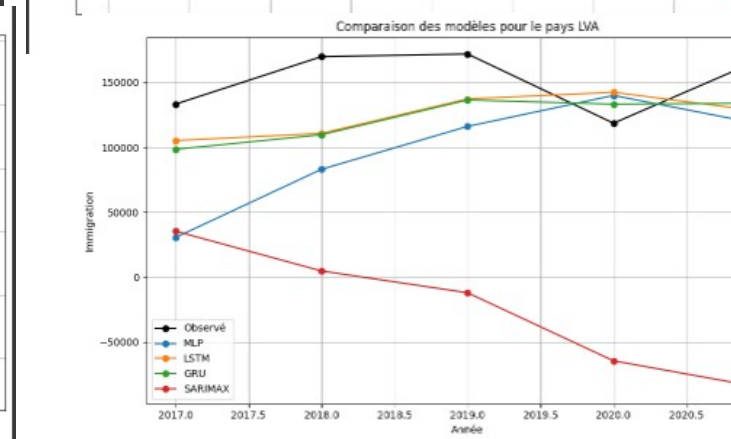
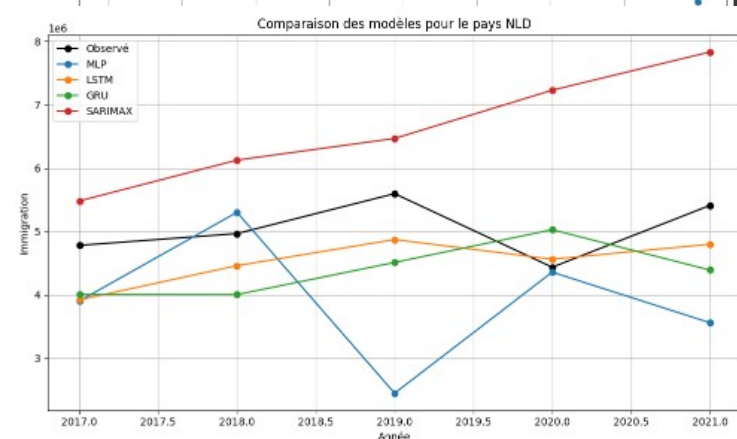
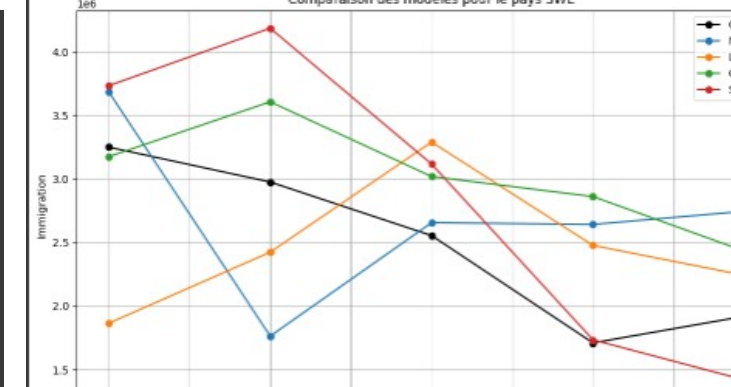
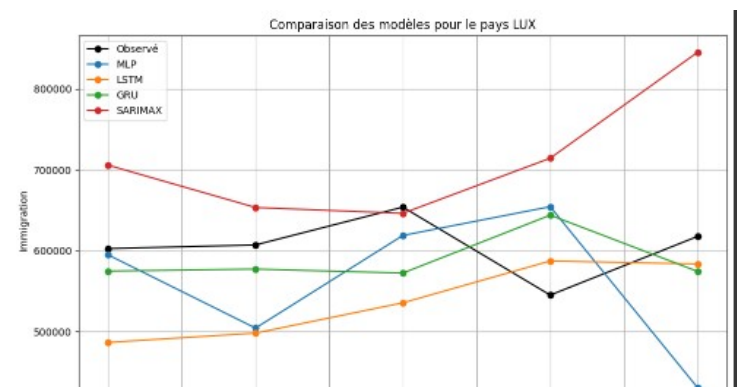
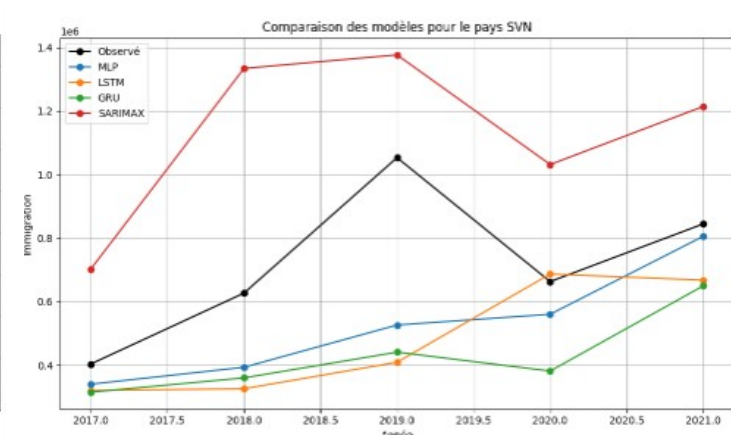
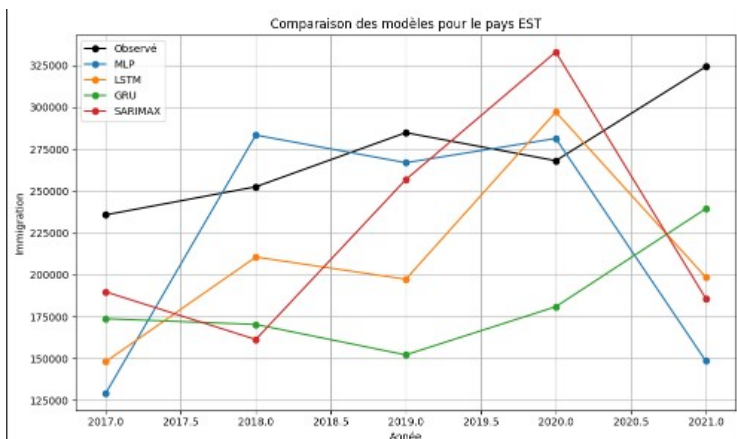
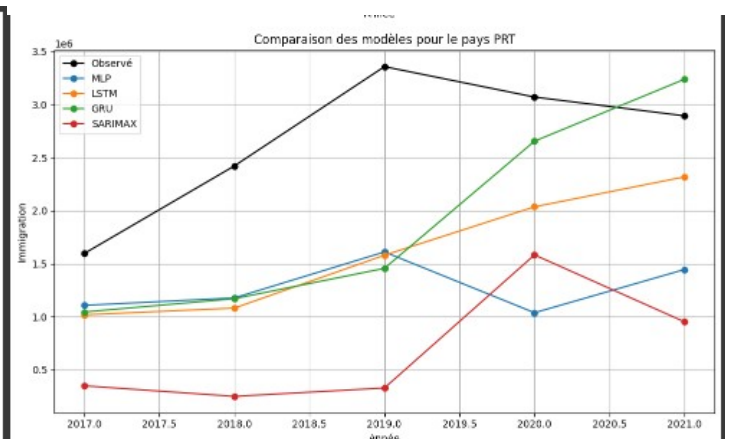
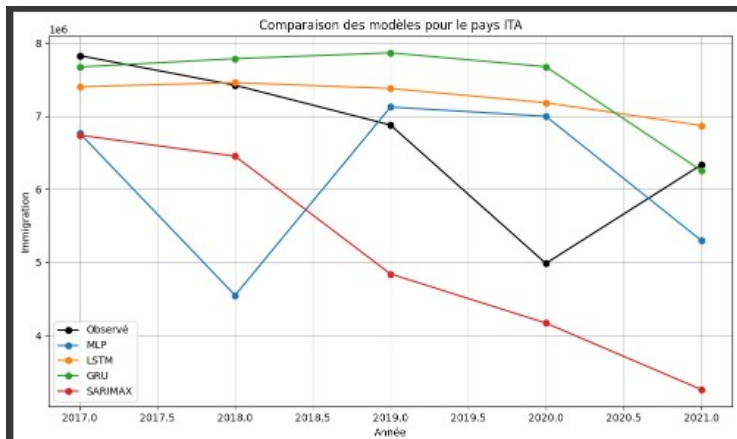
En somme, les résultats obtenus soulignent la pertinence scientifique et opérationnelle des outils de machine learning pour l'analyse des migrations contemporaines. Ils invitent à poursuivre l'intégration de ces méthodes dans les travaux en économie appliquée, en sciences sociales et en appui aux politiques publiques.

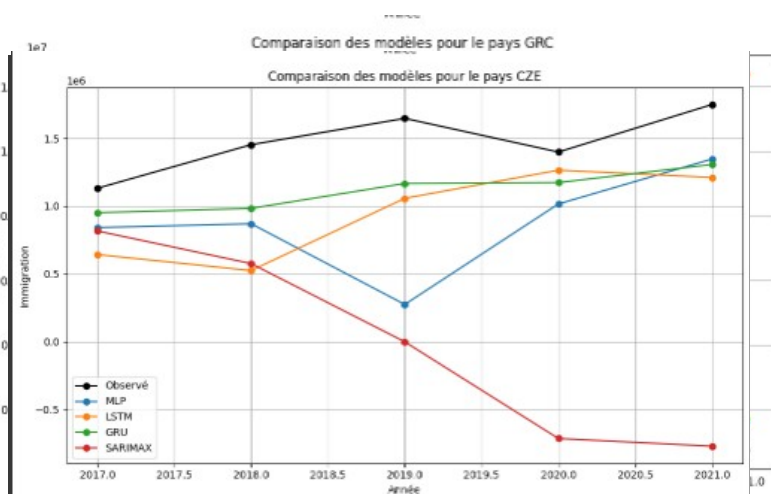
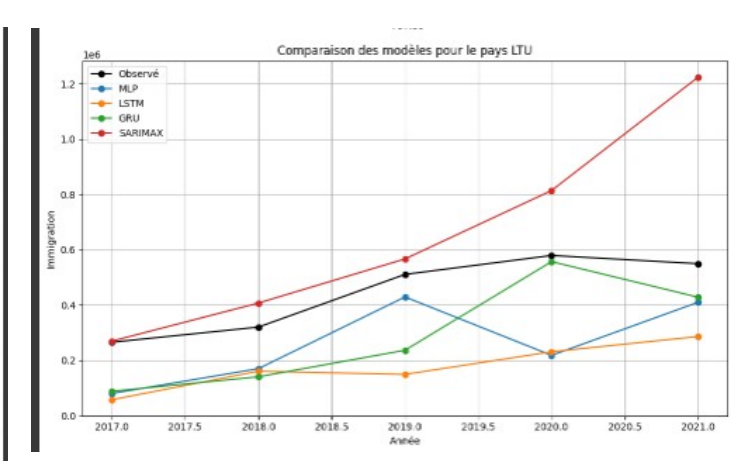
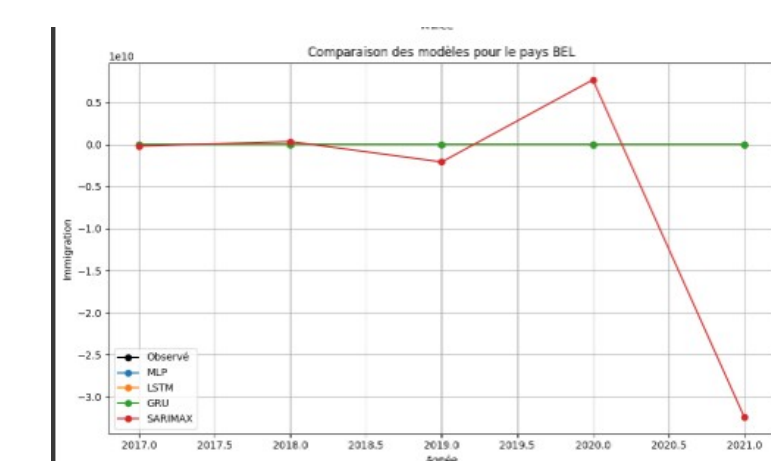
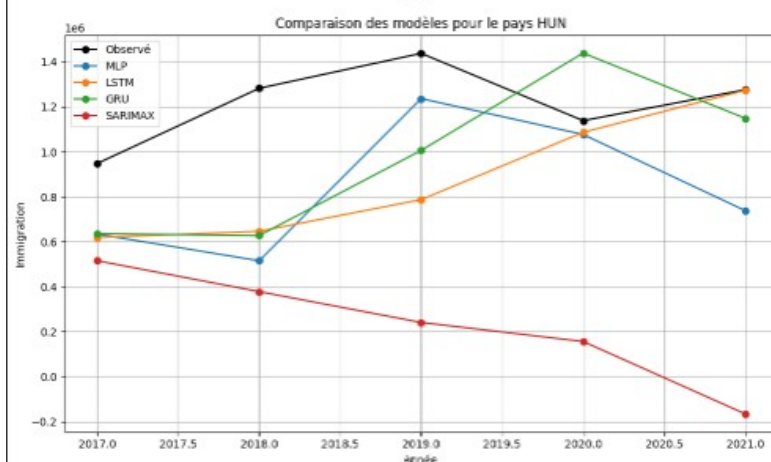
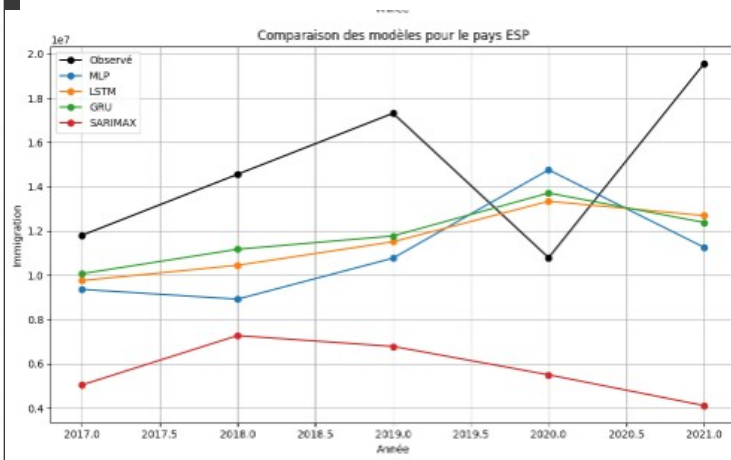
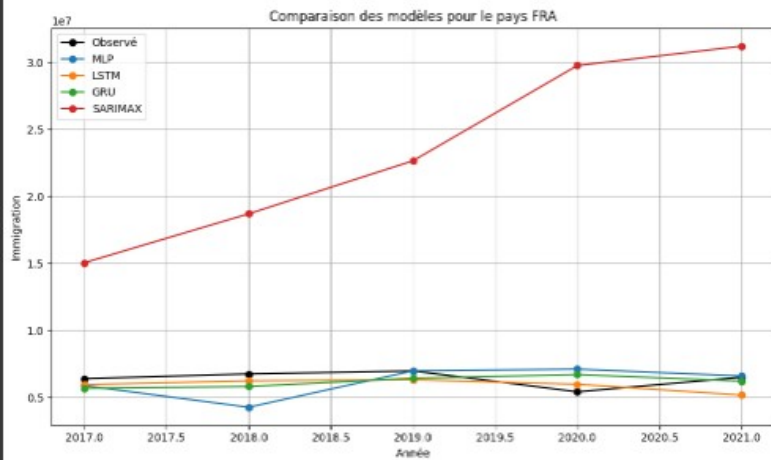
BIBLIOGRAPHIE

1. Albis, H., Boubtane, E., & Coulibaly, D. (2018). *Immigration and Public Finances in OECD Countries*. Economics Letters, Elsevier. <https://www.sciencedirect.com/science/article/pii/S0165188918303920>
2. Boubtane, E., Dumont, J.-C., & Rault, C. (2014). *Immigration and Economic Growth in the OECD Countries, 1986–2006*. IZA Discussion Paper No. 8681. Institute for the Study of Labor. <https://www.iza.org/en/publications/dp/8681>
3. Mohammad, R. A., Iman, S. W., Attico, Y. T. A., Antohi, V. M., Fortea, C., & Zlati, M. L. (2025). *International immigration and its effects on native labor market: Evidence from OECD countries*. *Frontiers in Human Dynamics*, 7, 1577022. <https://doi.org/10.3389/fhumd.2025.1577022>
4. Yaya, A. (2024). *Analyse de sentiment sur la théorie du remplacement*. Université de Pau et des Pays de l'Adour. (Mémoire de recherche non publié). [Analyse de la perception idéologique et médiatique du “grand remplacement” à travers les données textuelles issues des médias.]

ANNEXE

1. modele





ANALYSE DESCRIPTIVE)

```

Moyenne :
year          2.011316e+03
pop_d         1.376520e+04
Tauxnatalit_d 1.028912e+01
Tauxf_d       1.524170e+00
Chomage_d     8.848259e+00
Gini_d        3.133652e+01
Education_d   9.638428e+01
Etat_Droit_d  1.047178e+00
Corruption_d  9.417917e-01
Inflation     2.141405e+00
PIB           5.054110e+14
Immigration   2.415238e+06
dtype: float64

Ecart-type :
year          6.348042e+00
pop_d         1.753107e+04
Tauxnatalit_d 1.430629e+00
Tauxf_d       2.071263e-01
Chomage_d     4.431919e+00
Gini_d        3.776751e+00
Education_d   5.858255e+00
Etat_Droit_d  6.120041e-01
Corruption_d  7.820006e-01
Inflation     2.950815e+00
PIB           9.907879e+15
Immigration   3.155880e+06
dtype: float64

Coefficient de variation :
year          0.003156
pop_d         1.273579
Tauxnatalit_d 0.139043
Tauxf_d       0.135894
Chomage_d     0.500880
Gini_d        0.120522
Education_d   0.060780
Etat_Droit_d  0.584432
Corruption_d  0.830333
Inflation     1.377980
PIB           19.603607
Immigration   1.306653
dtype: float64

```

```

Matrice de corrélation :
year      pop_d      Tauxnatalit_d      Tauxf_d      Chomage_d \
year      1.000000    0.010776      -0.232475    0.106460    -0.108489
pop_d     0.010776    1.000000      -0.063871    0.019921    0.187749
Tauxnatalit_d -0.232475 -0.063871    1.000000    0.752864    -0.242053
Tauxf_d     0.106460  0.019921    0.752864    1.000000    -0.336062
Chomage_d   -0.108489  0.187749    -0.242053 -0.336062    1.000000
Gini_d      0.013449  0.237605    -0.193548 -0.261619    0.381346
Education_d  0.025620 -0.046798    0.060961  0.081361    -0.166404
Etat_Droit_d 0.006556 -0.120251    0.477262  0.513701    -0.378247
Corruption_d -0.043998 -0.092124    0.475596  0.523496    -0.352114
Inflation    -0.316072  0.001864    -0.033548 -0.161467    -0.108110
PIB          -0.004620  0.113729    -0.029034 -0.016201    -0.003592
Immigration  0.080089  0.706250    -0.188051 -0.185507    0.181209

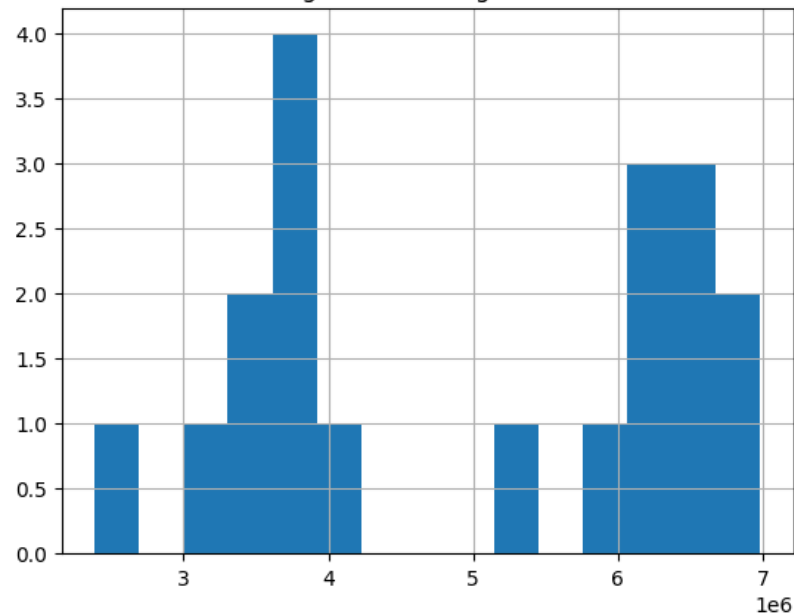
Gini_d      Education_d      Etat_Droit_d      Corruption_d      Inflation \
year      0.013449    0.025620    0.006556    -0.043998    -0.316072
pop_d     0.237605    -0.046798    -0.120251    -0.092124    0.001864
Tauxnatalit_d -0.193548    0.060961    0.477262    0.475596    -0.033548
Tauxf_d     -0.261619    0.081361    0.513701    0.523496    -0.161467
Chomage_d    0.381346    -0.166404    -0.378247    -0.352114    -0.108110
Gini_d       1.000000    0.069155    -0.465831    -0.448236    0.094225
Education_d  0.069155    1.000000    0.393571    0.416741    -0.139475
Etat_Droit_d -0.465831    0.393571    1.000000    0.946515    -0.252527
Corruption_d -0.448236    0.416741    0.946515    1.000000    -0.227659
Inflation    0.094225    -0.139475    -0.252527    -0.227659    1.000000
PIB          0.036509    0.032908    -0.037341    -0.031880    -0.012082
Immigration  0.192643    -0.250791    -0.121985    -0.078451    -0.041402

PIB      Immigration
year     -0.004620    0.080089
pop_d     0.113729    0.706250
Tauxnatalit_d -0.029034 -0.188051
Tauxf_d    -0.016201 -0.185507
Chomage_d -0.003592  0.181209
Gini_d     0.036509  0.192643
Education_d 0.032908 -0.250791
Etat_Droit_d -0.037341 -0.121985
Corruption_d -0.031880 -0.078451
Inflation   -0.012082 -0.041402
PIB         1.000000  0.139350
Immigration 0.139350  1.000000

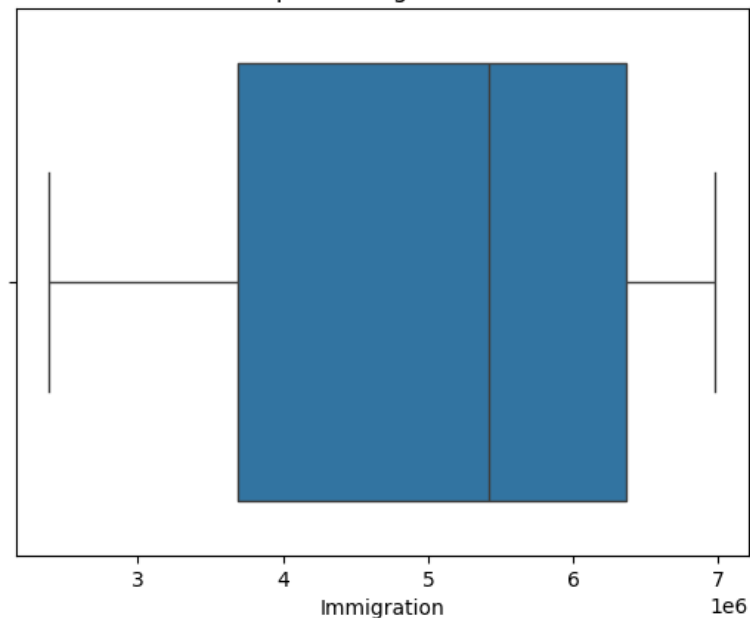
```

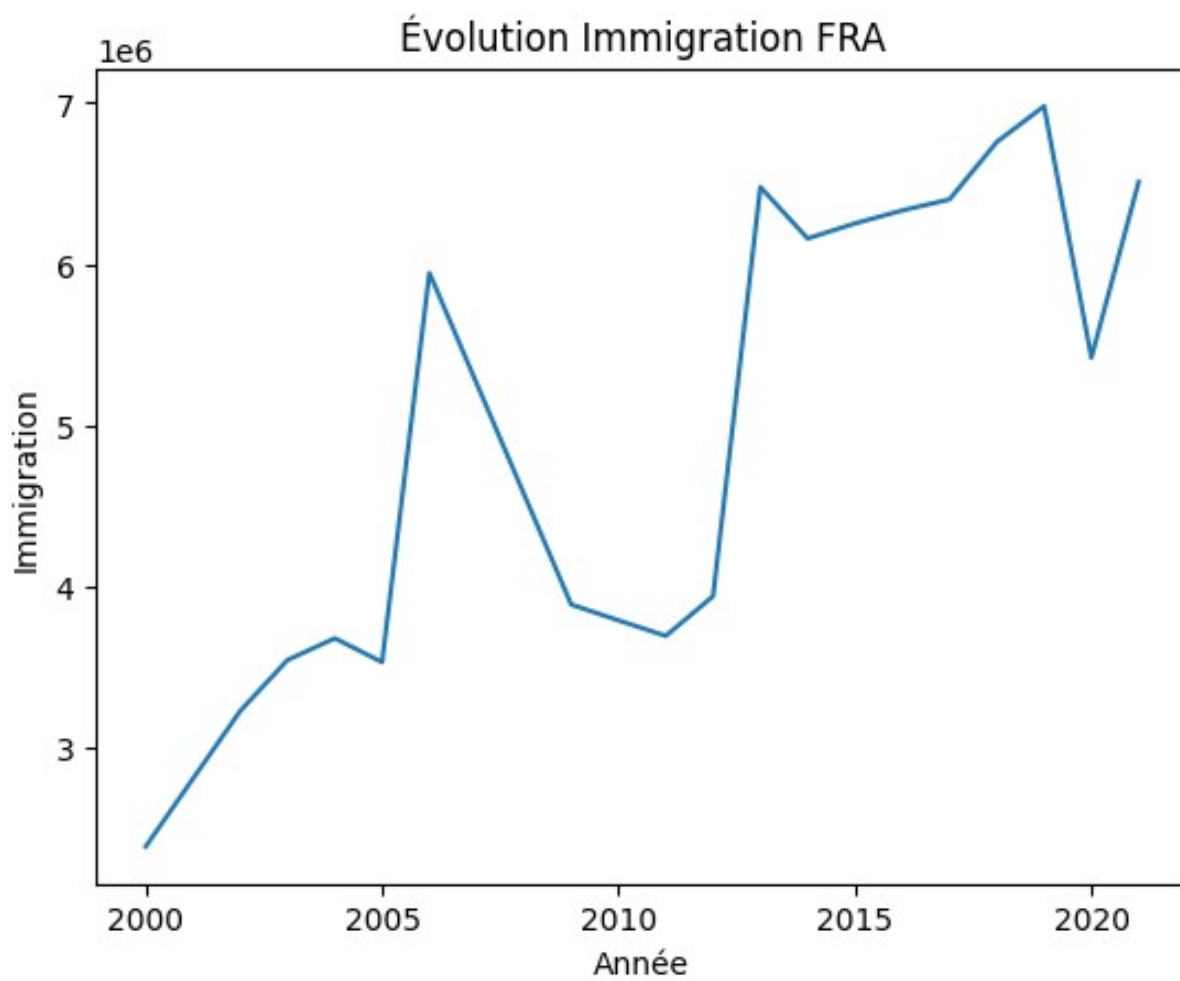
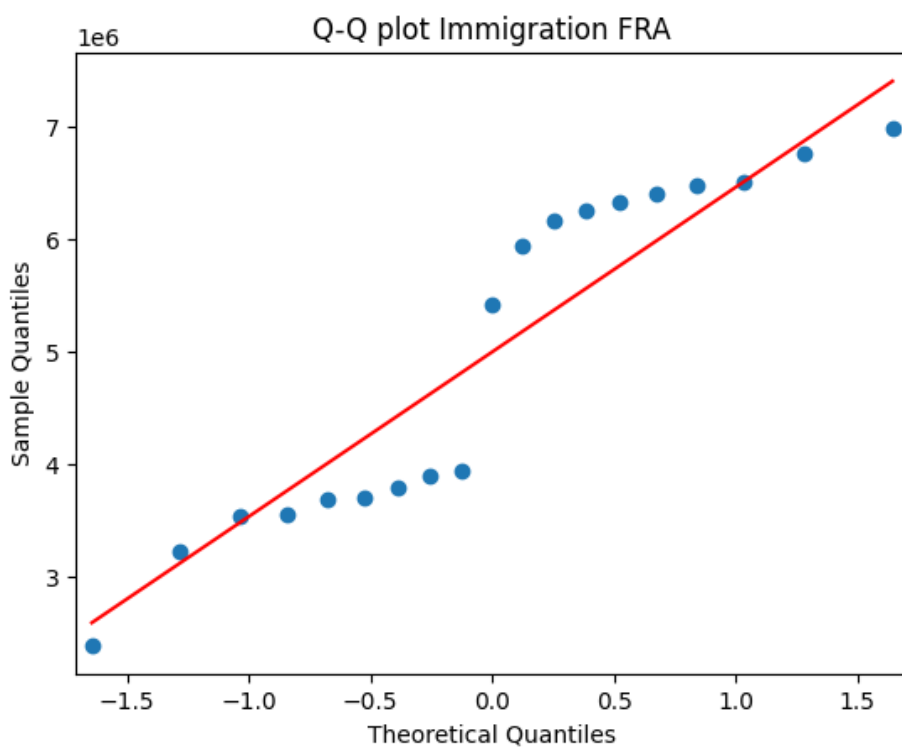
Distribution des variable

Histogramme Immigration FRA



Boxplot Immigration FRA





Test

Italie: Test de Shapiro-Wilk pour Immigration ITA : stat=0.9333, p-value=0.199

PRT: Test de Shapiro-Wilk pour Immigration PRT : stat=0.8643, p-value=0.01156

EST: Test de Shapiro-Wilk pour Immigration EST : stat=0.8021, p-value=0.0029

Lux: Test de Shapiro-Wilk pour Immigration LUX : stat=0.9130, p-value=0.084

NLD: Test de Shapiro-Wilk pour Immigration NLD : stat=0.9289, p-value=0.1653

SVN: Test de Shapiro-Wilk pour Immigration SVN : stat=0.8216, p-value=0.01246

BEL: Test de Shapiro-Wilk pour Immigration BEL : stat=0.9263, p-value=0.1479

SWE: Test de Shapiro-Wilk pour Immigration SWE : stat=0.9526, p-value=0.4365

FRA: Test de Shapiro-Wilk pour Immigration FRA : stat=0.8630, p-value=0.01101

ESP: Test de Shapiro-Wilk pour Immigration ESP : stat=0.8922, p-value=0.03517

DNK: Test de Shapiro-Wilk pour Immigration DNK : stat=0.8945, p-value=0.03874

HUN: Test de Shapiro-Wilk pour Immigration HUN : stat=0.7742, p-value=0.0004935

LTU: Test de Shapiro-Wilk pour Immigration LTU : stat=0.7407, p-value=0.0002474

GRC: Test de Shapiro-Wilk pour Immigration GRC : stat=0.9015, p-value=0.1005

POL : Test de Shapiro-Wilk pour Immigration POL : stat=0.8073, p-value=0.001463

AUT : Test de Shapiro-Wilk pour Immigration AUT : stat=0.9621, p-value=0.6141

IRL : Test de Shapiro-Wilk pour Immigration IRL : stat=0.9516, p-value=0.4201

CZE : Test de Shapiro-Wilk pour Immigration CZE : stat=0.9601, p-value=0.5749

FIN : Test de Shapiro-Wilk pour Immigration FIN : stat=0.8918, p-value=0.03473

LVA : Test de Shapiro-Wilk pour Immigration LVA : stat=0.9283, p-value=0.2291

Modélisation par régression linéaire multiple

OLS Regression Results							
Dep. Variable:	Immigration	R-squared:	0.433				
Model:	OLS	Adj. R-squared:	0.431				
Method:	Least Squares	F-statistic:	371.8				
Date:	Sat, 14 Jun 2025	Prob (F-statistic):	7.50e-49				
Time:	13:54:26	Log-likelihood:	-3445.3				
No. Observations:	218	AIC:	6895.				
DF Residuals:	216	BIC:	6901.				
DF Model:	1						
Covariance Type:	nonrobust						
	coef	std err	t	P> t	[0.025	0.975]	
Intercept	78.8491	20.797	3.791	0.000	37.857	119.841	
C(country_id_d)[T.BEL]	8.2272	0.182	3.791	0.000	0.849	1.186	
C(country_id_d)[T.CZE]	4.9864	1.315	3.791	0.000	2.394	7.579	
C(country_id_d)[T.DMC]	4.2957	1.133	3.791	0.000	2.062	6.529	
C(country_id_d)[T.ESP]	-0.3440	0.401	-3.792	0.000	-0.524	-0.165	
C(country_id_d)[T.EST]	5.9994	1.583	3.791	0.000	2.880	9.119	
C(country_id_d)[T.FIN]	5.0594	1.335	3.791	0.000	2.429	7.690	
C(country_id_d)[T.FRA]	-3.1380	0.826	-3.790	0.000	-4.758	-1.582	
C(country_id_d)[T.GRC]	3.9946	1.054	3.791	0.000	1.918	6.071	
C(country_id_d)[T.HUN]	2.5419	0.671	3.791	0.000	1.221	3.866	
C(country_id_d)[T.ITA]	2.9763	0.785	3.791	0.000	1.429	4.524	
C(country_id_d)[T.LIT]	-2.4634	0.650	-3.790	0.000	-3.744	-1.182	
C(country_id_d)[T.LTU]	3.8489	1.015	3.791	0.000	1.848	5.850	
C(country_id_d)[T.LUX]	4.8887	1.200	3.791	0.000	2.547	7.430	
C(country_id_d)[T.LVA]	3.1515	0.831	3.791	0.000	1.513	4.790	
C(country_id_d)[T.POL]	0.4836	0.186	3.794	0.000	0.194	0.613	
C(country_id_d)[T.PRT]	0.9402	0.248	3.791	0.000	0.451	1.429	
C(country_id_d)[T.SWE]	4.7143	1.244	3.791	0.000	2.263	7.165	
C(country_id_d)[T.SWE]	3.8184	1.007	3.791	0.000	1.833	5.804	
PIB	1.794e-11	0.32e-12	2.158	0.032	1.155e-12	2.43e-11	
pop_d	109.0950	5.758	18.948	0.000	97.747	120.443	
Tauxnatalit_d	571.1262	150.641	3.791	0.000	274.211	868.041	
Taux_d	86.4285	22.757	3.791	0.000	41.486	131.361	
Chomage_d	418.0871	110.273	3.791	0.000	200.737	635.437	
Gini_d	1628.3822	429.481	3.791	0.000	781.791	2474.813	
Education_d	5350.0090	1431.230	3.791	0.000	2569.652	8120.568	
Etat_droit_d	74.2412	19.582	3.791	0.000	35.644	112.838	
Corruption_d	69.6158	18.362	3.791	0.000	33.424	105.888	
Inflation	83.7584	22.690	3.791	0.000	40.210	127.291	
Omnibus:	127.684	Durbin-Watson:	2.345				
Prob(Omnibus):	0.000	Jarque-Bera (JB):	127.252				
Skew:	2.060	Prob(JB):	3.00e-180				
Kurtosis:	14.168	Cond. No.	2.31e+17				

Notes:

[1] Standard errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 2.31e+17. This might indicate that there are strong multicollinearity or other numerical problems.

Modèle (Quantile) (Annexe 4) : Modèle Quantile

```

QuantReg Regression Results
=====
Dep. Variable:      Immigration      Pseudo R-squared:      -0.2562
Model:              QuantReg         Bandwidth:              1.665e+06
Method:             Least Squares    Sparsity:                5.105e+06
Date:               Sat, 14 Jun 2025  No. Observations:         218
Time:               20:00:32          Df Residuals:           216
                                Df Model:              1
=====
               coef      std err          t      P>|t|      [0.025      0.975]
-----
Intercept      5.137e-27          0      inf      0.000      5.14e-27      5.14e-27
C(country_id_d)[T.BEL]      0          0      nan      nan          0          0
C(country_id_d)[T.CZE]      2.778e-34      7.98e-34      0.348      0.728      -1.29e-33      1.85e-33
C(country_id_d)[T.DNK]      2.651e-34      1.02e-33      0.261      0.795      -1.74e-33      2.27e-33
C(country_id_d)[T.ESP]      1.064e-34      3.09e-33      0.034      0.973      -5.98e-33      6.2e-33
C(country_id_d)[T.EST]      5.187e-34      8.91e-35      5.822      0.000      3.43e-34      6.94e-34
C(country_id_d)[T.FIN]      3.226e-30      6.43e-30      0.502      0.616      -9.45e-30      1.59e-29
C(country_id_d)[T.FRA]      6.25e-34      9.31e-33      0.067      0.947      -1.77e-32      1.9e-32
C(country_id_d)[T.GRC]      2.529e-34      7.6e-34      0.333      0.740      -1.25e-33      1.75e-33
C(country_id_d)[T.HUN]      1.19e-34      2.93e-34      0.407      0.685      -4.58e-34      6.96e-34
C(country_id_d)[T.IRL]      1.115e-34      4.32e-34      0.258      0.796      -7.4e-34      9.63e-34
C(country_id_d)[T.ITA]      2.393e-28      5.43e-29      4.404      0.000      1.32e-28      3.46e-28
C(country_id_d)[T.LTU]      3.361e-34      1.21e-34      2.767      0.006      9.67e-35      5.76e-34
C(country_id_d)[T.LUX]      1.238e-34      1.92e-34      0.646      0.519      -2.54e-34      5.02e-34
C(country_id_d)[T.LVA]      1.723e-34      6.13e-35      2.811      0.005      5.15e-35      2.93e-34
C(country_id_d)[T.POL]      2.164e-34      1.09e-33      0.199      0.843      -1.93e-33      2.36e-33
C(country_id_d)[T.PRT]      2.127e-35      1.8e-34      0.118      0.906      -3.34e-34      3.77e-34
C(country_id_d)[T.SVN]      1.261e-34      1.49e-34      0.846      0.399      -1.68e-34      4.2e-34
C(country_id_d)[T.SWE]      2.259e-34      1.62e-33      0.140      0.889      -2.96e-33      3.42e-33
PIB      5.141e-11      1.18e-11      4.351      0.000      2.81e-11      7.47e-11
pop_d      1.42e-23      3.26e-24      4.361      0.000      7.78e-24      2.06e-23
Tauxnatalit_d      2.303e-27      5.75e-28      4.005      0.000      1.17e-27      3.44e-27
Tauxf_d      3.542e-28      8.9e-29      3.982      0.000      1.79e-28      5.3e-28
Chomage_d      2.029e-27      5.1e-28      3.977      0.000      1.02e-27      3.03e-27
Gini_d      8.393e-27      2.06e-27      4.069      0.000      4.33e-27      1.25e-26
Education_d      2.421e-26      6.07e-27      3.990      0.000      1.23e-26      3.62e-26
Etat_Droit_d      1.154e-28      3.78e-29      3.056      0.003      4.1e-29      1.9e-28
Corruption_d      7.481e-29      2.93e-29      2.551      0.011      1.7e-29      1.33e-28
Inflation      3.675e-28      8.78e-29      4.187      0.000      1.95e-28      5.41e-28
=====

The smallest eigenvalue is -124. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
/usr/local/lib/python3.11/dist-packages/statsmodels/regression/linear_model.py:1966: RuntimeWarning: invalid value encountered in sqrt
return np.sqrt(eigvals[0]/eigvals[-1])

```

VIF

```

Colonnes utilisées pour VIF :
Index(['Tauxnatalit_d', 'Tauxf_d', 'Chomage_d', 'Etat_Droit_d',
      'Corruption_d'],
      dtype='object')
VIF de chaque variable explicative :
      variable      VIF
0  Tauxnatalit_d  116.137041
1      Tauxf_d    111.821957
2      Chomage_d    4.895899
3  Etat_Droit_d   37.802078
4  Corruption_d   22.443530

```

```

      variable      VIF
0      pop_d    1.012405
1  Tauxnatalit_d  0.927008
2      Chomage_d  1.044622
3      Gini_d    0.901075
4  Education_d   0.165493
5  Corruption_d   1.050435
6      Inflation  1.004148
7      PIB       1.027478

```

Sarimax

```

IRL: OK
CZE: OK
FIN: OK
LVA: OK

Aperçu du dataframe SARIMAX :
      Country_id  year      obs  sarimax_pred  sarimax_abs_error
0      BGR  2017      0.0  0.000000e+00      0.000000e+00
1      BGR  2018      0.0  0.000000e+00      0.000000e+00
2      BGR  2019      0.0  0.000000e+00      0.000000e+00
3      BGR  2020      0.0  0.000000e+00      0.000000e+00
4      BGR  2021      0.0  0.000000e+00      0.000000e+00
5      ITA  2017  7827846.0  6.739863e+06  1.087983e+06
6      ITA  2018  7423000.0  6.452202e+06  9.707976e+05
7      ITA  2019  6878846.0  4.842475e+06  2.036371e+06
8      ITA  2020  4985916.0  4.169897e+06  8.160190e+05
9      ITA  2021  6333782.0  3.261478e+06  3.072304e+06

MAE global SARIMAX : 197578620770
RMSE global SARIMAX : 1971438266493

```

MLP

```
Prévisions MLP (5 dernières années par pays) :
pays  année  obs      mlp_pred  mlp_abs_error
0  ITA  2017  7827846.0  7.282831e+06  5.450150e+05
1  ITA  2018  7423000.0  7.365902e+06  5.709800e+04
2  ITA  2019  6878846.0  6.572492e+06  3.063535e+05
3  ITA  2020  4985916.0  7.160950e+06  2.175034e+06
4  ITA  2021  6333782.0  6.859642e+06  5.258600e+05
..  ...  ...      ...      ...
95 LVA  2017  133250.0  6.426128e+04  6.898872e+04
96 LVA  2018  169884.0  5.552461e+04  1.143594e+05
97 LVA  2019  171912.0  1.163287e+05  5.558327e+04
98 LVA  2020  118560.0  5.534327e+04  6.321673e+04
99 LVA  2021  167310.0  7.083659e+04  9.647341e+04

[100 rows x 5 columns]

Tableau récapitulatif :
Modèle \
0 MLP (Réseau de neurones, rolling forecast)

Variables exogènes      AIC \
0 Aucune (univarié, seulement immigration passée) Non applicable

BIC      RMSE      MAE \
0 Non applicable  1.841784e+06  827710.586662

Commentaires / Note
0 Le modèle MLP prédit sur les 5 dernières année...
```

LSTM

```
LSTM (rolling forecast):
pays  année  obs      lstm_pred  lstm_abs_error
0  ITA  2017  7827846.0  7.186732e+06  6.411145e+05
1  ITA  2018  7423000.0  7.202312e+06  2.206875e+05
2  ITA  2019  6878846.0  7.542776e+06  6.639305e+05
3  ITA  2020  4985916.0  7.677680e+06  2.691764e+06
4  ITA  2021  6333782.0  6.988549e+06  6.547670e+05
..  ...  ...      ...      ...
95 LVA  2017  133250.0  1.054768e+05  2.777319e+04
96 LVA  2018  169884.0  1.100476e+05  5.983640e+04
97 LVA  2019  171912.0  1.289035e+05  4.300852e+04
98 LVA  2020  118560.0  1.200180e+05  1.457977e+03
99 LVA  2021  167310.0  1.089037e+05  5.840630e+04

[100 rows x 5 columns]

Tableau récapitulatif LSTM :
Modèle  AIC  BIC      RMSE      MAE  Commentaires / Note
0 LSTM (univarié) N/A  N/A  1.264998e+06  697305.417766
```

GRU

```
GRU (rolling forecast):
pays  année  obs      gru_pred  gru_abs_error
0  ITA  2017  7827846.0  7.318676e+06  5.091700e+05
1  ITA  2018  7423000.0  7.766459e+06  3.434590e+05
2  ITA  2019  6878846.0  7.939225e+06  1.060379e+06
3  ITA  2020  4985916.0  7.841662e+06  2.855746e+06
4  ITA  2021  6333782.0  7.442084e+06  1.108302e+06
..  ...  ...      ...      ...
95 LVA  2017  133250.0  1.016704e+05  3.157956e+04
96 LVA  2018  169884.0  1.192716e+05  5.061244e+04
97 LVA  2019  171912.0  1.409692e+05  3.094280e+04
98 LVA  2020  118560.0  9.517984e+04  2.338016e+04
99 LVA  2021  167310.0  1.400463e+05  2.726366e+04

[100 rows x 5 columns]

Tableau récapitulatif GRU :
Modèle  AIC  BIC      RMSE      MAE  Commentaires / Note
0 GRU (univarié) N/A  N/A  1.254146e+06  684138.654641
```