Analysis of the Employment and Socio-Cultural data

William Guilbaud, Théo Lauverjat, Pierre La Rocca

12 octobre 2021

Résumé

L'objectif de ce compte rendu est de présenter nos résultats et analyses concernant le jeu de données : **Employment and Socio-Cultural data**. À partir de ce dernier, nous avons choisi d'analyser les disparités existantes entre les différents secteurs d'activités de l'étude et les salaires hebdomadaires médians, en fonction de facteurs sociaux culturels (genre, ethnie, âge) au cours du temps.
Suite à une approche exploratoire, nous montrerons comment le genre ou l'ethnie d'une personne permettent d'expliquer la nature de son emploi au travers d'une ACP.
Puis, nous étudierons différents modèles de régression et leur efficacité à prédire le salaire médian à partir d'autres variables.

1 Introduction

1.1 Présentation du jeu de données

Notre étude porte sur le jeu de données **Employment and Socio-Cultural data**. Celui-ci rassemble des informations sur l'ethnie, le sexe, l'âge ou encore le revenue de personnes employées dans différentes industries et secteurs d'activités *américains* selon les années. Ces données ont été récoltées par l'*U.S Bureau of Labor Statistics* et sont regroupées en deux tables : *employed.csv* et *earn.csv*.

La table *employed* contient des informations relatives à la répartition de personnes au sein de différentes industries et identifiées selon leur sexe, leur ethnie et leur profession :

- industry (char) : définie à quel secteur industriel se rapportent les informations de la ligne. Le jeu de données présente 19 industries différentes ;
- major_occupation (char) : correspond à l'emploi principal dans *industry*. Le jeu de données contient 5 occupations majeures ;
- minor_occupation (char) : correspond à la spécialisation de l'activité dans *industry* : à chaque *major_ occupation* sont rattachées plusieurs *minor_ occupation*. Le jeu de données contient 11 occupations secondaires différentes ;
- race_gender (char) : permet de préciser quelles données sont représentées sur la ligne : soit le sexe/genre : Hommes ("Men") ou Femmes ("Women"), soit les ethnies : "White", "Asian" ou "Black or African American", soit le "Total" ;
- industry_total (double) : nombre total d'employés relatif au secteur d'activité de la ligne ;
- employ_n (double) : le nombre d'employés relatif à la *minor_ occupation* ;
- year (int) : année pour laquelle les données ont été enregistrées (entre 2015 et 2020).

La table *earn* présente le salaire hebdomadaire médian et le nombre de personnes employées regroupées par ethnie, genre et tranche d'âge en fonction du temps :

- sex (char) : sexe de la personne ;
- race (char) : ethnie de la personne ;
- ethnic_origin (char) : personne d'origine hispanique ou non. Nous ferons le choix de rattacher cette dernière à *race*, les colonnes ne s'interceptant pas et présentant des résultats en lien sur les autres paramètres ;
- age (char) : tranche d'âge ;
- year (int) : année pour laquelle les données ont été enregistrées (entre 2010 et 2020).
- quarter (int) : semestre de l'année pour lequel les données ont été enregistrées ;
- n_persons (double) : nombre de personnes employées par groupes
- median_weekly_earn (int) : revenue hebdomadaire médian en US dollars.

Ces deux tables n'étant pas interopérables, l'étude du jeu de données passera par une problématique générale à laquelle nous répondrons par deux approches complémentaires correspondant aux tables à notre disposition.

Étant donné les discriminations observées dans le monde professionnel, peut-on identifier et prédire l'influence que l'origine ethnique, l'âge et le genre peuvent avoir sur le salaire et le poste occupé par un individu ?

Notre travail sur la table *employed* visera à montrer l'évolution de la diversité dans les secteurs industriels en fonction du temps en années ;
L'étude de la table *earn* permettra d'illustrer l'influence des facteurs de genre, d'âge et d'ethnie sur le revenu pour chercher à mettre en exergue un facteur discriminatoire principal.

1.2 Transformation et nettoyage des données

Les données à notre disposition, réparties en deux tables, sont des résumés d'informations. Il n'existe ainsi pas d'individu en leur sein, seulement des valeurs synthétiques. Sur notre table *earn.csv* par exemple, les valeurs de la ligne correspondant au genre 'Both sex' sont égales à la somme des valeurs des lignes correspondant au genre 'Men' et 'Women'. Dans *employed*, pour le critère ethnique, il est néanmoins intéressant de remarquer que les valeurs associées à "TOTAL" diffèrent légèrement de l'addition des valeurs de `Asian`, `Black or African American` et `White`. Cela est lié au fait que d'autres ethnies ont pu être comptabilisées dans le total, comme les hispaniques, considérés comme une *ethnical_ origine* plutôt qu'une *race*.

Une première approche du nettoyage a consisté à construire une nouvelle table en extrayant les lignes des attributs que nous souhaitions analyser. Face à l'effet de taille observé sur les effectifs des industries, nous avons ensuite choisi de passer par des pourcentages pour avoir des valeurs plus représentatives de la répartition des genres et des ethnies.
Enfin, certaines données étaient incohérentes, comme la présence de facteurs ethniques dans la colonne `industry` de *employee*. Nous avons fait le choix de supprimer les enregistrements présentant de telles anomalies.
Pour la table *earn*, l'intersection des classes d'âges n'étant pas toujours vide mais globalement intègre – on avait par exemple des enregistrements pour les "16 - 24 ans", qui se recoupaient avec d'autres pour les "16 - 19 ans" et les "20 - 24 ans" –, nous avons opté pour l'élagage des catégories plus générales pour ne garder que les intervalles exclusifs minimaux.

Suite à ces transformations nous avons obtenu de nouvelles tables, comme *employed*, présentant les variables suivante :

- industry ;
- major_occupation ;
- minor_occupation ;
- industry_total ;
- year ;
- White_pourcentages : le pourcentage de personnes blanches dans la *minor_ occupation* donnée ;
- Black_pourcentages : le pourcentage de personnes noires" dans la *minor_ occupation* donnée ;
- Asian_pourcentages : le pourcentage de personnes asiatiques dans la *minor_ occupation* donnée ;
- Men_pourcentages : le pourcentage d'hommes dans cette dans la *minor_ occupation* donnée ;
- Women_pourcentage : le pourcentage de femmes dans la *minor_ occupation* donnée.

La table *employee* a également été scindée en deux sous-tables *Industry* et *NbEmployee*. La première donne le nombre total et le pourcentage relatif d'employés d'une industrie en fonction de l'année et de critères sociaux ethniques. La seconde, le nombre et le pourcentage relatif d'employés en fonction de l'année et de critères sociaux-ethniques pour une industrie, ses *major_ occupations* et *minor_ occupations*.

2 Approche exploratoire

Nous nous intéresserons ici à dresser un premier portrait des employés américains. Nous quantifierons leur nombre, et leur répartition au sein des différents secteurs d'activités. Nous mettrons également en dialogue les répartitions ethnique et genrée face au paramètre du salaire.

2.1 Earn

L'enjeu est ici d'expliquer l'existence d'écarts salariaux en fonction de l'âge, du genre et de l'origine ethnique.
Nous allons commencer par étudier l'identité des groupes présents dans la table *earn*. La table 2.1 montre que les Américains déclarant des revenus sur la période de 2010 à 2020 sont à majorité d'origine caucasienne (race "White").
Ensuite, les deux groupes ethniques suivants sont les Afro-Américains qui représentent environs 13 % de la population d'individus et les Asiatiques qui forment près de 6 % de la population étudiée. Concernant la répartition genrée, le jeu présente légèrement plus d'hommes – 54 % –, que de femmes – 46 %.

	Proportion d'individu [Minimum , Maximum]
Homme	[54.8%,56.2%]
Femme	[43.8%,46.2%]
White including Hispanic	[76.7%,81.1%] [14.7%,18.3%]
Black	[11.6 %,13.4%]
Asian	[4.9%,6.8%]

TABLE 1 – Composition des individus de la table *earn*

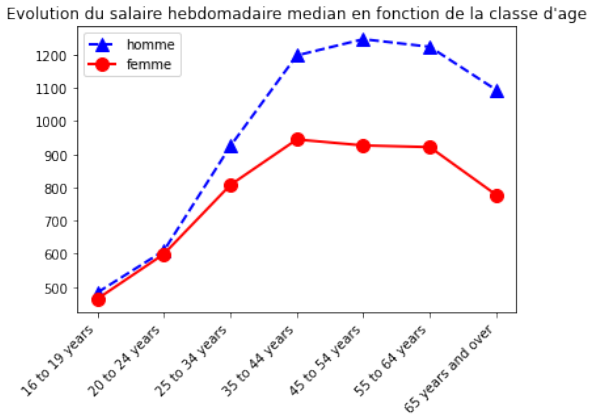

FIGURE 1 – Évolution du salaire hebdomadaire médian selon le genre en fonction de l'âge.

La figure 1 montre l'évolution du salaire des hommes et des femmes en fonction de leur âge. La tendance observée, à écart de salaire près, est la même. D'abord une progression du salaire entre 16 et 34 avant l'apparition d'un palier, suivi d'une décroissance progressive à l'âge de la retraite, à plus de 64 ans.
Si la disparité salariale entre hommes et femmes observée sur la tranche 16 - 24 ans semble faible, celle-ci se creuse progressivement entre 24 et 34 ans jusqu'à atteindre un écart stable à partir de 35 ans. De cet âge jusqu'à plus de 65 ans, les femmes touchent en moyenne 24% moins que les hommes.

La figure 2 montre l'évolution du salaire moyen d'une ethnie en fonction des tranches d'âges disponibles. S'il faut cependant davantage raisonner en termes de générations que d'évolution – d'autres facteurs sociaux non présents dans noter jeu de données étant susceptibles d'expliquer ces manifestations, à l'instar l'histoire de l'immigration des populations asiatiques aux États-Unis –, on remarque trois catégories différentes :

— Les salaires globalement faibles auxquels sont rattachés les hispaniques et les afro-américains ;
— Les salaires constamment croissants au fil des générations relatifs aux blancs ;
— Les salaires très élevés pour la génération 25 - 54 ans mais moins élevés pour la génération des 55 ans et plus, rattachés aux personnes d'origine asiatique ;

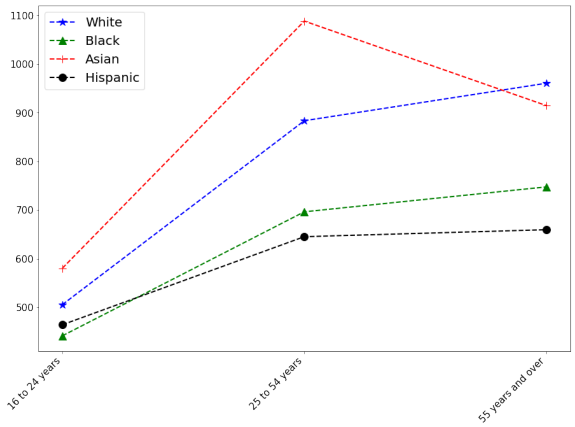

FIGURE 2 – Différences du salaire hebdomadaire médian selon l'ethnie en fonction de l'âge.

Ces deux graphiques mettent en avant une évolution des salaires différente en fonction du temps, à la fois liée à des facteurs de sexe et à des facteurs d'origine ethnique.

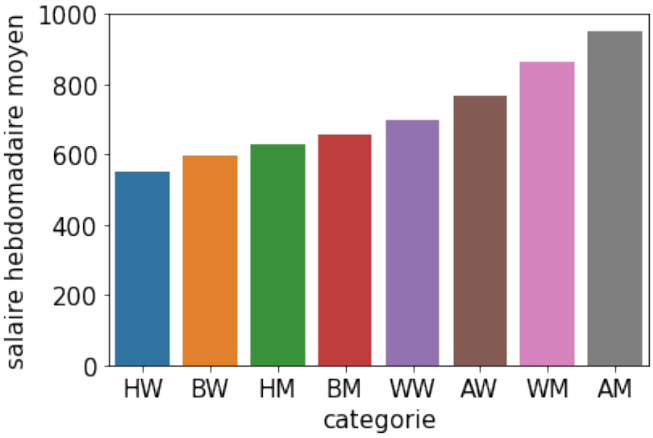

FIGURE 3 – Écarts de salaire hebdomadaire médian observés en fonction du genre et de l'ethnie.

Ce comportement est confirmé par la figure 3 [1]. Le diagramme qui y figure croise, sans tenir compte de l'âge cette fois-ci, le salaire hebdomadaire moyen avec les critères de sexe et d'ethnie. On peut distinguer graphiquement trois catégories :

— Les groupes au salaire élevé (supérieur à 800 dol-

1. Signification des abréviations en annexe A.1.

lars) : des hommes blancs ou asiatiques ;
— Les groupes possédants un salaire moyen (entre 600 et 800 dollars) : des femmes asiatiques ou blanches et des hommes afro-américains ;
— Les groupes possédants un salaire inférieur à 600 dollars : des femmes afro-américaines et des hispaniques, sans distinction de sexe, bien que les hommes touchent ici encore plus que les femmes.

Les différentes modélisations effectuées sur la table *earn* mettent en avant l'existence d'écarts de salaires relatifs à l'âge, le genre et l'origine ethnique. La suite du travail sur cette table visera à identifier l'influence plus précise des facteurs sociaux sur le salaire. On passera notamment par le recours à une ACP.

2.2 Employed

Les disparités entre sexes et ethnies, misent en avant en 2.1 au niveau des salaires, existent également dans la répartition des salariés au sein des différents secteurs d'activités et des différentes industries de la table *employed*.

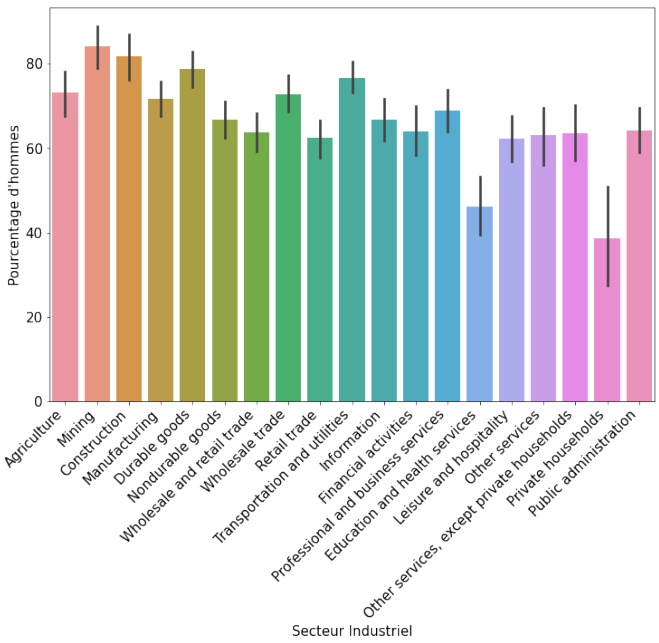

FIGURE 4 – Pourcentage d'hommes dans différents secteurs industriels

La figure 4 représente le pourcentage d'hommes dans différents secteurs industriels. Par différence, il renseigne également sur le pourcentage de femmes dans ces mêmes milieux. On y remarque une certaine répartition sexuée des secteurs industriels. Si de manière globale, les hommes sont majoritaires dans la plupart de ces secteurs, on observe tout de meme une certaine forme de genrification des activités. La construction et les mines sont des secteurs davantage masculins, là où *Private households* et *Education and health services* sont plus féminins.

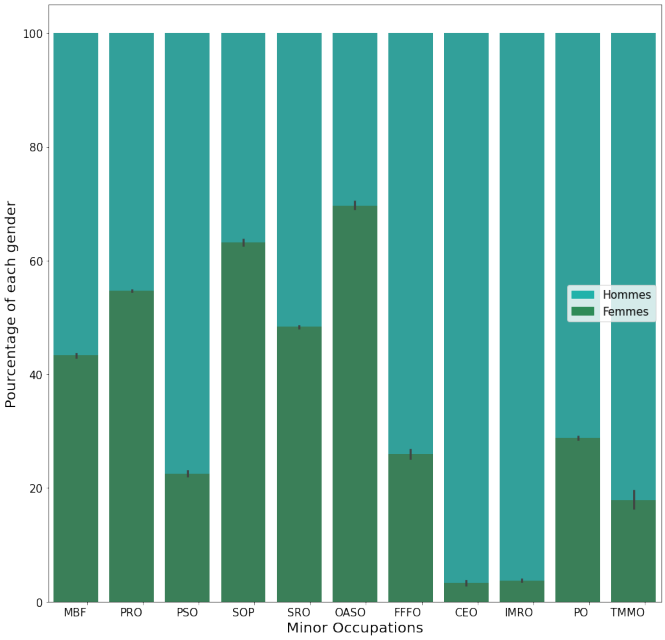

FIGURE 5 – Répartition des hommes et des femmes dans différentes spécialités

La figure 5, axée sur l'emploi au sein d'un secteur industriel, étend les observations faites sur 4 à la répartition genrée des emplois et de leurs spécialisations. Certaines activités comme "Construction and extraction occupations" (CEO) sont monopolisées par les hommes alors que d'autres comme "Office and administrative support occupations" (OASO) sont en majorité féminins.

La figure 6 est un diagramme cumulé des pourcentages ethniques présents dans les différentes industries de la table. On émet ici l'hypothèse que la différence observée entre chaque barre et le 100% vient de la population hispanique, la colonne *ethnic_ origin* de *earn* n'étant pas présente dans *industry*. La majorité des emplois sont occupés par des personnes blanches, ethnie majoritaire aux États-Unis. Relativement à l'ethnie, on observe que les *African American* sont davantage répartis dans les transports et l'administration publique que dans des secteurs comme l'agriculture. La population asiatique est quant à elle davantage présente dans les domaines rattachés à la manufacture, le commerce et les services.

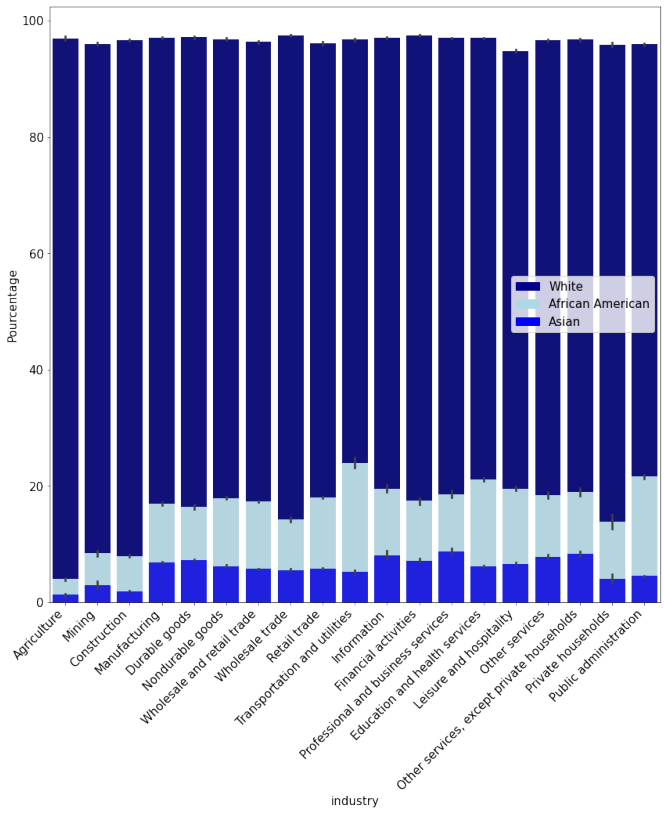

FIGURE 6 – Pourcentage de la répartition ethnique dans différents secteurs industriels

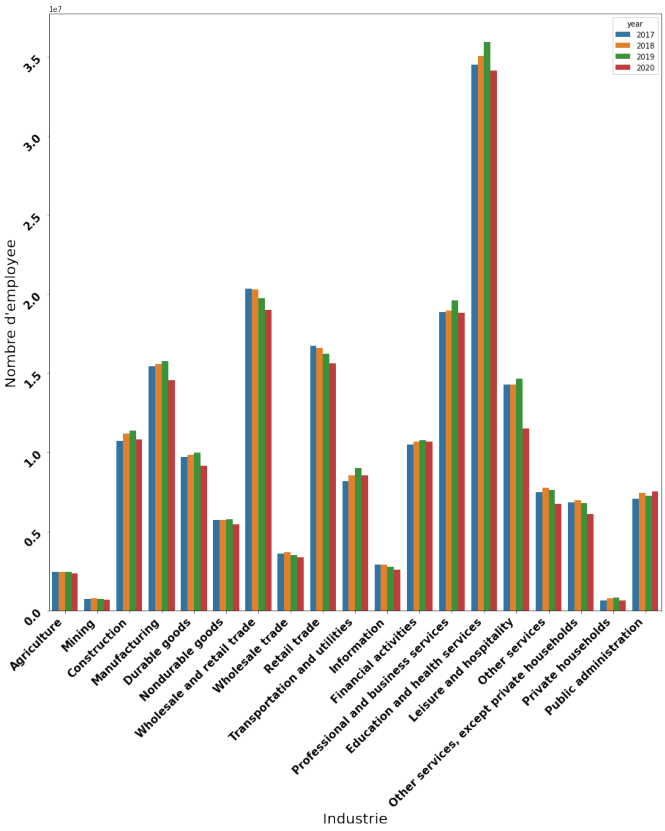

FIGURE 7 – Évolution du nombre total d'employés dans différentes industries au cours du temps

Avant de conclure sur cette première partie, la figure 7 illustre l'évolution du nombre total d'employés pour chaque industrie entre 2017 et 2020. Celle-ci est intéressante dans la mesure où l'on remarque que pour une majorité d'industries, l'année 2020 présente une rupture avec les années précédentes. Elle servira notamment à justifier les difficultés de nos modèles à prédire les paramètres de 2020 en connaissant les années précédentes.

Ainsi, l'analyse exploratoire effectuée sur les deux tables appuie l'idée qu'il existerait une répartition genrée et ethnique des activités industrielles, mais aussi des écarts de salaire pour ces mêmes critères. D'après 1 par exemple, les différences de salaires semblent également être générationnelles.
Dans l'idéal, il aurait été riche de lier les deux tables si cela avait été techniquement possible. Nous aurions pu émettre alors des hypothèses vérifiables, comme l'existence de différences de répartitions ethniques dans les secteurs industriels, en fonction des générations. D'autres liens similaires auraient pu être fait pour expliquer les différences salariales observées entre les sexes 2.1.
Ces liaisons auraient permis d'émettre l'hypothèse qu'une partie des écarts de salaires observés dans *earn* puissent être corrélés à la répartition genrée et ethnique des différents secteurs d'activités de *employee*. Cependant le jeu de données étudié ne contenant pas d'informations permettant de lier ces deux tables, il nous est impossible d'approfondir davantage cette hypothèse.

Nous nous proposons à présent de développer les premières réponses fournies par l'approche exploratoire à l'aide de différentes analyses des données.

3 Analyse des données et représentation

3.1 Classification

À partir de la table *employed* nous avons essayé de montrer comment le genre ou l'origine d'une personne permettaient d'expliquer la nature de son emploi. Pour cela nous avons réalisé une ACP.

3.1.1 ACP

L'analyse en composantes principales (ACP) consiste à transformer des variables corrélées en nouvelles variables décorrélées les unes des autres. L'ACP permet ainsi de réduire le nombre de variables, de rendre l'information moins redondante et de la visualiser plus facilement.

Les individus de cet ACP sont les couples *Industries* et *Minor_ Occupation* et les valeurs de l'ACP sont le pourcentage d'hommes, le pourcentage de femmes, le pourcentage de caucasiens, le pourcentage d'Afro-Americain et le pourcentage d'asiatiques travaillant pour ces couples d'industries et de secteurs d'activité. Toutes nos variables étant des variables quantitatives, l'utilisation d'une ACP est préconisée à d'autres méthodes de réduction de variables.

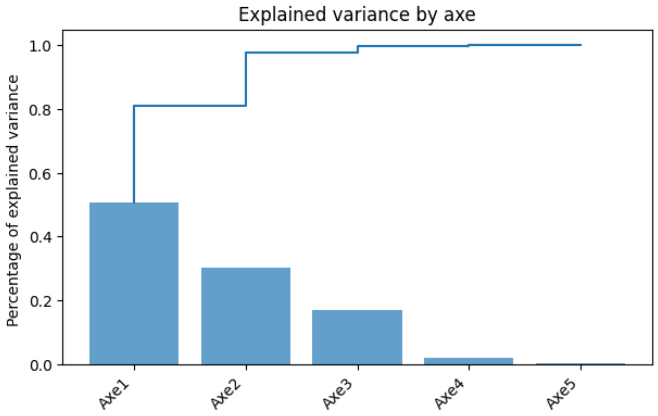

FIGURE 8 – Variance expliqué par chaque Axes de l'ACP

En affichant la somme cumulée de l'argument *explained_ variance_ ratio*, on observe que 81 % de la variance est expliquée par les deux premiers axes et 98 % par les trois premiers comme le montre la figure 8. Nous allons considérer que ces 3 premières variables résument suffisamment nos données.

Pour analyser le résultat de l'ACP nous allons afficher les variables en fonction des trois premiers axes.

D'après la figure 9, le premier axe sépare distinctement les hommes et les caucasiens d'un côté et les femmes, les Afro-Américains et les asiatiques de l'autre, alors que le deuxième axe sépare les caucasiens et les femmes d'un côté et les hommes et les Afro-Américains de l'autre, cependant la proportion d'asiatiques n'est

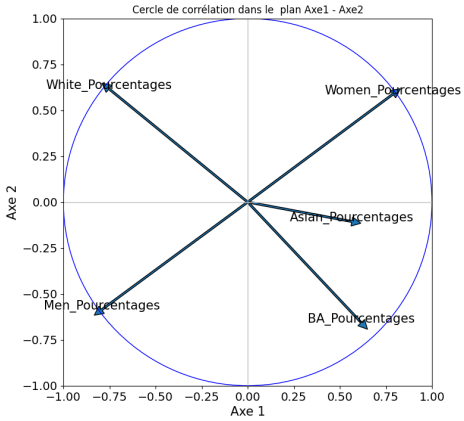

FIGURE 9 – Variables expliquées par l'Axe 1 et 2

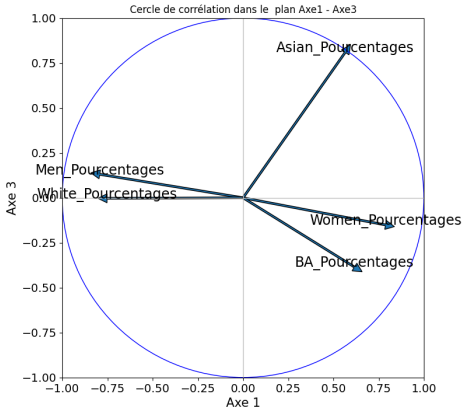

FIGURE 10 – Variables expliquées par l'Axe 1 et 3

presque pas expliquée par cet axe.

D'après la figure 10, le troisième axe sépare particulièrement les asiatiques du reste de la population ainsi que dans une moindre mesure les Afro-Américains dans la direction opposée.

Nous allons ensuite représenter nos individus selon les deux premiers axes pour analyser les disparités d'emploi. Se concentrer sur les deux premiers axes permet de visualiser en deux dimensions 80 % de l'information de la table, mais ne pas utiliser le troisième axe nous fait perdre une partie de l'information sur le critère de la proportion d'asiatiques dans les couples *Minor_ Occupations*, et *Industries*.

La figure 11 permet d'attirer l'attention sur le secteur d'activité. On observe que c'est un bon indicateur de la répartition ethnique et homme / femme des couples *Minor_ Occupations* et *Industries*. Particulièrement pour

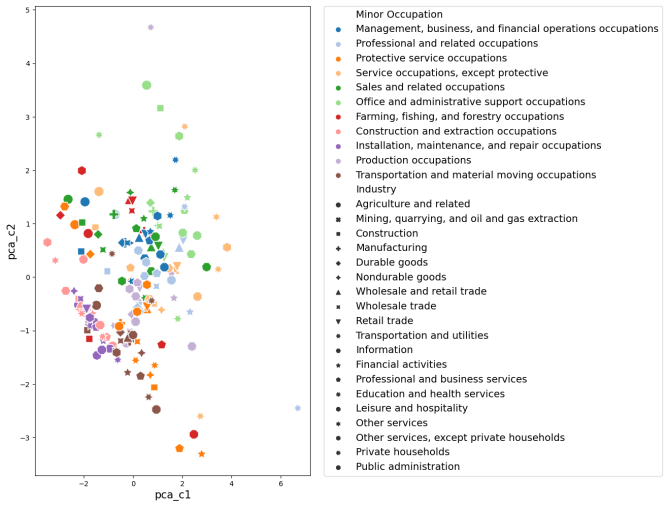

FIGURE 11 – Visualisation des disparités dans les secteurs d'activités en fonction des deux premiers composantes principales

les secteurs : "Installations maintenance and repair occupations" (en violet) et "Construction and extraction occupations" (en rose claire). Ces derniers se retrouvent tous les deux dans la partie en bas à gauche, qui correspond à un grand pourcentage d'hommes d'après la figure 9. À l'opposé, en haut à droite, des secteurs comme "Office and administrative support occupations" (en vert clair) correspondent à un grand pourcentage de femmes dans toutes les industries.

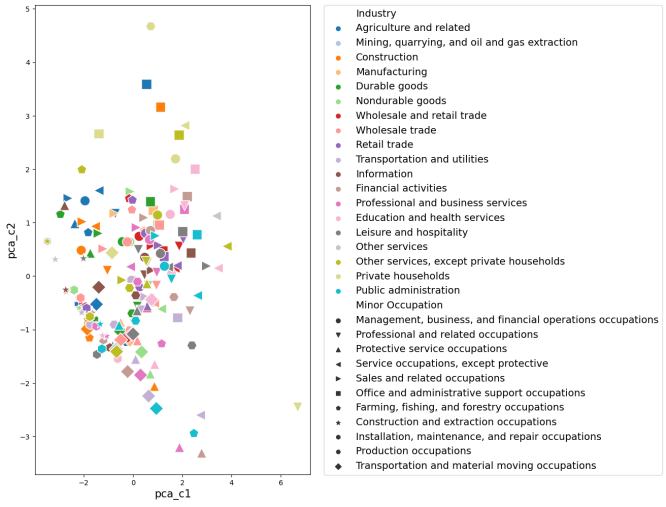

FIGURE 12 – Visualisation des disparités dans les industries en fonction des deux premières composantes principales

La figure 12 représente la même information que la figure 11 en se concentrant sur les industries. On observe que les disparités sont moins accentuées que dans la figure précédente. Cependant nous pouvons identifier certains *patterns* tels que l'industrie "Agriculture and related" (en bleu foncé) qui est à majorité en haut à gauche du graphique, ce qui y montre une prévalence de caucasiens. L'industrie "Leisure and hospitality" se trouve quant à elle majoritairement à droite du graphe, ce qui montre une prévalence de femmes et de minorités ethniques en son sein.

3.2 Regression

À partir de la table *earn* nous avons tenté d'expliquer le salaire médian en fonction des autres variables de la table : l'ethnie, le sexe et l'âge. Nous avons pour cela appliqué la régression que nous avons déclinée sur différentes méthodes afin de pouvoir comparer leurs performances sur notre jeu de données.
Pour appliquer ces méthodes, nous avons tout d'abord transformé les variables qualitatives de notre table en variables binaires selon l'encodage *oneHot*, puis nous avons séparé notre ensemble de données en un ensemble d'apprentissage et un ensemble de test.

3.2.1 Cross Validation temporelle

Sur notre ensemble d'apprentissage, nous avons entraîné nos différents modèles en utilisant la méthode de validation croisée [2].

Cette méthode consiste à séparer notre ensemble d'apprentissage en parties de longueurs égales et de les utiliser tour à tour comme ensemble de validation tandis que les autres sont utilisés pour l'entraînement. La performance du modèle peut ainsi être obtenue en moyennant les performances de chaque itération.

Au cours de ce projet, nous avons choisi d'utiliser la **cross validation temporelle**. En effet, la cross validation 'classique' sélectionne des échantillons de validation et d'entraînement aléatoirement dans notre jeu de données. Néanmoins, nos données étant regroupées par années, il serait incohérent d'utiliser des valeurs d'années futures pour pouvoir prédire une valeur du passé. Comme le montre la figure 13, pour chaque itération n, les données des années $X_1, ..., X_n$ sont donc utilisées pour l'entraînement et sont validées sur l'année $X_(n+1)$.

Pour implémenter cette séparation des données, nous avons utilisé `TimeSeriesSplit(n_splits=X)` que nous avons ensuite passé à la méthode `GridSearchCV()` afin de trouver le ou les hyperparamètres qui nous donnaient le meilleur résultat pour chacune de nos méthodes.

Dans la suite du rapport, nous utiliserons l'abrévia-

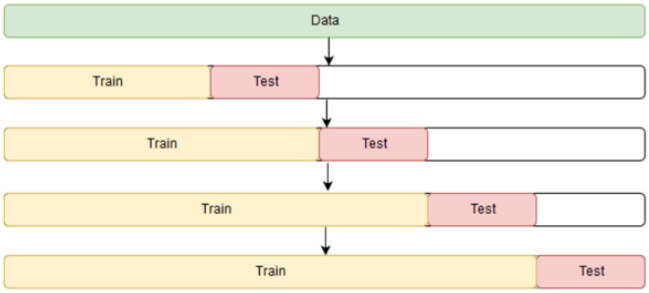

FIGURE 13 – Séparation des données d'entraînement et de validation pour chaque itération de la cross validation temporelle

tion **CV** pour faire référence à la cross validation et **CVT** pour parler de cross validation temporelle.

3.2.2 Regression lineaire et polynomiale

Dans un premier temps, nous avons réalisé une régression linéaire. Cette méthode cherche à prédire la valeur de la variable à expliquer comme une combinaison linéaire des autres variables, dites explicatives. C'est un modèle très simple, mais l'on peut choisir de modéliser des tendances plus complexes en ajoutant des variables polynomiales aux données. Le modèle appris sera dès lors une combinaison polynomiale des variables. Pour ce modèle, nous présenterons les résultats de la régression polynomiale de degré 2 qui nous a donné les meilleurs résultats.

Dans la suite du rapport, nous utiliserons l'abréviation **Lin** pour faire référence à la régression linéaire et **PolX** pour la régression polynomiale de degré X.

3.2.3 Les K plus proches voisins

L'algorithme des K plus proches voisins est un algorithme d'apprentissage supervisé utilisé pour la classification de données. Il existe néanmoins une version de l'algorithme utilisable pour la régression : on considère alors la valeur moyenne des k plus proches voisins au lieu de considérer leur classe. Ce nombre de voisins doit être choisi avec un compromis entre la précision de notre modèle et un risque d'erreur acceptable.

Dans la suite du rapport, nous utiliserons l'abréviation **KPP** pour désigner cette méthode.

3.2.4 Random forest

Une forêt aléatoire est une méthode qui vise à construire plusieurs arbres de décision, à les entraîner sur un modèle et à prendre la décision que la majorité des arbres considérera comme étant la meilleure. Dans le cas d'une régression, on prendra la valeur moyenne ou médiane des décisions de chaque arbre.

Un arbre de décision, dans le cas d'une régression, est un modèle construit à partir d'un ensemble d'apprentissage, qui va chercher à découper un espace en plusieurs régions où les valeurs à prédire sont proches. Chaque nœud interne décrit alors un test sur une des variables d'apprentissage, chaque branche représente un résultat du test et chaque feuille une valeur de la variable à expliquer (une valeur de classe lors d'une classification et une valeur numérique dans le cas d'une régression). Pour cela, l'espace est récursivement divisé à chaque étape en deux régions en choisissant parmi les variables explicatives celles qui minimisent un critère donné (exemple : le critère de GINI). Un processus d'élagage peut ensuite être appliqué pour limiter le sur-apprentissage.

Dans la suite du rapport, nous utiliserons l'abréviation **RF** pour désigner cette méthode.

3.2.5 Choix des métriques de comparaison

Afin de comparer les performances de nos différents modèles sur le jeu de test, nous avons choisi de calculer pour chacun d'entre eux :

- **RMSE** : l'erreur quadratique moyenne (MSE 1) d'un estimateur $\hat{\theta}$ et d'un paramètre θ est une mesure caractérisant la « précision » de cet estimateur. Le but est de mesurer la proximité d'une droite de régression par rapport à un ensemble de points. La différence entre la prédiction et la valeur réelle étant au carré, cet estimateur est sensible aux valeurs aberrantes. Nous avons choisi de travailler avec le RMSE 2, qui est la racine carrée du MSE, afin d'amener cette valeur à la même échelle que celle de l'erreur de prédiction.
- **MAE** : l'erreur absolue moyenne (MAE 3) calcule la moyenne des résidus entre la valeur réelle et de la valeur prédite. Cette dernière est moins sensible aux valeurs aberrantes que le RMSE et nous semblait plus facile à interpréter pour mesurer les écarts à la réalité.

$$MSE = \mathbb{E}\left[(\theta - \hat{\theta})^2\right] = \frac{1}{n} * \sum_{i=1}^{n}\left[(\theta_i - \hat{\theta}_i)^2\right] \quad (1)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} * \sum_{i=1}^{n}\left[(\theta_i - \hat{\theta}_i)^2\right]} \quad (2)$$

$$MAE = \frac{1}{n} * \sum_{i=1}^{n}\left|\theta_i - \hat{\theta}_i\right| \quad (3)$$

3.2.6 Résultats et interprétation

En appliquant chacune des méthodes présentées à notre jeu de données, nous avons pu observer une différence de performances selon les deux métriques.

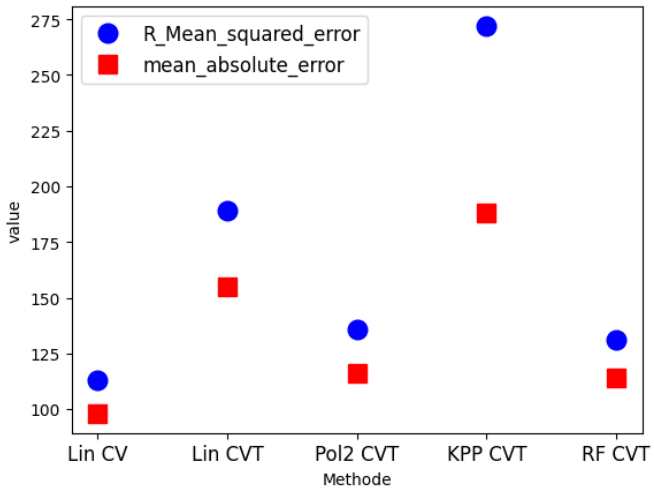

FIGURE 14 – Valeur des RMSE et MAE pour les différentes méthodes de régression utilisées.

Comme nous pouvons le voir sur la figure 14, nos meilleures méthodes de régression avec apprentissage par cross validation temporelle sont *random forest* (**RF**) et la régression linéaire de second degré (**Pol2**). Les salaires médians de notre jeu de données se situent dans une fourchette de 400 et 1500$:l'erreur absolue moyenne avoisinant les 130 pour les deux méthodes, nous pouvons dire que nos modèles ne semblent pas très performants.
Néanmoins, notre jeu de données de test pour ces régressions était composé des années 2019 et 2020. Or, lors de l'analyse exploratoire de nos données, nous avions remarqué que l'année 2020 différait des années précédentes sur plusieurs caractéristiques dans les deux tables *employed* (7) et *earn* (18). Ces disparités sur le jeu de test ont donc eu un impact sur les performances mesurées et nous avons donc recommencé nos régressions en omettant cette année. Ainsi, en apprenant notre modèle avec les années 2010 à 2017 et en le testant sur 2018 et 2019, nous obtenons des meilleures performances pour les métriques RMSE et MAE (respectivement 75 et 68) pour lesquelles nous pouvons dire que notre modèle est plutôt bon.

De plus, au vu de ces résultats, il semble que notre modèle soit plus facilement estimable par des méthodes non linéaires. Le fait que la méthode RF nous donne de meilleurs résultats est également logique puisque notre jeu de données contient beaucoup de variables qualitatives (transformées selon la méthode one Hot lors de son implémentation).
Il est néanmoins intéressant de remarquer qu'en moyenne, la méthode de régression linéaire avec apprentissage par cross validation classique semble être la plus performante, bien d'incohérente dans notre étude compte tenu de l'importance du temps dans l'agencement de nos données.

Finalement, afin de comprendre les résultats d'une de nos méthodes de régression les plus performante, la *random forest*, nous nous sommes intéressés à l'importance des variables dans ce calcul.

Pour cela nous avons utilisé la **"permutation feature importance"**. L'idée de cette méthode est de regarder l'importance d'une caractéristique en mesurant de combien la performance du modèle (R2, MSE, etc) diminue lorsque cette dernière n'est pas disponible. Pour mettre en place cela, les valeurs de cette caractéristique sont mélangées aléatoirement pour que la colonne ne contienne plus d'informations utiles. On recalcule alors de nouveau la performance de ce modèle et on recommence l'opération sur les autres variables explicatives. Ainsi, on peut classer les caractéristiques selon leurs impactes sur la performance du modèle.

Cette méthode à l'avantage d'être efficace pour un jeu de données dont le nombre de caractéristiques est faible et de pouvoir être applicable sur n'importe quel type de modèle prédictif.

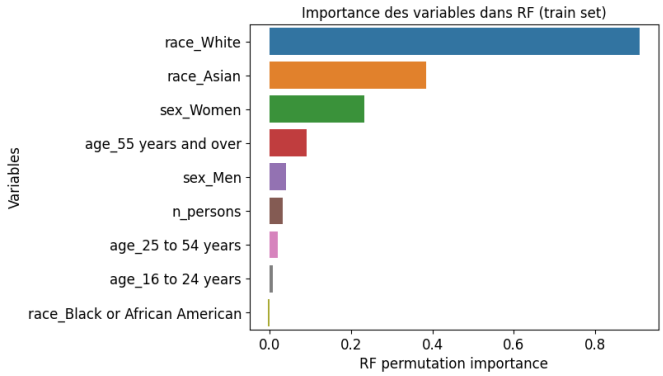

FIGURE 15 – Permutation d'importance des variables pour la méthode *random forest*.

À partir de 15, nous pouvons voir que les variables qui ont eu le plus d'importance pour estimer le salaire médian sont les ethnies caucasienne et asiatique ainsi que le genre des populations.

3.3 Méthodes ARMA

Pour prédire au mieux le revenu futur des différentes populations, nous avons décidé d'envisager une approche prédictive population par population au lieu de générer un modèle général en utilisant les méthodes ARMA.
Les méthodes de la famille ARMA sont des modèles permettant de comprendre et prédire les variations d'une seule série temporelle à la fois, en utilisant les données passées de cette même série. Ainsi nous pouvons nous demander si prédire chaque série indépendamment les unes des autres peut nous permettre d'obtenir de meilleurs résultats que nos méthodes de régression. Les modèles ARMA contiennent deux composantes :

— *Autoregressive model* (AR(p))4, un modèle qui analyse les observations passées pour prédire la valeur au temps t

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \tag{4}$$

— *Moving-average model* (MA(q))5 , un modèle qui analyse le bruit passé pour prédire la valeur au temps t

$$X_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{5}$$

Ainsi un modèle ARMA(p,q) a pour équation 6

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{6}$$

Le modèle ARMA nécessite que la série soit stationnaire, c'est-à-dire que sa covariance et sa moyenne n'évoluent pas dans le temps.
Pour cela on étend le modèle ARMA(p,q) avec *Autoregressive integrated moving average*, ARIMA(p,d,q) avec d l'ordre de différenciation de la série. Pour différencier à l'ordre n+1 on applique $\Delta y_t = y_t - y_{t1}$ à l'ordre n.

Nous allons prendre pour exemple la série des *Median weekly earning* pour un homme caucasien entre 25 et 54 ans.

D'après la figure 16, en décomposant la série, on observe que la série à une *trend*, elle est donc non stationnaire. Ainsi il existe une saisonnalité annuelle dans cette série.
Pour prendre en compte la saisonnalité, nous avons besoin d'utiliser une méthode SARIMA, *Seasonnal Auto-Regressive Integrated Moving Average*.

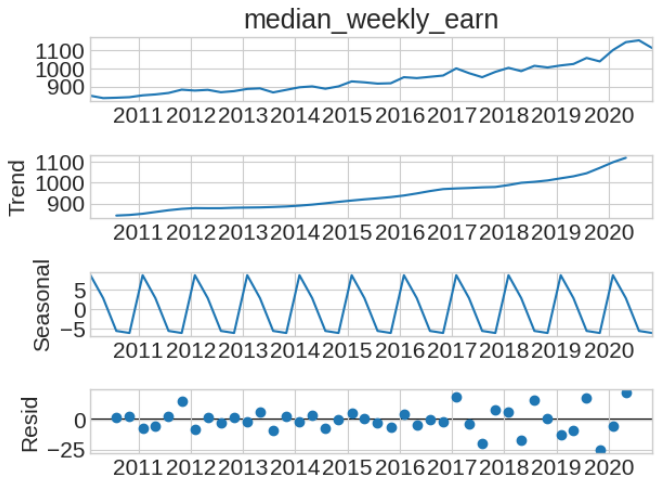

FIGURE 16 – Décomposition des composantes de la série

3.3.1 Choix des paramètres

La figure 16 nous montre une saisonnalité de fréquence annuelle (4 quarters), on peut en déduire que s=4.
Pour déterminer le paramètre d, nous avons différencié la série jusqu'à ce qu'elle passe le test de Dickey–Fuller, on trouve ainsi d=2.
Pour déterminer les paramètres p et q [1], il faut soit les déduire respectivement de la fonction d'autocorrélation partielle (PACF) et de la fonction d'autocorrélation (ACF), soit utiliser `pmdarima.auto_arima` sur une partie du jeu d'entraînement avec le paramètre d=2, les graphes d'ACF et PACF montrent peu de valeur significative ce qui tend à penser que les valeurs p et q sont faibles (<4). Nous avons choisi d'utiliser la première moitié des données pour déterminer les hyperparamètres et la seconde pour la CVT. La fonction converge vers les hyperparamètres ARIMA(3,2,0)(0,0,0)[4], il faut noter que les hyperparamètres de la saisonnalité sont de 0 ainsi la saisonnalité n'a pas été pris en compte. Cela provient saisonnalité du faible nombre de données ainsi que de la faible amplitude de la saisonnalité pour déterminer les hyperparamètres, en donnant plus de données on trouve une saisonnalité pour cette même série.

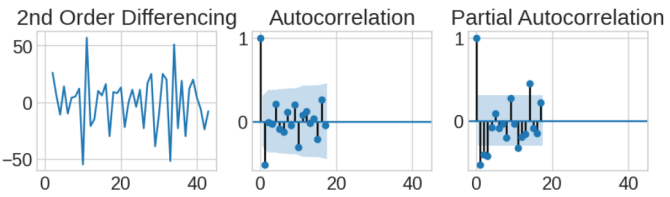

FIGURE 17 – ACF et PACF

Au décalage k, l'ACF est la corrélation entre les valeurs de séries séparées par k intervalles et la PACF est la corrélation entre les valeurs de séries séparées par k intervalles, compte tenu des valeurs des intervalles intermédiaires.

3.3.2 Résultat

On va ensuite essayer d'estimer la qualité de l'estimation en utilisant la CVT sur la deuxième partie des données. Le but à chaque itération est de prédire les 4 trimestres suivant les données d'entraînement. On trouve une RMSE de 18.3 et une MAE de 15.5. A noter qu'en changeant légèrement les hyperparamètres de SARIMA, les résultats sont faiblement changés. Ainsi, cette méthode semble être la méthode étudiée qui nous permet de prédire le plus efficacement les inégalités de revenus. Cependant à la différence des autres méthodes de régressions le modèle est très sensible à l'absence de certaines données dans les times series et ne permet pas de prédire des times series en entier.

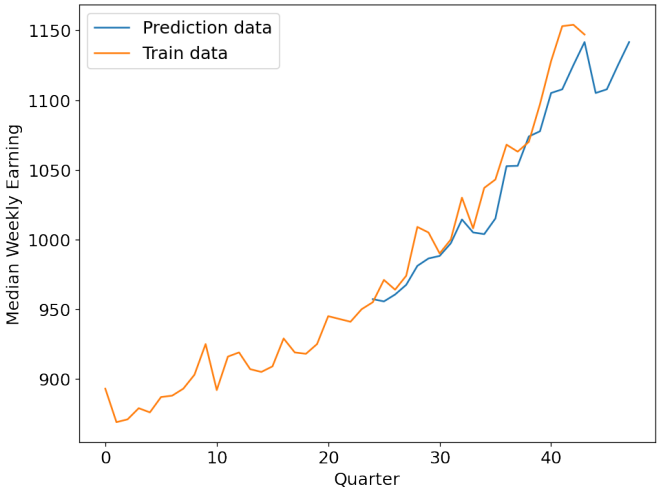

FIGURE 18 – Prédiction des revenus d'un homme Caucasien entre 25 et 54 ans avec la méthode SARIMA

4 Conclusion

En conclusion, nous avons présenté dans ce rapport notre analyse du jeu de données **Employment and socio-cultural data**. Après nettoyage de ces données, nous avons pu mettre en évidence l'existence de disparités salariales, genrées et ethniques dans le monde du travail par le biais d'une analyse exploratoire. Nous avons par la suite tenté de les expliquer et de les prédire grâce à une ACP et différentes méthodes de régression.

L'ACP sur *employed* nous a permis de mettre en avant des axes principaux permettant d'analyser que le genre et l'origine d'une personne permettent d'expliquer la nature de son emploi. Il existe une répartition genrée et ethnique de certaines industries.

Nous avons ensuite comparé plusieurs méthodes de régression sur la base de la RMSE et de la MAE dans le but d'expliquer le revenu hebdomadaire médian par rapport aux autres caractéristiques sociaux culturelles. La méthode de régression la plus efficace dans notre cas fut *random_forest*. Elle nous a permis de voir que les ethnies "White" et "Asian" et le genre avaient le plus d'influence sur l'apprentissage d'un modèle sur l'estimation d'un salaire moyen.

Finalement, nous aurons poussé l'usage de modèles un peu plus loin à l'aide des méthodes ARMA, ARIMA et SARIMA. Elles nous auront permis d'estimer la variation salariale avec plus de précision que les méthodes précédentes. Seulement, ces méthodes s'appuient sur des séries qui ne prennent pas en compte l'influence d'autres variables que celle considérée, ce qui limite leur comparaison avec nos autres modèles.

Une limite fondamentale du jeu de données étudié fut l'absence de possibilité de lier les données de la table *earn* aux données de la table *industry*. Cela rend par exemple impossible la caractérisation de la différence entre les populations d'études de chacune des tables. Un meilleur dialogue entre ces deux tables aurait pu permettre une analyse plus fine.

A Liste des abréviations utilisées

A.1 Figure 3

La première lettre de la catégorie correspond à l'ethnie, la seconde au genre.
Pour l'ethnie :
- W : White ;
- B : Black or African American ;
- A : Asian ;
- H : Hispanic.

Pour le genre
- M : Men ;
- W : Women.

A.2 Figure 5

Les abrévations pour les minor_occupation :
- MBF : Management, business, and financial operations occupations ;
- PRO : Professional and related occupations ;

— PSO : Protective service occupations ;
— SOP : Service occupations, except protective ;
— SRO : Sales and related occupations ;
— OASO : Office and administrative support occupations ;
— FFFO : Farming, fishing, and forestry occupations ;
— CEO : Construction and extraction occupations ;
— IMRO : Installation, maintenance, and repair occupations ;
— PO : Production occupations ;
— TMMO : Transportation and material moving occupations.

Références

[1] Selva Prabhakaran. Arima model – complete guide to time series forecasting in python. https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/, 2018.

[2] Soumya Shrivastava. Cross validation in time series. https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4, 2020.