Imputation de valeurs manquantes avec des réseaux de neurones

Prédiction des salaires dans l'enquête Emploi

Damien Babet

Journées de Méthodologie Statistique, 30 mars 2021

Table of contents

Insee
Mesurer pour comprendre

Introduction : les réseaux de neurones

Réseaux de neurone et statistique publique

Cette présentation est extraite d'un document de travail méthodologique de l'Insee (à paraître) sur des cas d'usages de réseaux de neurones pour la statistique publique, écrit avec Stéphanie Himpens, Quentin Deltour et Thomas Faria.

Les RN permettent d'envisager l'exploitation statistique des images, des textes, et ouvrent la voie à d'autres méthodes innovantes.

Cette présentation explore un autre usage : peut-on utiliser des RN pour une activité d'imputation classique du statisticien, sur des données classiques (mais relativement abondantes) ? Quels en sont les avantages et les inconvénients ?

Pourquoi les réseaux de neurones ?

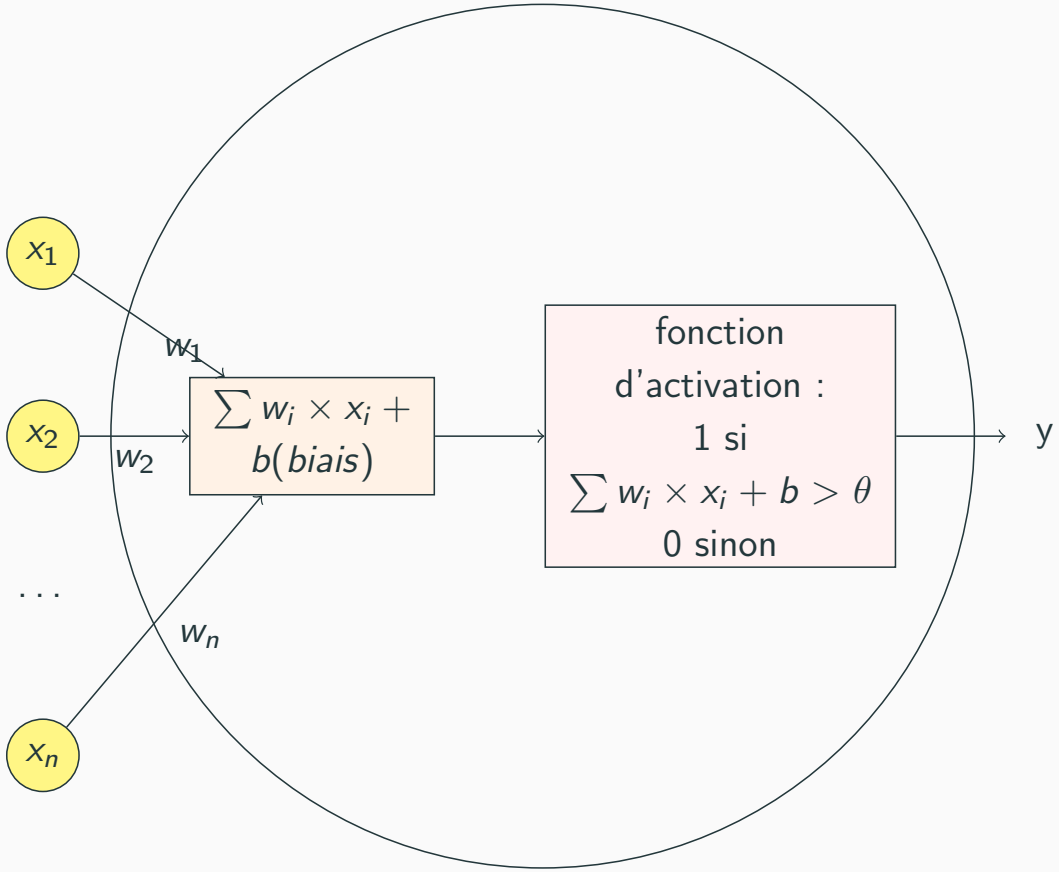
Les réseaux de neurones présentent les atouts et inconvénients du machine learning de manière accentué.

1. Grand nombre de paramètres impossibles à interpréter : problème de la boite noire
2. Grand nombre d'hyper-paramètres demandant un important travail de réglage et de cross-validation
3. Grandes performance prédictives en particulier quand les données sont abondantes
4. ... et quand les données sont complexes et très non linéaires (images, textes, etc.)

Insee
Mesurer pour comprendre

Qu'est-ce qu'un réseau de neurones ?

L'élément de base d'un réseau de neurones : *Le perceptron*



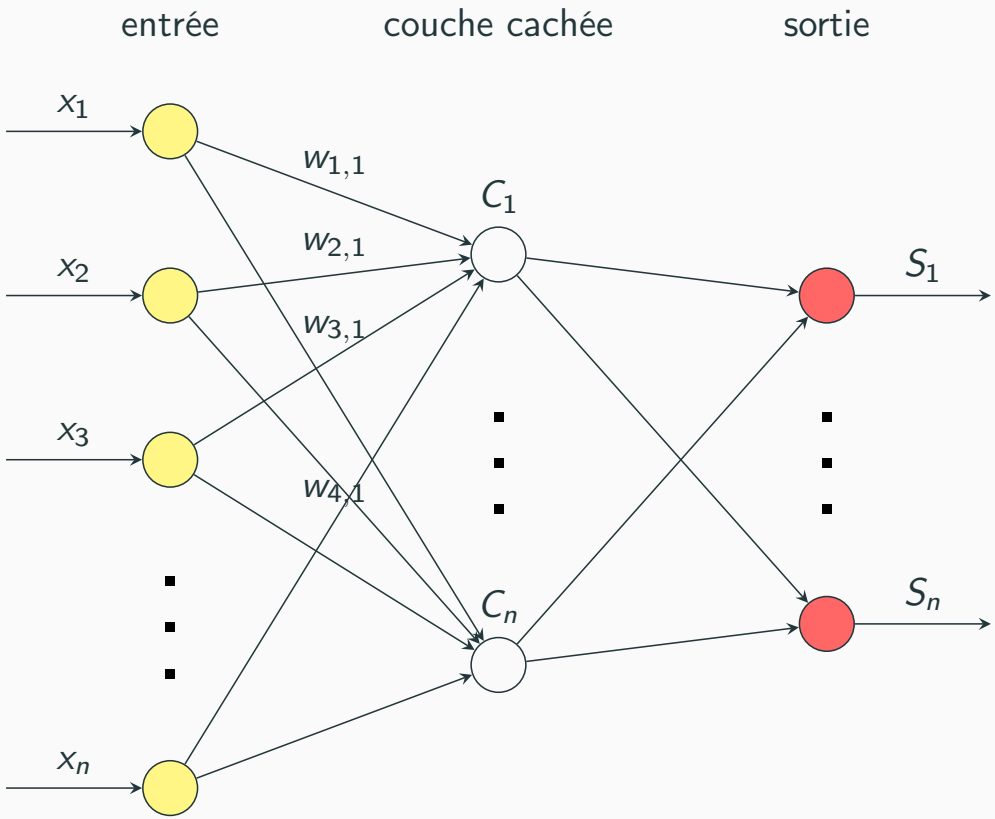
Ci-dessus : Le neurone formel de W. Mc Culloch et W. Pitts.

Entraîner le réseau consiste à estimer les poids w_i et le paramètre b.

Insee
Mesurer pour comprendre

Le perceptron multicouche

D. Rumelhart et simultanément Y. LeCun mettent au point la *rétropropagation* permettant d'estimer des systèmes multi-couches en 1986. Le *perceptron multi-couche* contourne certaines limitations du perceptron :



L'algorithme d'apprentissage

L'apprentissage se fait par descente de gradient et rétropropagation

- Descente de gradient : algorithme itératif, un gradient de la fonction de perte est calculé à chaque prédiction sur l'échantillon d'entraînement, et les paramètres sont légèrement modifiés dans la direction de diminution de la perte. Il n'y a pas de garantie d'atteindre un optimum global
- Rétropropagation : à partir du gradient en sortie du réseau, on calcule de proche en proche le gradient pour chaque neurone en remontant vers l'entrée.
- Le RN peut se voir comme une fonction composée de ses paramètres. Pour que l'apprentissage fonctionne, la fonction doit être (presque partout) dérivable.

Un grand nombre d'hyperparamètres

Un très grand nombre d'hyperparamètres à sélectionner

- Fonction de perte
- Structure du RN, nombre de couches, nombre de neurones à chaque couche
- Fonction d'activation (éventuellement différentes selon les couches)
- Optimiseur (taux d'apprentissage et sa mise à jour, inertie, etc.)
- etc.

Cette sélection peut se faire par exploration et cross-validation, ou s'appuyer sur les pratiques usuelles :

Type de problème	Activation	fonction de perte
classification binaire	sigmoid	binary_crossentropy
classification multiclasse	softmax	categorical_crossentropy
regression (dans $\Re$)	linéaire	mse

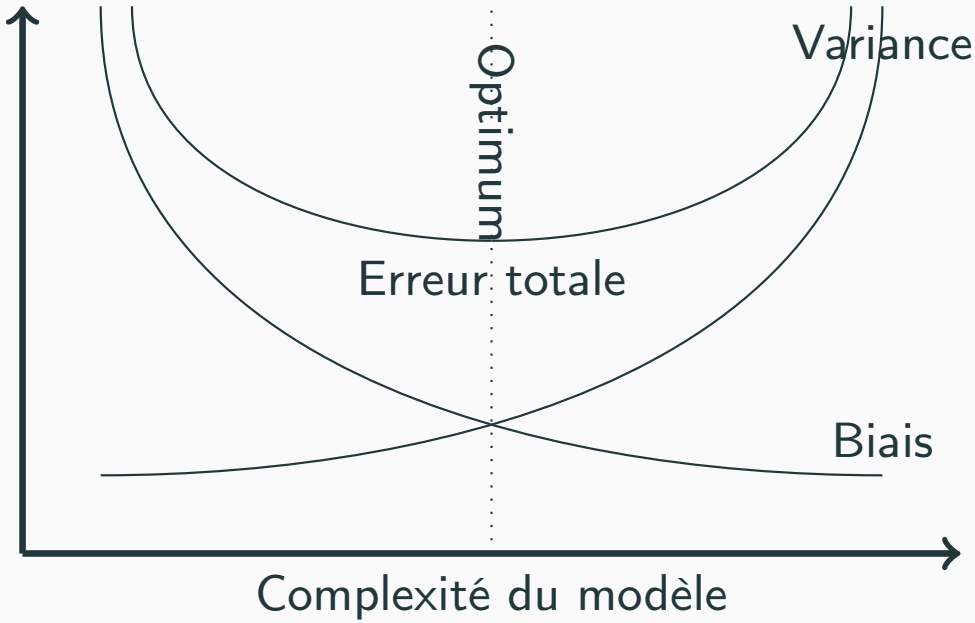
Figure 1: Choix de l'activation de la dernière couche et de la fonction de perte

Insee
Mesurer pour comprendre

Un grand nombre de paramètres à choisir

Stratégie pour choisir l'architecture du réseau :

1. Ajouter des couches et des neurones jusqu'à ce que le modèle sur-apprenne
2. Traiter le sur-apprentissage (simplifier le réseau, ajout de *dropout*, régularisation)



Il est également possible de tester différentes combinaisons de paramètres notamment à l'aide de librairies dédiées telles que *HyperOpt* etc.

Insee
Mesurer pour comprendre

Les outils

Les librairies les plus populaires sont sans doute *Tensorflow* (via *Keras*) et *pytorch*. Le logiciel python contient les fonctionnalités les plus avancées mais les librairies sont accessibles en R.

A l'Insee, ces environnements sont disponibles sous AUS V3. La plateforme (datalab.sscloud.fr) permet de paralléliser les traitements et donne accès à de la GPU mais non homologuée pour les données confidentielles.

La mise à disposition des codes du document de travail est prévue.

Prédire pour imputer

ML pour la non-réponse partielle

L'imputation de la non-réponse partielle se fait sous l'hypothèse habituelle d'ignorabilité (MAR : missing at random, Rubin 1976)

$$\mathbb{P}(r_i = 1|x_i, y_i) = \mathbb{P}(r_i = 1|x_i)$$

Cette hypothèse ne concerne pas le modèle d'imputation, et ne suffit pas à compenser un modèle mal spécifié. Le ML est un bon candidat ici :

1. Un modèle d'imputation n'a pas pour objectif d'être interprétable
2. Problème "de type $\hat{y}$" bien posé (Mullainathan et Spiess, 2017)
3. Les exploitations économétriques sur données imputées souffriront moins de l'imputation si le modèle est bien spécifié et riche par rapport au modèle économétrique

Insee
Mesurer pour comprendre

L'imputation du salaire

Dans l'enquête Emploi, questions sur le salaire depuis 1983 (par tranche) et 1989 (en clair). On reconstruit un fichier 1993 - 2018 avec environ 1,2 millions de salaires.

Question avec le plus fort taux de non-réponse : 4% ne répondent pas du tout, 15% répondent seulement par tranche.

Variable EEC qui n'est pas utilisée pour les statistiques sur le salaire, mais très utiles pour les exploitations économétriques de l'EEC

Imputation actuelle : variable SALRED par tirage aléatoire dans une distribution paramétrée par estimations de 5 modèles linéaires du log-salaire (pour 5 sous-populations différentes), après exclusion des valeurs extrêmes, conditionnellement à la tranche quand elle est connue.

Insee
Mesurer pour comprendre

Comparaison de performances

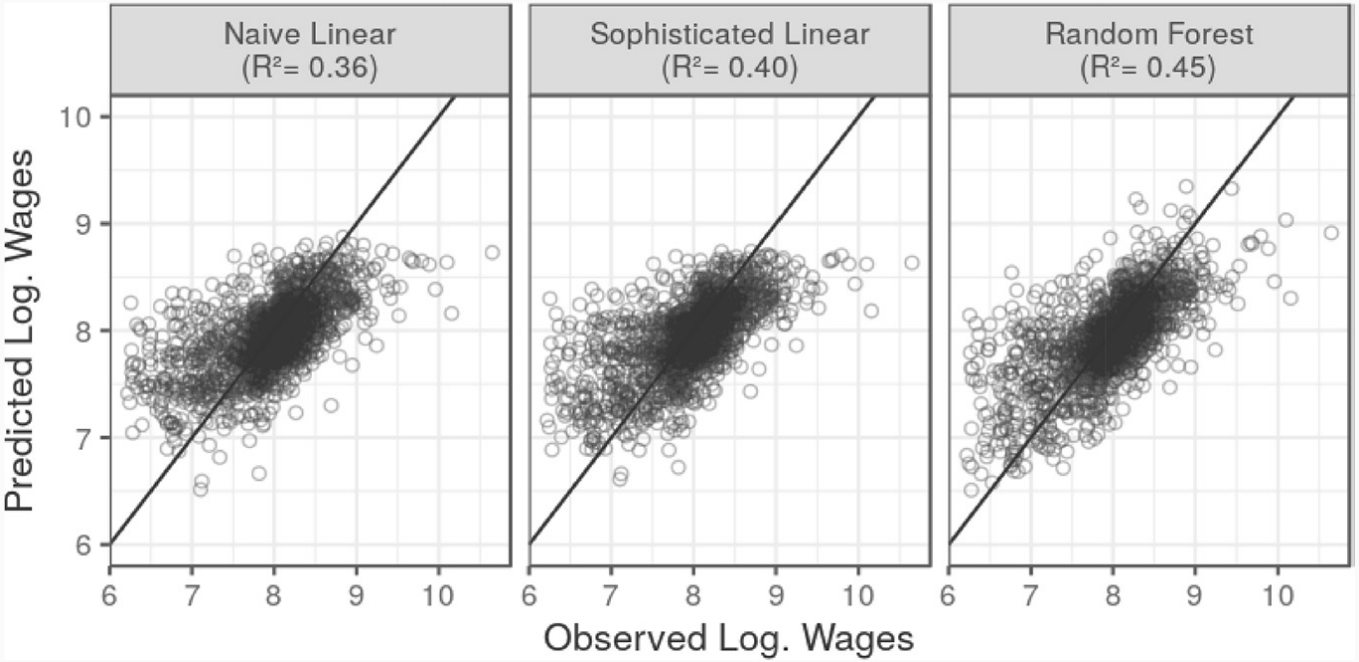
On prend comme métrique des performance le R^2, ou réduction de l'erreur quadratique moyenne par rapport à la variance de l'échantillon test. Sur le log-salaire, le RN réduit l'erreur quadratique résiduelle de 10 à 20% par rapport aux modèles linéaires.

Table 1: R^2 des modèles sur différents échantillons tests

	Echantillon test			
	log-salaire		salaire	
Modèle :	2018	Nés en mars	2018	Nés en mars
RN	68,3	71,9	33,0	27,6
OLS sur l'input RN	64,8	62,8	29,1	23,4
Mincer enrichi	63,4	62,9	28,7	27,6
Mincer secteur détaillé	63,7	63,3	29,1	27,9
Mincer basique	51,9	45,0	21,5	20,5
SALRED	43,2		18,6	

Insee
Mesurer pour comprendre

Prédire le salaire : un problème classique et difficile



Boelaert, Julien, and Étienne Ollion. "The Great Regression." Revue française de sociologie 59.3 (2018) : 475-506.

Insee
Mesurer pour comprendre

Réseau pour l'imputation du salaire

Le RN est implémenté avec Keras et Tensorflow sous python, et entraîné sur AUS, les serveurs de calcul de l'Insee. Sans parallélisation, chaque époque dure quelques dizaines de secondes à quelques minutes.

Type de couche	**Dimensions**
Vecteur en entrée	214
Dense + SeLu	500
Dropout	0.5
Dense + SeLu	100
Dense + SeLu	50
Dense + SeLu	20
Dense + SeLu	8
Dense + SeLu	4
Dense + SeLu	2
Dense + Linear	1

Loss : MSE sur le log-salaire

Variables quantitatives : valeurs manquantes imputées par leur minimum, et variable indicatrice de valeur manquante

Variables qualitatives de moins de 20 modalités : "one-hot encoding" (valeurs manquantes encodées comme une modalité)

Variables qualitatives de grande cardinalité : encodage via word2vec
L'input est normalisé (scikit learn scaler ou batch-normalisation)

Insee
Mesurer pour comprendre

La vectorisation des nomenclatures

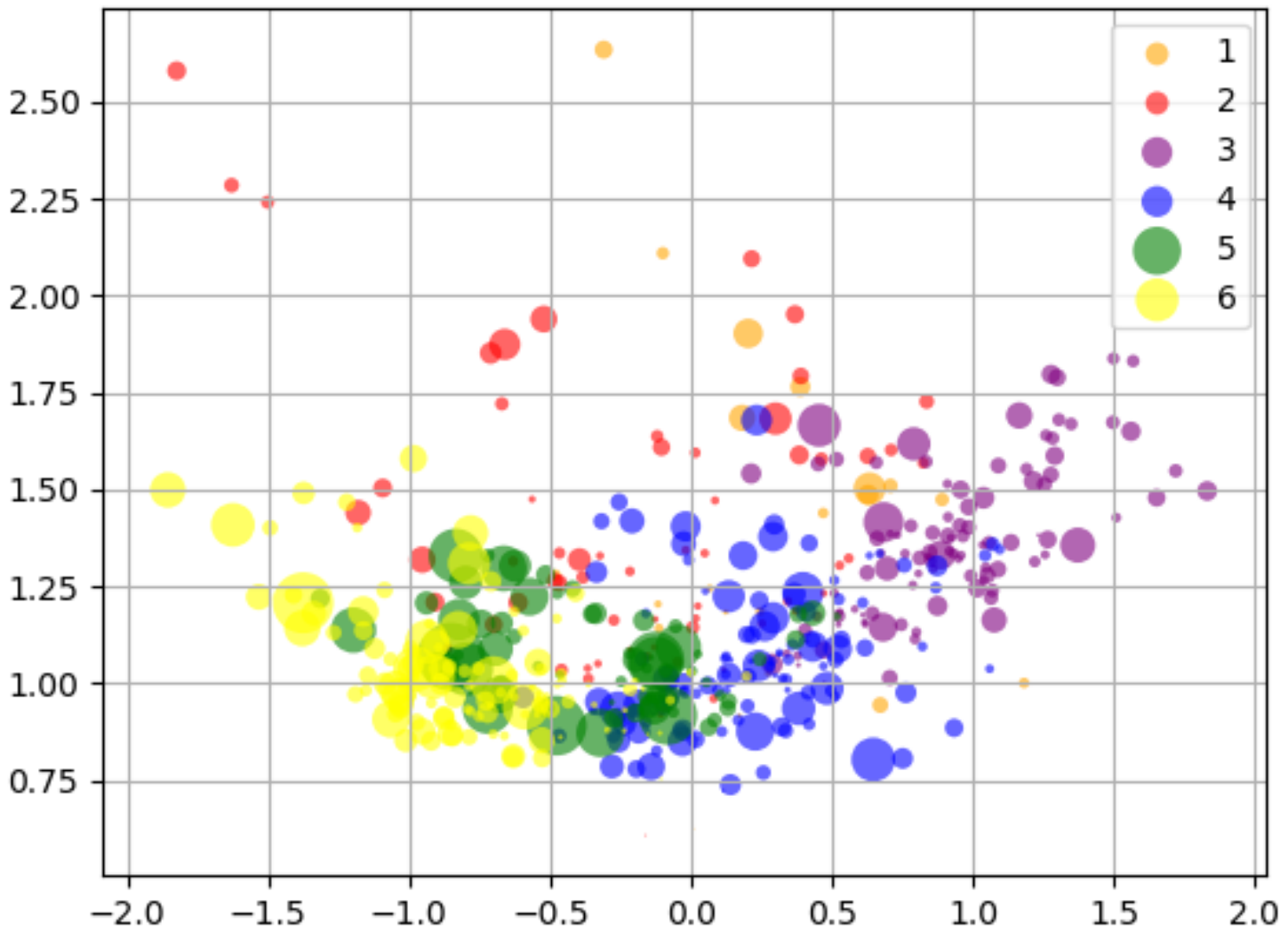
Les variables qualitatives de grande cardinalité posent problème au RN comme à l'OLS : le fichier est trop lourd.

Pour les exploiter malgré tout, on utilise l'algorithme Word2Vec (Mikolov et al. 2013), implémenté de manière performante dans la package Gensim notamment. Il sert à encoder les mots d'un lexique en un petit nombre de dimensions numériques, à partir d'un corpus de phrases.

L'idée consiste à traiter les proximités entre observations (réinterrogations, individus du même ménage, etc.) comme des "phrases" ou des contextes, et les modalités des variables comme les mots d'un lexique. Le résultat est comparable à une ACP non-linéaire.

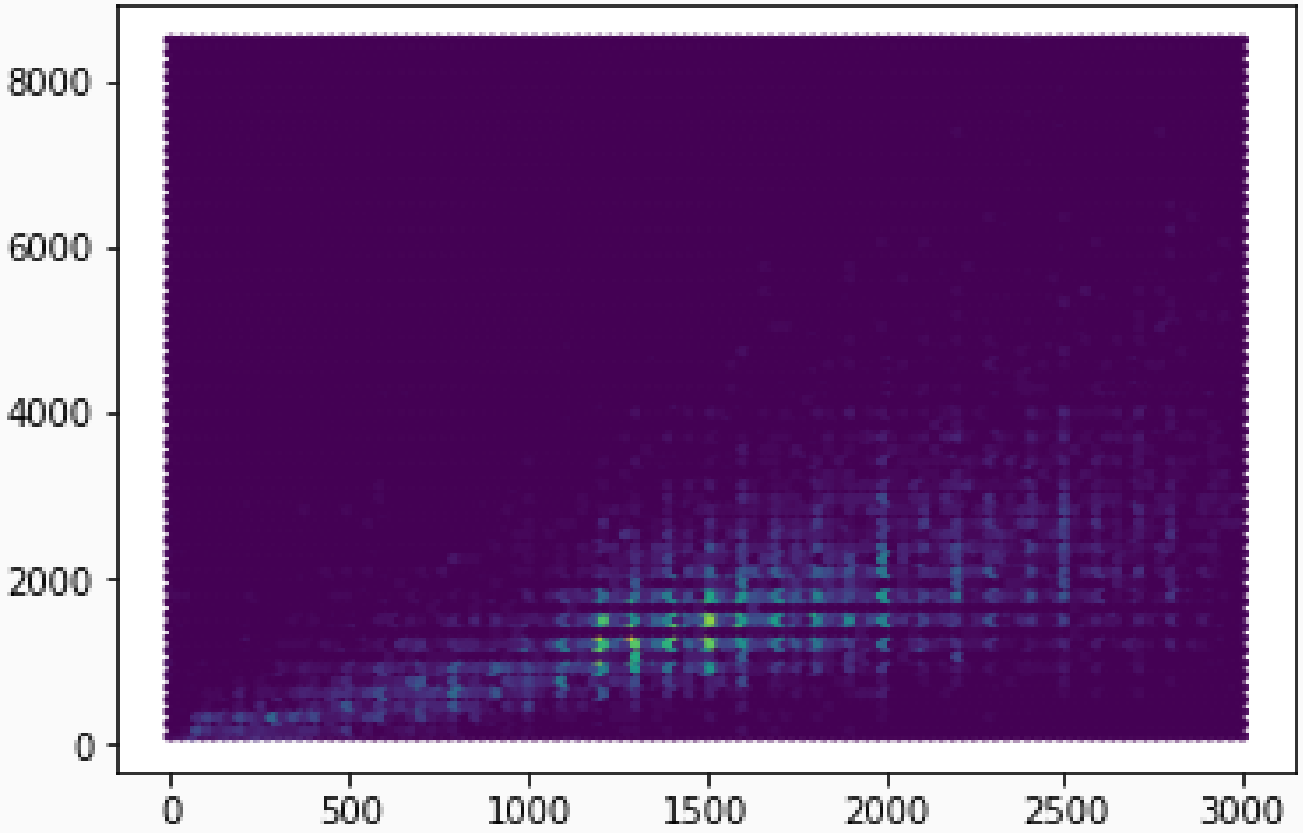
Cette méthode, qui peut s'appliquer à d'autres types de variables catégorielles, présente un intérêt en elle-même (Voir Doutreligne, Leduc, Nguyen, Vuagnat, 2020 pour une application aux données de santé).

Exemple : l'encodage des professions



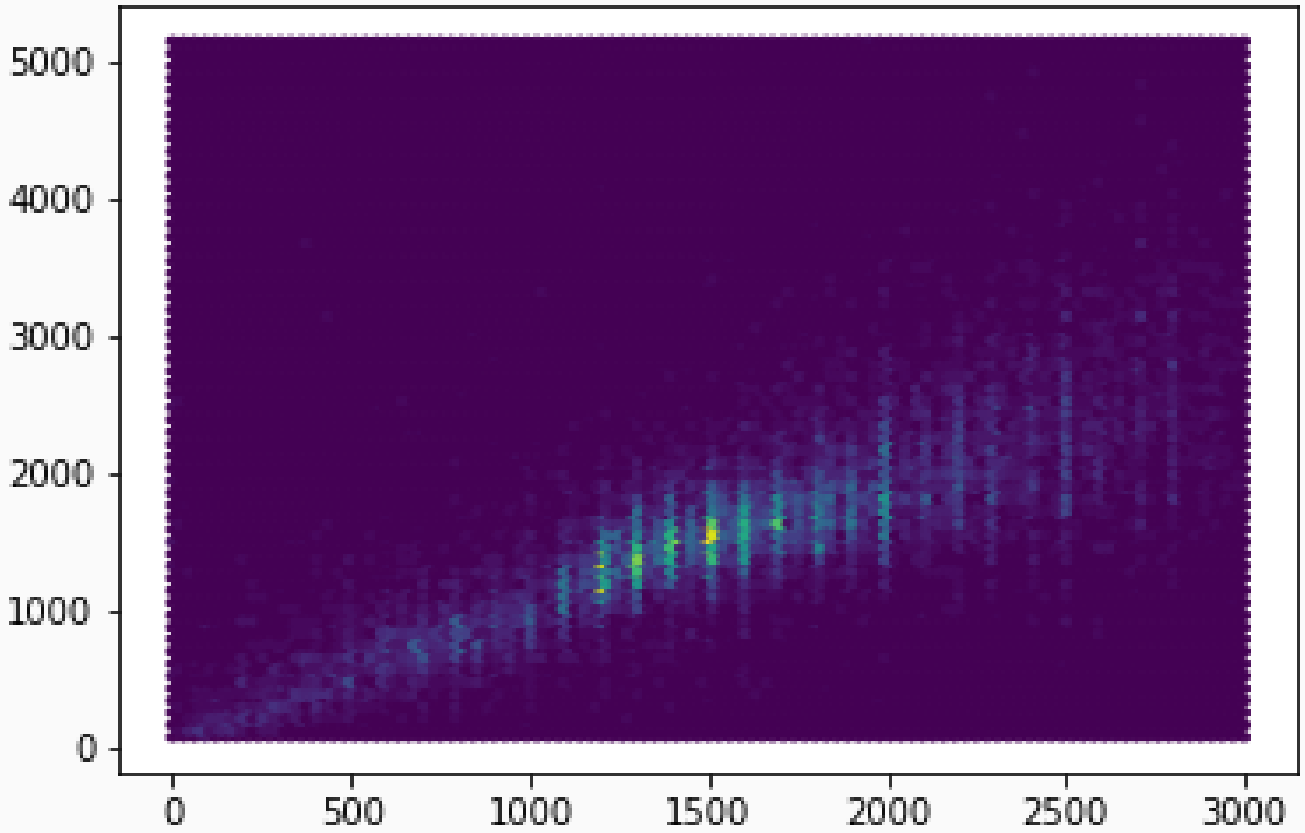
Insee
Mesurer pour comprendre

Performances de l'imputation actuelle



$R^2 = 0.186$

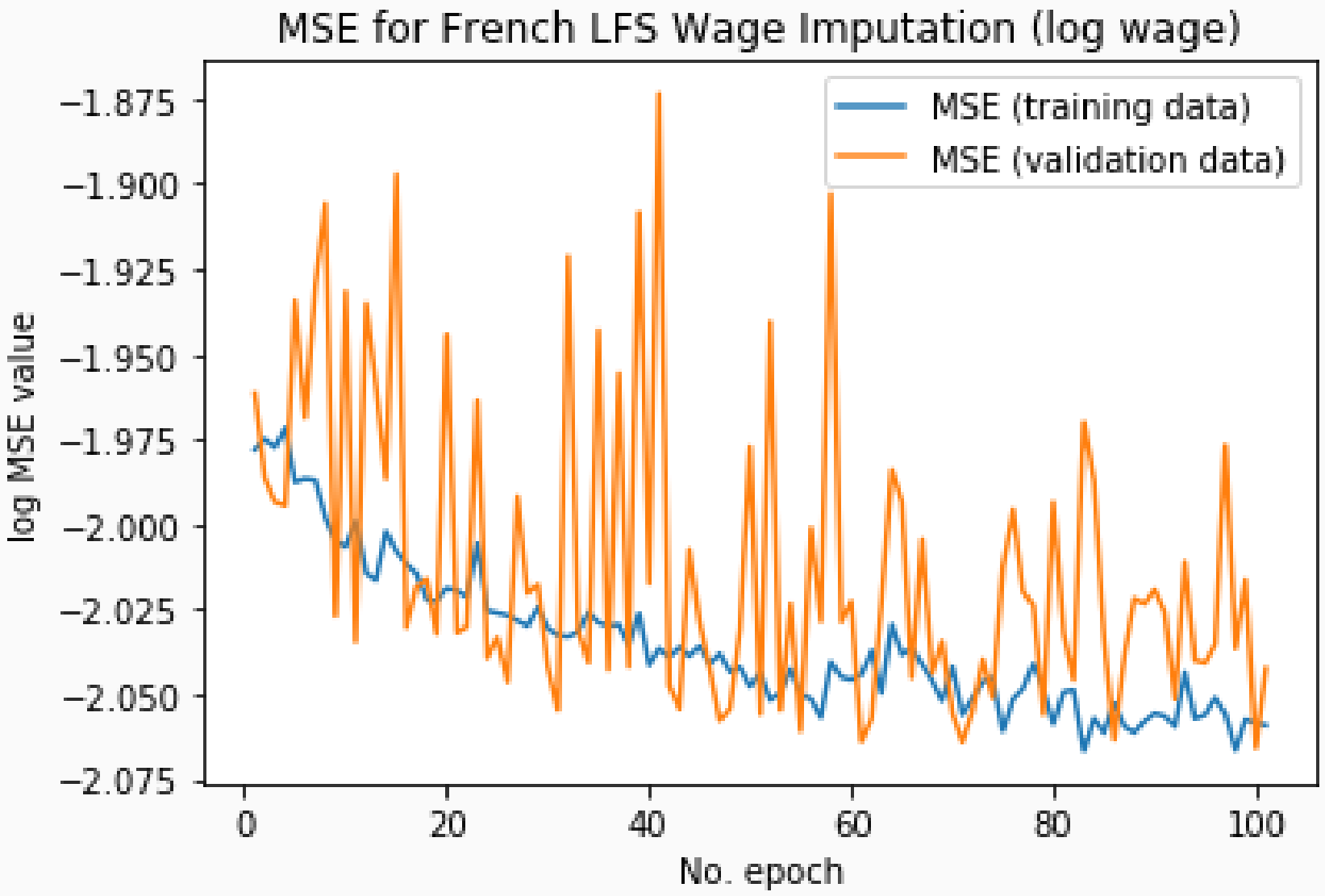
Performances de l'imputation par RN



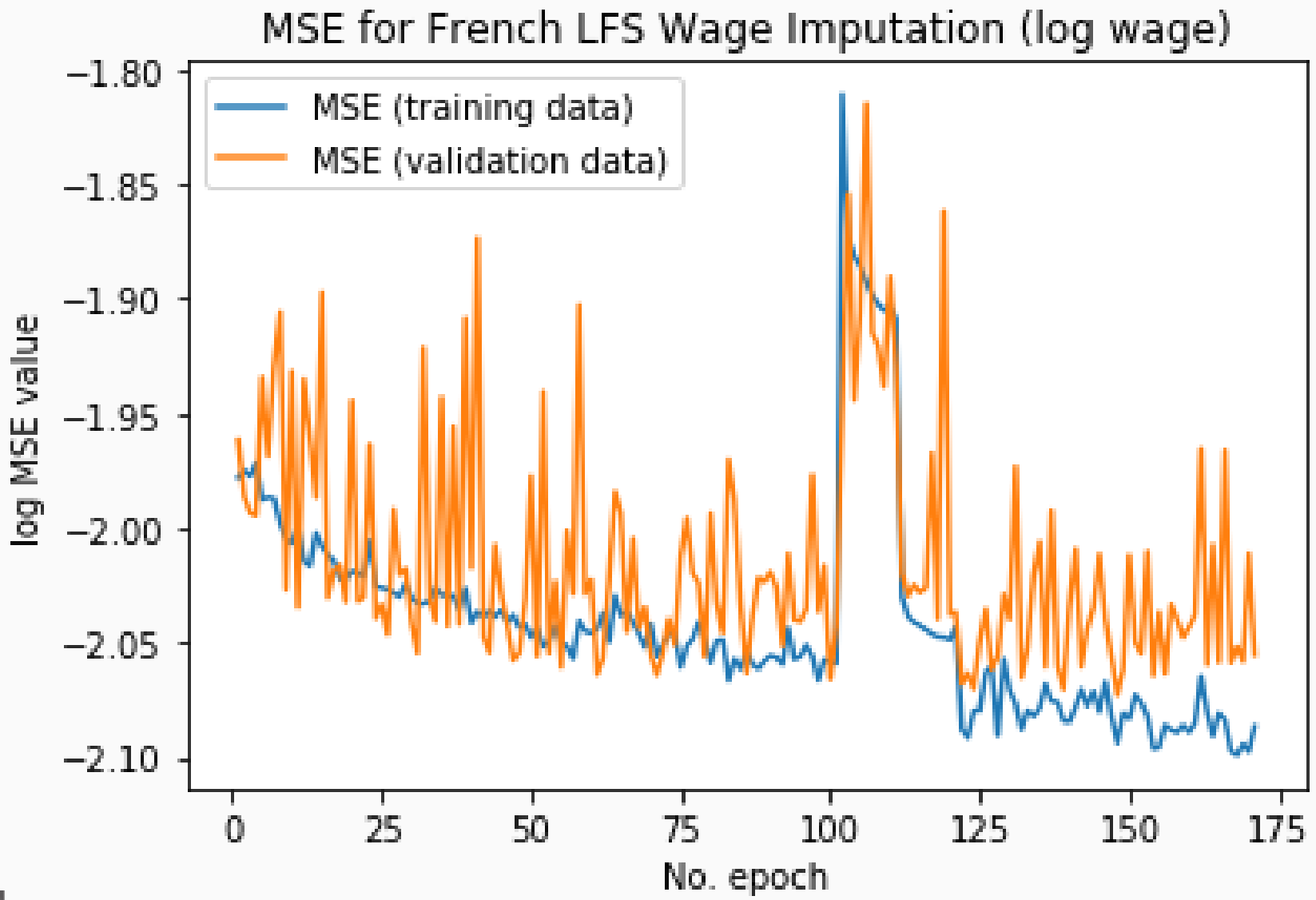
$R^2 = 0.381$ sur le premier trimestre 2018

Insee
Mesurer pour comprendre

Stabilité de l'apprentissage

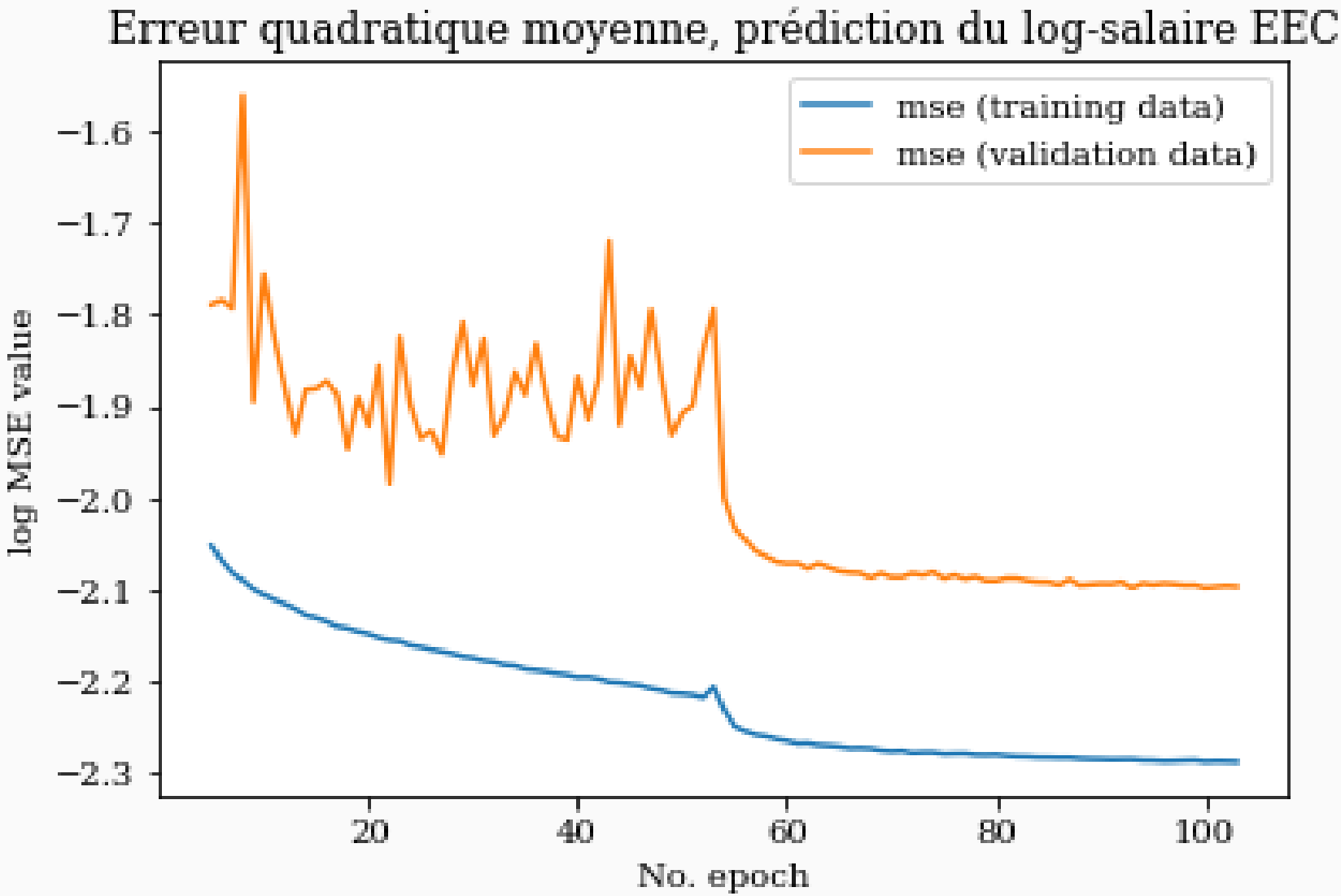


Stabilité de l'apprentissage



Insee
Mesurer pour comprendre

Stabilité de l'apprentissage



Insee
Mesurer pour comprendre

Stabilité de l'apprentissage

Quelques éléments heuristiques sur la stabilité de l'apprentissage

- Les grands batchs sont plus stables, les petits batchs permettent d'apprendre en moins d'époques, alterner les taille de batch semble intéressant.
- Autre fonctions de perte (mse sur le salaire, etc.) : il est facile de générer de l'instabilité !
- le dropout est un facteur d'instabilité
- On peut "aider" le RN à trouver le modèle linéaire, en ajoutant les inputs en entrée du neurone de sortie (en plus des autres entrées générées par tout le reste du réseau).

Une réduction du biais de non-réponse ?

On peut classer les salaires observés et prédits selon les tranches de l'enquête. On peut faire de même en cas de non-réponse sur le salaire, lorsque les enquêtés ont tout de même donné leur tranche de salaire

Figure 2: Distribution du salaire prédit par tranche, par tranche de salaire déclaré

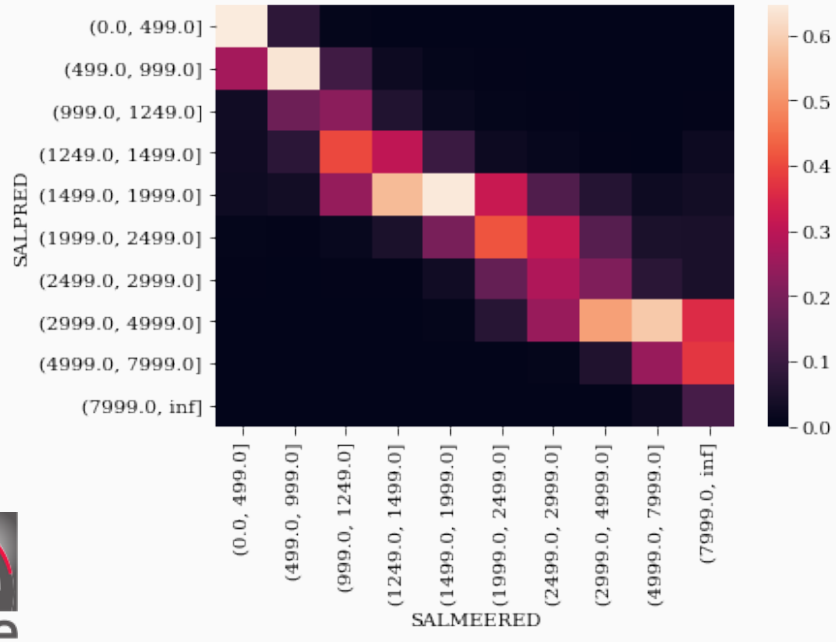
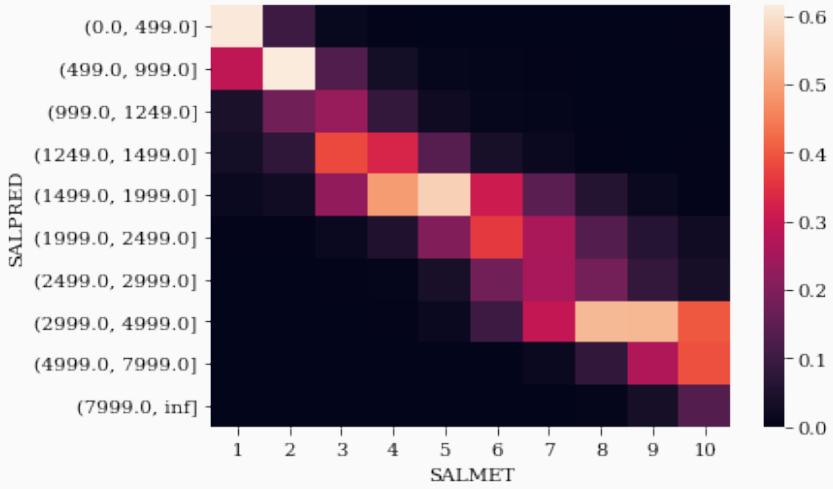


Figure 3: Distribution du salaire prédit par tranche, par tranche déclarée, pour les non-répondants sur le salaire

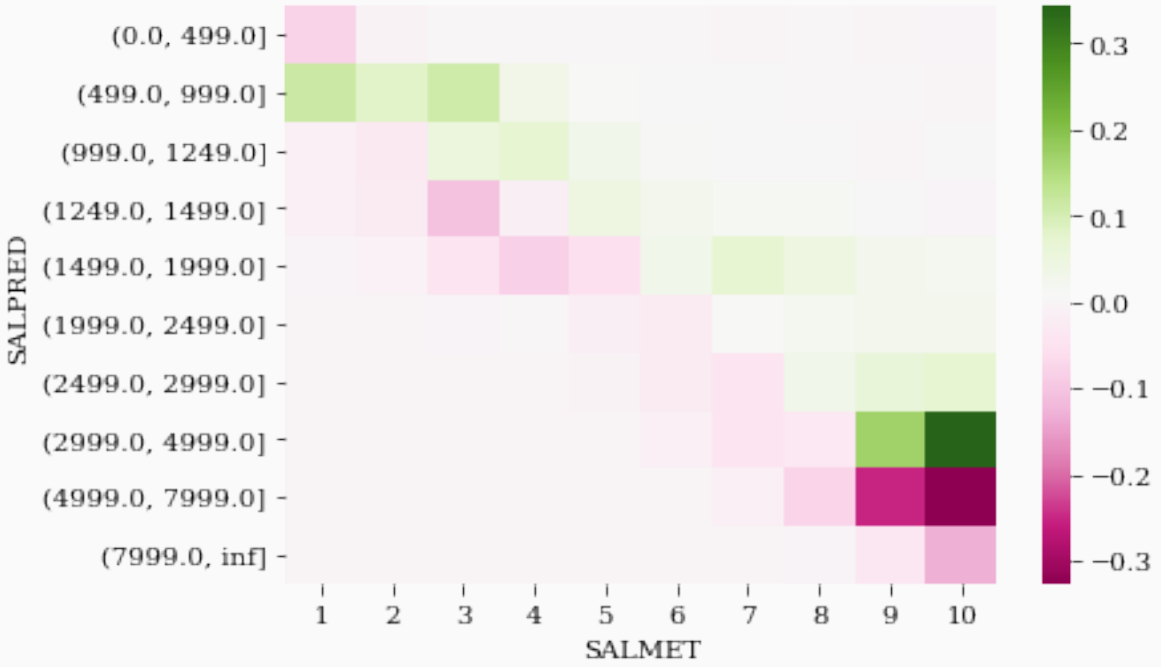


Insee
Mesurer pour comprendre

Une réduction du biais de non-réponse ?

On fait la même analyse pour un prédicteur linéaire, on regarde la différence

Figure 4: Différence OLS - RN des distribution des prédictions par tranche, pour les non-répondants



Atouts et limites

- Prédiction plus précise
- Possible réduction du biais
- Tirera meilleur parti de la dimension des données
- Coûteux en travail et en temps de calcul
- Adaptation plus compliquée aux évolutions de l'enquête

Insee
Mesurer pour comprendre

Discussion

Des prolongements possibles :

- Exploiter les réponses par tranches (pour enrichir l'entraînement, et prédire un salaire dans une tranche connue)
- Modèles doublement robustes (prédicteur de salaire couplé à un prédicteur de non réponse)
- Première étape pour une imputation par hot-deck
- Intégration dans un traitement plus générale de l'enquête (encodage de tout le fichier, traitement global de la non-réponse, des valeurs manquantes et de la correction d'erreur, etc.)
- Exploitations économétriques (notamment pour modéliser des facteurs de nuisance)

Des questions ouvertes :

- Des enjeux éthiques nouveaux ? Représentativité, causalité et confidentialité
- Quel coût de maintenance en cas de mise en production ?

Insee
Mesurer pour comprendre