Data challenge CDC – données carrière

# Description du sujet

Parmi les régimes de retraite dont la Caisse des Dépôts a la gestion, la CNRACL se distingue comme étant le plus gros régime de retraite de base, en termes d’effectifs de cotisants et pensionnés comme de masses de pension versées. Les enjeux de pilotage de ce régime sont donc particulièrement importants. Les prévisions et autres projections du régime participent à ce pilotage, puisqu’elles permettent d’identifier les enjeux, mais également d’étudier les répercussions de réformes passées ou de mesures à venir. Ces projections s’appuient majoritairement sur les données de gestion du régime (informations administratives sur les personnes, données de carrière, …) transmises lors des déclarations des employeurs. D’autres sources de données peuvent toutefois venir compléter ces informations.

La qualité des données transmises, récupérées et traitées est donc un enjeu particulièrement important pour la direction des études et des statistiques, en charge de la réalisation des projections. Par ailleurs, le modèle de projection du régime de la CNRACL est un modèle de microsimulation – nommé Canopée – qui va simuler les trajectoires individuelles et non à un niveau agrégé. A ce titre, il est important de pouvoir disposer, pour chaque personne présente dans le modèle (individu présent dans les effectifs initiaux (cotisants et pensionnés) ou projeté), de ses caractéristiques. Ainsi, pour chacune des variables retenues, il faudra disposer de l’information souhaitée, ou, à défaut, pouvoir lui en assigner une à l’aide d’une complétion ou d’une estimation la plus précise possible se basant sur son historique ou sur l’étude des autres individus.

La CNRACL gère les agents affiliés à la fonction publique territoriale (FPT) ainsi que ceux affiliés à la fonction publique hospitalière (FPH). D’autres fonctionnaires, en détachement dans l’une de ces deux entités, mais issus de la fonction publique d’Etat (FPE) ou d’autres administrations peuvent également être présents dans les effectifs et donc dans les données.

Chaque fonctionnaire appartient à un corps ou un cadre d'emplois classé en catégorie A, B ou C. Dans ce corps ou cadre d'emplois, il est titulaire d'un grade. Les filières regroupent quant à elles les cadres d’emplois / corps d’un même secteur d’activité.

Un exemple :

**M. Dupont**, né le **01/01/1970**, travaille dans la commune de **Talence**. Il est fonctionnaire territorial.

Profession : Eboueur

Indice brut de rémunération : **374**

Echelon : 5

Code grade : **TTH1**

Libellé du grade : **Adjoint technique**

Cadre/corps : **Agents techniques territoriaux**

Filière : **Technique**

Fonction publique : **FPT**

Parmi les informations ci-dessus, toutes celles en vert sont transmises dans les déclarations de son employeur au régime de retraite. Toutefois, si certaines données sont (en général) bien normées comme les noms, date de naissance, … ou soumises à des contrôles limitant leur saisie, ce n’est pas le cas de toute. En outre, l’évolution des bases de données induit une non-homogénéité dans la disponibilité de toutes les variables dans le temps. Ainsi, le code permettant d’identifier le grade n’a été mis en place qu’en 2011 et n’est donc jamais renseigné pour les périodes antérieures. Or, justement, la connaissance du grade est déterminante dans le modèle Canopée, puisque chaque individu devra être replacé, sur la dernière période précédant la projection, sur un grade donné, afin que la carrière puisse être simulée. L’autre information permettant d’identifier le grade étant le libellé, qui est en saisie libre donc non normé.

Enfin, les données réelles comportent toujours des données manquantes.

L’enjeu ici va être de pouvoir **affecter pour chaque période et chaque individu, un grade donné**, notamment quand le code grade est manquant. Il faudra également fournir des éléments d’explication de la manière dont on contrôle la qualité de la complétion effectuée.

# Problématiques

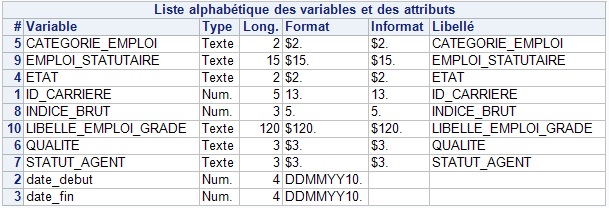
* Les grades évoluent au fil du temps : code, libellé
* Les grilles de rémunération évoluent au fil du temps, et donc les IB associés.
* Les agents évoluent durant leur carrière (échelon, mais aussi promotions, …)
* Les IB se chevauchent entre grades hiérarchisés d’un même cadre d’emplois, selon l’échelon.
* Il peut y avoir présence d’un code et d’un libellé avec information contradictoire. Idem avec l’indice brut qui peut ne pas correspondre => NB dans ce cas on privilégiera toujours le code.
* Pas de codes dans les données avant 2011

# Données à disposition pour ce challenge

* Nomenclature des emplois territoriaux (NET) : fait la correspondance entre le code et le libellé du grade « normé ».
  + *N4DS-NET-V01X16.csv*
  + *NET 2021\_avec\_codification\_v2.pdf* vient en complément pour expliquer la construction du code NET.
* Nomenclature des emplois hospitaliers (NEH)
  + *nehmed\_29032022.xlsx*
  + *2022\_10\_06\_nehpnm2022.xls*
* Nomenclature des emplois grade (NEG)
  + *NEG\_ta\_element.csv*
  + *NEG\_ta\_sup\_remplaces\_elements\_remplacement.csv*
* Base de données Hackhaton\_Full.sas7bdat (150 125 450 lignes et 10 colonnes)

# Dictionnaire des variables

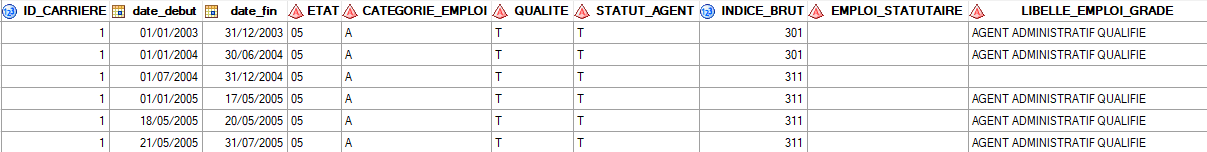
*Hackhaton\_Full\_v3.sas7bdat* ou *hackhaton\_100k\_alea\_b.csv* ou *hackhaton\_020k\_alea\_b.csv*



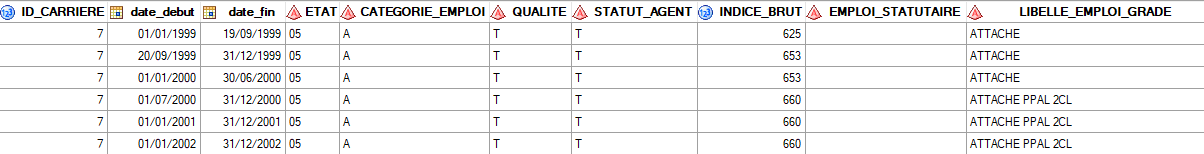
* *Categorie\_emploi* = catégorie d’emploi exercé par l’agent (voir annexe 3)
* *Emploi\_statutaire* = code grade de l’agent (rempli depuis 2011 seulement). Voir nomenclatures NET et NEH
* *Etat* = code état de la période (voir annexe 3)
* *ID\_carriere* = numéro d’identification d’un agent
* *Indice\_brut* = indice brut de rémunération d’un agent, correspondant à sa grille (voir NEG)
* *Libelle\_emploi\_grade* = libellé du grade de l’agent
* *Qualite* = statut juridique liant l’agent à son employeur pendant la période (voir annexe 3)
* *Statut\_agent* = fonction publique où exerce l’agent (voir annexe 3)
* *Date\_debut* = date de début de la période sur laquelle les autres informations sont valides
* *Date\_fin* = date de fin de la période sur laquelle les autres informations sont valides

**Annexe 1 : exemples de complétion possibles**

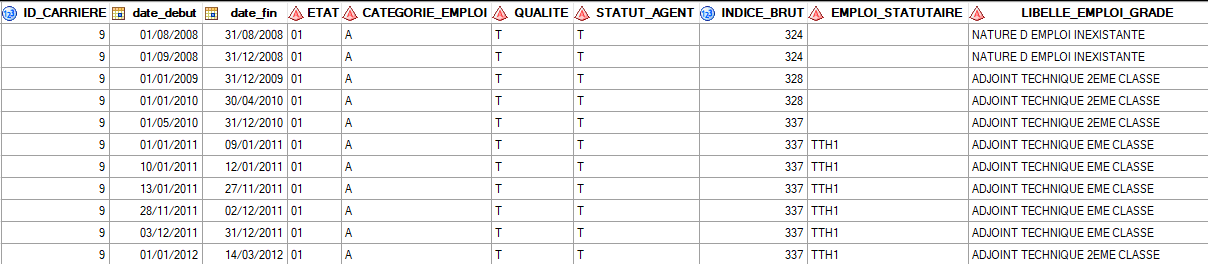
1. Complétion d’une modalité manquante sur une période quand les deux périodes l’entourant ont la même modalité



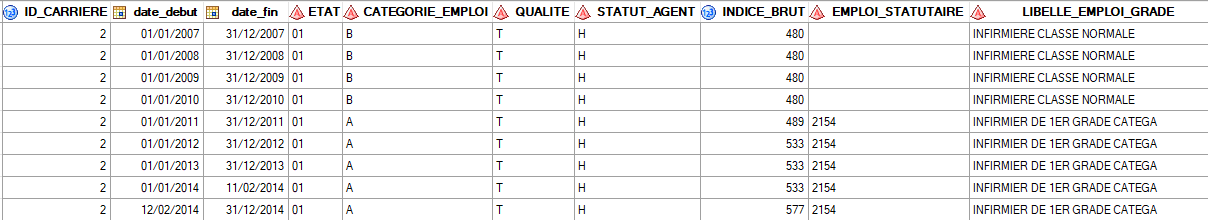
1. Complétion d’un code à partir d’un libellé assez bien normé et de l’indice brut qu’on doit pouvoir retrouver dans la grille(NEG) de la période



1. Complétion de périodes dont la modalité du libellé est similaire à certaines périodes suivantes qui ont la variable code renseignée

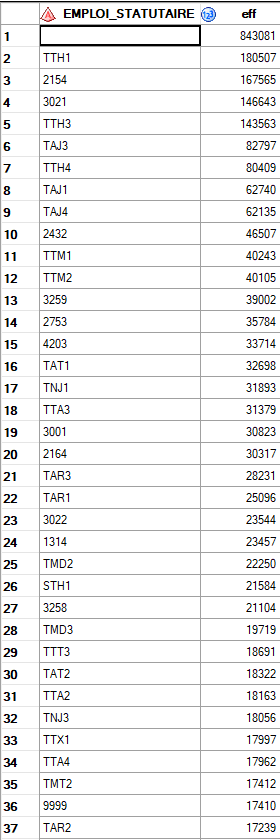


1. Complétion d’une information quand on connaît le code de la période suivante et que la modalité du libellé se retrouve dans la NEG (cas compliqué car changement de code et de libellé même si cohérence de l’emploi)



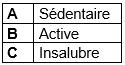
**Annexe 2 : statistiques par grade (code) de la dernière période de chaque individu**

(par ordre décroissant, 37 lignes avec les effectifs les plus élevés seulement, fait sur la table exhaustive)



**Annexe 3 : modalités des variables**

* **CATEGORIE\_EMPLOI**



* **ETAT**



NB : les codes ‘02’, ‘03’ et ‘20’ ont été écartés de la base en amont.

* **QUALITE**



* **STATUT\_AGENT**

