Traces 01

Collectif open data UP13

March 15, 2017

Traces 01 les jeux de données

Description du jeu avant anonymisation

L'université Paris 13 a enregistré dans son système d'information (dans Apogée), les données sur l'inscription des étudiant · e · s pour chaque année universitaire entre 2006(-2007) et 2015(-2016). Ces données portent sur les diplômes préparés, les étapes pour y parvenir, les composantes (UFR, IUT, etc.) concernées. Plus précisément chaque donnée occupe une ligne dont les colonnes sont les suivantes.

Colonne	Explication
CODE_INDIVIDU	Donnée masquée
ANNEE_INSCRIPTION	Année d'inscription : 2006 pour 2006-2007, etc.
LIB_DIPLOME	Nom du diplôme
NIVEAU_DANS_LE_DIPLOME	1, 2,... pour master 1, licence 2, etc.
NIVEAU_APRES_BAC	1, 2,... pour Bac+1, Bac+2,...
LIBELLE_DISCIPLINE_DIPLOME	Rattachement du diplôme à une discipline
CODE_SISE_DIPLOME	Code du système d'information sur le suivi de l'étudiant
CODE_ETAPE	Code interne d'une étape (année, parcours) de diplôme
LIBELLE_COURT_ETAPE	Nom court de l'étape
LIBELLE_LONG_ETAPE	Nom plus intelligible de l'étape
LIBELLE_COURT_COMPOSANTE	Nom de la composante (UFR, IUT etc.)
CODE_COMPOSANTE	Code numérique de la composante (inutilisé)
REGROUPEMENT_BAC	Type de Bac (L, ES, S, techno STMG, techno ST2S,...)
LIBELLE_ACADEMIE_BAC	Académie du Bac (Créteil, Versailles, étranger,...)
CONTINENT	Déduit de la nationalité (donnée masquée)

Colonne	Explication
LIBELLE_REGIME	Formation initiale, continue, pro, apprentissage
NIEME_INSCRIPTION	Nombre d'inscriptions à Paris 13

D'autres données telle que l'adresse des étudiant · e · s sont en cours d'anonymisation pour une prochaine ouverture.

Toilettage du jeu de donné initial

Le jeu de données initial contient des données *singulières* au sens ou des valeurs apparaissent moins de 5 fois en tout et pour tout. Pour le dire autrement, il y a des colonnes pour lesquelles des lignes prennent des valeurs rares. On commence par trouver toutes ces lignes, puis on les supprime, et on recommence jusqu'à avoir une donnée sans valeurs rares.

Le nombre initial de lignes, dans la donnée brute était de 213 289. En une première passe d'anonymisation voici le nom des colonnes et le nombre de valeurs rares (moins de 5 occurences) par colonne :

Colonne	valeurs rares
LIB_DIPLOME	1
LIBELLE_DISCIPLINE_DIPLOME	27
CODE_SISE_DIPLOME	41
CODE_ETAPE	319
LIBELLE_COURT_ETAPE	299
LIBELLE_LONG_ETAPE	362
NIEME_INSCRIPTION	3

Après suppression des lignes contenant des valeurs rares, une seconde passe donne :

Colonne	nombre de valeurs rares
CODE_ETAPE	2
LIBELLE_COURT_ETAPE	1
LIBELLE_LONG_ETAPE	1

Une dernière passe confirme que la donnée est 5-anonymisée par cellule au sens où, pour chaque colonne, chaque valeur apparait au moins 5 fois.

Le nombre de lignes supprimées est de 795, partant de 213 289, il reste **212 494 lignes** concernant **105 747 étudiant · e · s**. La répartition des lignes supprimées par année n'est pas homogène, comme le montre le tableau `up13_perte.csv`.

Année	Donnée brute	Donnée anonyme	Perte
2006	20 040	19 995	45
2007	19 914	19 868	46
2008	19 897	19 856	41
2009	21 277	21 253	24
2010	21 022	20 972	50
2011	21 497	21 444	53
2012	22 355	22 292	63
2013	22 223	22 127	96
2014	22 423	22 274	149
2015	22 641	22 413	228

Toutes les données suivantes sont produites en partant du tableau obtenu après toilettage.

Projection et k-anonymisation

Quatre jeux de données sont produits à l'aide d'une méthode d'anonymisation par suppression des lignes trop singulières. On sélectionne un certain nombre de colonnes du tableau et on supprime les lignes qui ne sont pas répétées au moins 5 fois. C'est un compromis entre la possibilité de croiser des données et l'exhaustivité. En effet, plus il y a de colonnes plus il faut supprimer de lignes car les lignes sont de plus en plus spécifiques.

Un tout premier jeu de données `up13_anonyme.csv` fait le choix de conserver toutes les colonnes sauf l'identifiant de l'individu. On perd ainsi un maximum de lignes au moment de la 5-anonymisation : 129 742 lignes doivent être supprimées (soit 61% de la donnée initiale).

Pour les trois autres jeux de données on a choisi de ne pas aller au delà de 5% de pertes dans les lignes du tableau (en partant de la donnée avant toilettage initial, donc en tenant compte des 795 lignes déjà perdues). Pour cela, on a choisi

une ou deux colonnes particulières et on a étendu ce nombre de colonnes, en choisissant systématiquement la colonne suivante comme occasionnant le moins de nouvelles pertes.

Le tableau `up13_etapes.csv` concerne les étapes de diplôme, il contient les colonnes "CODE_ETAPE", "LIBELLE_COURT_ETAPE", "LIBELLE_LONG_ETAPE", "NIVEAU_APRES_BAC", "LIBELLE_COURT_COMPOSANTE", "LIB_DIPLOME", "LIBELLE_DISCIPLINE_DIPLOME", "CODE_SISE_DIPLOME", "NIVEAU_DANS_LE_DIPLOME" et son anonymisation occasionne une perte supplémentaire de seulement 130 lignes.

Le tableau `up13_Academie.csv` concerne l'Académie du Bac et il contient les colonnes "LIBELLE_ACADEMIE_BAC", "NIVEAU_APRES_BAC", "NIVEAU_DANS_LE_DIPLOME", "CONTINENT", "LIBELLE_REGIME", "LIB_DIPLOME", "LIBELLE_COURT_COMPOSANTE" et son anoymisation implique la perte supplémentaire de 6 737 (soit une perte totale de 7532 lignes c'est à dire 3,5% de la donnée initiale).

Le tableau `up13_Bac.csv` concerne le type de Bac et le niveau atteint après le Bac, il contient les colonnes "REGROUPEMENT_BAC", "NIVEAU_APRES_BAC", "LIBELLE_REGIME", "CONTINENT", "LIBELLE_COURT_COMPOSANTE", "LIB_DIPLOME", "NIVEAU_DANS_LE_DIPLOME" et son anonymisation occasionne la perte supplémentaire de 3 145 lignes, donc 3 940 au total soit moins de 2% de la donnée initiale.

D'autres tableaux extraits de la même donnée initiale et construits selon la même méthode d'anonymisation, peuvent être fourni sur demande (préciser les colonnes souhaitées).

Calcul des traces

Une trace s'obtient en suivant le parcours d'un individu dans l'Université via ses inscriptions successives et en oubliant les annnées auxquelles ont eu lieu ces inscriptions et les autres données sur l'individu (à l'exception de son type de Bac).

Deux types de traces sont déduites de la donnée. Les traces débutants par le type de Bac, dans le fichier `up13_traces_bac.csv` et les traces débutant immédiatement par la première inscription à l'université, dans le fichier `up13_traces.csv`.

Les traces identiques sont regroupées et dénombrées. Leur nombre est indiqué en première colonne et la trace occupe ensuite autant de colonnes qu'il y a eu d'années d'inscription par ordre croissant des années d'inscription (sans tenir compte de l'année de départ, des interruptions éventuelles, ou de la réussite au diplôme).

Par exemple, on trouve parmi les traces avec Bac la ligne suivante :

Nombre	Trace		
460	Bacs généraux ES	IUTSD.DUT.1.S1TC	IUTSD.DUT.2.S2TC

Elle signifie que sur les 10 années de 2006-2007 à 2015-2016, il y a eu 460 bachelier · e · s ES qui ont été inscrit · e · s à l'université Paris 13 en première année du DUT "S2TC" de l'IUT de Saint-Denis, puis en seconde année du même DUT sans autre inscription à l'université avant ou après (on ne sait pas s'il y a eu interruption entre les deux années de DUT, s'il a volonté de poursuite des études en 2016-2017, qui a réussi sa deuxième année et qui a échoué, on peut juste inférer une certaine réussite en première année).

Les traces singulières, c'est à dire concernant moins de 10 individus, sont publiées mais en mettant à 1 le nombre de personnes concernées (le nombre réel est entre 1 et 9).

Par exemple, la ligne suivante nous informe que entre un · e étudiant · e et 9 étudiant · e · s se sont inscrit · e · s en master 1 d'informatique puis en master 2 recherche programmation et logiciel sûr et uniquement ces deux inscriptions entre 2006 et 2015, en ayant une équivalence du Bac en poche.

Nombre	Trace		
1	Equivalences	IG.Master.4.G4INF	IG.Master Rec.5.G5PLS

La présence ou l'absence d'une ligne correspondante dans les autres tableaux peut révéler des informations supplémentaires. Par exemple, s'il y avait la même ligne avec 8 autres Bac dans ce fichier et une ligne équivalente avec un nombre à 1 (donc de moins de 9 individus) dans le fichier de traces sans les Bacs, on pourrait déduire qu'une seule personne correspond à ce parcours. On pourrait aussi découvrir que depuis quelques années ce master est *indiférencié* (recherche et pro), et donc réduire la période à laquelle ce parcours a eu lieu. Il est donc important de ne pas diminuer trop les seuils d'anomysation (k = 5 ou k = 10 selon les cas) et éviter de permettre trop de croisements de données.

Il se peut aussi tout simplement qu'une valeur décrivant une inscription (par exemple IG.Master Rec.5.G5PLS), soit singulière en tant que ligne dans ce qui serait le tableau des inscriptions, en particulier si l'on tient compte de la donnée sur le Bac. Il y a 6 665 lignes singulières dans le tableau des inscriptions avec Bac (de colonnes "REGROUPEMENT_BAC", "LIBELLE_COURT_COMPOSANTE", "LIB_DIPLOME", "NIVEAU_APRES_BAC", "CODE_ETAPE"), et seulement 6 si on oublie la colonne "REGROUPEMENT_BAC". Mais ces tableau ne peuvent pas être complètement dérivés des tableaux des traces, car on n'a pas informé le nombre exacte de traces rares.

En tenant compte du type de Bac, il a 25 725 traces différentes : - 42 460 étudiant · e · s laissent des traces singulières et produisent 23 849 traces différentes (par souci d'anonymat on comptabilise faussement une seule personne par trace de chaque sorte, il pourrait y en avoir jusqu'à 9) - et surtout **63 28 étudiant · e · s** laissent une trace parmi 1 876 traces différentes et l'anonymat étant respecté, on les dénombre précisément.

En ne tenant pas compte du Bac, on obtient 16 077 traces différentes dont : - 24 583 étudiant · e · s laissant une trace singulière parmi 14 493 traces différentes (on compte faussement une seule personne par trace) - et surtout **81 164 étudiant · e · s** laissent une trace parmi 1 584 traces différentes.