Collecte et analyse des données PACES

Muriel Ney – Septembre 2015

QUESTION

Questions de recherche :

1) existe-t-il des variables permettant de prédire l'évolution d'apprenant ? En particulier des piques (il monte puis il baisse ou il baisse puis il monte) ou des descentes (régulier puis il baisse), ou des montés (régulier puis il monte..). = scénario 3 Hubble
2) peut-on trouver des types d'évolution et trouver des indicateurs qui permettent de classer les étudiants avec cette typologie ? = scénario 2 Hubble
3) Comment tenir compte de ces variables (contexte, ,..) pour proposer des rétroactions adaptés (type, moment, ..) à chaque type d'étudiant ?

Quelle est l'évolution de l'activité des étudiants au cours de l'année :

L'activité sera mesurée par (1) la présence et les résultats aux tests QCM (séances de tutorat) et (2) la participation aux FLQ (formulation en ligne des questions).

On regardera aussi l'impact du résultat au premier concours (au s1) sur l'activité du s2.

--

Suites possibles :

1) On peut aussi faire des analyses TAL (traitement des langues naturelles) sur les questions posées par les étudiants et sortir de typologies de questions pour caractériser les questions posés.
2) Il y a d'autres variables : satisfaction des étudiants (voir PACES pour récupérer ces données), par questionnaire (s'ils ont été inscrits dans une prépa, temps de travail, travail en groupe, …).

DONNES ET VARIABLES

Variable de Temps : discrète en 12 dates, chaque date correspond au début d'une séquence (de 1 à 12)

Remarque : chaque étudiant ne passe pas les matières dans le même ordre à l'intérieur d'une séquence, et on attend un effet matière important. C'est pourquoi, le pas de temps est une séquence.

Variable ID étudiant : numéro UJF (8 chiffres) et numéro PACES (4 chiffres), avec un tableau de correspondance.

Variables d'intérêt (à expliquer) :

- Résultats aux QCM (les tests des 12 séquences, avec 4 ou 3 matières par séquence)

Il faut rendre comparable des résultats qui ne le sont pas (la longueur des tests et la difficulté des tests sont très différentes d'une matière à l'autre).

Laura Dupuis a proposé cette variable : Ecart de la note (ramenée à une note sur 20) à la moyenne de la promo (moyenne de tous les étudiants ayant passé ce test, dans cette matière et cette séquence). => position par rapport à la valeur centrale

Autre possibilité, pour tenir compte de la distribution des notes (et pas seulement de la moyenne) :

Présence de la note dans un quartile (ou un décile) de la distribution des notes autour de la médiane. => position à la distribution
Variable quantitative discrète, avec 4 (si quartile) ou 10 (si décile) valeurs.
Deux variables quantitatives : la valeur pour chaque matière et la moyenne (sur 3 ou 4 matières) des valeurs pour chaque séquence.

- Participation aux FLQ

Nombre de questions posées par un étudiant sur chaque matière normalisée par le nombre de questions posées par tous les étudiants dans cette matière dans cette séquence.
Deux variables quantitatives, avec des valeurs entre 0 et 1 : le ratio pour chaque matière et la moyenne des ratios (sur 3 ou 4 matières) pour chaque séquence.
comptage des occurrences de mots spécifiques => voir si un logiciel d'analyse lexicale (sphinx, sur R) pourrait aider à la création de ces variables

Nombre de votes = questions choisies par l'étudiant dans une matière, normalisé par le nombre de votes sur cette matière par tous les étudiants
Deux variables quantitatives, avec des valeurs entre 0 et 1 : le ratio pour chaque matière et la moyenne des ratios (sur 3 ou 4 matières) pour chaque séquence.

Variables explicatives :

- Le sexe

Hypothèse : les filles fournissent un effort plus régulier et les garçons travaillent plus vers la fin

Variable qualitative nominale : codée MME. et M.
Distribution des valeurs : 501 (32,5%) hommes, 1040 (67,5 %) femmes
Fichier : PACES_brut_inscription_etudiant_2012-2013

- Le bac : deux variables = le type de bac et la mention au bac

Hypothèse : ceux qui ont un bac S spécialité Math savent fournir un effort régulier, idem si mention TB

Variable qualitative nominale : S (spécialité du bac S pas toujours précisée), STL, STG, ES, agricole, étranger, DAEU (Diplôme d'Accès aux Etudes Universitaires)
Distribution des valeurs : 1378 (89,5%) bac S, …
Variable qualitative ordinale : S (sans mention), P (passable), AB (assez bien), B (bien), TB (très bien), Vide (non précisé)
Distribution des valeurs : 325 (21%) S, 71 (4,6%) P, 497 (32,2%) AB, 415 (26,9%) B, 216 (14%) TB
Fichier : PACES_brut_inscription_etudiant_2012-2013

- Le cursus choisis (le choix se fait à l'issue du concours du s1)

Hypothèse : on peut voir un comportement spécifique chez ceux qui ont choisi Maïeutique, Odon ou Pharma…

Variable qualitative nominale : combinaisons de M (médecine), O (odontologie = dentiste), P (pharmacie) et Sf (sage-femme = maïeutique)
Distribution des valeurs : 211 abandons et 442 M, 138 MP, 125 P, 118 MSf, 108 MO, 88 MOP, 82 MOPSf, 55 Sf, 52 MPSf, 17 PSf, 16 MOSf, 15 O, 8 OP, 5 OSf, 3 OPSf
Fichier : PACES_transforme_tutorat-concours 2012-2013, onglet : tutorat_notes

- La profession du parent chef de famille

Hypothèse : ceux qui sont mieux soutenu financièrement et intellectuellement, on un effort plus régulier, moins d'abandon.

Problème : on n'a pas la profession de l'autre parent, on ne peut que supposer que le chef de famille a la profession la plus importante sur les plans financier et intellectuel.

Variable qualitative nominale : CSP (Catégories socioprofessionnelles) professions classées en 8 catégories : (1) agriculteurs, (2) artisans-commerçants et chefs d'entreprises, (3) cadres et professions intellectuelles supérieures, (4) professions intermédiaires, (5) employés, (6) ouvriers, (7) retraités et (8) divers (chômeurs n'ayant jamais travaillé, étudiants, femmes (et les hommes) au foyer, etc.). On a les sous-catégories (code et intitulé) …
Distribution des valeurs : 13 (1), 128 (2), 669 (3), 240 (4), 178 (5), 69 (6), 56 (7), 45 (8), 19 (inconnu)
Fichier : PACES_brut_inscription_etudiant_2012-2013

- Les cours particuliers

Hypothèse : si cours particulier, alors un effort plus régulier et moins d'abandon

Non disponible : très peu d'étudiants donnent l'info en début d'année…

- La matière :

Deux hypothèses : (a) on réussit mieux dans certaines matières que d'autre (b) l'activité n'est pas la même pour les deux catégories : matières spécifiques aux études médicales (M), matières scolaires car déjà entrevues au lycée (S)

Variable qualitative nominale :

Premier semestre :

Cycle	Date début				
C1	24/09/2012	BCH	HBDD	BPH	MAT
C2	08/10/2012	BCH	HBDD	BPH	BCE
C3	22/10/2012	BCH	HBDD	BPH	BCE
C4	05/11/2012	BCH	HBDD	BPH	BCE
C5	19/11/2012	BCH	BCE	BPH	BSTAT
C6	03/12/2012	BCH	BCE	BPH	BSTAT

- Second semestre :

Cycle	Date début				
C1	28/01/2013	ANT	ICM	PHS	SSH
C2	11/02/2013	ANT	ICM	PHS	SSH
C3	04/03/2013	ANT	ICM	PHS	SSH
C4	18/03/2013	ANT	ICM	PHS	
C5	02/04/2013	ANT	PHAR	PHS	
C6	15/04/2013	MAIEU	ODON	PHAR	PHS

Fichier : PACES_calendrier_2012-2013

Abréviations et proposition de catégorisation (à vérifier avec un médecin) :

BCH	Biochimie	S
HBDD	Histoire et Biologie du Développement	S
BCE	Biologie Cellulaire	S
BPH	Biophysique	S
MAT	Mathématiques	S
BSTAT	Bio Statistiques	S
PHS	Physiologie	M
ANT	Anatomie	M
ICM	Initiation à la connaissance du médicament	M
SSH	Santé, Société, Humanité	S
PHSS	Physiologie Spécifique	M
PHAR	Pharmacie	M
MAIEU	Maïeutique	M
ODON	Odontologie	M

- Résultats aux concours :

Hypothèse : on voit un changement d'activité après le premier concours, en fonction du résultat

4 matières au s1 (et 10 matières au s2), pour chacune, on a la note sur 20 et le rang de l'étudiant

Variables quantitatives : notes entre 0 et 20, rangs entre 1 et 1542
Dossier : PACES brut corrections concours -> un fichier .xls par matière

TRAITEMENTS POUR CONSTRUIRE CE FICHIER

Le fichier final doit avoir (en colonnes) les variables suivantes

Date
Séquence
ID étudiant
Matière

Dist_note
Dist_note_moy
Ratio_Q
Ratio_Q_moy
Ratio_vote
Ratio_vote_moy
Sexe
Bac
Mention_bac
CSP
Choix_cursus
Note_concours_s1
Rang_concours_s1

Date : à chaque séquence correspond une date de début (voir correspondance ci-dessus)

Séquence : les 6 cycles du s1 (c1 à c6) et les 6 cycles du s2 (c1 à c6) sont transformés en 12 séquences (s1 à s12)

ID_étudiant : ID_PACES

Matière : pour chaque séquence il y a 3 ou 4 matières (voir correspondance ci-dessus)

Dist_note : choisir l'option quartile ou centile, puis récupérer les notes par fusion de fichiers (un fichier par matière). A partir de la distribution des notes dans une matière, calculer la position de la note de chaque étudiant (son quartile ou son centile).

Dist_note_moy : position de la moyenne d'une étudiant dans une séquence (moyenne sur 3 ou 4 matières)

Ratio_Q : calculer le nombre de questions posées par un étudiant sur chaque matière divisé par le nombre de question total

Ratio_Q_moy : calculer la moyenne des ratios pour une séquence

Sexe, bac, mention bac, csp : il faut écrire un programme qui fait correspondre le ID_UJF avec le ID_PACES pour pouvoir utiliser ces données.

Choix_cursus : on l'a

Note_concours et rang_concours : fusionner les fichiers de chaque matière du concours au s1

VERIFICATION QUALITE DES DONNEES

Contrôle qualité des données selon le processus proposé par Laura Dupuis :

Rapport de Laura Dupuis (section 5 page 21) :

- doublons (un même étudiant/date avec deux résultats)

- données manquantes

- perte lors de la fusion des fichiers avec concordance des ID étudiant (UJF et PACES)

- cohérence entre fichiers brut et transformé (par Pierre Gillois)

ANALYSES

Evolution temporelle (Scénario 3 Hubble)

Avec UnderTracks visualisations des données de QCM et de FLQ en fonction du temps :

- Pas de temps = la séquence (ou la matière)

- Lissage (remplacer par une courbe) pour voir la tendance

- Auto-corrélation pour voir des périodicités

Analyse complémentaire (Scénario 2 Hubble) : Typologie d'étudiants obtenue par des analyses statistiques (Voir rapport Laura Dupuis)
- analyse factorielle des correspondances multiples (ACM)
- classification ascendante hiérarchique (CAH).

RESULTATS ATTENDUS