Démarrer avec TDA : analyse de séquences, *event history analysis*

Claire Lemercier
version du 18 août 2008
remarques bienvenues : Claire.Lemercier@ens.fr

Les données du fichier opttc.txt ont été utilisées dans : Claire Lemercier, « Les carrières des membres des institutions consulaires parisiennes au xixe siècle », *Histoire & Mesure*, XX-1/2, 2005, p. 59-95. Elles sont librement réutilisables à condition de citer la source.
Attention : les données du fichier cl.txt sont en partie fictives (des dates de naissance manquantes ont été complétées un peu arbitrairement). Il s'agit simplement d'une base pour un exercice réalisé lors d'un stage (dont les enjeux et résultats sont décrits, en mauvais anglais, dans le fichier careers.pdf) : ne pas les réutiliser pour des travaux d'histoire.

Installation de TDA

Il faut récupérer d'abord TDA pour Windows :
http://steinhaus.stat.ruhr-uni-bochum.de/binaries.html
Dézipper le fichier, le mettre de préférence dans un répertoire créé à cet effet, où on placera de préférence aussi tous les fichiers afférents (programmes, données…) au moins sous forme de copie. Sinon, à chaque fois que vous vous référez à un fichier dans un programme, il faudra en spécifier le chemin (c:\mondossier\donnees.txt par exemple). Si c'est dans le même dossier que le programme, pas besoin de chemin, il suffit du nom (donnees.txt).
Puis télécharger winTDA (interface plus cool) :
http://www.tufts.edu/~kschmi04/research/ tout en bas de la page.
Dézipper le fichier, ouvrir le programme, aller dans Fichier>Préférences et lui indiquer où l'on a placé TDA.exe à l'étape précédente.
Le manuel (prendre la version « expérimentale » en pdf) est ici :
http://www.stat.ruhr-uni-bochum.de/tman.html

NB : tous les fichiers manipulés ci-dessous sont des fichiers texte (on peut les ouvrir avec le bloc-notes, calepin, un traitement de textes, etc.), qu'ils n'aient pas d'extension ou aient une extension propre à TDA comme .cf.

Un exemple d'analyse de séquences

Les fichiers de données :

- données dans *opttc.txt*
- description des données dans *opttc.cf* (au début du fichier)
- explication des codes numériques ci-dessous

Description pure des séquences demandée dans *opttc.cf* :

- durée et nombre d'épisodes dans chaque état pour chaque individu (résultat : *tcseqgc*, légende du résultat : *tcseqgcd*)
- mesure d'« entropie » à chaque moment, cf. p. 586 du manuel (résultat : *tcent*, légende du résultat : *tcentd*)
- recherche d'une sous-séquence particulière, par exemple ici : juge une année, CC et juge l'année suivante (résultat : *tcpat*, légende du résultat : *tcpatd*)

Analyse de séquences proprement dite demandée dans *opttc.cf.* Calcul des distances entre séquences, avec comme option pour les coûts : coûts fondés sur les nombres de transitions observés. Résultat dans *tcom,* légende du résultat : *tcomd.*

Différentes possibilités de classification automatique (*cf.* manuel de TDA ; à noter qu'on peut aussi exporter les données de distances et réaliser la classification dans un autre logiciel, ce qui ouvre d'autres possibilités encore) :

- par exemple : *tchd.cf* calculant le « minimal diameter » (résultat dans *tchd2*), et un autre algorithme pour le même prix dans le fichier de commande !
- un exemple d'un autre type de format de sortie : un algorithme de classification hiérarchique descendante appliqué dans *tcclu.cf*, incluant la création d'un fichier pour graphe *tctree.df* qui permet un autre type de représentation, demandé dans *tcparti.cf* et dont on voit le résultat dans *tcparti*.

Codes utilisés pour l'analyse de séquences :

1 Suppléant au TC
2 Régent ou censeur BdF
3 Juge au TC
4 Année hors institutions
5 Membre CC
6 Conseiller mun ou général
7 Conseiller d'escompte
8 Membre CC, Régent ou censeur BdF
9 Membre CC, Juge au TC
10 Membre CC, Conseiller mun ou général
11 Membre CC, Conseiller d'escompte
12 Régent ou censeur BdF, Conseiller mun ou général
13 Juge au TC, Conseiller d'escompte
14 Membre CC, Président du TC
15 Conseiller d'escompte, Suppléant au TC
16 Membre CC, Suppléant au TC
17 Président du TC
18 Membre CC, Régent ou censeur BdF, Suppléant au TC
19 Conseiller d'escompte, Président du TC
20 Membre CC, Régent ou censeur BdF, Conseiller mun ou général

21 Juge au TC, Conseiller mun ou général
22 Membre CC, Conseiller mun ou général, Président du TC
23 Conseiller mun ou général, Conseiller d'escompte
24 Régent ou censeur BdF, Suppléant au TC
25 Membre CC, Régent ou censeur BdF, Juge au TC
26 Membre CC, Juge au TC, Conseiller d'escompte
27 Membre CC, Juge au TC, Conseiller mun ou général
28 Conseiller mun ou général, Suppléant au TC
29 Régent ou censeur BdF, Juge au TC
30 Membre CC, Conseiller d'escompte, Suppléant au TC
31 Membre CC, Régent ou censeur BdF, Président du TC
32 Membre CC, Conseiller d'escompte, Président du TC
33 Membre CC, Régent ou censeur BdF, Conseiller mun ou général
34 Conseiller mun ou général, Président du TC
35 Membre CC, Conseiller mun ou général, Conseiller d'escompte
36 Régent ou censeur BdF, Président du TC
37 Membre CC, Conseiller mun ou général, Conseiller d'escompte, Suppléant au TC
38 Membre CC, Régent ou censeur BdF, Conseiller mun ou général, Suppléant au TC

Un exemple d'*event history analysis*

À noter : pour vraiment aborder l'*event history analysis* avec TDA (explications à la fois sur la méthode et le logiciel), utiliser Hans-Peter Blossfeld et Götz Rohwer, *Techniques of Event History Modeling: New Approaches to Causal Analysis*, Lawrence Erlbaum Associates, 2001.

Les fichiers de données :
- données dans *cl.txt*
- description des données au début de *exCL1.cf*
- explication de quoi il retourne dans *careers.pdf*.

Estimation non paramétrique (*product limit estimation*) :
- sur l'ensemble des données : début du programme *exCL1.cf*, résultats dans *cl.ple.*
- pour en tirer un graphique, deux possibilités :
    - un programme du type *clgraph.cf* (donnée sans garantie, il faut sans doute revoir les coordonnées) qui produit un graphe Postcript du type *cl.ps* [à ouvrir dans GSview ou assimilé, pas dans TDA !], joli mais peu flexible.
    - au prix d'un petit copier-coller à partir de *cl.ple*, réaliser le graphe « à la main » sous Excel peut offrir plus de flexibilité.
- en différenciant des groupes selon diverses variables : suite du programme *exCL1.cf.* Exemple de « banquier ou non » : résultat dans *clbank.ple*. Exemple de « entrée en dernière période ou non » : résultat dans *clcohort.ple*. Exemple de « impliqué ou non » : résultat dans *clinv.ple*. Exemple de « isolé ou non » : résultat dans *clnet.ple*.

Estimations paramétriques :

- modèle « piecewise exponential » sans paramètres : fin du programme *exCL1.cf*. Allure de la fonction de survie estimée avec un modèle en trois périodes dans *clpiece.prs.*
- graphe correspondant : *clpiece.ps* [ouvrir comme un Postscript, pas dans TDA], produit par le programme *clpiecegr.cf*. Ou bien, là encore, passer par Excel.
- introduction d'une variable changeant dans le temps (« devenir président de la Chambre »), d'où « splittage » des données et donc création d'un nouveau fichier de données. Programme : *exCL2.cf*, données « splittées » : *clsplit.dat*, légende correspondante : *clsplit.tda.*
- introduction d'une seconde variable changeant dans le temps (« passer l'âge de 65 ans »), d'où « splittage » des données à la fois selon le passage président et le passage d'âge, et création d'un nouveau fichier de données. Programme : *exCL4.cf*, données « splittées » : *clsplit2.dat*, légende correspondante : *clsplit2.tda.*
- résultats pour une version paramétrique du modèle en trois périodes : coefficients, significativité, risque relatif. Plusieurs essais :
    - modèle avec nombreuses variables mais sans « passer l'âge de 65 ans » : programme *exCL3.cf,* résultats dans *cloutput*
    - modèle avec l'âge et plus parcimonieux : programme *exCL6.cf,* résultats dans *cloutput2*
    - modèle final : programme *exCL14-age.cf* (dernières instructions), résultats dans *output age 2* [un fichier assez en bazar je vous l'accorde : vous feriez mieux de faire tourner *exCL14-age.cf* et comparer ce que vous obtenez avec les résultats qui sont dans *careers.pdf*].

Pour comparaison, modèle de Cox (semi-paramétrique) : programme *exCL14-age.cf* (dernières instructions), résultats dans *output age 2.*

Différentes simulations d'individus à partir du modèle retenu : programme *exCL14-age.cf*, résultats dans *clsim1/2/3/4/5.dat.*