1. Cartographie via QGIS

J'ai utilisé la base de données géoréférencée ROUTE 500, j'ai filtré les autoroutes et j'ai ensuite créé des variables (société propriétaire de la concession, date d'extinction du contrat, première et dernière ouverture de tronçon). Ensuite j'ai suivi le tutoriel, ce qui donne une carte purement descriptive en l'occurrence. Pour une partie analytique, il faudrait que j'arrive à sectionner les différents axes autoroutiers en tronçons, ce que je peux faire à la main mais c'est un travail monstrueux et je n'ai pas encore trouvé le moyen d'automatiser ça.

La Base Route 500 est construite autour des axes autoroutiers, elle ne donne pas accès au détail des différents tronçons constituant chaque axe. Je peux faire une base qui recense les différents tronçons et créer un tableau de variable qui permettrait des traitements statistiques classiques, mais c'est vraiment sur l'opération de géocodage que j'ai échoué. Comme il y a des milliers de tronçons, je n'envisage pas de faire ça à la main et faute de réussir à les récupérer par un quelconque moyen qui m'est pour l'heure inconnu, je m'en tiendrais à des représentations graphiques du réseau intégral (ce qui empêche de facto beaucoup des analyses que je souhaitais faire et dont la représentation cartographique constituait une synthétisation visuelle de la donnée qui me semblait pourtant satisfaisante).

2. Analyse de réseau via R

Après une première analyse infructueuse réalisée avec les outils proposés lors du séminaire, je me suis résolu à employer un logiciel que je maîtrise mieux. Je me suis appuyé sur les Webin-R proposés par Joseph Larmarange (le meilleur pour l'apprentissage R, c'est une vraie pépite) et son site analyse-R. Pour ce qui concerne l'analyse de réseau, il conseillait les tutoriels de Katerine Ognyanova qui m'ont été très utiles. Je me suis servi des données d'entraînement qu'elle proposait.

Il s'agit d'une représentation des citations et des renvois entre 17 titres de presse américaine, classés selon leur type (journal, sites de « news » et télévision) et pondéré par leurs échanges. Si on regarde comment sont construits les deux fichiers qui forment la source de cette représentation, l'interprétation est assez simple.

Un premier fichier « edges » est un data.frame de 4 variables avec 49 individus. Il décrit les relations entre chaque binôme de média (n=17, on a 49 parce qu'on a pas intégré quand il n'y avait pas de relation), caractérise le type de relation (lien hypertexte ou citation) et donne le nombre de ces relations unilatérale (on une ligne où A cite B et une autre si B cite A). Un second dataframe résume

les informations utiles sur chaque média : un identifiant, un label explicite, une typologie du média (TV, journal ou internet) et un indicateur d'audience.

On a donc deux data.frame, l'un qui concentre les informations relatives aux sommets de notre réseau et un qui décrit les relations entre ces sommets. Comme sur R la mise en relation est très simple (une simple ligne de code qui vient spécifier qui est quoi suffit à calculer le réseau), on a ensuite surtout des considérations de présentation graphique de l'information (de base, les graphiques fournit par R sont assez austères et pas très propre). Le reste du code sert à ça, ainsi qu'à rajouter de l'information présente dans nos data.frame que la fonction plot sur le simple calcul du réseau de rend pas.

On obtient alors un graphique qui permet de situer les relations de citations entre les principaux médias aux Etats-Unis.

En regardant le graphique de loin (façon de voir les tendances lourdes), on voit que les médias « Journal » et « TV » sont principalement tournés vers eux-mêmes (à l'exeption notable de la BBC), tandis que les sites d'informations en ligne usent beaucoup du contenu des journaux papiers. La BBC apparaît comme une source pour les médias en ligne et les journaux, mais déconnectées des autres chaînes TV. Fox News qui représente la plus grande audience du panel ne se réfère elle qu'à ABC.

3. Régression logistique via R

J'ai suivi les cours en ligne de Joseph Larmarange et de son site « analyse-R » qui constitue, je trouve, la meilleure ressource pour se former à R. Ici on propose une étude de régression sur un des jeux de donnée de base du logiciel R « questionR » avec la base « hdv2003 » (enquête INSEE « Histoire de Vie » de 2003).

Le code a été fournit dans un autre document, je ne vais pas revenir ici dessus, simplement dire ce que signifie les différentes sorties obtenues.

La régression va mesurer l'incidence des variables explicatives (sexe, groupe age, niveau d'étude, heures passées devant la télévision et la pratique du sport) sur la variable à expliquer (ici la lecture de bandes dessinées).

```
reg <- glm(lecture.bd ~ sexe+ grpage + etud + relig +heures.tv +sport,
        data = hdv2003, family = binomial(logit))
tbl_regression(reg, exponentiate = TRUE)
```

Caractéristique	OR1	95% CI1	p-valeur
sexe			
Homme	—	—	
Femme	1,83	0,98 – 3,57	0,065
grpage			
[16,25)	—	—	
[25,45)	0,72	0,19 – 3,43	0,7
[45,65)	1,55	0,40 – 7,72	0,6
[65,99]	0,69	0,12 – 4,35	0,7
etud			
Primaire	—	—	
Secondaire	0,88	0,24 – 3,08	0,8
Technique/Professionnel	0,67	0,18 – 2,39	0,5
Sup?rieur	4,83	1,80 – 14,7	0,003
manquant	2,32	0,35 – 14,8	0,4
relig			
Pratiquant regulier	—	—	
Pratiquant occasionnel	0,67	0,26 – 1,74	0,4
Appartenance sans pratique	0,38	0,15 – 1,00	0,044
Ni croyance ni appartenance	1,16	0,48 – 2,96	0,7
Rejet	1,00	0,21 – 3,60	>0,9
NSP ou NVPR	1,08	0,06 – 6,33	>0,9
heures.tv	0,96	0,77 – 1,17	0,7
sport			
Non	—	—	

Oui	1,31	0,68 – 2,51	0,4

1 OR = rapport de cotes, CI = intervalle de confiance

Les clefs de lecture du tableau précédent :

- Les __ signifient qu'il s'agit de la modalité de référence sur la variable
- L'odd ratio (OR1) c'est le rapport des côtes : on prend la probabilité sur la modalité de référence et on la multiplie par l'odd ration pour avoir l'effet (donc si <1, la modalité à un effet négatif sur la variable à expliquer par rapport à la modalité de référence, si >1 elle a un effet favorable et si =1 absence d'effet)
    - Ici ça donnerait 1,83 x intercept = probabilité de lecture de BD par une femme à toutes les autres modalités de références.
- 95% CI1 donne l'intervalle de confiance du calcul de l'odd ration à 95%
- La p valeur est une mesure de la fiabilité du calcul de l'odd ratio

Ensuite on a produit une représentation graphique des odds ratio et de leurs intervalles de confiance, c'est la même chose mais c'est peut-être plus « reader friendly »

La représentation graphique des effets permet de visualiser l'effet des différentes modalités de chaque variables sur la variable à expliquer, avec les intervalles de confiance. C'est pas mal je trouve, mais il pas présentable parce que les échelles ne correspondent pas entre les différents graphiques (ça pourrait induire en erreur). Ensuite la même chose en plus jolie

Ensuite j'ai juste regardé la matrice de confusion du modèle pour voir si il tenait un peu la route. C'est un tableau croisé des valeurs observées et des valeurs prédites par le modèle en l'appliquant aux valeurs d'origine. Avec ça, on obtient pour chaque individu la probabilité qu'il est vécu le phénomène étudié, ici qu'il est lu des BD. Comme c'est du oui/non, on va simplement dire que quand la proba est supérieure à ½ on met OUI, sinon on met NON.

On fait une manip : table(lecturebd.pred >0.5, hdv2003$lecture.bd)

On obtient False 1948 (Non) et 47 (Oui) ➔ 47 erreurs sur 1993 observations soit 2, 35% d'erreur, ce qui est plutôt pas mal du tout, on est content le modèle marche bien.

4. Classification des capitalismes d'Asie du Sud-Est via Analyse en composante principale et classification ascendante hiérarchique.

J'ai repris un article d'une dizaine d'année qui essayit de faire une classification de type régulationniste des différents capitalismes de l'Asie du Sud-Est, 20 ans plus tard. Ils donnaient les valeurs employées en fin d'article, j'ai essayé de les retrouver quand c'était disponible et en prenant des proxys quand ça ne l'était pas.

D'abord la première étape ça a été de construire des data.frame exploitables pour un calcul d'analyse de correspondance, ça m'a pris deux semaines (les listes étaient pas toutes de la même tailles, pas rangées de la même façon selon les différentes pays etc… c'était pas drôle). Une fois que j'y suis parvenu, la partie calcul analyse commençait (cf. fichier htlm)

Je vais simplement commenter les graphs obtenus sur la première période, la méthode est symétrique pour la seconde.

1/ D'abord quand on calcul l'ACP, on regarde le graphique des valeurs propres via screeplot

- On lit que l'axe 1 de l'ACP a une eigenvalue de 9.738 et l'axe 2 de 4.945 ce qui correspond à respectivement 36 et 18% de l'inertie totale du jeu de données. Sur notre ACP avec les Axes 1 et 2, on condense ainsi environ 54% totale de la variance totale de notre data.frame, ce qui est pas mal du tout.

2/ ensuite on a envie de savoir comment les différentes variables contribuent à la formation des axes.

- Via fviz_contrib on va avoir un histogramme de la contribution des différentes variables à la constitution d'un axe.
- Ensuite pour savoir l'effet de la variable, on regarde le graphique suivant (si valeur négative, alors la variable contribue à positionner l'individu sur la gauche (si axe 1) ou le bas (axe2) de la représentation, et inversement). C'est un sorte de projection sur un axe de ce qu'on observait de manière peu lisible via s.cocircle

3/ ensuite on représente les individus (ici les pays) dans le plan factoriel

4/ une fois qu'on a eu ça, on va s'appuyer sur ce calcul pour produire une clusterisation via classificvation ascendante hiérarchique

- Calcul d'une matrice de distance
- Choix du méthode de clusterison

- Production d'un dendogramme
- Choix du nombre de sous-groupes à créer, par critère du saut d'inertie
- On a alors des groupes

5/ injecter la clusterisation dans le plan factoriel

Maintenant qu'on a tout ça, on va regarder ce que ça a trouvé et si ça a du sens

1/ les deux axes construits

- L'Axe 1 est constitué de :
  - Time required to start a business (days) – A (8%)
  - Domestic general government health expenditure (6.5%)
  - Literacy rate (6%)
  - Lending interest rate (6%)
  - Pupil-teacher ratio, primary (6%)
  - Depth of credit information index (6%)
  - Imports of goods, services and primary income (5.7%)
  - Exports of goods, services and primary income (5.7%)
  - Business extent of disclosure index (5.4%)
  - Avec Vert : variables relatives à l'activité économique privé, Jaune : proxy du niveau de richesse du pays et Bleu : développement humain
  - Soit un Axe 1 qui serait lié au niveau de développement du pays → quand on regarde la répartition des pays sur l'Axe 1 c'est cohérent
- L'Axe 2 est constitué de :
  - Portfolio Investment, net (BoP, current US$) – A (12.5%)
  - total tax rate (% of commercial profits) – A (10%)
  - Start-up procedures to register a business (8.5%)
  - Market capitalization of listed domestic companies (7.9%)
  - Current health expenditure (% of GDP) (7.5%)
  - Other taxes payable by businesses (7.5%)
  - Tariff rate, applied, weighted mean, all products (6.5%)
  - Stocks traded, turnover ratio of domestic shares (4.5%)
  - Avec une anomalie sur les dépenses de santé (globalement elles ont souvent posées problèmes, trop de variables prises pour les mesurer) sinon on a

libéralisation des marchés financiers en vert kaki et imposition de l'activité économique réelle en turquoise

- Soit un Axe 2 structuré autour du « laissez-faire » économique, pareil ça fait sens au vu de la répartition des pays sur l'axe.

Pareil pour la clusterisation, on voit le Vietnam se déplacer vers les pays les plus développés entre les deux périodes, ce qui a du sens en ce que la décennie 2010-2020 a vu ce pays connaître un important développement.