Sullivan Honnet
Jules Vittone

17/12/2019

Rapport des résultats du projet de Data Science

Sujet : ACP : bilan des entreprises françaises en 2013.

Table des matières

Table des matières

Introduction

Le sujet de data science que nous avons choisi est de faire une Analyse en Composante Principales du bilan des entreprise française en 2013. Notre groupe est constitué de Sullivan Honnet et de Jules Vittone, le travail a été organisé en journée de travail, lorsque l'on se réunissait pour travailler on définissait les axes de travail à réaliser durant la séance et on travaillait ensemble.

- Jules a principalement fait de l'élaboration de code ainsi de la recherche pour nous aider à traiter du sujet.
- Sullivan a principalement réalisé le rapport ainsi que du test et de l'analyse des résultats.

Dans ce rapport nous allons présenter dans un premier temps le sujet que nous avons choisi, puis expliquer les choix techniques que nous avons mis en œuvre pour réaliser notre analyse qui seront présentés par la suite pour enfin faire un bilan des données.

Si vous souhaitez lancer notre projet, il vous faut créer des dossiers « Image », « Image1 » et ainsi de suite jusqu'à « Image8 ». Dans chacun de ces dossiers, vous devrait inclure un sous-dossier nommé « Correlation ». Il suffit ensuite de rentrer la commande py projet.py pour lancer le projet.

I) Explication du sujet

Le but de notre sujet est de réaliser une Analyse en Composantes Principales, ACP, sur un jeu de données contenant les informations relatives aux entreprises français en 2013. Les données ont été récupérées à cette adresse : http://www.insee.fr/fr/themes/detail.asp?reg_id=0&ref_id=esane-2013, qui contient plusieurs types de données pour faire plus d'analyse.

L'intérêt de cette étude est d'en apprendre plus sur les différents liens entre les 25 caractéristiques comptable qui permettent de faire des analyses financières d'une entreprise.

Les caractéristiques sont :

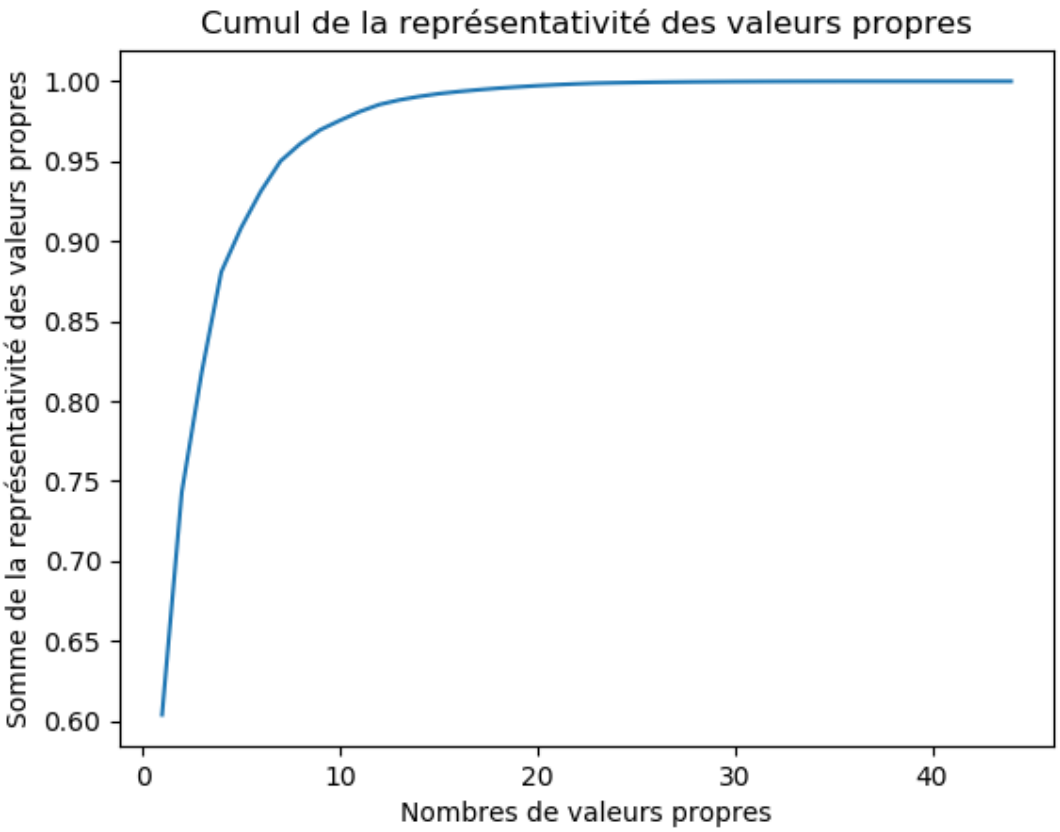
- (a)Nul Nombre d'unités légales
- (b) Cna Capital souscrit non appelé
- (c) Imi Immobilisations incorporelles
- (d) Imc Immobilisations corporelles
- (e) Ter Terrains
- (f) Con Constructions
- (g) Itm Installations techniques, matériel et outillage industriels
- (h) Aic Autres immobilisations corporelles
- (i) Mdt dont matériel de transport
- (j) Ime Immobilisations en cours
- (k) AeA Avances et acomptes
- (l) Imf Immobilisations financières
- (m) Tai Total de l'actif immobilisé
- (n) Smp Stocks - Matières premières approvisionnement et en cours
- (o) Sdm Stocks de marchandises
- (p) Aav Avances et acomptes versés sur commandes
- (q) Ccr Clients et comptes rattachés
- (r) Acr Autres créances
- (s) Vmp Valeurs mobilières de placement
- (t) Dis Disponibilité
- (u) Cdr Comptes de régularisation - Charges constatées d'avances
- (v) Tac Total de l'actif circulant
- (w) Acr Autres comptes de régularisation
- (x) Tab Total actif brut
- (y) Tan Total de l'actif net des amortissements et provisions inscrits à l'actif.

II) Explication technique

Pour ce projet nous avons décidé d'utiliser python pour l'extraction et le traitement des données récupérées, les raisons qui font que nous avons choisi ce langage est principalement car nous maîtrisons le langage, nous avons aussi trouvé de nombreux tutoriaux qui traitent de l'ACP en python.

Nous avons commencé par extraire du fichier les noms des différents secteurs, des caractéristiques et les valeurs associées, ensuite, nous avons supprimé toutes les lignes incomplètes des fichiers et toutes les lignes en double. En entrée, nous avions un fichier d'environ 2000 lignes, en sortie, le tableau obtenu en faisait plus que 800 lignes.

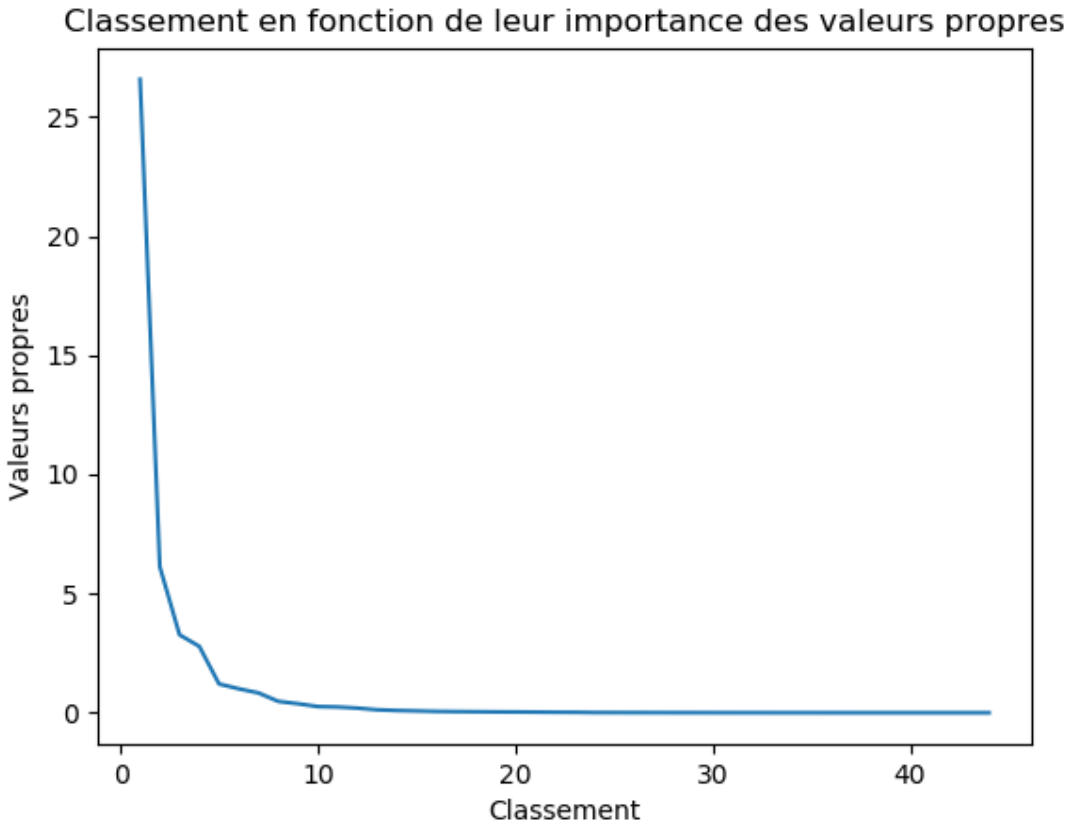
Ce tableau, nous lui faisons subir des transformations en fonction du traitement que l'on souhaite ou des valeurs que nous voulons.



Nous avons fini par inclure dans notre étude les valeurs du passif car nous pensons qu'étudier uniquement l'actif ne permet pas d'avoir des résultats vraiment significatifs. Nous avons donc créé des fichiers contenant toutes les données du passif et nous les avons ajoutées au tableau mentionné ci-dessus. Nous

obtenons donc 600 lignes valides à la fin de ce traitement ce qui nous a semblé suffisant pour pouvoir obtenir des résultats concluants.

Par la suite nous devions décider du nombre de caractéristique à prendre en compte pour notre analyse, nous avons donc réalisé un éboulis des valeurs propres et un graphique du cumul de variance restituée selon le nombre de facteurs pour faire cette décision, ensuite nous avons utilisé la règle du coude pour identifier le nombre de facteurs K* à retenir.



Dans notre cas 2 facteurs est ce qui semble être le choix à faire.

Par la contribution des variables aux axes nous avons pu déterminer les caractéristiques qui pèsent le plus dans la définition des axes, ce qui nous permet de faire par la suite un cercle des corrélations mais aussi la représentation des individus dans le premier plan factoriel.

III) Résultats obtenus et analyses

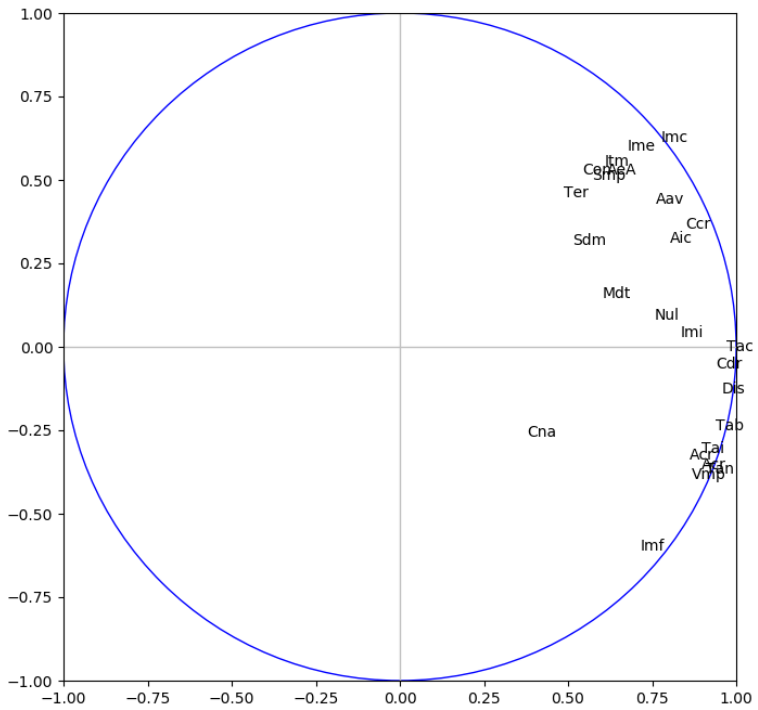


Figure 1 : Cercle de corrélation des variables comprenant toutes les variables

On trouve dans nos analyses sur l'actif des entreprises qu'un des critères déterminants est la quantité d'immobilisation corporelle, en effet, les secteurs ayant beaucoup d'immobilisation corporelle ressortent plus dans notre étude.

Les secteurs concernés sont surtout ceux du transport et de la gestion de patrimoine immobilier, l'industrie lourde est aussi représentée. Ceci s'explique probablement par la forte valeur du matériel utilisé dans leurs activités.

Les immobilisations financières sont pour leur part complètement décorrélées des immobilisations corporelles. On peut l'expliquer par une différence de secteurs. Une entreprise qui travaille avec des immobilisations financières est une société de service comme une banque, une société de gestion de fond ou de portefeuille qui utilise peu de matériel coûteux qui entrerait en compte dans les immobilisations corporelles.

Enfin, les immobilisations incorporelles sont elles-mêmes très peu corrélées avec les immobilisations corporelles et financières. L'explication est à chercher du côté des entreprises qui possèdent beaucoup d'immobilisation incorporelle, ce

sont généralement des entreprises de très haute technologie qui se consacrent à la recherche et n'ont que peu de patrimoine en-dehors des brevets.

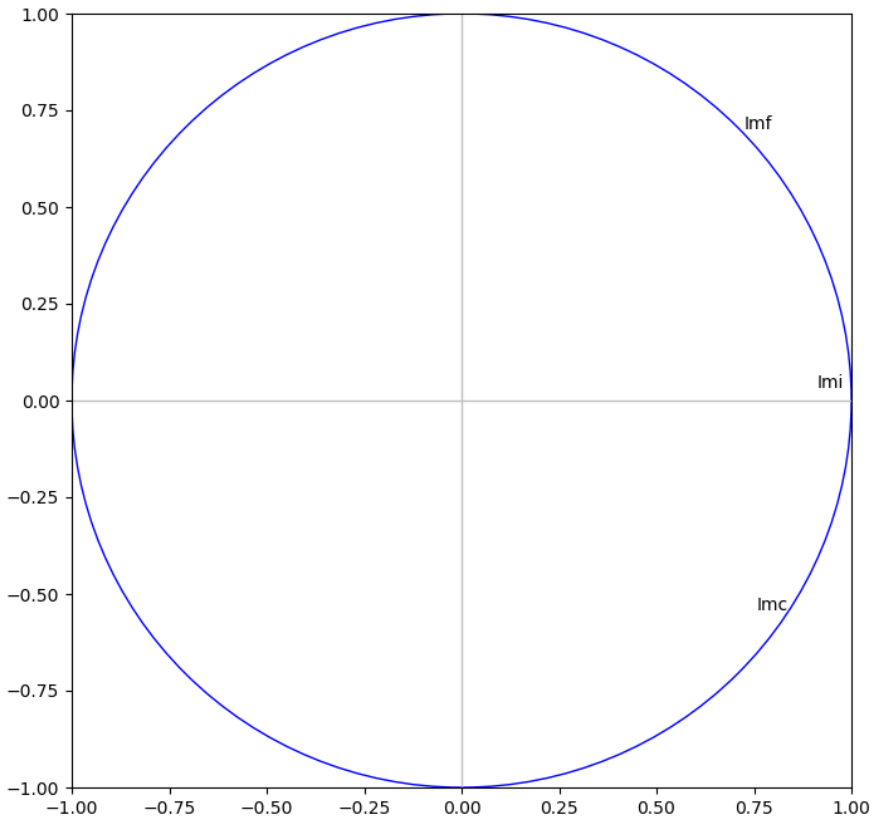


Figure 2 : Cercle de corrélation des immobilisations financières, corporelles et incorporelles

Pour conclure cette partie sur les différentes variables, nous nous sommes intéressés à une comparaison entre le total des actifs et le total des passifs et les résultats correspondent bien à nos attentes, en effet, les deux sont très fortement corrélés, cela peut s'expliquer assez simplement par le fait que dans un bilan comptable l'actif et le passif doivent toujours être égaux. La petite variation entre les deux s'explique par la présence de quelques arrondis sur les totaux.

Nous nous sommes par la suite intéressés aux secteurs et nous avons voulu identifier ceux dont l'inertie entraient le plus et le moins en jeu dans les résultats. Ces résultats ont été compilés dans un fichier texte qui liste les secteurs par ordre décroissant avec leur inertie. Nous nous retrouvons donc avec les secteurs d'activités spécialisées, scientifiques et techniques ainsi que les activités liées au commerce et aux sièges sociaux qui ont un impact très important dans les calculs suivi par les services de distribution et transport et de gestion de patrimoine immobilier.

Nous avons ensuite choisi de supprimer tous les secteurs ayant eu la plus grande inertie dans notre première étude pour nous concentrer sur les secteurs restants et nous n'avons pas relevé de points d'intérêt particulier. Eventuellement, on pourrait s'attacher maintenant à analyser un secteur précis ou alors ne choisir de garder que les secteurs ayant une inertie très faible pour vérifier qu'ils suivent les mêmes règles de répartition que le reste de l'économie.

Nous avons pu remarquer que les secteurs dont les corps de métier qui sont assez proche, par exemple les différents secteurs de l'agroalimentaire ou de métier de service sont souvent regroupés dans les mêmes zones ce qui permet de créer des points qui représentent ces secteurs sous une même appellation. Le fait que ces secteurs soit très proche les uns des autres est expliqué par les besoins de ces mêmes secteurs, en effet, en reprenant l'exemple de l'agroalimentaire, plus les entreprises qui transforment les matières premières ont besoin de matériaux plus le secteur qui les produits sera gros.

Pour conclure nos analyses, la prochaine étape si nous poursuivions le projet serait d'ajouter à nos analyses le fichier relatif aux amortissements et essayer de compléter les cases vides dans les lignes ce que nous n'avons pas pu faire en faisant le choix de nous concentrer sur l'ajout de nouvelles données. L'autre chose que nous aurions souhaité faire sans en avoir la possibilité aurait été d'inclure un code couleur continu pour pouvoir représenter une troisième dimension sur nos figures.

IV) Bilan et conclusion

A la fin du projet, nous avons pu constater que grâce à la méthode de l'ACP nous avons réussi à analyser de grand jeu de données et bien que nous n'ayons pas abordé tous les aspect possible du sujet qui est très vaste et qui pourrait amener à bien des analyses et conclusions, le monde de l'entreprise française dépend de beaucoup de facteur et que ces différents facteurs ont des dépendances entre eux, il aurait été intéressant de faire des comparatif entre les différentes années pour voir les évolutions des secteurs dans le temps et de voir leurs croissances ou leur décroissances ainsi que leurs impacts sur les autres secteurs. Pour donner un exemple, y a-t-il un lien entre développement du secteur du transport de marchandises et croissance du commerce.

L'ACP est donc une méthode rapide d'utilisation et relativement simple à implémenter pour des jeux de données spécifiques, il s'agit d'une méthode qui permet une analyse rapide d'un sujet et de voir à l'aide de graphiques les liens et les corrélations entre différents facteurs/individus.