L'apport des logiciels de statistique « open source » dans les études quantitatives des sciences de l'information et de la communication

Par : **Dr. Adlane HAFFAR**
Maître de conférences
Ecole Nationale Supérieure de Statistique et d'Economie Appliquée Koléa
Et

Par : **Dr. Khaled LALAOUI**
Maître Assistant
Ecole Nationale Supérieure de Journalisme et des Sciences de l'Information

Mots clé: ***Logiciel, open source, étude quantitative, méthodologie***.

ملخص:

يتناول هذا المقال إشكاليةاستغلال وتثمين البيانات الإحصائية المتعلقة ببحوث علوم الإعلام والاتصال. وعليه ولأسباب عملية، نقترح في دراستنا هذه استخدام برنامجا معلوماتيا حرا(مفتوح

المصدر في الأنترنت: تحميل وتفعيل مجانيين)كونه يلبي جميعالمتطلبات الأكاديمية في البحث العلمي، كسهولةالتعامل والتعلم، إلى جانب أمور أخرى تتعلق بالمتطلبات التقنية للبرنامج، بما في ذلك موثوقيةوثراءبرامجالرياضياتالمقترحة.قصد توضيح ما تم سرده أعلاه، قمنا بتقسيم مقالنا هذاإلىقسمين،خصص قسمه الأول لتقديم لمحة عن الأسلوب العلمي فيجمع البيانات، أما القسم الثاني فقد حاولنا تطبيق هذا البرنامج علىبعض البيانات المالية وأخرى متعلقة بعلم الإجراماإلى جانب تحليل وتفسير النتائج.

- **Préambule :**

Dans le champ des sciences de l'information et de la communication (SIC), il est devenu habituel de considérer que les chercheurs mettent en œuvre un pluralisme méthodologique ; une position régulièrement affirmée, notamment dans la littérature grise (mémoires et thèses). En fait, comme nous allons essayer de le montrer, ce pluralisme s'exerce dans un cadre particulier où l'approche quantitative est largement dominante.

Bien évidemment, la question de la méthodologie est associée à d'autres, de nature épistémologique, qui se rejoignent dans l'interrogation suivante : comment les chercheurs peuvent créer de la valeur par le traitement de l'information en utilisant des méthodes d'analyse quantitative ? Et quels sont les outils d'analyse statistique qui leurs facilites l'administration et l'exploitation de leur base de données ?

De cette question centrale découlent les questions secondaires suivantes :

- Quelle est l'importance de l'échantillonnage dans une étude quantitative ?
- Comment procéder à une enquête par questionnaire ?
- Quels sont les logiciels de statistique les plus utilisés dans les sciences de l'information et de la communication ?
- Quel est l'outil informatique d'analyse statistique de données qui soit en même temps performant, facile à utiliser tout en étant en open source ?

Le champ des SIC est résolument interdisciplinaire. Les méthodes mises en œuvre par les études qui en relèvent peuvent être diverses, mais chaque étude doit reposer sur une (des) méthodologie(s) bien identifiée(s). Est donc du ressort des SIC, l'étude des processus d'information ou de communication relevant d'actions contextualisées, finalisées, prenant appui sur des techniques, sur des dispositifs, et participant à des médiations sociales et culturelles[(1)]. Pour sa part, la méthode expérimentale – appliquée au domaine des SIC par exemple – cherche à étudier l'Homme en général, mettant l'accent sur les interactions entre son comportement et les situations qu'il vit concrètement, aussi bien au sein d'une organisation qu'à la suite d'une influence d'un ou plusieurs médias[(2)].

I. Choisir son échantillon :

Le chercheur dans le domaine des sciences humaines et sociales s'intéresse à l'étude des ensembles sociaux (par exemple une société globale ou des organisations concrètes dans la société globale) comme des totalités différentes de la somme de leurs parties. Au premier chef, il s'intéresse aux comportements d'ensemble, les structures et les systèmes de

relations sociales qui les font fonctionner et changer, non pour eux-mêmes, les comportements des unités qui les constituent. Mais même dans ce type de recherches spécifiquement sociologiques, les informations utiles ne peuvent souvent être obtenues qu'auprès des éléments qui constituent l'ensemble. Ainsi pour étudier l'idéologie d'un journal, il faudra analyser les articles publiés, même si ces analyses ne constituent pas en eux-mêmes, l'objet de l'analyse[(3)].

La totalité de ces éléments ou des " unités " constitutives de l'ensemble considéré est appelée "Population mère ou de recherche" ; ce terme pouvant désigner aussi bien un ensemble de personnes, d'organisations ou d'objet de quelque nature que ce soit[(4)].

Une population étant délimitée (par exemple, la population active d'une région, l'ensemble des entreprises d'information et de communication ou les articles publiés dans la presse sur un sujet donné au cours d'une année) il n'est pas pourtant toujours possible, ni d'ailleurs utile, de rassembler des informations sur chacune des unités qui la composent. De nos jours l'usage fréquent des sondages d'opinion a fini de prouver que l'on peut obtenir des informations fiables relatives à une population de plusieurs dizaines de millions d'habitants en n'interrogeant que quelques milliers d'entre eux.

Toutefois, on peut avoir recours aux techniques d'échantillonnage pour les objets les plus variés. Par exemple, un auditeur dans une entreprise analysera un nombre N de factures pour en tirer des informations relatives à la totalité des factures envoyées ou reçues par l'entreprise.

Un bibliothécaire examinera un échantillon représentatif des ouvrages possédés afin d'estimer leur état général de conservation.

Cependant et en dépit des nombreux avantages qu'elles présentent, les techniques d'échantillonnage sont loin de constituer un remè de en recherche sociale. Qu'en est-il exactement ?

Lorsqu'il a circonscrit son champ d'étude, trois possibilités s'offrent au chercheur :

1. Il recueille des données et porte ses analyses sur la totalité de la population couverte par ce champ

2. Il étudie un échantillon représentatif de cette population

3.Il étudie exclusivement certaines composantes très typiques, bien que non strictement représentatives de cette population.

II. L'enquête par questionnaire :

a. *présentation*

Elle consiste à poser à un ensemble de répondants, le plus souvent représentatifs d'une population, une série de questions relatives à leur situation sociale, professionnelle ou familiale, à leurs opinions, à leur attitude à l'égard d'options ou d'enjeux humains et sociaux, à leurs attentes, à leur niveau de connaissance, ou encore sur tout autre point qui intéresse les chercheurs. A la différence du sondage d'opinions, l'enquête par questionnaire vise la vérification d'hypothèses théoriques et l'examen des corrélations que suggèrent ces hypothèses. Compte tenu du grand nombre de personnes concernées et du traitement quantitatif des informations, les réponses aux questions sont pré codées

pour conduire les répondants à choisir leurs réponses parmi celles qui leur sont proposées[5].

***b**. objectifs à atteindre à partir de l'enquête par questionnaire*

On peut citer entre autres :

- ✓ La connaissance d'une population en tant que telle : ses conditions et ses modes de vie, ses comportements, ses valeurs ou ses opinions ;
- ✓ L'analyse d'un phénomène social que l'on pense mieux cerner à partir d'informations portant sur les individus de la population concernée (ex. impact d'une stratégie de communication ou de l'introduction de la Télévision Nationale Terrestre (TNT)dansles foyers algériens).

***c**. avantages*

Elle offre :

- ✓ La possibilité de quantifier des données et de procéder à de nombreuses analyses de corrélation,
- ✓ Réaliser l'objectif d'une réelle représentativité de l'ensemble des répondants

***d.** les limites et problèmes de cette méthode*

Il y a lieu de prévoir :

- ✓ La lourdeur et le coût généralement élevé du dispositif ;
- ✓ Le caractère souvent superficiel de certaines réponses empêche d'analyser à fond des phénomènes évolutifs tels que le travail au noir par exemple. Aussi et dans bien des cas, les résultats se présentent bien souvent comme de simples descriptions dépourvues d'éléments de compréhension pénétrante ;

- ✓ Le risque d'individualisation des répondants considérés indépendamment de leurs réseaux de relations sociales ;
- ✓ La fiabilité du travail peut souffrir d'une formulation peu claire des questions, du manque de confiance entre enquêteur et enquêté, ou simplement de l'inconscience professionnelle des enquêteurs.
- ✓ La non maitrise des logiciels statistiques d'analyse de données complique l'analyse des résultats.

III. Les logiciels statistiques les plus utilisés :

Dans cette partie, nous allons présenter les logiciels statistiques les plus populaires dans le monde académique et professionnel, les raisons qui ont fait leurs succès, leurs modalités d'enseignement, leurs retours d'expérience ainsi que le bilan et perspective de leurs adoptions comme logiciels de référence par les étudiants ainsi que les professionnels, à savoir[(6)] :

a. *Excel :*

Le logiciel Excel de la suite Office de Microsoft est, comme chacun le sait, un tableur. Il possède un certain nombre de fonctions statistiques, et on peut étendre ses capacités sur ce plan en lui adjoignant des macros mises à disposition par Microsoft, telles que « Utilitaire d'analyse » et « Utilitaire d'analyse-VBA », ou disponibles sur internet[(7)].

Raisons de l'utilisation de ce logiciel

- Le logiciel est présent quasiment partout dans le monde professionnel, entreprises ou administrations ;

- Le logiciel est parfois le seul disponible dans l'entreprise pour effectuer des traitements statistiques ;
- Les données circulent souvent au format Excel ou sont disponibles dans une base de données comme Access. Le traitement statistique peut s'effectuer sans faire passer les données par une phase de conversion. De même la saisie directe des données est facile ;
- Les étudiants sont en général à l'aise avec Excel, ou le deviennent très rapidement. Le logiciel est facile à utiliser, produit des graphiques attractifs de façon immédiate. Lorsque les étudiants commencent à avoir un peu d'expérience, le lien direct avec VBA leur permet de produire des rapports automatisés ;
- D'un point de vue pédagogique, Excel présente l'intérêt de permettre une décomposition fine des procédures statistiques, qui peut être exploitée pour l'acquisition des méthodes et résultats de base de la statistique. L'étudiant voit en même temps les données, les calculs intermédiaires, les graphiques.

Modalités d'enseignement

Le logiciel est utilisé pour l'apprentissage de la statistique par expérimentation, l'étudiant redécouvre par lui-même, par le biais de simulations, tel ou tel résultat de probabilités ou de statistique. On a recours à Excel pour illustrer, sur des exemples simples de données, des procédures présentées en cours. Ce logiciel permet très simplement de tirer un échantillon aléatoire dans une base de sondage, selon un plan de sondage, en population finie. Il permet aussi d'effectuer, dans un cadre limité, un traitement sur des données réelles. Enfin, il sert aussi de boîte à outils rapide

(lecture de tables, calculs d'indicateurs statistiques, construction de graphiques…).

Retour d'expérience

Les atouts d'Excel sont clairement le fait que ce logiciel est présent quasiment partout dans le secteur professionnel, qu'il est facile d'utilisation et qu'il présente un grand intérêt comme outil d'apprentissage basé sur la simulation ;

Le logiciel est avant tout un tableur et pas un logiciel dédié à la statistique. On observe des lacunes à des niveaux divers, pas de possibilité de construire de véritables histogrammes avec des classes d'amplitude différentes, pas de boîtes à moustaches. Les traitements statistiques offerts restent assez limités, même si des macros peuvent être trouvées pour enrichir les fonctionnalités. On peut être conduit à effectuer des tâches particulièrement répétitives. L'aide a été traduite en français de façon très approximative et la dénomination des fonctions semble souvent peu naturelle.

Les étudiants apprécient en général que la statistique soit enseignée en ayant recours à Excel. Ils sont souvent très à l'aise dans l'utilisation de ce logiciel. Leur savoir-faire dans l'utilisation de ce logiciel couplé avec leurs connaissances en statistique est apprécié des entreprises dans lesquelles ils font leur stage.

Bilan et perspectives

Dans la plupart des universités, on n'envisage pas de réduire le recours à Excel dans l'enseignement de la statistique. Il a un rôle bien identifié dans le dispositif d'enseignement de la statistique, aux côtés de logiciels du monde professionnel spécifiquement dédiés à la statistique tels que SAS, SPSS,

ou de logiciels du « monde libre » tels que R, pour ne citer que les plus utilisés.

b. *SPSS*

Ce logiciel est un logiciel généraliste, bien qu'à l'origine orienté vers la statistique en sciences sociales. Aux Etats-Unis et en Grande-Bretagne, son utilisation est largement répandue dans le monde professionnel et ce dans les secteurs les plus variés. Il rencontre un succès moins large ailleurs mais occupe cependant une place importante. Il couvre tous les champs de la statistique et il est basé sur l'utilisation de menus déroulants, sa prise en main est rapide[(8)].

Raisons de l'utilisation de ce logiciel

- Du point de vue pédagogique, ce logiciel offre l'avantage d'une utilisation facile, sans apprentissage informatique préalable, ce qui en autorise l'utilisation dès le début de la formation après une prise en main rapide. Il permet de mettre en application les méthodes statistiques et de servir de support à leur interprétation sur des données concrètes. Pour l'automatisation des tâches, ou la réalisation de celles qui ne figurent pas dans le pack de base, il peut être utilisé en programmant grâce à un éditeur de syntaxe ;
- Les sorties graphiques sont de très bonne qualité et l'éditeur de graphiques permet des mises en forme personnalisées très appréciées pour les restitutions audiovisuelles ou imprimées.

Modalités d'enseignement

SPSS est plus utilisé pour l'enseignement des modules de statistique descriptive, séries chronologiques, estimation et tests, régression linéaire et analyse de la variance, et analyse de données, du fait de ses grandes capacités de calcul.

Retour d'expérience

Les aspects positifs sont la facilité d'apprentissage, la possibilité pour l'enseignant d'axer le commentaire sur les aspects statistiques sans parler d'informatique, possibilité d'enregistrer son travail pour le reprendre ultérieurement.

Les aspects négatifs sont la difficulté au début de se rappeler le chemin dans les menus déroulants, la documentation en français est mal traduite et sources d'erreurs, la licence est chère et il n'existe pas de licence gratuite pour les étudiants.

Les étudiants doivent réfléchir sur les sorties sans avoir à passer du temps sur une étape « programmation », ce qui a l'avantage de permettre une pédagogie plus concentrée sur les aspects statistiques ;

Dans le cadre de la formation en apprentissage, les étudiants peuvent se trouver en situation d'avoir à utiliser en entreprise ce logiciel alors que son enseignement n'est pas encore très avancé. L'expérience montre que les étudiants parviennent généralement à s'autoformer sans trop de difficultés.

***c.** SAS :*

Le logiciel SAS est incontournable dans le domaine de la statistique et de l'informatique décisionnelle. Il permet d'appréhender l'ensemble des étapes d'un processus décisionnel, de l'importation des données à partir d'un

système dédié jusqu'à la mise en forme et la diffusion des résultats[9].

Raisons de l'utilisation de ce logiciel

- SAS est très répandu dans le monde professionnel, particulièrement dans le monde de la santé, mais aussi dans le milieu bancaire, les assurances, les sociétés d'enquête, d'études de marché, les grandes industries, les administrations ;
- Le traitement statistique offert par SAS est particulièrement riche et complet, voir même exhaustif. Il couvre la totalité des domaines particuliers de développement des méthodes statistiques de l'économétrie à la biostatistique, en passant par le contrôle de qualité, le marketing ;
- SAS est particulièrement « robuste » et peut traiter des jeux de données très volumineux, qui peuvent atteindre plusieurs millions d'individus ;
- Ayant son langage propre, il est parfois difficile pour des étudiants peu formés à ce logiciel de créer des macros SAS en stage et/ou en situation d'emploi. Par contre, il permet de mieux maîtriser les calculs effectués, ce qui est un plus pédagogique ;
- SAS met à disposition des enseignants, et des étudiants pendant leurs études, une version personnelle. Les étudiants peuvent travailler chez eux avec cette version comme s'ils étaient dans une salle informatique. Enfin SAS propose aussi un dispositif pour qu'un étudiant puisse utiliser le logiciel en stage dans une société qui n'en dispose pas.

Modalités d'enseignement

L'enseignement a très souvent lieu en salle d'ordinateurs, sous une forme qui combine apprentissage et pratique. Les cours sont effectués par des mathématiciens, des statisticiens ou des informaticiens. Une partie de l'enseignement peut être faite dans le cadre des modules « logiciels spécialisés ». Cela peut concerner par exemple la gestion de données, la programmation, l'édition de rapports, sujets qui pourront être enseignés par des informaticiens, ou des mathématiciens qui ont investi dans le domaine. Les procédures statistiques peuvent être ensuite présentées par les enseignants de statistique dans le cadre des enseignements concernés.

Retour d'expérience

Les étudiants montrent parfois au départ un intérêt limité pour ce logiciel, mais qui devient grandissant lorsqu'ils comprennent son importance. Toutefois, c'est un enseignement difficile du fait du langage spécifique et différent des langages de programmation classiques. Les étudiants ont parfois du mal à intégrer certains aspects.

Par contre, les entreprises ayant recours à ce logiciel sont en grande majorité très satisfaites des compétences acquises par l'utilisation de ce logiciel. SAS est très apprécié du monde professionnel, qui exige souvent ce profil pour des emplois de « programmeurs statistiques ».

Bilan et perspectives

Malgré quelques difficultés dans l'enseignement et la réticence des étudiants au début, les enseignants sont globalement satisfaits de ce logiciel. Le principal problème est le manque de temps pour montrer l'ensemble des possibilités de SAS. Mais, ayant acquis les bases et compris

la logique, les étudiants arrivent aisément à se former par eux-mêmes lorsqu'ils sont confrontés à des aspects qu'ils n'ont pas étudiés en cours.

Cependant, en considérant le prix élevé des licences SAS, un logiciel gratuit comme le R rivalise de plus en plus avec SAS sur le plan de l'exhaustivité des méthodes statistiques proposées, en particulier dans les développements méthodologiques récents.

d. R :

R est un logiciel statistique qui permet la lecture, la manipulation et le stockage de données. La grande majorité des méthodes statistiques actuelles y sont présentes par défaut ou au sein de « packages » dont la liste est en constante évolution. Les principaux facteurs qui expliquent son importance actuelle sont sa gratuité, sa fiabilité et sa disponibilité sous la plupart des systèmes d'exploitation. Initialement conçu pour illustrer l'enseignement de la statistique, R a connu une croissance exponentielle pendant les quinze dernières années dans le monde académique. Son développement actuel lui permet de rivaliser avec la plupart des logiciels payants utilisés dans les entreprises[(10)].

Raisons de l'utilisation de ce logiciel

- Il ne présente aucun problème d'accessibilité puisqu'il peut être téléchargé gratuitement sur n'importe quelle machine ;
- La programmation permet aux étudiants de mieux appréhender les différentes étapes des méthodes statistiques employées contrairement aux logiciels de type « clique-bouton » qui apparaissent parfois comme des boîtes noires ;

- De plus, compte tenu de son évolution actuelle, on peut penser que sa maîtrise va faire rapidement partie des compétences utiles à l'insertion professionnelle des étudiants.

Modalités d'enseignement

L'enseignement est principalement dispensé par des statisticiens pour illustrer les notions étudiées dans leur discipline. Par exemple, dans le module « technique de simulation », R permet de générer de nombreux types de variables aléatoires, de calculer facilement les indicateurs utiles et de faire les graphiques associés. Il peut aussi être utilisé par des informaticiens en cours de programmation objet.

Retour d'expérience

R est généralement moins apprécié par les étudiants que les logiciels à base de menus déroulants. D'une part, sa prise en main apparaît plus complexe du faite de l'apprentissage d'un langage dans un volume d'enseignement souvent insuffisant pour rendre les étudiants autonomes dans sa manipulation. D'autre part, le logiciel est encore peu implanté dans le monde professionnel et les entreprises sont encore peu demandeuses de cette compétence. Les étudiants ne considèrent souvent pas la connaissance de R comme un atout sur le marché du travail.

Cependant, quand les étudiants proposent l'utilisation de ce logiciel durant leur stage, les retours sont généralement très positifs. La grande diversité des méthodes disponibles suscite toutefois l'intérêt des étudiants. Les plus intéressés explorent ainsi de nouvelles fonctionnalités statistiques dans le prolongement des notions étudiées en cours en fonction des besoins rencontrés en stage.

Bilan et perspectives

Une fois l'apprentissage des bases dispensé, R peut être utilisé tout au long de la formation pour illustrer les différentes méthodes statistiques enseignées. Actuellement peu utilisé en milieu professionnel, on peut s'attendre à un développement rapide de son utilisation, notamment du fait de sa totale gratuité et du développement de modules qui le rendent plus accessible. Citons, par exemple, *Rcommander*, une interface à base de menus déroulants qui facilite la prise en main du logiciel, *Rexcel* qui simplifie la communication entre R et Excel ou encore *odfweave* qui permet l'automatisation de la production de rapports.

On peut s'interroger sur l'équilibre qui sera réalisé dans les années à venir entre l'enseignement de ce logiciel prometteur et celui des logiciels de statistique du monde professionnel.

IV. Outils statistiques d'analyse de données en *open source* :

Cette partie décrit un certain nombre de techniques statistiques pour analyser les données avec le logiciel R[(11)].« R » est un langagede programmation en mathématique et en statistique qui élargie son champ d'intervention au domaine de la finance, réseaux et télécom, biologie, développement mobile, sécurité et système d'exploitation entre autre. « R » est gratuit et donc ne nécessite pas l'obtention d'une licence, et disponible en *open source*, ce qui lui permet d'êtreen constante évolution grâce à une communauté d'utilisateurs de plus en plus large[(12)].

a. *Statistiques descriptives*

R propose plusieurs fonctions pour calculer les statistiques descriptives, notamment la fonction "mean" permet d'obtenir la moyenne d'un vecteur, tandis que les fonctions "min" et "max" renvoient, respectivement, la valeur minimale et maximale.

Pour illustrer nos propos, analysons les informations financières mensuelles sur l'action BNP, évoluant entre le 04 Janvier 2010 et le 02décembre 2013.Les actions constituent la classe d'actifs financiers la plus importante pour au moins deux raison : elles sont la source primordiale du financement des entreprises et elles représentent la part la plus grande dans les portefeuilles des investisseurs. L'action est la part représentative d'une fraction unitaire du capital social d'une société. Son détenteur possède quatre droits :

- Etre informé de la santé économique et financière de l'entreprise, en accédant librement au rapport annuel et en ayant la possibilité d'interroger un dirigeant par écrit ;
- Voter à l'assemblée générale des actionnaires, selon le principe 1 action = 1 voix, la majorité pour adopter une résolution étant fixée à 50 % des voix + 1 voix en assemblée générale ordinaire (AGO) et aux deux tiers des voix en assemblée générale extraordinaire (AGE) ;

- Recevoir le dividende, quote-part du résultat net de la société, sur décision de l'assemblée générale ordinaire ;
- Etre créancier résiduel, c'est-à-dire recevoir le produit de la cession des actifs de la société en cas de cessation ou de vente de l'activité, une fois les autres parties prenantes désintéressées (salariés, fisc, prêteurs et fournisseurs)[(13)].

Tous d'abord, nous chargeons les données à partir du site internet « yahoo finance » avec

Mots clé:

Logiciel, open source, étude quantitative, méthodologie.

la commande « read.csv » :

Date de cotation	Cours				
	Ouverture	Elevé	Bas	Fermeture	Volume de transaction
04/01/2010	56,11	60,38	50,55	52,15	4 105 400
01/02/2010	51,56	54,55	45,65	53,13	5 065 400
01/03/2010	53,9	59,34	53,38	56,86	3 689 700
01/04/2010	57,16	58,25	48,92	52,13	5 335 000
…	…	…	…	…	…
02/09/2013	48,16	51,56	47,42	50	3 661 700
01/10/2013	50,17	54,8	50,05	54,54	3 521 200
01/11/2013	54,6	55,82	52,34	55,25	3 290 300
02/12/2013	55,32	56,72	51,73	56,65	3 109 800

Tableau N°1 : Cotation de l'action BNP du 04 Janvier 2010 au 02 Décembre 2013[(14)].

On travail uniquement avec le cours de clôture de l'action :

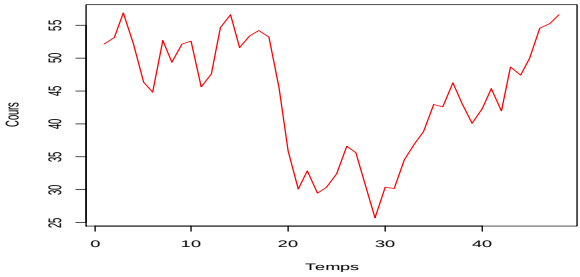
BNP.PA <- BNP.PA$Close

On inverse les données pour les avoir par classement chronologique :

BNP <- rev(BNP.PA)

On trace l'action sur un graphique pour faire une première analyse, en utilisant la commande « plot » :

plot(BNP, type = 'l', col = 'red', xlab = 'Temps', ylab = 'Cours')



Graphique N°1 : Evolution de l'action BNP du 04 Janvier 2010 au 02 Décembre 2013[(15)].

On constate deux tendances distinctes durant la période d'étude, une tendance baissière du premier au 29ième mois

qui est la conséquence de la crise financière de 2008, la tendance haussière vient relancer le cours de l'action à partir du 30ième mois et l'entreprise renoue avec la croissance.

Enfin, on procède à l'analyse à partir de la commande « summary » :

summary (BNP)

On obtient:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
25.73	36.44	45.54	44.20	52.26	56.86

Tableau N°2 : Statistique descriptive de l'évolution de l'action BNP du 04 Janvier 2010 au 02 Décembre 2013**(16)**.

Le tableau ci-dessus nous donne une idée sur la volatilité du titre, puisqu'il nous permet d'estimer l'écart entre les valeurs extrêmes atteintes par le cours de l'action.

***b**. Corrélation entre les variables*

Lorsque l'on analyse des données, il est très fréquent de vouloir savoir si deux variables sont corrélées. Informellement, la corrélation répond à la question « si

j'augmente (ou diminue) x, est-ce que y augmentera (ou diminuera), et de combien ? ». Formellement, elle mesure la dépendance linéaire de deux variables quelconques. Ses

valeurs varient de -1 à 1 ; 1 signifie que l'une des variables est une fonction linéaire positive de l'autre, 0, que les deux variables ne sont pas corrélées du tout et -1, que l'une des variables est une fonction linéaire négative de l'autre (les deux progressent dans des directions totalement opposées)[17].

La mesure de corrélation la plus utilisée est le coefficient de corrélation de Pearson :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Où $\bar{x}$ est la moyenne de x et $\bar{y}$ est la moyenne de y.

Nous allons illustrer cela en calculant la matrice de corrélation d'un portefeuille contenant 4 actions de l'indice boursier français CAC 40, à savoir BNP, Accor, Danone et Airbus.

```
BNP.PA <-
read.csv(paste("http://ichart.finance.yahoo.com/table.csv
?",
"s=BNP.PA&b=1&a=00&c=2010&e=31&d=11&f=201
3&g=m",sep=""))

BNP.PA <- BNP.PA$Close

BNP <- rev(BNP.PA)

AC.PA <-
read.csv(paste("http://ichart.finance.yahoo.com/table.csv
?",
"s=AC.PA&b=1&a=00&c=2010&e=31&d=11&f=2013
&g=m",sep=""))

AC.PA <- AC.PA$Close

AC <- rev(AC.PA)

AIR.PA <-
read.csv(paste("http://ichart.finance.yahoo.com/table.csv
?",
"s=AIR.PA&b=1&a=00&c=2010&e=31&d=11&f=2013
&g=m",sep=""))

AIR.PA <- AIR.PA$Close

AIR <- rev(AIR.PA)

BN.PA <-
read.csv(paste("http://ichart.finance.yahoo.com/table.csv
?",
"s=BN.PA&b=1&a=00&c=2010&e=31&d=11&f=2013
```

```
&g=m",sep=""))

BN.PA <- BN.PA$Close

BN <- rev(BN.PA)
```

Puis nous allons tracer les graphiques d'évolution du cours de chaque action:

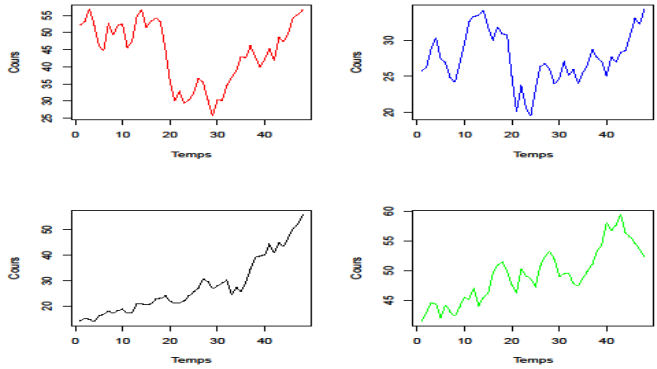
```
op <- par(mfrow = c(2,2))

plot(BNP, type = 'l', col = 'red', xlab = 'Temps', ylab = 'Cours')

plot(AC, type = 'l', col = 'blue', xlab = 'Temps', ylab = 'Cours')

plot(AIR, type = 'l', col = 'black', xlab = 'Temps', ylab = 'Cours')

plot(BN, type = 'l', col = 'green', xlab = 'Temps', ylab = 'Cours')
```



Graphique N°2: Evolution des cours des actions en portefeuille[(18)].

Les graphiques nous donnent une première idée sur la corrélation des cours des actions, les actions de BNP et Accor sont corrélées positivement, Airbus et Danone aussi, par contre, il n'y a pas de corrélations claires entre les cours de BNP/Accor et Air Bus/Danone.

Et enfin, nous calculons la matrice des corrélations, en utilisant la commande « cor »[1] :

```
Tableau_Cours_Action = data.frame(BNP,AC,AIR,BN)

Corr_Cours_Action<- cor(Tableau_Cours_Action)

Corr_Cours_Action
```

	BNP	AC	AIR	BN
BNP	1.0000000	0.74682878	0.0204997	-0.17436045
AC	0.7468288	1.00000000	0.2193192	0.08767219
AIR	0.0204997	0.21931919	1.0000000	0.85189276
BN	-0.1743604	0.08767219	0.8518928	1.00000000

Tableau N°3 : Matrice de corrélation des cours historiques des actions.

Le tableau ci-dessus vient confirmer notre analyse graphique, les actions de BNP et Accor sont corrélées positivement, Air Bus et Danone aussi, par contre le cours de BNP et celui de Danone sont corrélées négativement, la corrélation entre le cours de BNP et celui d'Air Bus est nulle, celle d'Accor et Air Bus est presque nulle.

c. *Analyse en composantes principales*

L'analyse en composantes principales est une autre technique d'étude des données. Elle consiste à transformer un ensemble de variables éventuellement corrélées en un ensemble de variables non corrélées.

On dispose d'un nuage de points, dans un espace de dimension élevée, dans lequel on ne voit pas grand-chose. L'analyse en composantes principales va nous donner un sous-espace de dimension raisonnable, tel que la projection sur ce sous-espace retienne le plus d'information possible,

tel que le nuage de points projeté soit le plus dispersé possible. Cela permet de réduire la dimension du nuage de points.

Nous illustrerons nos propos par un exemple sur la criminologie, en supposant qu'un journal ou des autorités locales veuillent communiquer les chiffres officiels concernant les crimes perpétrés durant une période déterminée, l'analyse en composantes principales leurs permettra d'affiner les données brutes en ressortant des informations plus précises. Nous commencerons par charger le package « MVA »**(19)** :

```
library(help=MVA)
```

On utilisera la base de données « USArrests », qui contient le nombre d'arrestations par 100 000 habitantsdans 50 Etats Américains durant l'année 1973.[1]

```
data(USArrests)

p <- princomp(USArrests)
```

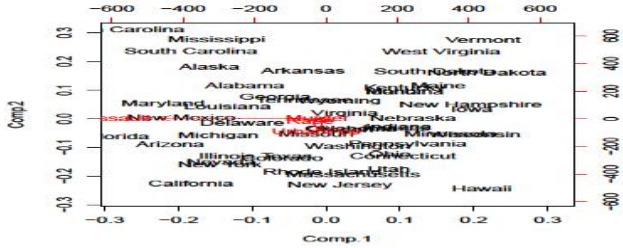
Puis ne faisons appel à la base de données en tapant son nom :

```
USArrests
```

	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	
Arizona	8.1	294	80	31.0
California	9.0	276	91	40.6
Connecticut	3.3	110	77	11.1
…	…	…	…	…
Rhode Island	3.4	174	87	8.3
South Carolina	14.4	279	48	22.5
Virginia	8.5	156	63	20.7
Washington	4.0	145	73	26.2

Tableau N°4 : Nombre d'arrestations par 100 000 habitants dans 50 Etats Américains durant l'année1973[(20)].

Nous utilisons la commande « biplot » pour créer un premier graphique[(21)]

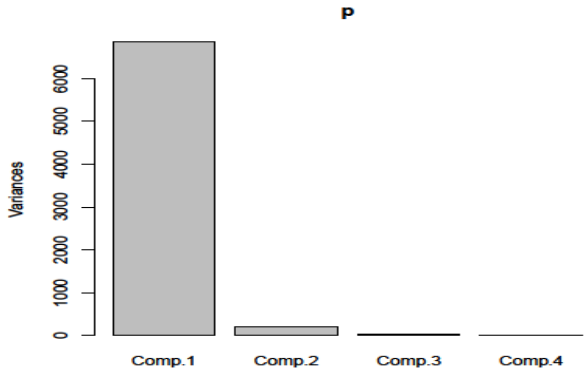


Graphique N°3 : Projection simultanée des variables et des individus[(22)].

Ce sur graphique ci-dessus, sont projetées les Etats avec les quatre variables d'étude, ce dernier va nous permettre de faire une classification avec une typologie de chaque classe, en les regroupant selon leurs coordonnées et aussi la distance entre les individus (Etats). A titre d'exemple, les Etats « Maryland », « Louisiana », « New Mexico », « Delaware » et « Michigan » constituent un premier groupe homogène caractérisés par la prédominance de la variable « Assault », en d'autres termes, ses Etats enregistrent un chiffre important des cas d'agression.

Voici les valeurs propres. On constate qu'il y a une chute brutale : cela signifie que l'analyse en composantes principales est pertinente et qu'on peut se limiter aux vecteurs propres dont les valeurs propres sont élevées.

plot(p)



Graphique N°4 : Valeurs propres et le choix des axes[(23)].

Le graphique ci-dessus, va nous permettre d'abord de choisir le nombre des axes retenus dans notre analyse, ainsi que le pourcentage d'inertie représenté dans ces axes (la qualité d'information expliqué par ses derniers) à partir de ce graphique, on remarque que le premier axe explique une partie très importante de l'information, et on retient le deuxième axe comme un axe complémentaire.

Dans la figure suivante, les anciens vecteurs de base ont été représentés en rouge. On les représente parfois sur un dessin isolé, dans un cercle, le cercle des corrélations

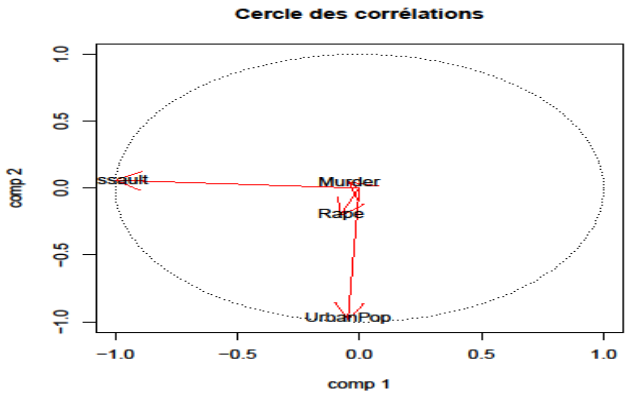
```
a <- seq(0,2*pi, length=100)

plot( cos(a), sin(a), type ='l', lty = 3, xlab = 'comp 1',
ylab = 'comp 2', main = "Cercle des corrélations" )

v <- t(p$loadings)[1:2,]

arrows(0,0, v[1,], v[2,], col='red')

text(v[1,], v[2,], colnames(v)
```

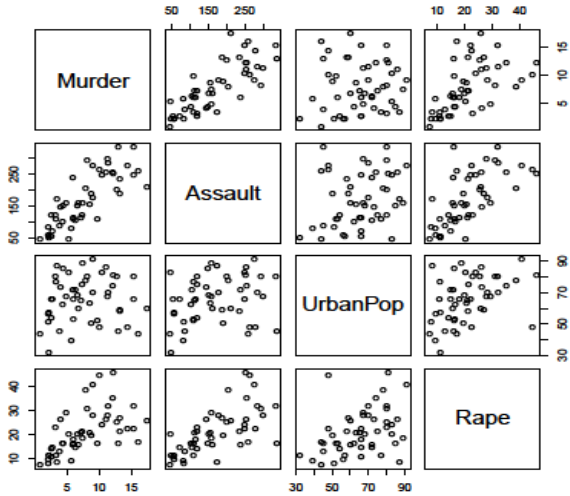


Graphique N°5 : Projection des variables[24].

Le graphique ci-dessus représente le cercle de corrélation des variables qui nous permet de déterminer les variables les plus significatives sur la base de leurs projections dans le cercle de corrélation. En se basant sur les principes suivants : Les variables les plus proches de l'unité (la limite du cercle) sont les variables les plus significatives, dans notre cas, sont les variables « UrbanPop » dans l'axe 2, et « Assault » dans l'axe 1. Par contre, les variables « Murder » et « Rape » sont très proches du centre de gravité, donc ne donne pas des informations utiles pour notre analyse.

Enfin, nous pouvons représenter les relations entre les crimespar un tableau de nuage de points, en utilisant la commande « pairs » :

pairs(USArrests)



Graphique N°6 : Relations entre les crimes perpétrés dans 50 Etats Américains durant l'année 1973.

Ce graphique illustre les corrélations entre les différents crimes perpétrés, si le nuage de point est dispersé, la corrélation est difficilement établi, si le nuage de point suit une tendance, alors la corrélation peut être positive ou négative selon la trajectoire du nuage de point

```
ibrary(help=MVA)
  data(USArrests)
  p <- princomp(USArrests)
  Tableau_Crimes_USA<-
  data.frame(USArrests$Murder,USArrests$Rape,USArr
  ests$Assault,USArrests$UrbanPop)
  Corr_Crimes_USA<- cor(Tableau_Crimes_USA)
  Corr_Crimes_USA
```

	USArrests. Murder	USArrests. Rape	USArrests. Assault	USArrests.UrbanPop
USArrests. Murder	1.00000000	0.5635788	0.8018733	0.06957262
USArrests. Rape	0.56357883	1.0000000	0.6652412	0.41134124
USArrests. Assault	0.80187331	0.6652412	1.0000000	0.25887170
USArrests. UrbanPop	0.06957262	0.4113412	0.2588717	1.00000000

Tableau N°5 : Matrice de corrélation des crimes perpétrés dans 50 Etats Américains durant l'année 1973[(25)].

Le calcul des corrélations entre les variables nous confirmera l'évolution des nuages de points du graphique ci-dessus :La corrélation entre le meurtre (Murder) et le viol (Rape) est positive, ce qui dénote d'une relation réciproque, la corrélation entre le meurtre et l'agression (Assault) est positive et proche de l'unité, ce qui peut être interpréter par le fait qu'un meurtre est souvent précédé par une agression, enfin, la corrélation entre le meurtre et la densité urbaine (UrbanPop) est nulle.

La corrélation entre le viol et l'agression est positif, ce qui nous semble tout à fait normal, la corrélation entre le viol et la densité urbaine est aussi positive, ce qui dénote d'une relation bilatérale entre ses deux variables, enfin la corrélation entre l'agression et la densité urbaine est positive mais reste faible, et donc la relation n'est pas déterminée.

Conclusion :

Le but de cet article est de mettre en évidence l'apport des outils d'analyse statistique dans le traitement des données propres au domaine de l'information et de la communication, en utilisant un logiciel qui soit à la fois pluridisciplinaire, gratuit et facile à assimiler, et dont la finalité pour le chercheur est la valorisation de son travail d'investigation. Pour arriver à notre objectif, nous avons commencé par mettre en évidence la démarche de sélection de l'échantillonnage, puis le type d'enquête, dont le choix s'est porté sur le questionnaire, par la suite, nous avons fait un baillage des logiciels statistiques les plus utilisés, enfin, nous avons mis en évidence la pertinence de l'utilisation du logiciel « R » comme outil d'analyse statistique sur des données financières en premier lieu, et sur des données relatives à la criminologie en second.

De par la qualité et l'exhaustivité des résultats obtenus, il nous sembleévident que le recours au logiciel « R » offre plusieurs avantages. Le logiciel « R » est multi-plates-formes fonctionnant sous Linux, Windows et Macintosh. Il possède un langage de programmation simple, efficace, interprété et non compilé. « R » est un logiciel dans lequel de nombreuses techniques statistiques modernes et classiques ont été implémentées. Il est particulièrement performant pour la manipulation de données, le calcul et l'affichage de graphiques vu les techniques statistiques modernes et classiques implémentées dedans de façon qu'il devient pendant les dernières années l'un des logiciels les plus populaire dans la fouille de données (en anglais *Data Mining*), qui est une discipline en plein essors.

« R » est devenu ces cinq dernières années un standard incontournable dans le « petit monde des grands logiciels d'analyse statistiques » : Sas®, Spss® et Statistica® proposent aujourd'hui des interfaces pour intégrer des calculs et des graphiques en « R » dans leurs logiciels. Il est même possible d'utiliser R dans Microsoft Excel®. « R » est riche en termes de statistique, mathématique et informatique. On y trouve un langage, des données, des exemples. Car « R » a été conçu pour être pédagogique et facile à apprendre : lorsqu'une fonction implémente un calcul particulier, elle doit avoir un fichier d'aide, des données et des exemples à consulter sans avoir à taper autre chose que les mots *help* ou *example.*

« R » propose des centaines (voire des milliers) d'implémentations d'algorithmes statistiques ou d'analyse des données. Aucun logiciel commercial ne peut rivaliser avec les fonctionnalités disponibles *via* les sites CRAN (*Comprehensive R Archive Network*) qui hébergent tous ces programmes. Il existe des centaines de milliers d'utilisateurs de R de par le monde. En adoptant « R », l'utilisateur peut être sûr qu'il utilise le même logiciel que ces collègues. Les performances de « R » sont comparables, voire meilleures, que la plupart des logiciels d'analyse commerciaux, « R » demande le chargement des données en mémoire avant de les traiter : si l'ordinateur contient assez de mémoire, R s'exécutera donc très rapidement. La mémoire étant une ressource peu onéreuse, acheter 32 Go de mémoire coûtera moins cher qu'une licence monoposte pour un logiciel commercial équivalent.

Références :

[1]Françoise Bernard et Robert-Vincent Joule, **Le pluralisme méthodologique en sciences de l'information et de la communication à l'épreuve de la communication engageante**, paru dans « Questions de Communication », N° 07, Edition 2005, p.185.

[2]Ibrahima Lo, **Méthodologie de la recherche en sciences sociales**, support de cours édité par l'Agence universitaire de la Francophonie (AUF), 2014, p.8

[3]Bernard LAMIZET & Ahmed SILEM**, Dictionnaire Encyclopédique des sciences de l'information et de la communication**, Editions ELLIPSES, Paris, France, 2009, p.209.

[4]Ibrahima Lo, **Méthodologie de la recherche en sciences sociales**, Op.cit, p.27.

[5]Ibrahima Lo, Ibid, pp.33-34.

[6]Grégoire .G, Jollois .F.X, Petiot .J.F, Qannari .A, Sabourin .S, Swertwaegher .P, Turlot .J.C, Vandewalle .V, Viguier-Pla .S. (2012). - **Les logiciels et l'enseignement de la statistique dans les départements statistique et informatique décisionnelle** -. Société Française de Statistique. France. p. 3.

[7] Monier. C, Le Guen. F (2013). – Maîtrisez Excel 2013 -. Edition Pearson. France. p. 7.

[8]Malhorta. N, Décaudin. J-M, Bouguerra. A (2007). – **Etudes marketing avec SPSS** -. Edition Pearson. France. p. 6.

(9) Ringuedé.S (2011). – SAS -. Edition Pearson. France. p. 4.

(10) Lafay. P, Drouilhet. R-M, Liquet. B (2011). – **Le logiciel R** -. Edition Springer. France. p. 2.

(11) http://www.cran.r-project.org/

(12) Adler J., - R, l'essentiel. Edition Pearson. France. 2011. p. 8.

(13) Rocchi J.M, Bertonèche M, Bouzou N, Gresse C. – MBA Finance -. Edition Eyrolles. France. p. 267.

(14) Elaboré par nous-même.

(15) Elaboré par nous-même.

(16) Elaboré par nous-même.

(17) Gujarati D.N., - **Econométrie. Edition de boeck**. 2004. Bruxelles, Belgique. p. 870.

(18) Elaboré par nous-même.

(19) Introduction to AppliedMultivariateAnalysiswith R: http://www.cran.r-project.org/.

(20) Elaboré par nous-même.

(21) Zoonekynd V., - **Cours sur les méthodes factorielles**, autour de l'analyse en composantes principales. 2004. p. 1.

(22) Elaboré par nous-même.

(23) Elaboré par nous-même.

(24) Elaboré par nous-même.

(25) Elaboré par nous-même.

Liste Bibliographique :

1- Ouvrages :

- Adler J., - R, l'essentiel. Edition Pearson. France. 2011.

- Bernard LAMIZET & Ahmed SILEM, Dictionnaire Encyclopédique des sciences de l'information et de la communication, Editions ELLIPSES, Paris, France, 2009.
- Françoise Bernard et Robert-Vincent Joule, Le pluralisme méthodologique en sciences de l'information et de la communication à l'épreuve de la communication engageante, paru dans « Questions de Communication », N° 07, Edition 2005.
- Grégoire .G, Jollois .F.X, Petiot .J.F, Qannari .A, Sabourin .S, Swertwaegher .P, Turlot .J.C, Vandewalle .V, Viguier-Pla .S. (2012). - Les logiciels et l'enseignement de la statistique dans les départements statistique et informatique décisionnelle -. Société Française de Statistique. France
- Gujarati D.N., - Econométrie. Edition de boeck. 2004. Bruxelles, Belgique
- Ibrahima Lo, Méthodologie de la recherche en sciences sociales, support de cours édité par l'Agence universitaire de la Francophonie (AUF), 2014
- Lafay. P, Drouilhet. R-M, Liquet. B (2011). – Le logiciel R -. Edition Springer. France.
- Malhorta. N, Décaudin. J-M, Bouguerra. A (2007). – Etudes marketing avec SPSS -. Edition Pearson. France
- Monier. C, Le Guen. F (2013). – Maîtrisez Excel 2013 -. Edition Pearson. France
- McNeil, D. R. (1977) *Interactive Data Analysis*. New York: Wiley.
- Planchet F., Thérond P., Kamega A., - Scénarios économiques en assurance, modélisation et simulation. Edition Economica. Paris, France. 2009

- Ringuedé. S (2011). – SAS -. Edition Pearson. France.
- Rocchi J.M, Bertonèche M, Bouzou N, Gresse C. – MBA Finance -. Edition Eyrolles. France
- Zoonekynd V., Cours sur les méthodes factorielles, autour de l'analyse en composantes principales. 2004

2- Webographie :

- http://www.cran.r-project.org
- Introduction to Applied Multivariate Analysis with R: http://www.cran.r-project.org