UNIVERSITE SIDI MOHAMED BEN ABDELLAH

FACULTE DES SCIENCES ET TECHNIQUES

DEPARTEMENT DES MATHEMATIQUES

Master Mathématiques et Application au Calcul Scientifique (MACS)

MEMOIRE DE FIN D'ETUDES

Pour l'obtention du Diplôme de Master Sciences et Techniques (MST)

Décomposition des données massives Big Data en Clustering et application de la régression linéaire multiple

Réalisé par: **FARFAR Assia**

Encadré par: **Pr. AMMOR Wafae**

Soutenu le 02 février 2017

Devant le jury composé de:

-Pr. AMMOR Wafae (FST-Fès)

-Pr. EL HILALI ALAOUI Ahmed (FST-Fès)

-Pr. EL KHOUKHI Fatima (FLSH –Meknès)

-Pr. Hilali Abdelmajid (FST-Fès)

Année Universitaire 2016 / 2017

Remerciements

En premier lieu, je tiens à exprimer ma profonde reconnaissance à Mme. Ammor Wafae, d'avoir accepté d'encadrer mon travail et m'initier à la recherche.

Je voudrais également lui témoigner ma gratitude pour son soutien afin de mener ce travail à bon port.

Mes vifs remerciements vont aussi aux membres du jury pour avoir accepté d'examiner ce travail et de l'enrichir par leurs propositions.

Enfin, Je voudrais remercier ma famille, mes enseignants et toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Dédicaces

Que ce travail témoigne de mes respects :

♥ A mes parents : Abdelkader et Meryem

Grâce à leurs tendres encouragements et leurs grands sacrifices, ils ont pu créer le climat affectueux et propice à la poursuite de mes études.
Aucune dédicace ne pourrait exprimer mon respect, ma considération et mes profonds sentiments envers eux.
Je prie le bon Dieu de les bénir, de veiller sur eux, en espérant qu'ils seront toujours fiers de moi.

♥ A mes sœurs et mon frère.

♥ A tous mes professeurs :

Leurs générosité et leur soutien m'oblige de leurs témoigner mon profond respect et ma loyale considération.

♥ A tous mes amis et mes collègues :

Ils vont trouver ici le témoignage d'une fidélité et d'une amitié infinie.

FARFAR Assia

Table des matières

Introduction

L'explosion quantitative des données numériques a obligé les chercheurs à trouver de nouvelles manières de voir et d'analyser ces données. Il s'agit de découvrir de nouveaux ordres de grandeur concernant la capture, la recherche, le partage, le stockage, l'analyse et la présentation des données, ainsi est né le « Big Data ».

Volontairement ou involontairement, l'humanité génère de plus en plus d'informations. Les progrès, en matière de capture et de stockage, autorisent la conservation de cette masse de données. Des outils et des méthodes permettent, aussi, de traiter ces volumes gigantesques d'informations. Ce phénomène est appelé le Big Data. Il s'agit d'un concept permettant de stocker un nombre indicible d'informations sur une base numérique.

Le terme Big Data se réfère aux technologies qui permettent aux entreprises d'analyser rapidement un volume de données très important. En mixant intégration de stockage, analyse prédictive et applications, le Big Data permet de gagner en temps, en efficacité et en qualité dans l'interprétation de données.

L'expression « Big data » a fait son apparition en octobre 1997 dans la bibliothèque numérique de l'ACM, au sein d'articles scientifiques qui pointent du doigt les défis technologiques à visualiser les « grands ensembles de données ». Le Big data est né, avec lui ses nombreux défis. Dans les années 2000, alors que l'exabytes entrent en jeu dans la quantification des données produites annuellement, la valeur du Big data est mise en avant, d'abord pour les bénéfices qui peuvent en tirer la recherche dans les secteurs de la physique, de la biologie ou des sciences sociales.

Depuis ces dix dernières années, les informations disponibles via Internet et les différents objets connectés ne cessent d'augmenter. Tout comme les énergies telles que l'électricité et le pétrole ont marqué la révolution industrielle au XXème siècle, le XXIème siècle sera celui de la donnée et de la manière de l'interpréter. Les données deviennent ainsi le carburant de l'économie numérique, Tous les secteurs économiques, du commerce au secteur automobile en passant par le secteur

énergétique, tous les domaines de la vie quotidienne (santé, éducation…) sont concernés. Le Big Data regroupe à la fois le traitement de ces grandes masses de données, leur collecte, leur stockage jusqu'à leur visualisation et leur analyse.

Comment mettre en relation toutes ces données ?

Comment les faire parler ?

Comment simplifier l'analyse de ces masses de données et aboutir à des résultats satisfaisants ?

Dans ce travail et afin de contribuer à la réponse de quelques questions on a pensé à une nouvelle méthodologie analytique basée sur la méthode de la régression linéaire qui permet de réduire la charge de calcul.

L'objectif principal de mon travail consiste à concevoir un modèle de régression linéaire qu'on a appelé « **modèle de régression linéaire divisée** » capable d'aboutir aux mêmes résultats avec moins de calcul et dans un temps réel.

La méthode de régression linéaire divisée consiste a divisé les grandes données de la population en quelques sous-ensembles de données de petite taille à l'aide de la méthode de classification.

Pour cela on a traité deux exemples avec un nombre de données varié et on a effectué des simulations et les résultats obtenus montrent que cette méthode est applicable et permet d'avoir de bon résultat avec moins des calculs.

I. Présentation générale de Big Data

1. Qu'est-ce que le Big Data ?

1.1 Définition

Le Big data signifie grosses données ou encore données massives. Ils désignent un ensemble très volumineux de données qu'aucun outil classique de gestion de base de données ou de gestion de l'information ne peut vraiment travailler.

En effet, nous procréons environ **30 fois plus de données** seront générées d'ici **2020**. Ces données sont baptisées Big Data ou volumes massifs de données.

Brièvement : Big data est un Énorme volume de données structurées et non structurées, difficilement gérables avec des solutions classiques de stockage et de traitement, ces données proviennent de sources diverses.

1.2 Les caractéristiques de Big Data

- **Volume**

Derrière le terme « Big » se cache un volume de données, jamais atteint jusqu'à aujourd'hui. Le flux de données étant continu, le volume ne fait que croître chaque jour. Le volume décrit la quantité de données générées par des entreprises ou des personnes. Le Big Data est généralement associé à cette caractéristique

- **Vitesse**

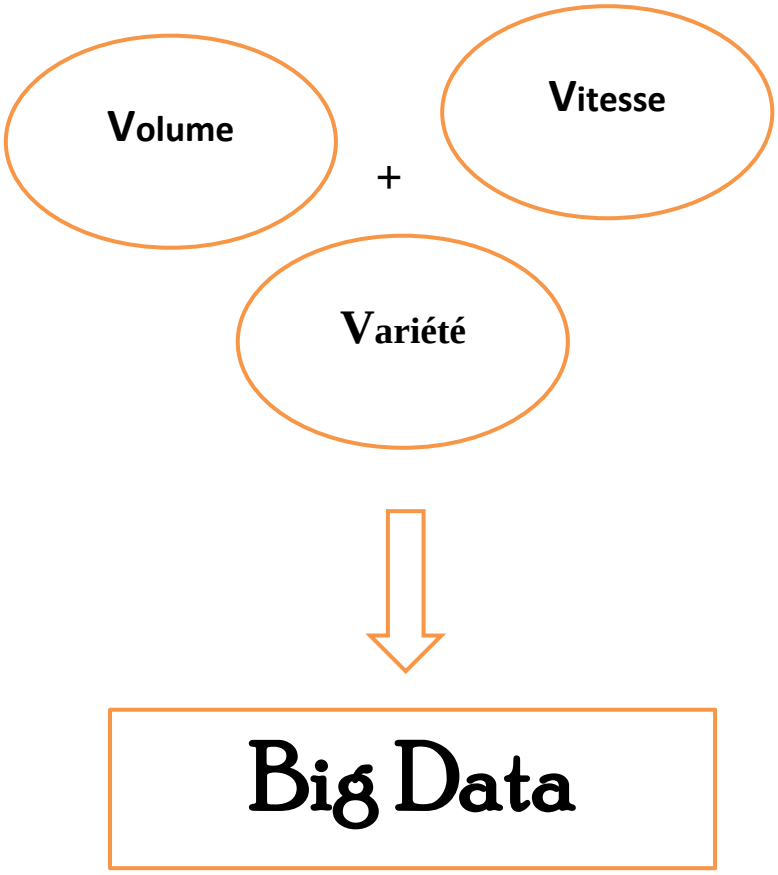
La vitesse décrit la fréquence à laquelle les données sont générées, capturées et partagées. Du fait des évolutions technologiques récentes, les consommateurs mais aussi les entreprises génèrent plus de données dans des temps beaucoup plus courts.

- **Variété**

La prolifération de types de données provenant de sources comme les médias sociaux, les interactions Machine to Machine et les terminaux mobiles, crée une très grande diversité.

Le Big Data se définit selon les trois ‘V’, volume, vitesse et variété.

A ces « 3V » on ajoute un quatrième ‘V’ qui est la valeur que l’on pourra attribuer à ce volume et cette variété de données.



Plusieurs outils arrivent aujourd’hui à maturité et permettent de traiter ces volumes variés de données. Dès le début des années 2000, les géants du Web tels que Google, Yahoo, Facebook, ont été les premiers confrontés à cette masse de données à traiter. Pour ce faire, ils ont imaginé des solutions répondant aux exigences suivantes :

- Traitement, stockage et intégration de données volumineuses, structurées ou non, en fonction des diverses sources de données existantes

- Architecture parallèle, basée sur la technique des clusters (grappe). Il s'agit de mettre ensemble plusieurs serveurs pour répartir la charge d'une tâche ou bien d'utiliser plusieurs serveurs, en même temps, pour effectuer des calculs et ainsi être plus rapide.
- Analyse, en temps réel, de données non structurées
- Performance et haute disponibilité des applications et de l'accès aux données.

La plupart de ces solutions font appel à un duo « Stockage – Traitement »

1.3 Les structures de données

On distinguera des données structurées et des données non structurées :

- ❖ données structurées : ce sont des données qui peuvent être organisées sous formes de tableau. Ces données peuvent être affichées par un tableur et contiennent des lignes et des colonnes de variables, variables dont l'ensemble des valeurs possibles peuvent être déterminés. Exemple : les âges d'une population. De plus, les bases de données structurées peuvent être aisément manipulés.
- ❖ données non structurées : les textes issus de PDF, documents textes, des fichiers audio, des images, des messages issus de discussions instantanées… Ce sont des données qui semblent plus difficiles à catégoriser.

Un exemple pour illustrer ce concept de données structurés et non structurés : Dans un mail, l'adresse mail du destinataire, la date sont des données structurées, tandis que le corps du message est une donnée non structurée.

« Le monde numérique contiendrait seulement 5% de données structurées pour 95% de données non structurées ».

L'intérêt du recours au Big Data est d'analyser les données non structurées car celles-ci sont de plus en plus produites par le développement des échanges sur internet. En effet, de plus en plus de données liées aux sites web, aux mails, aux réseaux sociaux

sont créés. C'est dans ce type de données qu'on peut identifier des informations pertinentes pour la sécurité, l'usage commercial, ou pour la recherche scientifique.

Pour mieux saisir les chiffres donnés par la suite, le tableau suivant donne les correspondances de volume :

Nom	Symbole	Conversion
Kilo-octet	Ko	1 Ko=1000 octet
Méga-octet	Mo	1 Mo=1000 Ko
Giga-octet	Go	1 Go=1000 Mo
Téra-octet	To	1 To=1000 Go
Péta-octet	Po	1 Po=1000 To
Exa-octet	Eo	1 Eo=1000 Po
Zetta-octet	Zo	1 Zo=1000 Eo

Les correspondances de volume pour l'unité octet

1.4 Big Data en chiffre

Par jour

- 144.8 milliards d'Email.
- **Facebook** génère **10 téraoctets** de données par jour
- 340 millions tweets
- 684 000 bits de contenu partagé sur Facebook.

Par minute

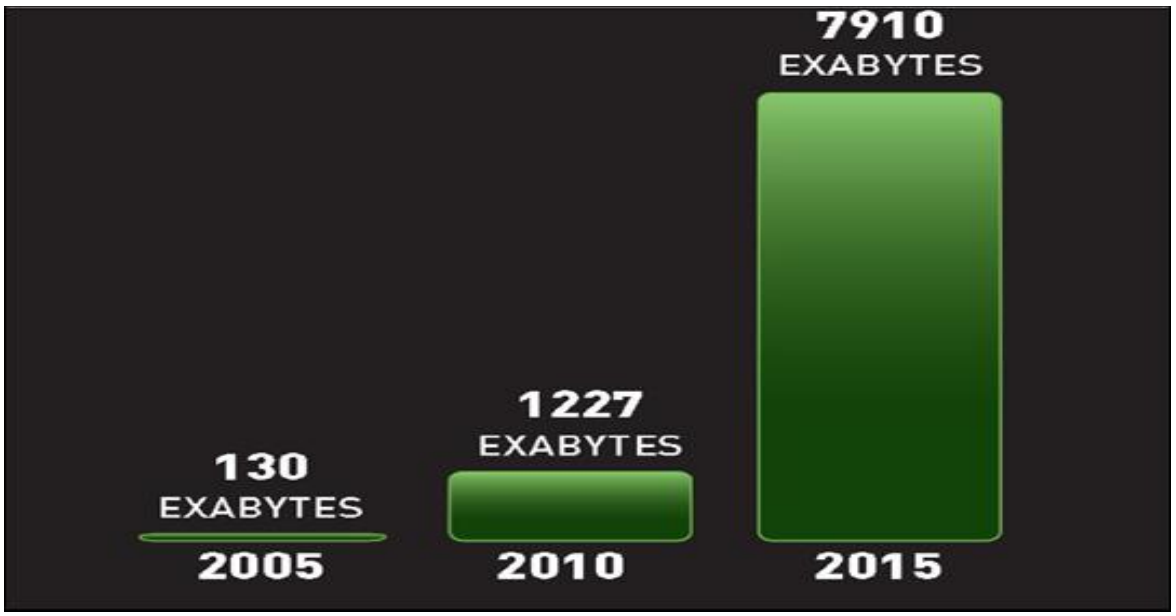
- 72 heures de vidéo sont partagées sur YouTube.

● 2 millions de recherches sur Google.

● 34 000 “likes” des marques sur Facebook.

● 3 600 nouvelles photos sur Instagram.

● 571 nouveaux sites web



Prévisions du volume mondial de données créées

Ce volume devrait atteindre entre 35 et 40 zéttaoctets en 2020.

1.5 Quand est ce qu’on peut Parler de Big Data ?

Bien évidemment, il n’existe pas de définition universelle, et la bonne réponse est “ça dépend”. En fait, d’un point de vue pratique, et dans la plupart des discussions relatives à cette thématique, les Big Data se caractérisent par des jeux de données très volumineux, de l’ordre de plusieurs giga-octets à quelques téraoctets.

2. Les secteurs d'utilisation des Big Data

2.1 Secteur de la santé

L'analyse efficace et en temps réel des Big Data a déjà fait ses preuves dans le domaine de la santé. En effet, plusieurs modèles ont été testés pour améliorer le service médical privé et public, de même que la qualité de vie des patients, et ce, dans différents pays. Big Data peut encore révolutionner le domaine de la santé, non seulement en soutenant l'optimisation des services opérationnels, mais aussi en offrant des outils d'aide à la décision plus efficaces et en diminuant les coûts importants de ce secteur. En bref, l'exploitation et l'intégration adéquate de larges sources de données médicales apportent plusieurs opportunités notables, en particulier:

-L'optimisation des services et des dépenses médicaux : L'analyse du Big Data aide les organismes œuvrant dans le secteur de la santé à mieux détecter les services nécessitant une réorganisation et à suivre en temps réel la qualité des services rendus et la performance des unités médicales, de même que leurs besoins en approvisionnement humain et matériel.

- La personnalisation des services médicaux : A titre d'exemple, en exploitant l'analyse des données en temps réel, des modèles médicaux permettent de suivre à distance l'état des patients pour ajuster les doses ou faire des recommandations selon les symptômes relevés.

-Une meilleure prévention : Grâce à l'analyse avancée des flux de données cliniques collectés dans le secteur public et privé, les modèles prédictifs du Big Data peuvent aider à mieux planifier les moyens de prévention et à soutenir la gestion des épidémies, en particulier la détection précoce des signes alarmants touchant la santé de la population. Cela aide les décideurs à élaborer des plans de réponses optimisés selon le besoin de chaque région et selon la gravité des symptômes des individus.

- L'intégration de plusieurs sources médicales distribuées et hétérogènes constitue un défi de taille, afin de réussir ce pari et de mieux exploiter les opportunités du Big Data dans le secteur de la santé.

2.2 Secteur d'agriculture

L'accès aux flux de données reliées à l'agriculture provenant de plusieurs sources (capteurs intelligents, caméras, agriculteurs, données sur le climat, etc.) permettrait d'améliorer la productivité des terres agricoles, de planifier des stratégies efficaces de protection ou d'approvisionnement et de mieux suivre la demande du marché par région et par type de clients. Par exemple, un projet japonais vise à développer un system avancé d'analyse afin de recommander aux utilisateurs finaux, selon leurs préférences ou symptômes, la meilleure combinaison de produits alimentaires, les restaurants offrant le menu répondant aux exigences, et les producteurs offrant les produits désirés (tels que les produits bio). Le système vise à interconnecter les parties prenantes à travers une plateforme commune intégrant les données provenant de plusieurs acteurs (utilisateurs, restaurants, producteurs agricoles). Ce système permettrait d'accéder aux informations utiles par profil et une interaction entre ces acteurs.

2.3 Secteur du tourisme

Plusieurs modèles Big Data ont vu le jour ou sont en émergence dans l'optique d'améliorer les activités touristiques et de mieux servir les touristes. Grâce, par exemple, aux données géographiques, il est possible d'extraire des informations pour mieux comprendre le comportement et les préférences des touristes et améliorer les services privés et publics associés. De plus, il est possible d'envoyer aux touristes en temps réels des recommandations de sites à visiter et des activités... Les recommandations sont le fruit d'une combinaison d'une part des résultats d'analyse de l'historique des tendances des touristes précédents et d'autre part des résultats d'analyse en temps réels du comportement du touriste, son profil, sa géolocalisation et les sites visités.

2.4 Secteur du transport

L'application des technologies du Big Data au domaine du transport apporte plusieurs avantages. En effet, en ayant la possibilité d'analyser efficacement et rapidement une panoplie de flux de transport, il est possible d'améliorer la satisfaction des passagers et d'offrir des outils d'aide en temps réels aux conducteurs, en considérant plusieurs facteurs : localisation et destination, prévisions climatiques, préférences des clients, historiques et tendances, etc. Cela permettrait d'optimiser les circuits des taxis entre les villes et les aéroports, réduire le temps d'attente, identifier les périodes de pénuries ou d'excès des moyens de transport en commun, créer un équilibre dans la répartition des moyens de transport selon les prédictions de l'offre et de la demande. Actuellement, les modèles émergents permettent de tester l'efficacité et l'impact des politiques de transport existantes et de formuler des recommandations aux différents acteurs de transports (managers, chauffeurs de taxis, voyageurs, etc.). Le but étant d'améliorer les politiques, optimiser les services et maximiser les profits. En résumé et vu le rôle important de ce secteur, l'exploitation des technologies du Big Data et des nouvelles méthodes d'analyse peut non seulement augmenter la satisfaction des voyageurs et soutenir la prise de décision mais aussi de maximiser les profits tout en créant une synergie optimale entre les différents acteurs de ce secteur.

2.5 Secteur des technologies

Les techniques et les méthodes d'analyse conventionnelles sont généralement coûteuses et ne sont pas bien adaptées pour analyser et supporter le traitement rapide de larges sources de données (terabytes à zetabytes, voir plus) en perpétuelle évolution. Face à cette réalité, la révolution du Big Data a poussé les industries à :

- développer des technologies plus performantes et moins coûteuses offrant de nouvelles possibilités pour nettoyer, sauvegarder, traiter et analyser efficacement et en temps réel de très grands volumes de flux de données. Ces technologies telles que les bases de données NoSQL, l'exploration des flux de données, le clustering et le traitement complexe des événements permettent de traiter des données hétérogènes, non structurées, et de différents formats.

- développer de nouvelles plateformes et applications plus performantes et moins coûteuses pour mieux gérer, contrôler et analyser les informations de la sécurité. En effet, alors que le traitement des événements de sécurité prenait de 20 min à 1h à travers les logicielles traditionnelles. Le même traitement prend environ 1 min grâce au système d'analyse de Hadoop. Auparavant, la majorité des événements de sécurité devaient être supprimés après 60 jours par faute de moyens efficaces de sauvegarde. Grâce aux technologies du Big Data, il est possible, dans une vision d'analyse à long terme, de sauvegarder un large volume des événements de sécurité pour une plus long durée. Cette option permet de pouvoir combiner les résultats d'analyse des flux d'information (données en mouvement) à ceux des historiques (données statiques). Par conséquent, les outils du Big Data améliorent la détection des failles et des menaces de sécurité de même que l'analyse rapide et efficace des événements provenant de plusieurs sources (Pare-feu, caméra de surveillances, événements d'achats par cartes Visa aux magasins, transactions en ligne, événements de processus d'affaires, etc.).

2.6 Secteur de Marketing

C'est tout le secteur qui se trouve renouvelé : le Big Data permet en effet aux professionnels du secteur de connaître leur client « à 360° », c'est-à-dire à la fois par son parcours internet mais également par ses achats en magasin ou ses préférences affichées sur les réseaux sociaux.

2.7 Secteur de commerce

C'est, sans doute le domaine, qui possède le plus de données personnelles sur ses clients et, donc le plus d'expérience, sur la segmentation et la personnalisation du besoin. La conservation des historiques de courses permet de faire des propositions d'articles similaires à ce que le client a l'habitude de consommer. C'est le principe des sites de commerce, qui personnalisent leurs propositions en fonction des articles que le client a déjà consultés ou achetés sur leurs sites.

2.8 Secteur industriel

Ce secteur utilise, énormément, de données dans différents domaines.

Le monde du digital et des objets communiquant va accroître ce volume de données et poser la question de leur utilisation.

Un exemple qui prend de l'ampleur, au niveau mondial, est celui des « smart grids ». Il s'agit de réseaux électriques « intelligents » publics, auxquels sont ajoutés des fonctionnalités issues des nouvelles technologies de l'information et de la communication. Le but est d'assurer l'équilibre entre l'offre et la demande d'électricité, à tout instant, et de fournir un approvisionnement sûr, durable et compétitif aux consommateurs. La pièce centrale de ce dispositif, étant un compteur, qui envoie des données, pour permettre d'ajuster cette fourniture en électricité.

3 Enjeux de Big Data

3.1 Enjeux techniques

Il existe essentiellement trois types de défis techniques autour du big data :

- Le stockage et la gestion des données massives, de l'ordre de la centaine de téraoctets ou du pétaoctet, qui dépassent les limites courantes des bases de données relationnelles classiques du point de vue du stockage et de la gestion des données.
- La gestion des données non-structurées (qui constituent souvent l'essentiel des données dans les scénarios Big Data), c'est-à-dire comment organiser du texte, des vidéos, des images, etc…
- L'analyse de ces données massives, à la fois pour le reporting et la modélisation prédictive avancée, mais également pour le déploiement.

3.2 Enjeux économiques

D'après le cabinet de conseil dans le marketing IDC, « le marché du Big Data représentera 24 milliards de dollars en 2016, avec une part de stockage estimée à 1/3 de ce montant ». Il va sans dire que la « donnée » est le nouvel or noir du siècle présent, les spécialistes s'accordent déjà sur le fait que le Big Data sera l'arme économique de demain pour les entreprises et se présentera comme un levier qui fera la différence.

Les entreprises collectent de plus en plus d'information en relation avec leurs activités (production, stockage, logistique, ventes, clients, fournisseurs, partenaires, etc), toutes ces informations peuvent être stockées et exploitées pour stimuler leur croissance.

Les Big Data permettent :

- D'améliorer les stratégies marketing et commerciale
- D'améliorer et entretenir la relation client
- De fidéliser la clientèle
- De gagner de nouvelles parts de marché
- De réduire les coûts logistiques
- De favoriser la veille concurrentielle
- Le client est un acteur majeur dans ce contexte. Jusqu'à présent, la vente consistait à se demander « J'ai un produit, à qui vais-je pouvoir le vendre? ». A l'ère du Big Data, nous devons changer le paradigme pour dire « J'ai un client, de quoi a-t-il besoin aujourd'hui ?». En connaissant mieux son public, à travers ses achats, ses activités sur Internet, son environnement, les commerçants peuvent améliorer l'expérience-client, exploiter la recommandation, imaginer le marketing prédictif (le marketing prédictif regroupe les techniques de traitement et de modélisation des comportements clients qui permettent d'anticiper leurs actions futures à partir du comportement présent).

3.3 Enjeux juridiques

Le principal enjeu juridique reste la protection de la vie privée.

II. Les techniques statistiques utilisées pour l'analyse de Big Data

1. Le statisticien et le Big Data

Les dix dernières années ont connu une modification importante du volume et de la nature des données auxquelles le statisticien est confronté, la statistique est un élément important dans les Big Data parce que de nombreuses méthodes statistiques sont utilisées pour l'analyse des données massives.

L'évolution rapide des systèmes d'information génère des données de plus en plus volumineuses à causer de profonds changements de paradigme dans le travail de statisticien.

Les techniques statistiques classiques fonctionnent souvent bien avec des volumes de données moins importants, cependant dès que le volume de données devient massif, il y a un certain nombre de problèmes qui apparaissent.

2. La classification

Classifier, c'est regrouper entre eux des objets similaires selon certains critères. Les diverses techniques de classification visent toutes à répartir n individus, caractérisés par p variables X_1, X_2, ..., X_p en un certain nombre m de sous-groupes aussi homogènes que possible, chaque groupe étant bien différencié des autres.

Brièvement, la classification est une méthode mathématique d'analyse de données : pour faciliter l'étude d'une population d'effectif important, on les regroupe en plusieurs classes de telle sorte que les individus d'une même classe soient le plus semblables possible et que les classes soient le plus distinctes possibles. Pour cela il y a diverses façons de procéder.

Domaines d'application

Le Clustering s'applique dans plusieurs domaines :

Domaine	Formes de données	Clusters
Text mining	Textes Mails	Textes proches Dossiers automatiques
Web mining	Textes et images	Pages web proches
BioInformatique	Gènes	Gènes ressemblants
Marketing	Infos clients, produits achetés	Segmentation de la clientèle
Segmentation d'images	Images	Zones homogènes dans l'image

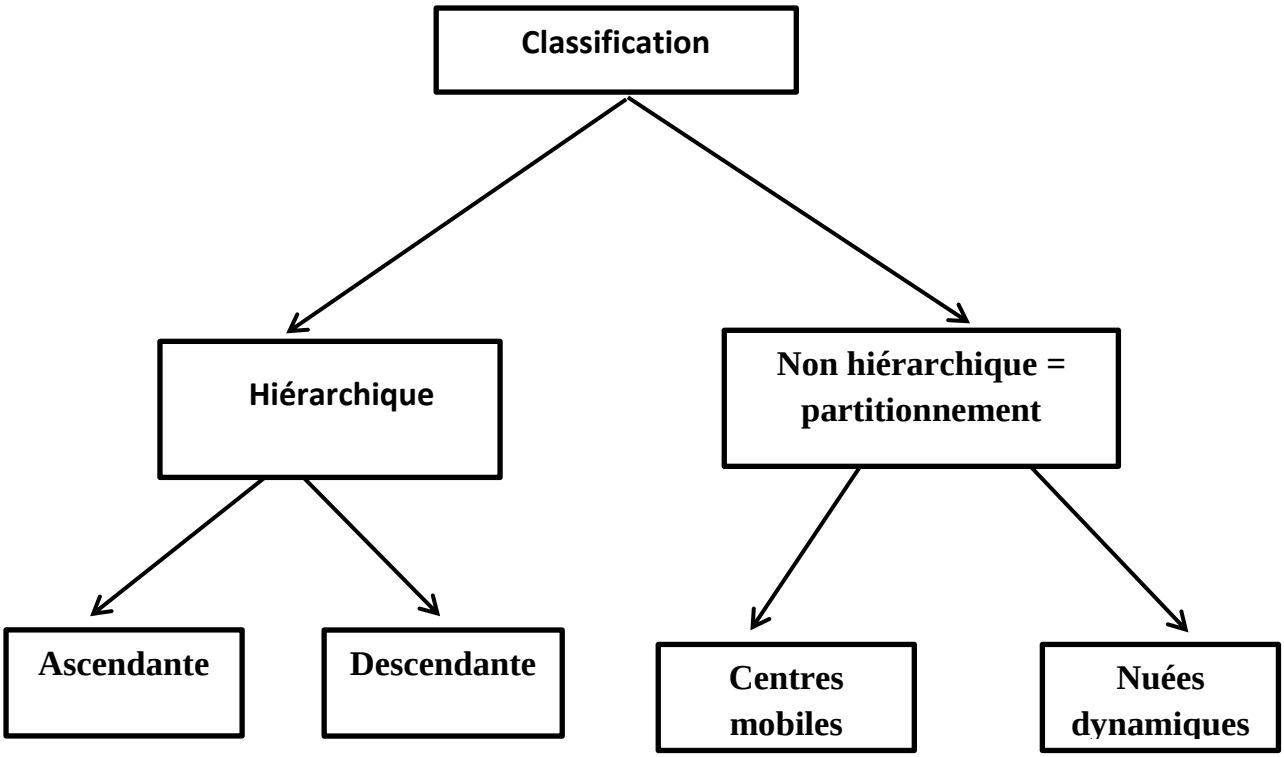
Objectif de la classification

L'objectif de la classification automatique est de former des groupes d'individus ou de variables a fin de structurer un ensemble de données. On cherche souvent des groupes homogènes c'est à dire que les objets d'une même classe doivent être « similaires » et les objets de deux classes différentes doivent être «distincts».

Les méthodes de classification se distinguent entre autre par la structure de classification obtenue (partition, hiérarchie…etc.).

Les méthodes de classification

- Méthodes Hiérarchiques (classification hiérarchiques)
- Partitionnement (k-means.)



2.1 Algorithme de classification hiérarchique ascendante(CAH)

CAH : L'idée de l'algorithme de Classification Ascendante Hiérarchique (CAH) est de créer, à chaque étape, une partition de T= $\{w_1, w_2, \ldots w_n\}$ en regroupant les deux éléments les plus proches. Le terme "élément" désigne aussi bien un individu qu'un groupe d'individus.

Objectif :

- On veut mettre en relief les liens hiérarchiques entre les individus ou groupe d'individus.
- Détecter les groupes d'individus qui se démarquent le plus.

L'algorithme de CAH est décrit ci-dessous :

- On choisit un écart. On construit le tableau des écarts pour la partition initiale des n individus de T : $P_0 = (\{w_1\}, \{w_2\} \dots \{w_n\})$
  Chaque individu constitue un élément.
- On parcourt le tableau des écarts pour identifier le couple d'individus ayant l'écart le plus petit. Le regroupement de ces deux individus forme un groupe A. On a donc une partition de T de n - 1 éléments : A et les n - 2 individus restants.
- On calcule le tableau des écarts entre les n - 1 éléments obtenus à l'étape précédente et on regroupe les deux éléments ayant l'écart le plus petit (cela peut être deux des n-2 individus, ou un individu des n-2 individus restants avec A). On a donc une partition de T de n-2 éléments.

On itère la procédure précédente jusqu'à ce qu'il ne reste que deux éléments.

On regroupe les deux éléments restants. Il ne reste alors qu'un seul élément contenant tous les individus de T.

Exemple :

On considère la matrice de données X dans $\mathbb{R}^2$ définie par :

$$X = \begin{pmatrix} 2 & 2 \\ 7.5 & 4 \\ 3 & 3 \\ 0.5 & 5 \\ 6 & 4 \end{pmatrix}$$

On va regrouper les individus avec l'algorithme CAH et la méthode du voisin le plus éloigné munie de la distance euclidienne.

Le tableau des écarts associé à $P_0 = (\{w_1\}, \{w_2\} \ldots \{w_5\})$ est :

	W_1	W_2	W_3	W_4	W_5
W_1	0	5.85	**1.41**	3.35	4.47
W_2	5.85	0	4.60	7.07	1.5
W_3	**1.41**	4.60	0	3.20	3.16
W_4	3.35	7.07	3.20	0	5.59
W_5	4.47	1.5	3.16	5.59	0

Les éléments (individus) w_1 et w_3 ont l'écart le plus petit : ce sont les éléments les plus proches. On les rassemble pour former le groupe : $A = \{w_1; w_3\}$. On a une nouvelle partition de T :

$$P_1 = (\{w_2\}, \{w_4\}, \{w_5\}, A).$$

Le tableau des écarts associé à P_1 est :

	W_2	W_4	W_5	A
W_2	0	7.07	1.5	5.85
W_4	7.07	0	5.59	3.35
W_5	1.5	5.59	0	4.47
A	5.85	3.35	4.47	0

On a

$$e(w_2;A) = \max(e(w_2; w_1); e(w_2; w_3)) = \max(5.85; 4.60) = 5.85$$

$$e(w4;A) = \max(e(w_4; w_1); e(w_4; w_3)) = \max(3.35; 3.20) = 3.35$$

Et

$$e(w_5;A) = \max(e(w_5; w_1); e(w_5; w_3)) = \max(4.47; 3.16) = 4.47$$

Les éléments (individus) w_2 et w_5 sont les plus proches. On les rassemble pour former le groupe : $B = \{w_2; w_5\}$.

On a une nouvelle partition de T : $P_2 = (\{w_4\}; A; B)$

Le tableau des écarts associé à P_2 est :

	W_4	A	B
W_4	0	3.35	7.07
A	3.35	0	5.85
B	7.07	5.85	0

On a

$$e(B; w_4) = \max(e(w_2; w_4);\quad e(w_5; w_4)) = \max(7.07; 5.59) = 7.07$$

Et

$$e(B;A) = \max(e(w_2;A)$$

$$e(w_5;A) = \max(5.85; 4.47) = 5, 85$$

Les éléments w_4 et A sont les plus proches. On les rassemble pour former le groupe :

$C = \{w_4;A\} = \{w_1; w_3; w_4\}$. On a une nouvelle partition de T :

$P_3 = (B;C)$

Le tableau des écarts associé à P_3 est :

	B	C
B	0	7.07
C	7.07	0

On a

$e(C;B) = \max(e(w_4;B); e(A;B)) = \max(7.07; 5.85) = 7.07$

Il ne reste plus que 2 éléments B et C ; on les regroupe. On obtient la partition

$P_5 = \{w_1, \ldots, w_5\} = T$. Cela termine l'algorithme de CAH.

Au final,

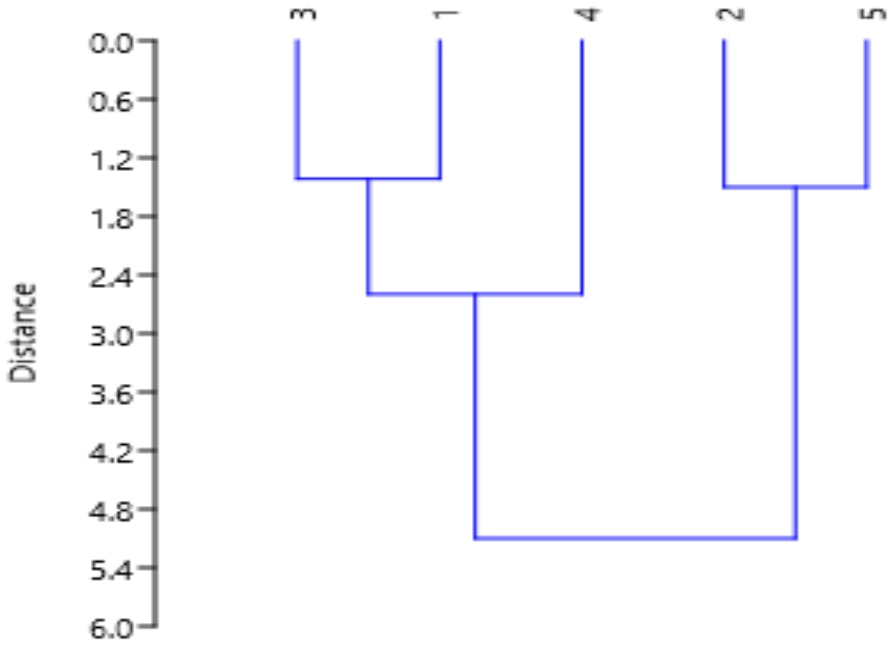
↝ Les éléments $\{w_1\}$ et $\{w_3\}$ ont été regroupés avec un écart de 1.41

↝ Les éléments $\{w_2\}$ et $\{w_5\}$ ont été regroupés avec un écart de 1.50

↝ Les éléments $A = \{w_1; w_3\}$ et $\{w_4\}$ ont été regroupés avec un écart de 3.35

↝ Les éléments $C = \{w_4;A\}$ et $B = \{w_2; w_5\}$ ont été regroupés avec un écart de 7.07

On peut donc construire le dendrogramme associé :



Comme le plus grand saut se situe entre les éléments B et C (on a 7.07-3.35 = 3.72), on propose les deux groupes : B et C.

Remarque :

Cet algorithme est inapplicable dans le cas des données massives.

2.2 Algorithme des centres mobiles (k means)

L'algorithme des centres mobiles vise à classer une population en k classes. Cela se fait de manière automatique ; il n'y a pas de lien hiérarchique dans les regroupements contrairement à l'algorithme CAH. Il est le mieux adapté aux très grands tableaux de données. L'algorithme des centres mobiles avec la méthode de Lloyd (la plus standard) est décrit ci-dessous :

-On choisit q points au hasard dans $\mathbb{R}^p$, Ces points sont appelés centres.

- On calcule le tableau de distances entre tous les individus et les q centres.

- On forme alors q groupes de la manière suivante : chaque groupe est constitué d'un centre et des individus les plus proches de ce centre que d'un autre. On obtient une partition P_1 de P.

- On calcule le centre de gravité de chacun des q sous-nuages de points formés par les q groupes. Ces q centres de gravité sont nos nouveaux q centres.

- On calcule le tableau de distances entre tous les individus et les nouveaux q centres.

- On forme alors q groupes, chaque groupe étant constitué d'un centre et des individus les plus proches de ce centre que d'un autre. On a une nouvelle partition P_2 de p.

- On itère la procédure précédente jusqu'à ce que deux itérations conduisent à la même partition.

Remarque :

La classification des individus dépend du choix des centres initiaux. Plusieurs méthodes existent pour choisir judicieusement ces centres.

Exemple : Dans une étude industrielle, on a étudié 2 caractères : X_1 et X_2, sur 6 individus w_1, w_2,…w_6.

Les données recueillies sont :

	X_1	X_2
W_1	-2	2
W_2	-2	1
W_3	0	-1
W_4	2	2
W_5	-2	3
W_6	3	0

1. Dans un premier temps, on fait une classification par l'algorithme des centres mobiles avec, pour centres initiaux, C_0^1de coordonnées (-1;-1) et C_0^2 de coordonnée (2 ; 3).

2. Dans un deuxième temps, on fait de même avec, pour centres initiaux, C_0^1 de coordonnées (-1, 2) et C_0^2 de coordonnées (1; 1).

- On considère les centres initiaux C_0^1 de coordonnées (-1;-1) et C_0^2 de coordonnées (2;3).

Le tableau des distances entre les individus et ces centres est :

	W_1	W_2	W_3	W_4	W_5	W_6
C_0^1	3.16	1	1	4.24	4.12	4.12
C_0^2	4.12	5.66	4.47	1	4	3.16

Exemple de calcul : d (w_1,C_0^1)=$\sqrt{(-2-(-1))2+(2-(-1))2}$=3.16

D'où les deux groupes : A= {w_1, w_2, w_3} et B= {w_4, w_5, w_6}

On considère deux nouveaux centres, C_1^1 et C_1^2, lesquels sont les centres de gravité des deux groupes A et B.

Donc C_1^1 a pour coordonnées $\left(\frac{(-2-2+0)}{3};\frac{(2-1-1)}{3}\right)=(-1.33;0)$ et C_1^2 a pour coordonnées $\left(\frac{2-2+3}{3};\frac{2+3+0}{3}\right)=(1;1.67)$.

Le tableau des distances entre les individus et ces centres est :

	W_1	W_2	W_3	W_4	W_5	W_6
C_1^1	2.11	1.20	1.66	3.88	3.07	4.33
C_1^2	3.02	4.02	2.85	1.05	3.28	2.61

D'où les deux groupes :

$$A= \{w_1, w_2, w_3, w_5\} \quad \text{et} \quad B= \{w_4, w_6\}$$

On considère deux nouveaux centres, C_2^1 et C_2^2, lesquels sont les centres de gravité des deux groupes A et B.

Donc C_2^1 a pour coordonnées $\left(\frac{-2-2+0-2}{4}; \frac{2-1-1+3}{4}\right) = (-1.5; 0.75)$ et C_2^2 a pour coordonnées $\left(\frac{2+3}{2}; \frac{2+0}{2}\right)$= (2.5 ; 1)

Le tableau des distances entre les individus et ces centres est :

	W_1	W_2	W_3	W_4	W_5	W_6
C_2^1	1.35	1.82	2.30	3.72	2.30	4.56
C_2^2	4.61	4.92	3.20	1.12	4.92	1.12

D'où les deux groupes :

$$A= \{w_1, w_2, w_3, w_5\} \quad \text{et} \quad B= \{w_4, w_6\}$$

On retrouve la même classification que l'étape précédente, on arrête l'algorithme.

3. Régression linéaire

En statistique un modèle de régression linéaire est un modèle de régression d'une variable expliquée sur une ou plusieurs variables explicatives dans lequel on fait l'hypothèse que la fonction qui relie les variables explicatives a la variable expliquée est linéaire dans ses paramètres.

On parle aussi de modèle linéaire ou modèle de régressions linéaire.
En général, le modèle linéaire de régression linéaire désigne un modèle dans lequel l'espérance conditionnelle de Y sachant X est une transformation affine de X. Cependant, on peut aussi considérer des modèles dans lesquels c'est la médiane conditionnelle de Y sachant X ou n'importe quel quantile de la distribution de Y sachant X qui est une transformation affine de X.

Le modèle de régression linéaire est souvent estimé par la méthode des moindres carrés mais il existe aussi de nombreuses autres méthodes pour estimer ce modèle.

Bien qu'ils soient souvent présentés ensemble, le modèle linéaire et la méthode des moindres carrés ne désignent pas la même chose. Le modèle linéaire désigne une classe de modèles qui peuvent être estimés par un grand nombre de méthodes, et la méthode des moindres carrés désigne une méthode d'estimation. Elle peut être utilisée pour estimer différents types de modèles

3.1 Régression linéaire simple

Un modèle de régression linéaire simple est de la forme :

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Où

- Y est la variable dépendante (une variable aléatoire.).
- β_0 et β_1 sont les coefficients (ordonnée à l'origine et pente).
- X est la variable indépendante (variable explicative).
- **ε** est une erreur aléatoire.

L'espérance de Y pour chaque X est le point sur la droite d'équation :

$$E(Y / X) = \beta_0 + \beta_1 X$$

Conditions d'application du modèle linéaire :

- La courbe de régression n'est autre que la courbe joignant les moyennes des distributions des Yi.
- Pour chaque valeur de X, E(X) = 0 et V (X) = σ^2
- **ε~** N (0; σ^2)
- Les erreurs " sont indépendantes (non corrélées).
- Les Yi ne sont pas corrélées.

On cherche à :

- Estimer les paramètres β_0, β_1 et σ^2
- Vérifier si le modèle est adéquat.

3.1.1 Estimation des paramètres β_0 et β_1

Supposons que n paires d'observations (X_1, Y_1); (X_2, Y_2);… ; (X_n, Y_n) ont été faites. Substituant dans le modèle linéaire, on obtient :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \Rightarrow \quad \varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

Les coefficients sont déterminés par la méthode des moindres carrés qui minimise la somme des carrés des erreurs :

L (b0, b1) = Min ($\sum_{i=1}^{i=n} e_i^2$) = $\sum_{i=1}^{i=n}(Y_i - b_0 - b_1 X_i)^2$

On résout le système de deux équations à deux inconnues ∇L(b0, b1)=0.

$$\nabla L(b_0, b_1) = 0 \quad \Rightarrow \quad \begin{cases} \frac{\partial L}{\partial b_0}(b_0, b_1) = \sum_{i=1}^{i=n}(Y_i - b_0 - b_1 X_i) \\ \frac{\partial L}{\partial b_1}(b_0, b_1) = \sum_{i=1}^{i=n} X_i(Y_i - b_0 - b_1 X_i)^2 \end{cases}$$

$$\Rightarrow \quad \begin{cases} b_0 = \overline{Y} - b_1 \overline{X} \\ b_1 = \frac{\sum_i^n X_i Y_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{i=n} X_i^2 - n\overline{X}^2} \end{cases}$$

Avec : $\overline{X} = \frac{1}{n}\sum_1^n x_i$ et $\overline{Y} = \frac{1}{n}\sum_1^n y_i$

Propriété de β_0 et β_1 :

La droite de régression estimée est $\hat{Y} = b_0 + b_1 X$

Les variables aléatoires b_0 et b_1 sont des estimateurs de l'ordonnée à l' origine β_0 et de la pente β_1.

Théorème

E (b_0) = β_0 et E (b_1) = β_1

Estimation du paramètre σ^2 :

La différence entre la valeur estimée $\hat{Y} = b_0 + b_1 X_i$ et la valeur observée Yi est appelée résidu et est dénotée $E_i = \hat{Y}_i - Y_i$

Théorème

- Si σ est connue on a V (b_0) = $\sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{i=n}(X_i - \overline{X})^2}\right]$ et V (b_1) = $\frac{\sigma 2}{\sum_{i=1}^{i=n}(X_i - \overline{X})^2}$
- Si σ est inconnue on a S (b_0) = $S^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{i=n}(X_i - \overline{X})^2}\right]$ et S (b_1) = $\frac{S2}{\sum_{i=1}^{i=n}(X_i - \overline{X})^2}$

Distributions pour b_0 et b_1

Théorème

- Si σ est connue on a : $\frac{b_0-\beta_0}{\sigma(b_0)} \sim N(0,1)$ et : $\frac{b_1-\beta_1}{\sigma(b_1)} \sim N(0,1)$
- Si σ est connue on a : $\frac{b_0-\beta_0}{S(b_0)} \sim T_{n-2}$ et : $\frac{b_1-\beta_1}{S(b_1)} \sim T_{n-2}$

Tests d'hypothèse pour β_0

La distribution

$$\mathbf{t_{exp}} = \frac{b_0-\beta_0}{S(b_0)} \sim T_{n-2}$$

Permet de tester des hypothèses du type

$H_0 : \beta_0 = 0$

$H_1 : \beta_0 \neq 0$

Soient α le seuil de signification du test et b_0 l'estimateur du test.

Sous l'hypothèse H_0 et les conditions d'applications du test :

Règle de décision :

- On rejette H_0 si $\frac{b_0}{S(b_0)} \notin \left[-T_{\frac{\alpha}{2};n-2}; T_{\frac{\alpha}{2};n-2}\right]$
- Sinon on accepte H_0 c-à-d que β_0 n'est pas significatif dans ce cas le modèle sans constante est $Y= \beta_1 X + \varepsilon$ ($Y_i= \beta_1 X_i +\varepsilon_i$). On a alors :

$$b_1 = \frac{\sum_1^n X_i Y_i}{\sum_{i=1}^{i=n} {X_i}^2}, \qquad S^2 = \frac{\sum_{i=1}^{i=n}(Y_i-\hat{Y}_i)^2}{n-1}, \qquad S^2(b_1) = \frac{S^2}{\sum_{i=1}^{i=n}(X_i-\bar{X})^2}$$

3.1.2 Inférence concernant la moyenne de la distribution conditionnelle Y à X = X_h

Distribution d'échantillonnage $\hat{Y}_h$

$\hat{Y}_h = b_0+b_1X_h$ est une distribution normale de moyenne $E(\hat{Y}_h) = \beta_0 + \beta_1 X_h$ et de variance

$$V(\hat{Y}_h) = \sigma 2\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum_{i=1}^{i=n}(X_i-\bar{X})^2}\right]$$

Une estimation de $\sigma^2(\hat{Y}_h)\ est\ S^2(\hat{Y}_h) = S^2\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum_{i=1}^{i=n}(X_i-\bar{X})^2}\right]$

Intervalle de confiance de E ($\hat{Y}_h$) à X = $\widehat{X}_h$: au seuil 1- α

⇝ Si σ est connue $\hat{Y}_h - Z_{\frac{\alpha}{2}}\sigma(\hat{Y}_h) \leq E(\hat{Y}_h) \leq \hat{Y}_h + Z_{\frac{\alpha}{2}}\sigma(\hat{Y}_h)$

⇝ Si σ est inconnue $\hat{Y}_h - T_{\frac{\alpha}{2};n-2}S(\hat{Y}_h) \leq E(\hat{Y}_h) \leq \hat{Y}_h + T_{\frac{\alpha}{2};n-2}S(\hat{Y}_h)$

3.1.3 Analyse de la variance

On détermine dans quelle mesure la droite de régression est utile à expliquer la variabilité existante dans les observations.

Décomposition de la variation :

$$\sum_{i=1}^{i=n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{i=n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{i=n}(Y_i - \hat{Y}_i)^2$$

↓ ↓ ↓

Var. Totale — Var. expliquée Par la droite — Var. inexpliquée Par la droite

On définit :

- La somme des carrés dûe à l'erreur par :

$$SC_{RES} = \sum_1^n e_i^2 = \sum_{i=1}^{i=n}(Y_i - \hat{Y}_i)^2$$

- La somme des carrés dûe à la régression par :

$$SC_R = \sum_{i=1}^{i=n}\left(\hat{Y}_i - \bar{Y}\right)^2 = b_1^2\sum_{i=1}^{i=n}(X_i - \bar{X})^2$$

- La somme des carrés totale par :

$$SC_T = SC_{RES} + SC_R$$

$$= \sum_{i=1}^{i=n} (Y_i - \overline{Y})^2$$

Le coefficient de détermination

$$r^2 = \frac{SC_R}{SC_T} = \frac{\sum_{i=1}^{i=n} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{i=n} (Y_i - \overline{Y})^2}$$

Tableau d'analyse de la variance : ANOVA

Source de variation	Somme de carrée	d.d.l	Carrée moyen
Régression	SC_R	1	$CM_R = \frac{SC_R}{1}$
Résiduelle	SC_{RES}	n-2	$CM_{RES} = \frac{SC_{Res}}{n-2}$
Totale	SC_T	n-1	

Signification de la régression

Il s'agit de tester les hypothèses :

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Accepter H_0 implique que l'on conclut qu'il n'y a pas cause à effet entre X et Y. Ceci peut signifier que

- La relation entre X et Y n'est pas linéaire.

- La variation de X influe peu ou pas sur la variation de Y.

Au contraire, rejeter H_0 implique que l'on conclut que la variation de X influe sur la variation de Y.

Le critère est : rejeter H_0 au seuil de signification α si $F_{exp} = \frac{CM_R}{CM_{RES}} > F_{\alpha;1;n-2}$.

3.2 Régression linéaire multiple

Un modèle de régression linéaire multiple est de la forme :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k + \varepsilon$$

Où :

- ✓ Y est la variable dépendante (une variable aléatoire).
- ✓ $\beta_0, \beta_1, \ldots \beta_k$ Sont les k + 1 paramètres du modèle.
- ✓ X_i représente l'ième valeur des k variables explicatives. On les considère comme des grandeurs certaines.
- ✓ ε est la fluctuation aléatoire non observable.

Hypothèse fondamentales du modèle de régression multiple

1. ε est une v.a de moyenne 0 et de variance constante = σ^2

2. Il n'existe aucune corrélation entre les ε_i ; cov $(\varepsilon_i, \varepsilon_j)$=0 $\quad \forall\, i \neq j$

3. $X_1, \ldots, X_n$ sont des grandeurs certaines.

4. Les ε_i sont distribuées normalement $\quad \varepsilon_i \sim N(0, \sigma^2) \quad \forall\, i$

3.2.1 Estimation des paramètres $\beta_i \ \forall\, i = 1, \ldots, n$

On se donne deux n-échantillons (X_n) et (Y_n) qui ne sont pas mutuellement indépendants où :

- (X_n) forme une suite de vecteurs de dimension p > 1
- Xi = $(X_{i1}, \ldots, X_{ik})$ est la i^{iem} composante de (X_n).

Le modèle s'écrit alors :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i \qquad i = 1, \ldots, n$$

Forme matricielle de régression multiple

$$Y = \beta X + \varepsilon$$

$$Y=\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix},\quad X=\begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ & & . & & \\ & & . & & \\ & & . & & \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix},\quad \beta=\begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{pmatrix},\quad \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}.$$

Nous obtenons une estimation de B à l'aide de l'égalité suivante :

$B=(X^TX)^{-1}X^TY$

Où X^T est la matrice transposée de X.

Une estimation de $E(Y_i) = b_0 + b_1X_{i1} + \cdots + b_kX_{ik}$

$$\hat{Y}_i = b_0 + b_1X_{i1} + \cdots + b_kX_{ik}$$

Et $V(Y_i)=\sigma^2$

Soit $e_i = Y_i-\hat{Y}_i$

En appliquant la méthode du moindre carrées qui consiste à minimiser la somme des carrés résiduelle.

3.2.2 Analyse de la variance en régression multiple

Calcule des sommes de carrés en régression multiple

Source de variation	Somme des carrés	d.d.l	Carré moyen
Expliquée par la régression	$SC_R = \sum_{i=1}^{i=n}(\hat{Y}_i - \bar{Y})^2$	K	$CM_R = \frac{SC_R}{k}$
Résiduelle	$SC_{RES} = \sum_{i=1}^{i=n}(Y_i - \hat{Y}_i)^2$	n-k-1	$CM_{RES} = \frac{SC_{RES}}{n-k-1}$
Totale	$SC_T = \sum_{i=1}^{i=n}(Y_i - \bar{Y})^2$	n-1	

3.2.3 Estimation de σ^2

CM_{RES} est une estimation ponctuelle de σ^2

$$CM_{RES} = S^2 = \frac{\sum_{i=1}^{i=n}(Y_i - \hat{Y}_i)^2}{n-k-1}$$

Le coefficient de détermination multiple R^2 :

$$R^2 = \frac{SC_R}{SC_T} = \frac{\sum_{i=1}^{i=n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{i=n}(Y_i - \bar{Y})^2}$$

Expressions matricielles pour les sommes des carrés et carrés moyens :

- ◊ La variation totale : $SC_T = Y^TY - n\bar{Y}^2$ avec n-1 d.d.l.
- ◊ Somme des carrées due à la régression : $SC_R = B^TX^TY - n\bar{Y}^2$ avec k d.d.l.
- ◊ Sommes des carrées résiduelle: $SC_{RES} = SC_T - SC_R = Y^TY - B^TX^TY$ Avec n-k-1 d.d.l.
- ◊ Le carré dû à la régression : $CM_R = \frac{SC_R}{k} = \frac{B^TX^TY - n\bar{Y}^2}{k}$

◊ Le carré résiduel : $CM_{Res} = \frac{SC_{RES}}{n-k-1} = \frac{Y^T Y - B^T X^T Y}{n-k-1} = S^2$

3.2.4 Test de signification de la régression dans son ensemble

Nous voulons tester si la régression est significative dans son ensemble

$$H_0 \quad : \quad \beta_0 = \beta_1 \quad ... \quad \beta_k = 0$$

$$H_1 \quad : \quad \text{au moins un des } \beta_j \text{ est différent de 0}$$

$F_{exp} = \frac{CM_R}{CM_{Res}} \sim$ Fisher à k et n - k - 1 d.d.l.

Au seuil de signification α, sous l'hypothèse H_0 et les conditions d'application du test

Règle de décision :

On rejette H_0 si $F_{exp} > F_{\alpha,k,n-k-1}$

Si on décide de favoriser H_1, nous concluons que la contribution de l'ensemble des variables pour expliquer les fluctuations de Y est significative.

3.2.5 Contribution marginale et estimation par intervalle

Test de signification de chaque variable :

Sous les hypothèses de normalité des erreurs, $b_j \sim N(\beta_j, \sigma^2)$

$$H_0 \quad : \quad \beta_j = 0$$

$$H_1 \quad : \quad \beta_j \neq 0$$

Soit α le seuil de signification du test, sous H_0 et les conditions d'application du test, on a :

$$\frac{b_j}{S(b_j)} \sim T_{n-k-1}$$

Règle de décision :

$$t_{exp} = t_j = \frac{b_j}{S(b_j)}$$

↝ Si $t_{exp} \in \left[-t_{\frac{\alpha}{2};n-k-1}, t_{\frac{\alpha}{2};n-k-1}\right]$ on accepte H_0.

↝ Sinon on rejette H_0 et on accepte H_1, dans ce cas on dit que l'effet de la variable X_j est significatif.

La variance S^2 (b_j) s'obtient de la matrice des variances covariances des coefficients de régression.

$$CM_{RES}(X^TX) = \begin{pmatrix} S^2(b_0) & cov(b_0,b_1) & \dots & cov(b_0,b_k) \\ cov(b_0,b_1) & S^2(b_1) & \dots & cov(b_1,b_k) \\ & & . & \\ & & . & \\ & & . & \\ & & . & \\ cov(b_k,b_0) & cov(b_k,b_1) & \dots & S^2(b_k) \end{pmatrix}$$

Intervalle de confiance des β_j : au seuil $1 - \alpha$

$$b_j - t_{\frac{\alpha}{2};n-k-1} S^2(b_j) \leq \beta_j \leq b_j + t_{\frac{\alpha}{2};n-k-1} S^2(b_j)$$

3.2.6 Estimation de $E(Y_h)$ par intervalle de confiance

$$\hat{Y}_h - t_{\frac{\alpha}{2};n-k-1} S(\hat{Y}_h) \leq E(Y_h) \leq \hat{Y}_h + t_{\frac{\alpha}{2};n-k-1} S(\hat{Y}_h)$$

Ou S ($\hat{Y}_h$) est l'écart type de $\hat{Y}_h$

Intervalle de prévision pour Y_h :

$$\hat{Y}_h - t_{\frac{\alpha}{2};n-k-1} S(d_h) \leq Y_h \leq \hat{Y}_h + t_{\frac{\alpha}{2};n-k-1} S(d_h)$$

$d_h = Y_h - \hat{Y}_h$ et $S^2(d_h) = S^2 + S^2(\hat{Y}_h) = S^2\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{i=n}(X_i - \bar{X})}\right]$

Variance $S^2(\hat{Y}_h)$ sous forme matricielle :

$$S^2(\hat{Y}_h) = (X_h^T((X^TX))^{-1}X_h)\, CM_{Res} \quad \text{ou} \quad X_h = \begin{pmatrix} 1 \\ X_{h1} \\ . \\ . \\ . \\ X_{hk} \end{pmatrix}$$

3.3 La régression linéaire divisée

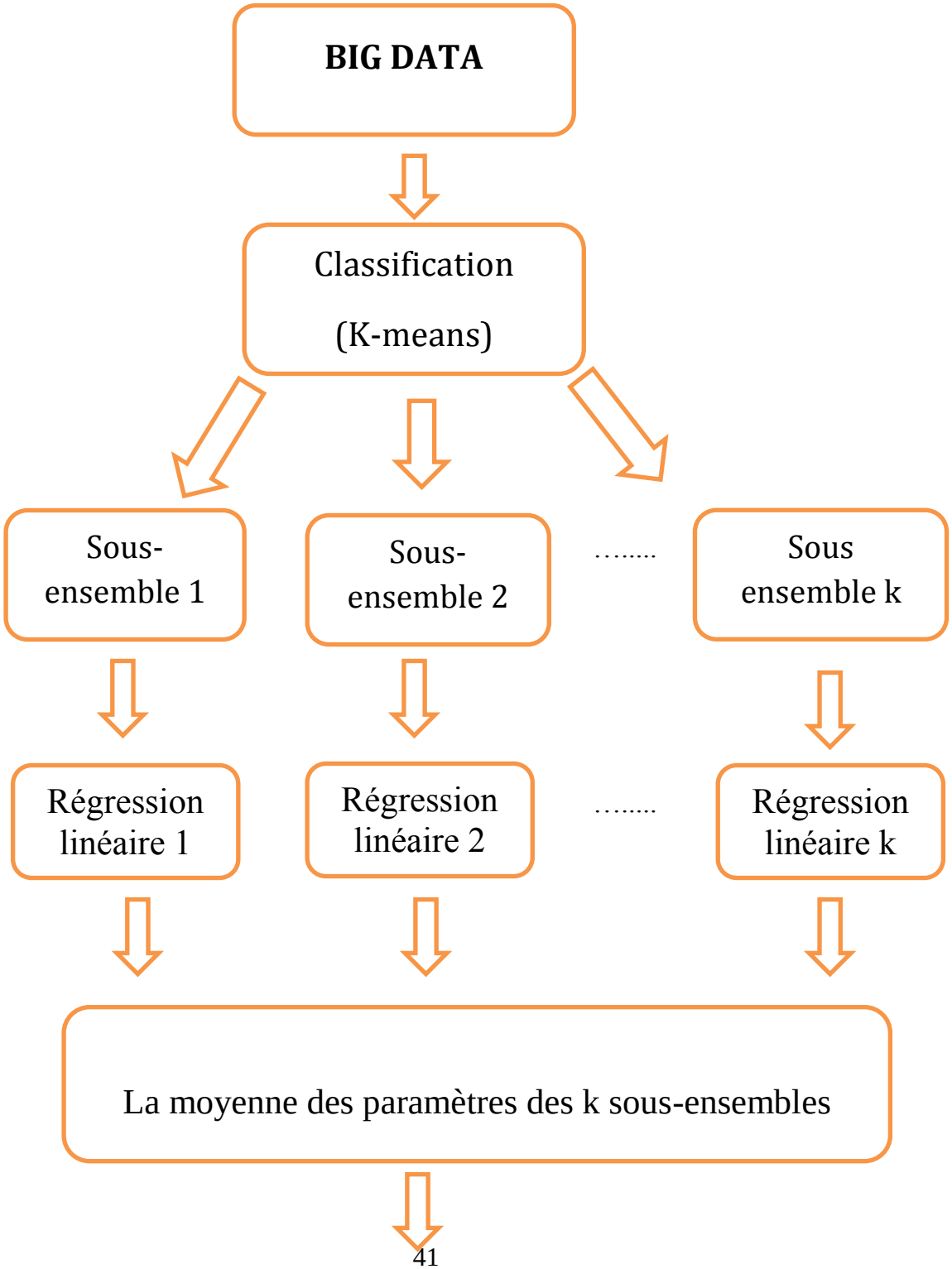
À l'ère des grandes données, nous pouvons obtenir des données énormes. Mais les statistiques traditionnelles ont besoin de beaucoup de temps pour l'analyse de ces grandes données. Pour résoudre ce problème, nous proposons une approche qui a comme principe de diviser les données de grande taille en des sous-ensembles de données de petite taille.

Dans ce travail, nous choisissons la régression linéaire comme modèle statistique pour l'analyse de données importantes.

Pour surmonter le problème de calcul dans la régression Big Data, nous proposons une analyse de régression divisée, qui divise les données globale en n ensembles de sous-données.

L'analyse de régression divisée, divise les données globales en n ensembles de sous-données. En outre, nous appliquons ce partitionnement de données pour estimer les paramètres dans le modèle de régression.

Pour divisé les donnée Big Data on a plusieurs techniques de partitionnement, dans notre travail en s'intéresse à la méthode de K-means).


BIG DATA
Classification
(K-means)
Sous-
ensemble 1
Sous-
ensemble 2
……..
Sous
ensemble k
Régression
linéaire 1
Régression
linéaire 2
……..
Régression
linéaire k
La moyenne des paramètres des k sous-ensembles

Résultats de la régression linéaire divisée

Procédure de la régression linéaire divisée de Big Data

III. Résultats expérimentaux de l'exemple 1

Une étude américaine concerne 237 enfants décrits par leurs âges en mois, leurs tailles en cm et leurs poids en Kg.

On veut expliquer le poids à partir de l'âge et la taille.

Pour surmonter le problème de calcul dans la régression Big Data, nous proposons une analyse de régression divisée.

enfant i	Age (mois) X1	taille (cm) X2	poids (kg) Y
1	143	143,002	38,25
2	191	158,75	50,625
3	160	157,48	42,525
4	157	163,83	55,575
5	191	165,862	48,15
6	141	156,972	38,25
7	185	160,782	45,45
8	210	166,37	63
9	149	163,322	49,725
10	169	158,242	44,775
11	173	159,512	46,125
12	150	155,702	42,3
13	144	151,13	42,075
14	146	152,	49,05
15	155	155,702	48,15
16	183	163,83	46,125
17	154	152,	51,3
18	152	153,67	47,25
19	148	153,67	38,025
20	164	165,862	44,1
21	177	155,702	36,45
22	183	168,91	50,4

23	182	166,37	59,85
24	165	140,97	30,15
25	163	143,51	37,8
26	171	160,02	37,8
27	193	151,892	51,75
28	169	156,21	38,25
29	155	158,242	47,25
30	171	158,75	50,4
31	140	136,652	30,825
32	149	148,082	41,85
33	150	151,13	35,325
34	140	135,89	36,45
35	166	156,21	46,575
36	146	143,002	37,575
37	139	146,05	43,2
38	177	156,972	64,125
39	166	150,622	40,275
40	184	158,242	48,6
41	177	155,702	50,4
42	145	149,86	41,175
43	167	158,242	41,625
44	185	152,	47,7
45	156	138,43	33,75
46	191	160,782	51,075
47	189	163,322	51,075
48	157	153,67	50,4
49	171	156,21	40,95
50	143	156,21	52,425
51	182	157,48	41,175
52	154	154,94	55,125
53	141	142,24	32,625
54	167	154,94	42,075
55	141	155,702	38,25
56	175	153,162	38,7
57	153	160,782	48,6
58	185	149,86	46,8
59	139	156,21	46,8
60	143	130,302	22,725
61	147	155,702	51,75
62	164	147,32	37,575
63	175	154,432	42,075
64	170	163,322	40,5
65	186	146,812	42,75

66	185	165,862	53,1
67	168	156,21	42,75
68	139	134,112	28,575
69	178	161,29	66,825
70	147	141,732	33,75
71	183	163,322	49,275
72	148	143,002	34,65
73	144	141,732	33,075
74	190	169,672	63
75	143	148,082	34,875
76	147	151,13	45,45
77	172	164,592	63,9
78	179	160,02	44,325
79	142	142,24	32,625
80	150	138,43	33,3
81	147	130,81	28,8
82	182	162,56	50,175
83	164	160,782	48,6
84	180	155,702	49,725
85	161	149,86	41,4
86	142	143,51	31,05
87	178	156,21	46,575
88	145	149,352	40,05
89	180	160,782	51,3
90	176	155,702	50,4
91	180	149,86	50,4
92	162	147,32	37,8
93	197	156,21	54,45
94	182	148,082	47,025
95	169	157,48	44,325
96	147	151,892	38,025
97	197	164,592	50,4
98	145	146,812	37,8
99	143	140,97	37,8
100	147	148,082	50,175
101	154	159,512	42,075
102	140	152,	34,65
103	178	168,91	52,875
104	148	149,86	42,75
105	190	144,272	44,325
106	186	144,78	37,575
107	165	155,702	47,925
108	155	167,64	65,025

109	210	157,48	52,2
110	144	154,94	41,4
111	186	161,29	48,6
112	157	153,67	47,25
113	139	153,67	39,15
114	146	146,05	40,5
115	151	168,402	52,65
116	153	152,	37,8
117	176	165,1	53,325
118	146	145,542	37,35
119	151	154,94	36,45
120	193	168,402	59,85
121	143	146,05	33,75
122	173	175,26	50,625
123	144	151,13	39,6
124	147	144,78	37,8
125	150	151,13	37,8
126	140	148,59	38,925
127	184	168,91	50,4
128	168	168,91	50,175
129	203	168,91	52,65
130	200	180,34	66,15
131	145	143,51	40,95
132	182	170,18	59,85
133	177	160,02	49,95
134	150	149,86	44,1
135	171	156,972	50,4
136	142	142,24	39,375
137	144	152,	40,05
138	193	182,88	67,5
139	139	139,7	33,075
140	196	163,83	44,1
141	153	146,812	35,775
142	164	168,91	50,4
143	151	150,622	39,15
144	144	145,542	34,425
145	189	170,18	57,6
146	160	153,67	37,8
147	141	135,382	37,8
148	206	173,482	60,3
149	140	151,13	42,525
150	183	167,64	47,475
151	144	159,512	42,3

152	162	163,83	53,55
153	175	162,56	41,4
154	170	162,052	50,625
155	156	168,402	47,7
156	188	170,942	50,4
157	193	172,212	57,375
158	156	148,082	41,625
159	156	173,99	51,3
160	149	133,35	36,45
161	142	149,352	37,8
162	152	151,13	47,25
163	143	146,05	45,45
164	173	167,64	50,4
165	177	153,67	50,4
166	150	156,972	53,1
167	162	160,02	40,95
168	148	153,67	53,1
169	206	176,53	77,175
170	194	165,862	60,525
171	186	168,91	50,4
172	164	147,32	37,8
173	155	145,542	36,225
174	189	165,1	51,3
175	150	151,13	37,8
176	183	164,592	49,95
177	156	156,972	50,4
178	150	149,86	44,775
179	250	171,45	77,175
180	185	167,64	47,25
181	140	143,51	37,8
182	160	150,622	35,325
183	164	153,67	42,75
184	175	172,72	50,4
185	174	167,64	48,6
186	149	144,78	41,4
187	169	157,48	45
188	157	147,32	36,225
189	156	156,21	48,825
190	144	144,78	37,8
191	142	139,7	31,5
192	189	168,402	50,4
193	174	177,292	53,775
194	163	165,862	52,875

195	155	156,972	41,175
196	175	166,37	51,3
197	188	160,782	51,975
198	141	146,05	38,25
199	140	144,272	37,575
200	159	160,782	50,4
201	152	154,432	43,65
202	161	144,272	33,75
203	159	159,512	44,55
204	178	161,29	46,125
205	164	156,21	63
206	165	164,592	44,1
207	150	154,432	57,6
208	147	128,27	35,55
209	173	155,702	41,85
210	164	146,812	42,75
211	176	162,052	44,325
212	180	156,972	46,8
213	151	148,082	38,7
214	178	170,942	53,775
215	186	167,64	50,4
216	175	161,29	44,325
217	164	161,29	48,6
218	144	152,4	52,875
219	172	165,1	50,4
220	168	152,4	42,075
221	158	165,1	54,45
222	176	156,21	36,45
223	188	180,34	63
224	188	167,132	67,725
225	166	158,75	37,8
226	166	170,942	54,45
227	162	152,4	47,25
228	166	157,48	40,95
229	163	167,64	50,4
230	174	160,02	50,4
231	160	162,56	52,2
232	149	143,002	32,4
233	146	139,7	32,175
234	153	164,592	57,6
235	178	162,052	50,4
236	142	139,7	34,2
237	167	157,48	48,375

Après l'application de l'algorithme de k-means on obtient le clustering suivant :

Cluster 1

enfant i	Age (mois)	taille (cm)	poids (kg)
2	191	158,75	50,625
5	191	165,862	48,15
7	185	160,782	45,45
8	210	166,37	63
10	169	158,242	44,775
11	173	159,512	46,125
16	183	163,83	46,125
20	164	165,862	44,1
21	177	155,702	36,45
22	183	168,91	50,4
23	182	166,37	59,85
26	171	160,02	37,8
27	193	151,892	51,75
28	169	156,21	38,25
30	171	158,75	50,4
35	166	156,21	46,575
38	177	156,972	64,125
40	184	158,242	48,6
41	177	155,702	50,4
43	167	158,242	41,625
44	185	152,4	47,7
46	191	160,782	51,075
47	189	163,322	51,075
48	157	153,67	50,
49	171	156,21	40,95
51	182	157,48	41,175
56	175	153,162	38,7
58	185	149,86	46,8
63	175	154,432	42,075
64	170	163,322	40,5

65	186	146,812	42,75
66	185	165,862	53,1
67	168	156,21	42,75
69	178	161,29	66,825
71	183	163,322	49,275
74	190	169,672	63
77	172	164,592	63,9
78	179	160,02	44,325
82	182	162,56	50,175
83	164	160,782	48,6
84	180	155,702	49,725
87	178	156,21	46,575
89	180	160,782	51,3
90	176	155,702	50,4
91	180	149,86	50,4
93	197	156,21	54,45
94	182	148,082	47,025
95	169	157,48	44,325
97	197	164,592	50,4
103	178	168,91	52,875
104	148	149,86	42,75
105	190	144,272	44,325
106	186	144,78	37,575
107	165	155,702	47,925
108	155	167,64	65,025
109	210	157,48	52,2
111	186	161,29	48,6
117	176	165,1	53,325
120	193	168,402	59,85
122	173	175,26	50,625
127	184	168,91	50,4
128	168	168,91	50,175
129	203	168,91	52,65
130	200	180,34	66,15
132	182	170,18	59,85
133	177	160,02	49,95
135	171	156,972	50,4
138	193	182,88	67,5
140	196	163,83	44,1
142	164	168,91	50,4
145	189	170,18	57,6
148	206	173,482	60,3
150	183	167,64	47,475

152	162	163,83	53,55
153	175	162,56	41,4
154	170	162,052	50,625
156	188	170,942	50,4
157	193	172,212	57,375
164	173	167,64	50,4
165	177	153,67	50,4
169	206	176,53	77,175
170	194	165,862	60,525
171	186	168,91	50,4
174	189	165,1	51,3
176	183	164,592	49,95
179	250	171,45	77,175
180	185	167,64	47,25
184	175	172,72	50,4
185	174	167,64	48,6
187	169	157,48	45
192	189	168,402	50,4
193	174	177,292	53,775
194	163	165,862	52,875
196	175	166,37	51,3
197	188	160,782	51,975
204	178	161,29	46,125
205	164	156,21	63
206	165	164,592	44,1
209	173	155,702	41,85
211	176	162,052	44,325
212	180	156,972	46,8
214	178	170,942	53,775
215	186	167,64	50,4
216	175	161,29	44,325
217	164	161,29	48,6
219	172	165,1	50,4
222	176	156,21	36,45
223	188	180,34	63
224	188	167,132	67,725
226	166	170,942	54,45
229	163	167,64	50,4
230	174	160,02	50,4
235	178	162.052	50.4
237	167	157.48	48.375

Cluster 2

Enfant i	Age (mois)	Taille (cm)	Poids (kg)
1	143	143,002	38,25
3	160	157,48	42,525
4	157	163,83	55,575
6	141	156,972	38,25
9	149	163,322	49,725
12	150	155,702	42,3
13	144	151,13	42,075
14	146	152,4	49,05
15	155	155,702	48,15
17	154	152,4	51,3
18	152	153,67	47,25
19	148	153,67	38,025
24	165	140,97	30,15
25	163	143,51	37,8
29	155	158,242	47,25
31	140	136,652	30,825
32	149	148,082	41,85
33	150	151,13	35,325
34	140	135,89	36,45
36	146	143,002	37,575
37	139	146,05	43,2
39	166	150,622	40,275
42	145	149,86	41,175
45	156	138,43	33,75
48	157	153,67	50,
50	143	156,21	52,425
52	154	154,94	55,125
53	141	142,24	32,625
54	167	154,94	42,075
55	141	155,702	38,25
57	153	160,782	48,6
59	139	156,21	46,8
60	143	130,302	22,725
61	147	155,702	51,75
62	164	147,32	37,575
68	139	134,112	28,575
70	147	141,732	33,75
72	148	143,002	34,65
73	144	141,732	33,075
75	143	148,082	34,875
76	147	151,13	45,45
79	142	142,24	32,625

80	150	138,43	33,3
81	147	130,81	28,8
85	161	149,86	41,
86	142	143,51	31,05
88	145	149,352	40,05
92	162	147,32	37,8
96	147	151,892	38,025
98	145	146,812	37,8
99	143	140,97	37,8
100	147	148,082	50,175
101	154	159,512	42,075
102	140	152,4	34,65
104	148	149,86	42,75
110	144	154,94	41,
112	157	153,67	47,25
113	139	153,67	39,15
114	146	146,05	40,5
115	151	168,402	52,65
116	153	152,4	37,8
118	146	145,542	37,35
119	151	154,94	36,45
121	143	146,05	33,75
123	144	151,13	39,6
124	147	144,78	37,8
125	150	151,13	37,8
126	140	148,59	38,925
131	145	143,51	40,95
134	150	149,86	44,1
136	142	142,24	39,375
137	144	152,4	40,05
139	139	139,7	33,075
141	153	146,812	35,775
143	151	150,622	39,15
144	144	145,542	34,425
146	160	153,67	37,8
147	141	135,382	37,8
149	140	151,13	42,525
151	144	159,512	42,3
155	156	168,402	47,7
158	156	148,082	41,625
159	156	173,99	51,3
160	149	133,35	36,45
161	142	149,352	37,8

162	152	151,13	47,25
163	143	146,05	45,45
166	150	156,972	53,1
167	162	160,02	40,95
168	148	153,67	53,1
172	164	147,32	37,8
173	155	145,542	36,225
175	150	151,13	37,8
177	156	156,972	50,
178	150	149,86	44,775
181	140	143,51	37,8
182	160	150,622	35,325
183	164	153,67	42,75
186	149	144,78	41,4
188	157	147,32	36,225
189	156	156,21	48,825
190	144	144,78	37,8
191	142	139,7	31,5
195	155	156,972	41,175
198	141	146,05	38,25
199	140	144,272	37,575
200	159	160,782	50,4
201	152	154,432	43,65
202	161	144,272	33,75
203	159	159,512	44,55
207	150	154,432	57,6
208	147	128,27	35,55
210	164	146,812	42,75
213	151	148,082	38,7
218	144	152,4	52,875
220	168	152,4	42,075
221	158	165,1	54,45
225	166	158,75	37,8
227	162	152,4	47,25
228	166	157,48	40,95
231	160	162,56	52,2
232	149	143,002	32,4
233	146	139,7	32,175
234	153	164,592	57,6
236	142	139,7	34,2

Résultats et discussion :

Le modèle de régression linéaire multiple est :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

On applique la méthode de k-means sur les données et ça nous a permis d'avoir un regroupement de deux clusters qui représentent deux sous-populations de la population initiale.

Chacun de ces clusters contient les éléments les plus homogènes entre eux.

Une fois que ces clusters sont constitués, on applique la régression linéaire pour chaque cluster.

Pour la population initiale le modèle de régression linéaire est le suivant :

$$\widehat{Y}_i = -57.51895 + 0.10812 X_{i1} + 0.54744 X_{i2}$$

Nous avons ensuite estimé les paramètres de la régression pour les deux clusters respectivement.

Le tableau suivant montre le résultat de la simulation de notre modèle de la régression divisée.

(Les calculs sont faits grâce au logiciel R et logiciel PAST 3)

Clusters	b_0	b_1	b_2	Erreur
Cluster 1	-66.6314	0.1907	0.5106	33,7528
Cluster 2	-50.5795	-0.0262	0,6375	20,9200
Moyen	-58.6055	0.0822	0.5740	27,3364
Globale	-57.5189	0.1081	0.5474	28,1738

Résultat des données de simulation pour le modèle de régression divisée de l'exemple 1

En faisant la moyenne de chaque paramètre respectif des deux clusters on obtient une moyenne presque similaire de paramètre de la population globale.

Et pour vérifier la validité de notre modèle on a calculé les intervalles de confiance à 95% pour les paramètres de régression.

On a obtenu les résultats suivant :

Intervalle de confiance au seuil 1 - α de β_0 du cluster 1 :

$$S(b_j) = \begin{pmatrix} s(b_0) \\ s(b_1) \\ s(b_2) \end{pmatrix} = \begin{pmatrix} 12.97 \\ 0.0434 \\ 0.0777 \end{pmatrix}$$

$T_{0.025\ ;\ 111} = 1.96$

$$-66.6314 - 1.96 \times 12.97 \leq \beta_0 \leq -66.6314 + 1.96 \times 12.97$$

$$-91.7826 \leq \beta_0 \leq -41.2102$$

Intervalle de confiance au seuil 1 - α de β_1 du cluster 1

$$0.1907 - 1.96 \times 0.0434 \leq \beta_1 \leq 0.1907 + 1.96 \times 0.0434$$

$$0.1056 \leq \beta_1 \leq 0.2757$$

Intervalle de confiance au seuil 1 - α de β_2 du cluster 1

$$0.5106 - 1.96 \times 0.0777 \leq \beta_2 \leq 0.5106 + 1.96 \times 0.0777$$

$$0.3583 \leq \beta_2 \leq 0.6628$$

Intervalle de confiance au seuil 1 - α de β_0 du cluster 2

$$S(b_j) = \begin{pmatrix} s(b_0) \\ s(b_1) \\ s(b_2) \end{pmatrix} = \begin{pmatrix} 9.4388 \\ 0.0576 \\ 0.0555 \end{pmatrix}$$

$T_{0.025\ ;\ 122} = 1.96$

$$-50.5795 - 1.96 \times 9.4388 \leq \beta_0 \leq -50.5795 + 1.96 \times 9.4388$$

$$-69.0795 \leq \beta_0 \leq -32.0794$$

Intervalle de confiance au seuil 1 - α de β_1 du cluster 2

$$-0.0262 - 1.96 \times 0.0576 \leq \beta_1 \leq -0.0262+1.96\times 0.0576$$
$$-0.1390 \leq \beta_1 \leq 0.090$$

Intervalle de confiance au seuil 1 - α de β_2 du cluster 2

$$0.6375 - 1.96 \times 0.0555 \leq \beta_2 \leq 0.6375 + 1.96 \times 0.0555$$
$$0.52872 \leq \beta_2 \leq 0.74628$$

On conclut pour un niveau de confiance à 95%, tous les intervalles contiennent les paramètres de régression β_0, β_1 et β_2 de la population initiale. Par conséquent, nous avons montré la validité de notre travail.

IV. Résultats expérimentaux et discussion de l'exemple 2

Le second exemple consiste à expliquer la consommation des véhicules (en L/100 km) à partir de p = 3 variables explicatives: la cylindrée (taille du moteur, en cm^3), la puissance (en kw) et le poids (en kg) à partir de n = 1000 observation.

Le nombre de données est assez grand pourra prend beaucoup de temps pour être analysé on a procédé alors au découpage de l'ensemble des données par la méthode des k-means qui nous a donné 10 sous-ensemble.

On applique de la même façon la procédure de régression linéaire effectuée dans l'exemple 1.

On a voulu s'assurer est ce que le modèle trouvé de la population globale est similaire à celui trouvé en divisant la population en dix sous-ensemble.

Le modèle globale de la régression linéaire multiple est donnée par :

$$\mathbf{Y_i=\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i} \qquad \textbf{pour i = 1 … 1000}$$

Le modèle estimé est:

$$\widehat{Y}_i = 1.1494 + 0.002272X_{i1} + 9.18E\text{-}05X_{i2} + 0.0030935X_{i3}$$

Le tableau suivant donne les valeurs des b_0, b_1, b_2, et b_3 pour les dix sous – populations (clusters).

L'avant dernière ligne du tableau donne la moyenne de chaque paramètre pour les dix sous –populations.

Les deux valeurs des b_0, b_1, b_2, b_3 et l'erreur respective sont très similaire entre le modèle globale et le nouveau modèle constitué de moyen des paramètres de régression.

Sous-ensembles	b_0	b_1	b_2	b_3	Erreur
Sous-ensemble 1	1,1193	0,0023764	-0,00014801	0,0029324	0,344862802
Sous-ensemble 2	2,4385	0,0049158	-0,00078449	-0,00097988	0,27365803
Sous-ensemble 3	-1,3284	0,003187	-0,0001128	0,0032157	0,732902
Sous-ensemble 4	0,30425	0,0010663	0,0012644	0,0052688	0,1971
Sous-ensemble 5	-0,98215	0,0030161	0,00026063	0,0032382	0,320901344
Sous-ensemble 6	1,4947	0,004942	0,0052834	4,43E-05	0,256423096
Sous-ensemble 7	1,0973	0,0024629	0,0004525	0,0028234	0,49430738
Sous-ensemble 8	2,4114	0,0011547	0,00046367	0,003078	0,20558108
Sous-ensemble 9	3,5275	0,0022641	0,0032582	0,0010143	0,41446321
Sous-ensemble 10	1,2697	0,0029881	0,0053781	0,0018979	0,66336698
Moyenne	1,13521	0,00283734	0,00153156	0,00225331	0,38833131
Globale	1,1494	0,002272	9,18E-05	0,0030935	0,43505706

Résultat des données de simulation pour le modèle de régression divisée de l'exemple 2

Nous savions qu'il existait des différences entre les paramètres de la régression des sous-ensembles et la population globale, mais la valeur moyenne des paramètres de la régression des sous-ensembles était semblable à la valeur du paramètre de la population initiale.

Ensuite, nous avons calculé les intervalles de confiance à 95% pour les paramètres de régression pour les dix sous-ensembles afin de valider notre modèle.

Le tableau suivant montre les intervalles de confiance des paramètres de régression pour les dix sous-ensembles.

Sous- ensembles	Intervalle de confiance de β_0	Intervalle de confiance de β_1
Sous-ensemble 1	-5,134276≤ β_0 ≤7,372876	-0,000194924≤ β_1 ≤0,004947724
Sous-ensemble 2	-9,77834912≤ β_0 ≤14,65534912	-0,003941962≤ β_1 ≤0,013773562
Sous-ensemble 3	-9,239352≤ β_0 ≤6,6874932	-0,000112938≤ β_1 ≤0,006486938
Sous-ensemble 4	-0,7730052≤ β_0 ≤1,3815052	0,000493451≤ β_1 ≤0,001639149
Sous-ensemble 5	-9,3066217≤ β_0 ≤7,3423217	-0,0004163≤ β_1 ≤0,0064485
Sous-ensemble 6	0,72942≤ β_0 ≤2,25998	0.002212≤ β_1 ≤0,00631
Sous-ensemble 7	0,6741948≤ β_0 ≤1,5204052	0,00210834≤ β_1 ≤0,00281746
Sous-ensemble 8	1,14957014≤ β_0 ≤3,57322986	0,00037257≤ β_1 ≤0,0023683
Sous-ensemble 9	-0,73≤ β_0 ≤7,78501	-0,0004≤ β_1 ≤0,00495
Sous-ensemble 10	-2,70850214≤ β_0 ≤5,24790214	0,000925793≤ β_1 ≤0,005050407

Sous- ensembles	Intervalle de confiance de β_2	Intervalle de confiance de β_3
Sous-ensemble 1	-0,00240122 $\leq \beta_2 \leq$0,002105206	0,002196557$\leq \beta_3 \leq$0,003668243
Sous-ensemble 2	-0,02280591$\leq \beta_2 \leq$0,021236933	-3,46E-03$\leq \beta_3 \leq$3,50E-03
Sous-ensemble 3	-0,00286579$\leq \beta_2 \leq$0,002640193	0,002238528$\leq \beta_3 \leq$0,004192872
Sous-ensemble 4	-0,00147372$\leq \beta_2 \leq$0,00400252	0,00355186$\leq \beta_3 \leq$0,00698574
Sous-ensemble 5	-0,0027276$\leq \beta_2 \leq$0,0032488	0,002216679$\leq \beta_3 \leq$0,004259721
Sous-ensemble 6	-0,023697208$\beta_2 \leq$0,013130408	-0,0013$\leq \beta_3 \leq$0.0038121
Sous-ensemble 7	-0,0009687$\leq \beta_2 \leq$0,0018737	0,00224937$\leq \beta_3 \leq$0,00328731
Sous-ensemble 8	-0,00254035$\leq \beta_2 \leq$0,00346769	0,00095119$\leq \beta_3 \leq$0.00339743
Sous-ensemble 9	-0,0063$\leq \beta_2 \leq$0,0128	-0,0002$\leq \beta_3 \leq$0,00319
Sous-ensemble 10	-0,007126441$\leq \beta_2 \leq$0,017882641	-0,000125417$\leq \beta_3 \leq$0,003921217

Intervalles de confiance à 95% des $\beta_0, \beta_1, \beta_2$ et β_3

Tous les intervalles de confiance des sous-ensembles comprennent les paramètres de régression de la population globale (β_0, β_1 , $\beta_2$et β_3). Par conséquent, nous pouvons vérifier la performance de notre approche de régression.

Conclusion et discussion

Dans ce mémoire, on a essayé de trouver une approche qui pourrait simplifier le traitement des données de grand masse et surmonter aussi la charge des calculs surtout quand on a des moyens limités.

L'idée de cette approche consiste à diviser les données globales de la population en sous-ensembles de données par la méthode de k-means.

On a appliqué le processus sur les clusters obtenus.

L'objectif est de vérifier si les paramètres de la régression pour la population globale et la moyenne des paramètres des clusters sont similaires.

Pour cela on a traité deux exemples et on a trouvé que la valeur moyenne des paramètres de la régression des sous-ensembles était semblable à celles de la population globale, ce qui signifie que la méthode de la régression divisée est bien performante.

Il serait aussi intéressant de vérifier cette approche pour des données encore plus grandes, et avoir si elle reste toujours applicable.

On pourrait aussi voir s'il serait possible de faire des subdivisions à partir des sous ensemble.

Références

[1] Bai J., and Ng, S. Large dimensional factor analysis. Fundations and Trends(R) in Econometrics 3, 2 (2008), 89_163.

[2] Besse, P., Garivier, A., and Loubes, J.-M. Big Data Analytics - Retour vers le Futur 3 ; De Statisticien à Data Scientist. Mar. 2014.

[3] Besse, P., and Villa-Vialaneix, N. Statistique et big data analytics, volumétrie - l'attaque des clones, 2014.

[4] Biemer, P. Dropping the to se : applying the paradigm to big data. In The 2014

International Total Survey Error Workshop (2014), National Institute of Statistical Science.

[5] Fan, J., and Fan, Y. High Dimensional Classification Using Features Annealed Independence Rules. Annals of statistics 36, 6 (2008), 2605_2637.

[6] J. J. Berman, "Principle of Big Data", Morgan Kaufmann, **(2013)**.

[7] J. Luand D. LiBias, "Correction in a Small Sample from Big Data", IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 11,**(2013)**,pp. 2658-2663.

[8] M. Riondato, "Sampling-based Randomized Algorithms for Big Data Analytics", PhD dissertation in the Departement of Computer Science at Brown University, **(2014)**.

[9] S. Junand D. Uhm, "A Predictive Model for Patent Registration Time Using Survival Analysis", Applied Mathematics & Information Sciences, vol. **7**, no.5,**(2013)**,pp. 1819-1823.

[10] S. M. Ross, "Introductory Statistics", McGraw-Hill, **(1996)**.

[11 S. Jun, "A Technology Forecasting Method Using Text Mining and Visual Apriori Algorithm", Applied Mathematics & Information Sciences-An International Journal, vol. 8, no.(1L),**(2014)**,pp. 35-40.

[12]P. Vincent, L. Badriand M. Badri, “Regression Testing of Object-Oriented Software: Towards a Hybrid Technique”, International Journal of Software Engineering and Its Applications, vol. 7, no.4,**(2013)**,pp. 227-240.

[13] S. Ha, S. Lee and K. Lee, “Standarization Requirements Analysis on Big Data in Public Sector based on Potential Business Models”, International Journal of Software Engineering and Its Applications, vol. 8, no.11,**(2014)**,pp. 165-172.

[14] Statistical Commission of the Unece. Big data and modernization of statistical systems. In Conference of European Statisticians (2014), United Nations Economic Commission for Europe.

[15] Tibshirani, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58 (1994), 267_288.