

Projet de Fin d'Etudes  
\*\*\*\*\*

Enriching ECCBC's internal data with Geospatial information and leveraging clustering analysis to determine an optimal and dynamic segmentation of its customers and their demand models.

Préparé par : M. Salif SAWADOGO

Sous la direction de : M. Mohammed El Haj TIRARI (INSEA)  
M. Jimmy RICO (ECCBC)

*Soutenu publiquement comme exigence partielle en vue de l'obtention du  
Diplôme d'Ingénieur d'Etat en Statistique*

*Devant le jury composé de :*

- M. Mohammed El Haj TIRARI (INSEA)
- M. Ahmed DOGHMI (INSEA)
- M. Jimmy RICO (ECCBC)



# Résumé

Dans ce présent projet réalisé à ECCBC, il est question d'enrichir la base de données internes avec des informations externes à caractère spatiale propres à chaque point de ventes (PdV) afin de réaliser une segmentation dynamique de ces derniers et modéliser leurs fonctions de demande. Marrakech et Tizi Ouzou sont les zones pilotes.

D'abord, nous avons enrichi la base de données grâce à des techniques d'analyse spatiale et un algorithme de correspondance entre la source de données interne et la source externe, construit selon un principe d'apprentissage supervisé.

Ensuite, nous présentons dans ce rapport, la segmentation du groupe Hôtel-Restaurants-Café réalisée grâce à des techniques de clustering (K-mean et CAH). 8 segments ont été découverts pour les Café-Restaurant et 3 pour les hôtels. Les PdV n'ayant pas été candidats à l'exercice de clustering ont été assignés aux segments grâce au Random Forest.

Enfin, des modèles de demande selon le cluster ont été construites par article en utilisant des modèles de régression multiple avec et sans contraintes de cannibalisation. Ce projet se termine par la proposition d'un algorithme de recommandation d'assortiment optimal pour les réfrigérateurs par segments en utilisant les modèles de demande.

**Mots-clés :** Zone de chalandise, Matching, Similarité, localisation, Clustering,

Modèle de demandes, assortiment

# Dédicace

*Avant de commencer, j'adresse toutes les louanges à Allah le miséricordieux qui nous a permis de réaliser ce travail dans de bonnes conditions malgré toutes les difficultés rencontrées.*

*Je dédie ce travail à ma maman, à mon père, à ma marâtre, à mon oncle Hamidou et à tous mes frères et sœurs pour leurs soutiens de toute nature. Je leur dois énormément, pour leur amour inconditionnel, leurs dévouements, ainsi que leur confiance.*

*Ce travail est également dédié à tous ceux qui ont contribué à sa réalisation de près ou de loin et qui s'y reconnaîtront*

# Remerciements

*Au moment où s'achève la rédaction de ce rapport, Je tiens à exprimer ma profonde et sincère gratitude à mon encadrant M. Jimmy RICO, Manager du groupe Data science et business intelligence de l'ECCBC pour m'avoir accordé l'opportunité de réaliser mon projet de fin d'études sur l'un des projets les plus innovants de l'entreprise et m'avoir fourni des conseils précieux tout au long de ce stage. Sa technicité, sa facilité à concilier méthodes d'apprentissage statistique et business, sa méthodologie de communications de ces travaux m'ont profondément inspiré. C'est un grand privilège et un honneur de travailler dans son équipe.*

*Je voudrais également exprimer mes sincères remerciements à mon encadrant de l'INSEA, professeur Mohammed El Haj TIRARI qui m'a offert l'honneur d'être mon encadrant durant ce stage. Je le remercie énormément pour ses différents enseignements, ses précieux conseils et ses différentes corrections du présent rapport.*

*Mes remerciements s'adressent également à M. Farouk Bencheikh, Senior Business Analyst pour ses conseils en matière de modélisation orienté business et Moad Charhbili, mon collègue stagiaire chargé du projet de Forecasting pour ses échanges d'idées. Mention spéciale à tout le personnel d'ECCBC pour leur soutien, leur recommandation et leur vif intérêt pour mener à bien ce projet. Ce dernier n'aurait pas eu lieu sans le bon environnement de travail offert et l'infrastructure mis à notre disposition malgré le contexte de crise sanitaire.*

*Par ailleurs, je remercie le professeur M. Ahmed DOGHMI d'avoir accepté de juger ce travail. Par cette occasion mes remerciements vont également à l'endroit de tout le corps professoral et administratif de l'INSEA pour les efforts qu'ils déploient jour et nuit pour nous assurer une formation de qualité.*

*Je ne pourrai pas non plus clore mon discours de remerciements sans remercier mes amis à l'INSEA et les autres, au pays ou ailleurs. Ils se reconnaîtront dans ce rapport. Ils ont fait de ma formation d'ingénieur l'un des moments inoubliables de ma vie.*

*Enfin, mes remerciements vont à toutes les personnes qui m'ont soutenu pour mener à bien le travail fascinant directement ou indirectement .*

# Tables des matières

Résumé .....	- 1 -
Dédicace .....	- 2 -
Remerciements .....	- 3 -
Tables des matières.....	- 4 -
Liste des abréviations : .....	- 9 -
Liste des tableaux .....	- 10 -
Liste des figures.....	- 11 -
Introduction générale.....	- 13 -
Cadre contextuel et Méthodologique du travail.....	- 15 -
Introduction .....	- 16 -
I.     Présentation de l'organisme d'accueil.....	- 16 -
I.1.   Equatorial Coca-Cola Bottling Company .....	- 16 -
I .1.1.   Presentation de Equatorial Coca-Cola Bottling Company .....	- 16 -
I.1.2.   Organigramme de Equatorial Coca-Cola Bottling Company.....	- 17 -
I.1.3.   Business de l'entreprise .....	- 18 -
a)   Système de production.....	- 18 -
b)   Système de distribution.....	- 19 -
I.2.   Le groupe Data Science et Business Intelligence .....	- 20 -
II.    Cadre conceptuel et Problématique.....	- 21 -
II.1.   Enoncé de la problématique .....	- 21 -
II.2.   Objectif du projet.....	- 23 -
II .3.   Principaux résultats attendus .....	- 23 -
II.4.   Questions de recherche .....	- 23 -
III.   Méthodologies de recherche :.....	- 24 -
IV.    Conduite et planification du projet .....	- 25 -
Conclusion.....	- 26 -
Introduction à la géomatique .....	- 28 -
Introduction .....	- 29 -
I.     Concepts généraux .....	- 29 -
I.1.   Définitions .....	- 29 -
I.2.   Composantes de l'information géographique .....	- 29 -
I.3.   Différents types de données spatiales .....	- 31 -
I.4.   Représentation numérique des données spatiales .....	- 32 -
II.    Les fondements de la localisation .....	- 33 -

II .1.	Système géodésique et Ellipsoïdes de référence .....	34 -
II.2.	Les systèmes de coordonnées géographiques .....	35 -
II.3.	Les projections cartographiques .....	36 -
II .4.	Distance géodésique.....	38 -
	Conclusion.....	39 -
	Eléments d'analyse spatiale.....	40 -
	Introduction .....	41 -
I.	Le comportement spatial du consommateur .....	41 -
II.	le « trading area » ou zone de chalandise.....	42 -
	II.1. Définitions .....	42 -
	II.2. Modèles de zone de chalandise .....	44 -
	II.2. Index spatial .....	47 -
	Conclusion.....	48 -
	Eléments d'analyse textuelle .....	49 -
	Introduction .....	50 -
I.	Mesure de similarité de données textuelles .....	50 -
	I.1. La distance de Levenshtein.....	51 -
	I.2. Distance Damerau-Levenshtein .....	52 -
	I.3. Distance Hamming .....	53 -
	I.4. Distance Jaro.....	53 -
	I.5. Distance Jaro-Winkler .....	54 -
	I.6. N-grammes.....	54 -
	I.7. Autres mesures de similarités .....	55 -
II.	La préparation de texte en Natural Language Processing .....	55 -
	II.1. Détection de la langue .....	56 -
	II.2. Suppressions des ponctuations et les caractères spéciaux .....	56 -
	II.3. Tekenisation .....	56 -
	II.4. Suppression des stopword ou mots vides .....	57 -
	II.5. le Stemming.....	57 -
	II.6. Le pliage ASCII .....	58 -
	Conclusion.....	58 -
	Elements d'apprentissage statistique .....	59 -
	Introduction .....	59 -
I.	Typologie d'apprentissage statistique .....	60 -
	I.1. Apprentissage supervisé .....	60 -
	I.2 Apprentissage non supervisé.....	60 -
	I.3Apprentissage semi supervisé .....	61 -
II.	Techniques d'apprentissage non-supervisée : Clustering.....	61 -

I.	Critères de clustering.....	- 62 -
I .1.	Le choix des variables.....	- 62 -
I.2.	Choix de la métrique.....	- 63 -
I.3.	La Forme des clusters.....	- 68 -
I.4.	Stabilité des clusters.....	- 70 -
II.	Techniques.....	- 71 -
II.2.	Clustering hiérarchique .....	- 71 -
II.2.	Les méthodes basées sur les centroïdes.....	- 74 -
II.3.	Cas particulier des centroïdes :k-mean.....	- 75 -
II.3.1.	Principe.....	- 76 -
II.3.2.	Forme des clusters du k-means.....	- 76 -
II.3.3.	K-mean++.....	- 77 -
II.4.	Les méthodes de densités : DBSCAN.....	- 77 -
	Tableau recapitualtif.....	- 80 -
III.	Techniques d'apprentissage supervisé.....	- 80 -
III.1	Généralités .....	- 80 -
III.2.	Évaluation des performances des modèles .....	- 86 -
III.2.1.	Les métriques régression .....	- 86 -
III.3.2.	Les métriques de classification.....	- 88 -
a)	<b>Matrice de confusion et ses métriques</b> .....	- 88 -
b)	L'aire sous la courbe ROC : AUC.....	- 90 -
	Conclusion.....	- 91 -
	Données et infrastructures et schéma de modélisation .....	- 93 -
	Introduction .....	- 94 -
I.	Données utilisées .....	- 94 -
I.1.	Données internes.....	- 94 -
I.1.1.	Les données de ventes .....	- 95 -
I.1.2.	Les données d'indicateur d'exécution de livraison (RED).....	- 95 -
I.1.3.	Les données sur notre force de vente.....	- 95 -
I.1.4.	Les données sur les réfrigérateurs .....	- 95 -
I.1.5.	Les données relatives au temps .....	- 95 -
I.1.6.	Les données relatives au calendrier .....	- 96 -
I.1.7.	Matrice de cannibalisation .....	- 96 -
I.1.	Données externes .....	- 96 -
I.1.1.	Les données d'Experian .....	- 97 -
I.1.2.	Les données d'Unacast .....	- 98 -
I.1.3.	Les données de WorldPop .....	- 99 -
I.1.4.	OpenStreetMap et de Pitney Bowes .....	- 100 -

I.1.5.	Les données scrapées de TripAdvisor .....	- 103 -
I.1.6.	Les données de Flickr.....	- 107 -
II.	Infrastructures du projet .....	- 108 -
II.1.	Cloud et architectures.....	- 108 -
II.2	Azure Blob storage.....	- 110 -
II.3.	Python .....	- 110 -
II .4.	Databricks, Apache Spark et Azure Databricks .....	- 111 -
II.5.	Azure Translator.....	- 113 -
III.	Schéma de la modélisation du projet.....	- 114 -
	Conclusion.....	- 115 -
	Trade area analysis et Matching .....	- 116 -
	Introduction .....	- 117 -
I.	Trade Area Analysis.....	- 117 -
I.1.	Préparation des composants de l'analyse de spatiale .....	- 118 -
I.2.	Détermination de la zone de chalandise.....	- 119 -
I.1.1	Calcul des scores d'urbanicité .....	- 119 -
I.1.2.	Feature engineering du score de densité : urbanicité.....	- 120 -
I.1.3	Calcul de distance des zones de chalandise.....	- 120 -
I.1.4.	Profil de la zone chalandise .....	- 122 -
II.	Matching.....	- 123 -
II.1.	Principe du matching.....	- 123 -
II.2.	Construction de l'échantillon de matching manuel .....	- 123 -
II.3.	Similarité des noms et des adresses .....	- 124 -
II.4.	Similarité des distances .....	- 126 -
II.5.	Seuil optimal de similarité et courbe ROC .....	- 126 -
II.5.	Algorithme de matching.....	- 127 -
III.	En route vers le clustering .....	- 129 -
	Conclusion.....	- 129 -
	Clustering .....	- 130 -
I.	Exécution de l'algorithme pour les restaurants et café .....	- 131 -
I.1	Transformations et distance.....	- 131 -
I.2.	Variables de segmentation .....	- 131 -
I.3.	Nombre de cluster retenus et performances .....	- 134 -
I.4 .	Interprétation des clusters.....	- 136 -
I.5.	Structure hiérarchique des clusters .....	- 141 -
II.	Exécution de l'algorithme pour les Hotels.....	- 141 -
III.	Classifications des PdV à l'aide du Random Forest.....	- 142 -
	Conclusion.....	- 142 -

Modélisation de la demande et optimisation des assortiments .....	- 143 -
Introduction .....	- 144 -
I.     Modèles de demandes.....	- 144 -
I.1. Variables et forme des modèles de demandes .....	- 144 -
I.2.     Aperçu des modèles de demandes.....	- 145 -
II. Optimisation des assortiments .....	- 148 -
Conclusion.....	- 150 -
Conclusion générale.....	- 152 -
Annexes .....	- 154 -
Bibliographie .....	- 155 -

# Liste des abréviations :

---

*AG : Agro-alimentaire/ Alimentation générale*

*API : Interface de Programmation d'Applications*

*BU : Business Unit*

*CAH: Classification Ascendante Hiérarchique*

*CMD : Customer master database (Base de données clients)*

*DBSCAN: Density-based spatial clustering of applications*

*DL: Dépôts de lait*

*ECCBC: Equatorial Coca-Cola Bottling Company*

*EPSG : Geodetic Parameter Dataset*

*FPR: Taux de faux positif*

*HORECA: Hotels-Restaurants-Café*

*INSEA: Institut National de Statistique et d'Economie Appliquée*

*MAE: Mean Absolute Error*

*MAPE: Mean Absolute Percentage Error*

*OSM: OpenStreetMap*

*PdV : Point de Ventes*

*PET: Polyéthylène Téréphthalate*

*POI: Point d'Intérêt*

*RED: Right Execution Daily (Indicateur d'execution correcte des ventes)*

*SIG: Système d'Information géographique*

*SKU: Stock Keeping Unit*

*TA : TripAdvisor*

*TPR : Taux de vrai positif*

*UC: Unit Case*

*WGS 84: World Geodetic System 1984*

# Liste des tableaux

---

Tableau 1 : Exemple d'ellipsoïdes de référence .....	- 35 -
Tableau 2: Les Distance utilisées en clustering .....	- 66 -
Tableau 3: Stratégies d'agrégation .....	- 73 -
Tableau 4: Couverture des données de TripAdvisor .....	- 107 -
Tableau 5: Vérifications des rayons de zones de chalandise .....	- 122 -
Tableau 6 :Exemple de similarité de texte.....	- 125 -
Tableau 7: Exemple de l'algorithme de matching.....	- 127 -
Tableau 8: Variables actives.....	- 132 -
Tableau 9: Variables inactives.....	- 133 -
Tableau 1 : liste des variables inactives avec une variabilité interclasses négligeable.....	145
Tableau 10: Evolution du nombre de classes et performances .....	- 135 -

# Liste des figures

---

Figure 1: Pays d' implantation d'ECCBC .....	- 17 -
Figure 2: Organigramme ECCBC .....	- 17 -
Figure 3: Les produits de la multinationale .....	- 19 -
Figure 4: Domaines de spécialité de groupe .....	- 20 -
Figure 5: Segmentation actuelle .....	- 21 -
Figure 6: Diagramme de Grant .....	- 25 -
Figure 7: Formes de représentation de l'information géographique .....	- 30 -
Figure 8 : modèles de données .....	- 33 -
Figure 9 : géoïde la plus précise à ce jour .....	- 34 -
Figure 10: Types de projection .....	- 36 -
Figure 11: Zone de chalandise .....	- 43 -
Figure 12: La loi de Reilly et point de partage .....	- 45 -
Figure 13: Délimitation de la zone de chalandise .....	- 46 -
Figure 14: Les types de grilles couramment utilisées .....	- 47 -
Figure 15: Découpage de la wilaya de Tizi Ouzou en grille hexagonale .....	- 48 -
Figure 16: Exemple de similarité des adresses .....	- 50 -
Figure 17: Mapping des techniques d'apprentissages non supervisé .....	- 61 -
Figure 18: Distance euclidienne et Manhattan .....	- 65 -
Figure 19: structure naturelle à trois segments .....	- 71 -
Figure 20 Illustration de l'algorithme de DBSCAN par rapport aux CAH et K-means .....	- 78 -
Figure 21: Illustration de l'epsilon voisinage .....	- 79 -
Figure 22: Mapping des techniques d'apprentissage supervisé .....	- 81 -
Figure 23: Weak learners et Strong learners .....	- 83 -
Figure 24: courbe ROC .....	- 90 -
Figure 25 : Données d'OpenStreetMap .....	- 101 -
Figure 26: Données de TripAdvisor .....	- 103 -
Figure 27: Etapes de scraping .....	- 104 -

Figure 28: Données de Flickr .....	- 108 -
Figure 29: Architecture du projet .....	- 109 -
Figure 30: Logo Python.....	- 110 -
Figure 31: Logo Databricks .....	- 112 -
Figure 32: Architecture d'Azure Databricks .....	- 113 -
Figure 33: Schéma de segmentation par pilote.....	- 114 -
Figure 34: Carte de la ville de Marrakech .....	- 115 -
Figure 35: Etapes de modélisation du projet .....	- 115 -
Figure 36: Distribution de PdV .....	- 117 -
Figure 37: Hexagonalisation du territoire de Marrakech.....	- 118 -
Figure 38: Ruralité de la Ville de Marrakech .....	- 120 -
Figure 39: Courbe ROC et sélection du seuil optimal .....	- 126 -
Figure 40 : Choix du nombre de cluster .....	- 135 -
Figure 41: Structure hiérarchique des Restaurant et Café .....	- 141 -
Figure 42: Structure hiérarchique des Hotels .....	- 141 -
Figure 43: Exemple d'assortiment du réfrigérateur .....	- 148 -
Figure 44: Algorithme d'optimisation de l'assortiment du refroidisseur.....	- 149 -

# Introduction générale

---

Parmi les défis auxquels sont confrontés les commerciaux, responsables de vente ou responsables marketing, il y a la question de la pertinence de l'offre au regard de la demande. Autrement, ces responsables s'interrogent sur la manière de proposer les bons produits et services, au bon moment ; de sorte à intéresser les clients et les consommateurs potentiels. L'une des méthodes de prédilection que propose le marketing est la segmentation de la clientèle ; un processus selon un point de vue business, qui consiste à diviser la clientèle en groupes de clients potentiels qui partagent des caractéristiques et des besoins similaires.

Cependant, jusque-là, la segmentation de la clientèle réalisée par la plupart des entreprises est basée entièrement sur les données de ventes internes de l'entreprise en question. D'ailleurs, le plus souvent ces données sont incohérentes, incomplètes ou même incorrectes. De plus, les entreprises les plus ambitieuses, qui réalisent des enquêtes sur de petits échantillons s'heurtent à des problèmes de représentativité. Les données recueillies sont généralement des informations déclaratives et ne reflètent donc pas la réalité.

Comment alors contourner ces faiblesses de la segmentation traditionnelle en entreprise ? Telle est la question que s'est posée ECCBC, un des embouteilleurs de Coca Cola en Afrique. C'est dans cette perspective que notre projet de fin d'études se présente comme une contribution à l'élaboration d'une segmentation basée sur une approche moderne, différente de la segmentation actuelle de l'entreprise. La nouvelle segmentation sera le point de départ pour optimiser les ventes et revenus en proposant nos produits à la clientèle adaptée. Les clients d'ECCBC sont les points de ventes, la population n'achetant pas directement les boissons Chez ECCBC mais auprès des points de ventes.

Notre objectif est alors dans un premier temps d'utiliser la technologie et des méthodes d'analyses avancées pour améliorer la connaissance du client. Les connaissances concernées sont le profil socio-économique, la situation géographique et son potentiel et sa fréquentation. Dans un second temps, nous exploiterons les données recueillies pour implémenter une segmentation optimale de la clientèle. La suite de cet exercice de clustering étant de modéliser la demande des segments par articles et permettre de proposer les boissons adéquates par segment.

Vu la complexité du projet, nous avons décidé de diviser notre rapport en deux grandes parties. La première est consacrée à l'introduction des concepts utilisés pour mener à bien ce travail. Elle repose sur quatre (04) chapitres ordonnés selon la problématique du projet. La deuxième se concentre sur la modélisation et présente les résultats obtenus en trois (03) chapitres. A cet effet, nous présentons la méthodologie de l'enrichissement de la base de données, les résultats du clustering réalisé, les modèles de demandes ainsi que leurs implications dans la recommandation d'assortiment qui maximiseront les ventes et profits de l'entreprise.

## **Chapitre introductif :**

# **Cadre contextuel et Méthodologique du travail**

Introduction .....	- 16 -
I. Présentation de l'organisme d'accueil.....	- 16 -
I.1. Equatorial Coca-Cola Bottling Company .....	- 16 -
I.1.1. Presentation de Equatorial Coca-Cola Bottling Company .....	- 16 -
I.1.2. Organigramme de Equatorial Coca-Cola Bottling Company.....	- 17 -
I.1.3. Business de l'entrepise .....	- 18 -
a) Système de production.....	- 18 -
b) Système de distribution.....	- 19 -
I.2. Le groupe Data Science et Business Intelligence .....	- 20 -
II. Cadre conceptuel et Problématique.....	- 21 -
II.1. Enoncé de la problématique .....	- 21 -
II.2. Objectif du projet.....	- 23 -
II .3. Principaux résultats attendus .....	- 23 -
II.4. Questions de recherche.....	- 23 -
III. Méthodologies de recherche :.....	- 24 -
IV. Conduite et planification du projet .....	- 25 -
Conclusion.....	- 26 -

## **Introduction**

L'objectif de chapitre est de présenter le contexte général du projet. Nous commencerons par présenter l'organisme d'accueil qui est Equatorial Coca-Cola Bottling Company et particulièrement la division de la Data Science & Advanced Analytics, ensuite, nous citerons les motivations et objectifs du projet. Pour clôturer ce chapitre, nous présenterons la conduite et la planification du projet.

### **I. Présentation de l'organisme d'accueil**

#### ***I.1. Equatorial Coca-Cola Bottling Company***

##### ***I.1.1. Presentation de Equatorial Coca-Cola Bottling Company***

Equatorial Coca-Cola Bottling Company (ECCBC) est l'entreprise partenaire d'embouteillage de The Coca-Cola Company en Afrique du Nord et de l'Ouest, où ses équipes produisent, commercialisent et distribuent les marques les plus appréciées au monde et un large choix de boissons de haute qualité. Son projet en Afrique a débuté en 1989 avec une concession de Coca-Cola Company pour opérer en Guinée équatoriale. L'entreprise qui en a résulté a ensuite été rejointe par la Guinée Conakry, la Mauritanie, le Cap-Vert, la Guinée Bissau et la Gambie. ECCBC, grande multinationale, a été créée en 1997 pour combiner les opérations de tous ces pays et servir de plate-forme pour la croissance future. Pour ce faire, elle participe, avec la Fondation Coca-Cola pour l'Afrique dans des projets orientés sur l'amélioration de l'éducation, l'accès à l'eau potable, la santé, l'esprit d'entreprise et la durabilité. Au cours des deux dernières décennies, la compagnie a étendu ses opérations à de nouveaux territoires en Afrique, tels que le Ghana, le Maroc et l'Algérie. Son siège social se trouve en Espagne avec un autre siège annexe à Casablanca. Les pays où sont implantés les Business Unit d'ECCBC sont recensés dans la figure ci-dessous.

Figure 1: Pays d' implantation d'ECCBC

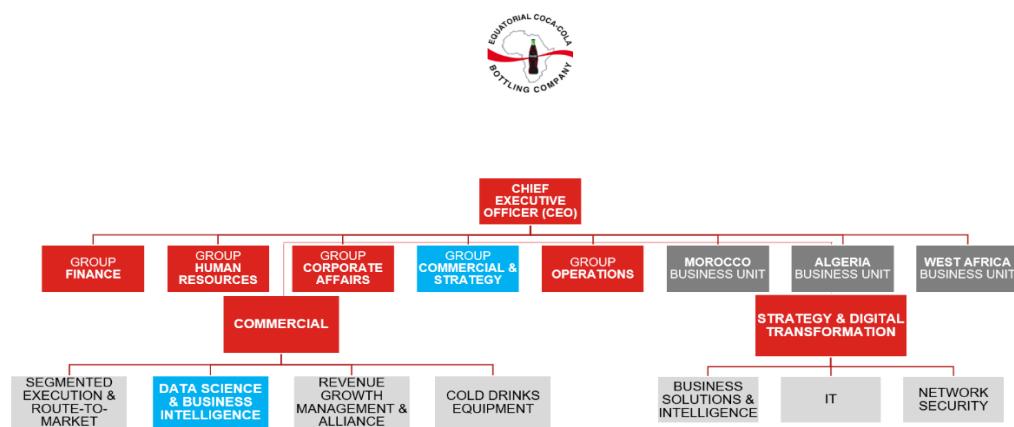


Source : rapport interne

### 1.1.2. Organigramme de Equatorial Coca-Cola Bottling Company

La multinationale est divisée en groupes et en Business Unit. Les Business Unit (B.U) sont les équipes locales. Au Maroc, il est connu sous le nom de ECCBC Morocco alors qu'en Algérie, on parle d'ECCBC Algeria. Quant aux groupes, ils sont divisés en équipe. Dans la figure ci-dessous, se trouve l'organigramme de l'entreprise avec un zoom sur la partie qui nous concerne qu'est la partie commerciale.

Figure 2: Organigramme ECCBC



Source : rapport interne

Les groupes ne dégagent pas des profits. Ils interviennent dans la gestion, la planification des stratégies et l'orientation des investissements des différents B.U.

### *1.1.3. Business de l'entreprise*

ECCBC produit, vend et distribue plusieurs types de produits grâce à ses BU.

#### a) Système de production

Les produits d'ECCBC sont en formats verre, PET<sup>1</sup> et les canettes.

Le marché des boissons est très concurrentiel. Nous pouvons diviser les produits d'ECCBC en deux grandes catégories que sont :

➤ Les boissons gazeuses :

- **Core Brand** : qui regroupe les marques phares des BU et surtout de Coca-Cola Company, et qui sont connues sous le nom : Coca-Cola, Sprite ;
- **Premium Brand** : qui comporte les marques qui ont moins d'impact en termes de rentabilité mais aussi en termes de symbolique, ces marques sont : Pom's, Hawaï et Schweppes ;
- **Mega Premium Brand** : regroupe des produits de luxe qui sont destinés à une clientèle spécifique. Ces produits sont : Coca-Cola Light et Coca-Cola Zero ;
- **B.Brand** : cette catégorie regroupe les marques dont le poids et l'impact viennent en dernier lieu, on cite Top's.

➤ Les boissons non gazeuses :

- Les eaux industrielles : Ciel et BONAQUA,
- Les jus de Fruits : Miami,
- Les boissons énergétiques et isotoniques : BURN, MONSTER, Acquarius .

---

<sup>1</sup> Les trois lettres PET forment l'abréviation de polyéthylène-téréphtalate. Le PET est utilisé dans la fabrication des bouteilles en plastique

**Figure 3: Les produits de la multinationale**



Source : rapport interne

### b) Système de distribution

La distribution des produits sur le territoire agréé d'un Business Unit fait appel à une logistique très complexes. Pour gérer les commandes et livrer les clients, toutes catégories confondues, ECCBC utilise les systèmes de distribution suivants.

- Le système conventionnel :

Les vendeurs visitent les points de ventes avec le camion chargé de boissons deux fois par semaine pour la distribution des produits.

- Le système de la prévente :

Les tâches de la prise des commandes et de la livraison sont séparées. Le pré-vendeur s'occupe de la collecte des commandes auprès des clients grâce à une application. Ces PdV sont livrés par camion le lendemain.

- Le système moderne :

Touche les grandes et moyennes surfaces, en l'occurrence des supermarchés.

- Le système indirect :

Concerne les semi-grossistes situés dans les zones où il y a très peu de points de vente (PdV), les semi-grossistes se chargent de livrer les produits à ces points de vente.

## I.2. Le groupe Data Science et Business Intelligence

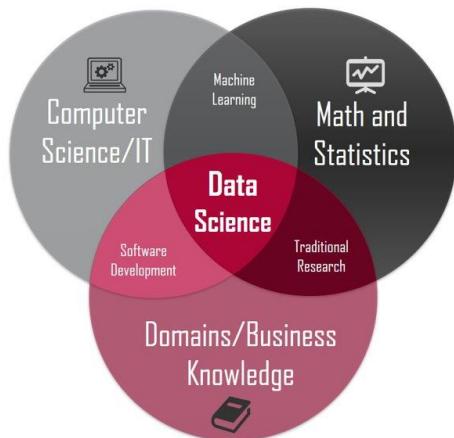
La Division de la Data Science et Business Intelligence (DS&BI) est l'entité responsable des systèmes intelligents et de l'Advanced Analytics de l'entreprise. L'organisation du DS&BI a été mise en place afin de servir différents pôles de l'entreprise et de ses différentes BU au niveau africain.

Le stage faisant objet du présent mémoire de fin d'études s'est déroulé plus précisément au sein de cette entité. La DS&BI a pour mission de faire des données un atout stratégique et les mettre au premier plan pour prendre des décisions éclairées, permettant ainsi une croissance durable et aidant ECCBC à devenir le premier embouteilleur du système Coca-Cola en Afrique.

Les principales missions de l'entité sont :

- La collecte des données des autres entités,
- La détection des limites et la validation de la fiabilité des modèles prédictifs en vue de leur automatisation et intégration dans les processus opérationnels,
- Les prédictions de ventes,
- La définition du champ d'usage des algorithmes d'analyse.
- Le traitement des données selon le besoin demandé,
- L'exposition des résultats dans des tableaux de bords interactifs,

**Figure 4: Domaines de spécialité de groupe**



## II. Cadre conceptuel et Problématique

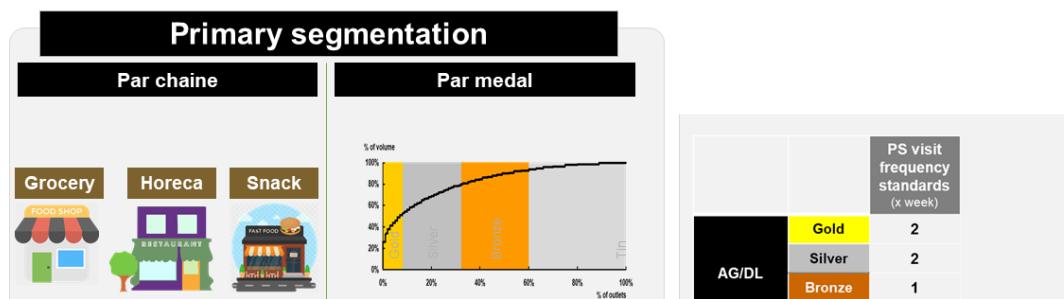
### II.1. Enoncé de la problématique

Ces dernières années ECCBC a mené des études pour mieux comprendre ses forces et ses faiblesses afin de mieux préparer une croissance durable dans les années à venir. Le cabinet Deloitte a eu à recommander trois grands champs dans lesquels l'entreprise devrait investir très prochainement si elle veut réaliser une croissance horizontale et croissance verticale. Une de ces trois recommandations est la segmentation dynamique de la clientèle en intégrant des informations externes. Une segmentation dynamique parce que chaque année, elle sera mise à jour.

La segmentation actuelle des BU repose actuellement sur 2 dimensions qui animent la politique de service client.

Figure 5: Segmentation actuelle

**La segmentation NABC repose actuellement uniquement sur les critères qui reposent sur la politique de service client**

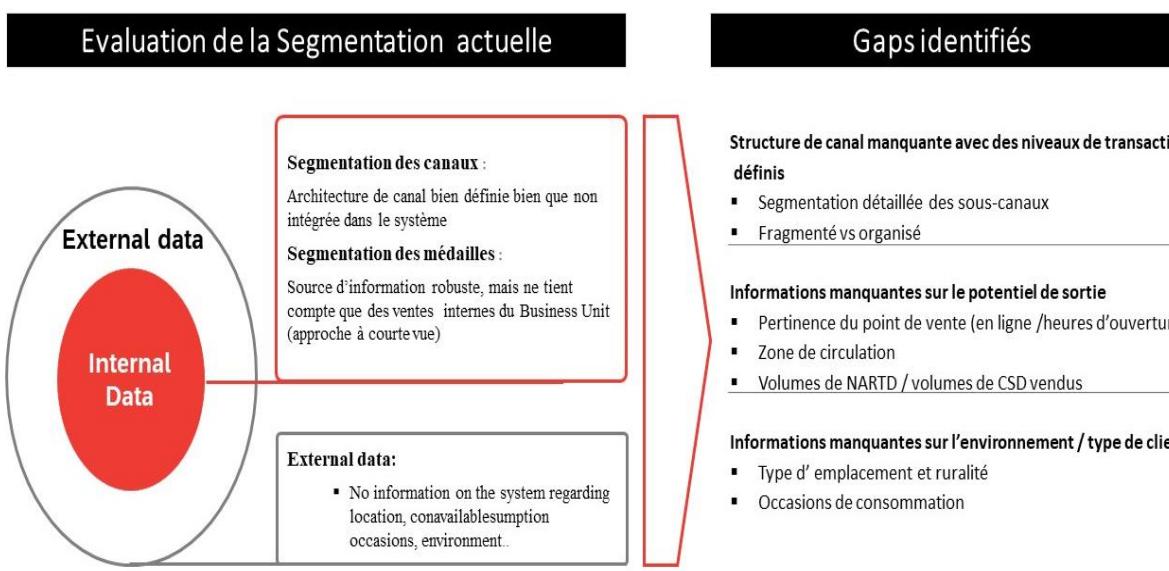


La segmentation par chaîne est basée sur les caractéristiques des points de ventes. Ce sont :

- *Les grosseries* : Il est composé des AG c'est-à-dire des points de ventes agroalimentaires (les épiciers) et des DL (dépôts de lait). Les DL représentent les points de ventes traditionnels de lait ;
- *Les HORECA* : c'est une abréviation d'Hôtel, Restaurant et Café ;
- La dernière classe est constituée des *Snacks*.

La segmentation par médaille est une segmentation simpliste qui est basées sur les performances de ventes des points de ventes. Cela est fait en représentant la courbe cumulée des ventes afin de déterminer les différentes classes. Il existe quatre (04) types de médaille à savoir Gold, Silver, Bronze and Tin. En fonction de ces deux segmentations, des services sont adaptés. Par exemple, Les client Or et de type AG aura des services de visite de ventes 2 fois par semaine.

La remarque importante dans ces deux simples segmentations est la non-utilisation d'informations extérieures à l'entreprise. La segmentation actuelle présente des lacunes évidentes qui ont une incidence sur la capacité d'exécution de la multinationale. Pour bien caricaturer les faiblesses de la segmentation actuelle, nous avons recensé les points suivants :



1

Il devient alors légitime de se poser les questions suivantes :

- Quels types d'informations externes utilisées ?
- En utilisant les techniques d'apprentissage statistiques, quelle nouvelle segmentation peut-on mettre en œuvre ?
- Quel service définir par segments ?

C'est dans cette dynamique qu'intervient notre projet de fin d'études afin de donner des réponses à ces questions.

## ***II.2. Objectif du projet***

L'objectif du présent projet s'articule autour des objectifs suivants :

- Enrichir la base de données avec des données externes (localisation, environnement, économie...)
- Définir la segmentation optimale
- Modélisation de la demande des produits par clusters
- Recommander l'assortiment optimal (les produits à placer dans le réfrigérateur des clients par selon le cluster d'appartenance) afin de maximiser les volumes vendus des PdV.

## ***II .3. Principaux résultats attendus***

- Enrichissement des données internes avec des informations externes sur les points de vente : *Occasion, Premiumness, Trendiness, POI Location, Touristiness* qui sont pertinentes pour un meilleur placement de produit.
- Un certain nombre de sous-segments dans la segmentation actuelle, homogènes en ce qui concerne les caractéristiques externes des points de vente.
- Liste des SKU<sup>2</sup> recommandés par sous-segment. La liste est classée en fonction de l'adéquation de chaque SKU aux segments.
- Documentation technique et algorithmes codés en Python.

## ***II.4. Questions de recherche***

Objectif 1 :

- ✓ Quelle est la structure de la base de données opérationnelle de l'organisme ECCBC  
?

---

<sup>2</sup> Un sku est identifiant propre à chaque type de produit. Autrement, un SKU correspond à un article.

- ✓ Quelles sont les différentes données externes à utiliser ?
- ✓ Comment fusionner les données externes et internes ?

Objectif 2 : Quelles informations du business inclure dans les modèles ?

Objectif 3 et 4 : Quels modèles du Machine Learning choisir pour dériver une segmentation clientèle optimale et en modélisant la demande et les planogrammes ?

### **III. Méthodologies de recherche :**

Les méthodes de recherche permettent la collecte des données et des informations utiles qui serviront après une analyse détaillée à la prise de décision. Pour cette étude nous avons choisi la méthode documentaire et les entrevues.

La méthode documentaire :

La méthode documentaire consiste au recueil des informations internes de l'organisme pour comprendre le fonctionnement et le besoin de différents services afin de réaliser une bonne segmentation. Grâce à cette méthode nous avons cerné les documents relatifs au sujet du Business Intelligence et des techniques statistiques afin de clarifier les idées et appréhender le sujet globalement pour élaborer nos modèles.

Les entrevues :

Cette approche consiste à réaliser un ensemble d'entrevues et de réunions avec le tuteur, les cadres de l'équipe DS&BI et également les cadres du département commercial afin de préciser le sujet et ses objectifs, et discuter les différentes méthodes de sa réalisation. Ces réunions ont permis de souligner les grandes lignes de ce projet et d'étudier les besoins informationnels. Ces réunions semi-officielles sont la source de toutes les données nécessaires pour le déroulement de ce projet afin de se projeter dans l'atteinte des résultats.

## IV. Conduite et planification du projet

Afin d'avoir une vision globale des étapes de déroulement du projet et de pouvoir gérer au mieux le temps qui est notre durée de stage, un planning de projet est donc primordial. Ce dernier permettra la structuration et l'organisation des différentes tâches journalières pour aboutir à un travail cohérent et efficace au sein de l'organisme.

Le planning du projet s'est fait en plusieurs phases qui se chevauchaient parfois.

Phase 1 : Compréhension de la problématique et business inputs

Phase 2 : Acquisition des données

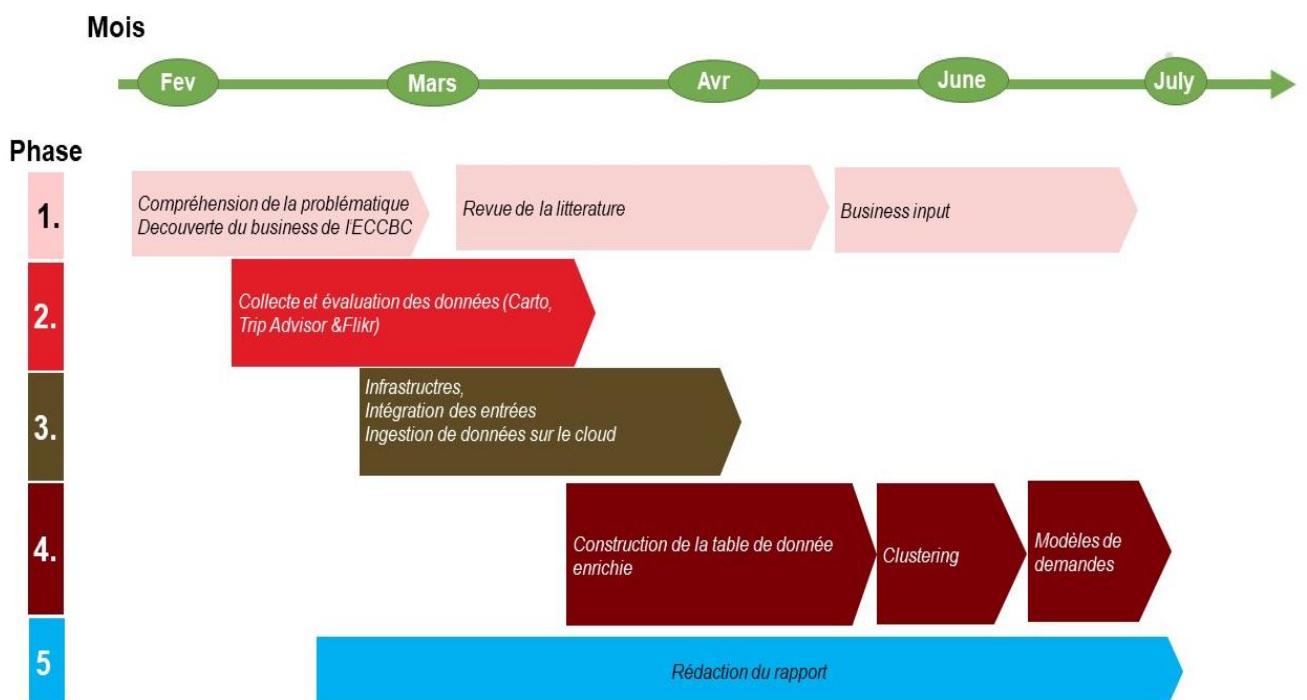
Phase 3 : Mise en place de l'infrastructure de travail sur le cloud

Phase 4 : Modélisation

Phase 5 : Rédaction du rapport de

Le diagramme de Gantt est comme suit :

Figure 6: Diagramme de Gant



## **Conclusion**

Dans ce chapitre nous avons présenté le cadre général de notre projet, puis nous avons déterminé la problématique, les questions de recherche et la méthodologie empruntée dans la réalisation du travail.

Nous enchaînons avec la partie théorique à travers laquelle nous essaierons de définir les concepts théoriques exploités dans le projet et les dessous de la littérature.

# Première partie

Eléments théoriques

## Chapitre 1 :

---

# Introduction à la géomatique

Introduction.....	- 29 -
I. Concepts généraux.....	- 29 -
I.1. Définitions .....	- 29 -
I.2. Composantes de l'information géographique.....	- 29 -
I.3. Différents types de données spatiales.....	- 31 -
I.4. Représentation numérique des données spatiales.....	- 32 -
II. Les fondements de la localisation .....	- 33 -
II .1. Système géodésique et Ellipsoïdes de référence.....	- 34 -
II.2. Les systèmes de coordonnées géographiques .....	- 35 -
II.3. Les projections cartographiques .....	- 36 -
II .4. Distance géodésique .....	- 38 -
Conclusion .....	- 39 -

# Introduction

Le but de chapitre est d'introduire les notions liées à l'information géographique. Avec la digitalisation, des données de localisation sont générées chaque jour par les milliers d'utilisateurs de web et d'applications dans le monde. Comment localiser un utilisateur sur la terre ? A quelle distance se trouve-t-il d'un autre utilisateur donné. Ce chapitre donne un aperçu des fondamentaux pour comprendre les données spatiales.

## I. Concepts généraux

### *I.1. Définitions*

Le terme « Géomatique » a été proposé pour la première fois en 1960 par Bernard Dubuisson. Il s'agit d'un néologisme qui est le résultat de combinaison de « géographie » et « informatique ». Le choix de ce terme a été motivé par les changements et les avancées qu' a apportée l'informatique. Actuellement ce terme est utilisé à travers le monde entier pour designer l'ensemble des outils et des méthodes permettant d'acquérir, de représenter, d'analyser et d'intégrer des données géographiques.

On appelle information géographique, toute représentation d'un objet ou d'un phénomène réel ou imaginaire, présent, passé ou futur, localisé dans l'espace à un moment donné.

### *I.2. Composantes de l'information géographique*

Généralement, on associe deux composantes de l'information géographique qui sont le niveau **sémantique** et le niveau **géométrique**.

Par niveau sémantique de l'informatique géographique, on parle de la nature ou de l'aspect l'objet ou encore son identité. Exemple : le nom de la route qui fait face au portail de l'INSEA, le nom d'une rivière, la taille de la population d'une Wilaya. Quant au niveau géométrique, il fait référence à la forme de la localisation de l'objet. C'est dans ce sens qu'on parle également de système de coordonnées afin de permettre cette localisation. Ce système de coordonnées peut être globale et valable partout sur terre

comme le système géodésique mondial WGS84. Ce système de coordonnées nous intéressera particulièrement pour la suite de ce travail. Le système de coordonnées peut être un système de coordonnées relatives par rapport à un point d'origine quelconque. On en veut pour preuve les relevés topographiques.

L'information géographique que nous utiliserons peut se représenter sous trois formes qui sont distinctes. En premier lieu, nous avons la représentation en image (photo satellitaire). Cette représentation à elle seule ne donne aucune information sur le niveau sémantique de l'objet. En deuxième lieu, nous avons la représentation cartographique qui est la représentation par excellence de l'information géographique pour sa clarté et son interprétation facile, cela grâce aux repères et légendes qui y sont associées. En dernier, lieu, nous avons la représentation que nous utiliserons lorsque nous passerons à la modélisation de notre projet. Il s'agit de la représentation par un texte (adresse, et autres attributs) ou un fichier de données littérales. Ces trois formes de représentation sont distinctes mais complémentaires (Figure 7) :

Figure 7: Formes de représentation de l'information géographique



Texte : sémantique sans géométrie

Etablissement	Adresse
Ecole Supérieure de Génie Biomédical	Zénith mellinium, Bâtiment 6, Sidi Maarouf , Casablanca
Ecole des Sciences de l'Information – Rabat	Avenue Allal El Fassi, cité Al Irfane, BP: 6204 Rabat-Instituts
Institut National de Statistique et d'Economie Appliquée INSEA	Avenue Allal El Fassi, B.P. 6217. Rabat-Instituts

### I.3. Différents types de données spatiales

« Une donnée spatiale est une observation dont on connaît non seulement la valeur, mais aussi la localisation ».<sup>3</sup> D'après la classification proposée par CRESSIE<sup>4</sup>, on distingue trois types de données spatiales à savoir les données ponctuelles, les données continues et les données surfaciques. Cependant, il rappelle que la distinction entre les données ne s'observe pas au niveau de la taille de l'unité géographique mais au niveau du processus générateur des données.

Les données spatiales ponctuelles se caractérisent par la distribution dans l'espace des observations. Le processus générateur des données génère les coordonnées géographiques associées à l'apparition d'une observation. On n'étudie pas de valeur associée à l'observation ; seule compte la localisation. Comme exemple, il y a la distribution des superettes ou des AG dans la ville de Rabat. L'analyse de ces types des données revient principalement à mesurer l'ampleur de l'écart entre la distribution spatiale des observations et une distribution complètement aléatoire dans l'espace. Dans le cas où l'on trouve que la distribution de ces points n'est pas aléatoire à travers la ville, on pourrait alors identifier des clusters et mesurer leur significativité. On parle de données continues lorsqu'il existe une valeur pour la variable d'intérêt en tout point du territoire étudié. Les données sont générées de façon continue sur un sous ensemble de  $R^2$ . En revanche, ces données sont mesurées uniquement en un nombre discret de points. Il s'agit, par exemple, de la composition chimique du sol (utile à l'industrie minière), de la qualité de l'eau ou de l'air (pour des études sur la pollution), ou encore de diverses variables météorologiques. Dans les faits, on appelle géostatistique l'analyse spatiale des données continue. Ce domaine cherche à prédire la valeur d'une variable en un point où elle n'a pas été échantillonnée. En plus, il intervient dans l'optimisation des plans

---

<sup>3</sup> Insee - Eurostat Manuel d'analyse spatiale, Théorie et mise en œuvre pratique avec R, *Insee Méthodes* n° 131- octobre 2018, p.4

<sup>4</sup> CRESSIE, Noel A.C. (1993b). « Statistics for spatial data : Wiley series in probability and statistics ». Wiley-Interscience, New York 15, p. 105–209.

d'échantillonnage des données. Dans ces groupes de données se trouvent également les rues, les fleuves, les rivières ...

Les données surfaciques ne désignent pas nécessairement les objets surfaciques comme un stade de football. En effet, pour une donnée surfacique, la localisation des observations est considérée comme fixe, mais les valeurs associées sont générées suivant un processus aléatoire. Ces données caractérisent le plus souvent une partition du territoire en zones contigües, mais elles peuvent également être des points fixes du territoire. Il s'agit, par exemple, du PIB par région, Les ventes annuels du BU au Maroc par Wilaya ou du nombre de réfrigérateurs Coca Cola à réparer par Commune à Marrakech. Ici, l'objet de l'analyse spatiale est d'étudier les relations entre les valeurs des observations voisines. C'est pourquoi, lorsqu'on s'intéresse à des données surfaciques, la structure de voisinage des observations est une étape à étudier préalablement avant toute recherche des influences potentielles d'un voisin à un autre. La zone de chalandise d'une supérette ou d'un AG est également un exemple de données surfacique.

A données ponctuelles, données continues et données surfaciques on associe respectivement point, lignes polygones pour des raisons de représentation numériques.

#### *I.4. Représentation numérique des données spatiales*

Lorsqu'on se penche dans une vue numérique, les géomaticiens et géographes considèrent deux modèles de représentation des données que sont le modèle raster ou maillée et le modèle vectoriel.

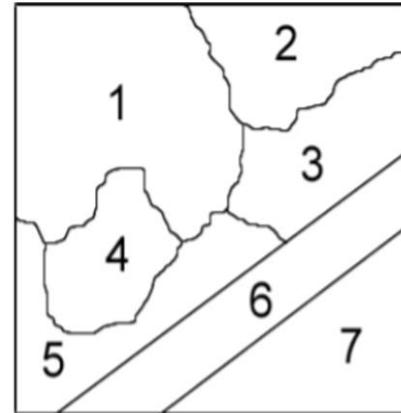
Un raster est composé au sens mathématique d'une matrice au de pixels arrangés en lignes et colonnes pour ainsi former une grille dans lequel est représenté, avec une relative précision, un ensemble d'informations. Les rasters peuvent avoir plusieurs origines, photographies aériennes numériques ou encore images satellites et numériques. Quant aux modèles vectoriels, la représentation est faite dans un premier temps par un ou plusieurs points grâce à leurs coordonnées qui composent la forme de l'objet. Dans un second temps, les points sont joints au point suivant par un segment de droite. Les cartes Raster sont adaptés aux calculs algébriques et également aux

données continues alors que les modèles vectoriels sont adaptés aux données discrètes. En plus, comparés aux rasters, ils sont plus manipulables et modifiables et de tailles de stockage plus faible.

Figure 8 : modèles de données



**Image**



**Vecteurs**

r

1	1	1	1	1	1	2	2	2	2
1	1	1	1	1	1	2	2	2	2
1	1	1	1	1	1	1	2	3	3
1	1	1	1	1	1	1	3	3	3
1	1	4	4	1	3	3	3	3	6
5	4	4	4	5	5	5	5	6	6
5	4	4	5	5	6	6	6	6	7
5	5	5	6	6	6	7	7	7	7
5	5	6	6	7	7	7	7	7	7

**Raster**

Source : nos travaux

Terrains de culture différentes (image à gauche). La représentations numériques nous permettent d'apercevoir les mêmes objets sur l'images.

## II. Les fondements de la localisation

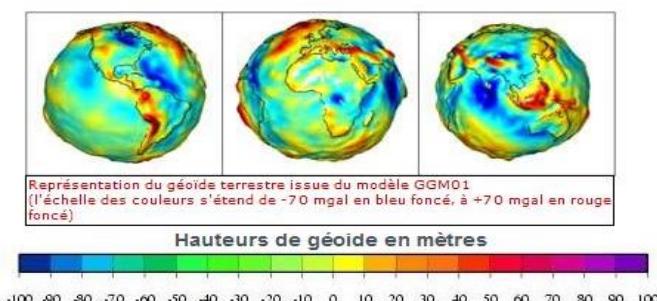
La définition du mode de localisation et de la projection cartographique sont des aspects qui seront abordées dans ce paragraphe. Avant d'aller plus loin, on rappelle que

nous pouvons nous repérer sur la terre de deux façons. La première est textuelle. Il s'agit là par exemple du nom de l'endroit où l'on veut se rendre, l'adresse postale, le numéro de la parcelle cadastrale. Ce référencement fait partie du quotidien des hommes. D'ailleurs, les agents des impôts ou également les agents d'abonnements à l'électricité utilisent principalement ce mode de localisation. Dans ce paragraphe, nous donnons plus d'attention au second mode qui est assez formalisé. Nous utilisons ce mode de localisation afin de mener à bien notre modélisation. On utilise dans ce cas des coordonnées d'un système de référence.

## *II .1. Système géodésique et Ellipsoïdes de référence.*

Au début, les navigateurs étant les premiers à utiliser ce mode de localisation avaient comme référence les étoiles. Mais lorsqu'on s'est rendu compte que la terre avait la forme d'une sphère aplatie, les mathématiciens ont contribué à la définition d'un système de coordonnées basées sur un ellipsoïde de révolution. Cependant, il y a problème car la forme de la terre n'est pas assez elliptique qu'on le modélisait si bien qu'on pourrait utiliser plusieurs ellipsoïdes pour la modéliser. Au cours du XIXème siècle, il s'est révélé que le géoïde est une représentation de la surface terrestre plus précise que l'approximation sphérique ou ellipsoïdale. Ce géoïde correspond à une equipotentielle du champ de gravité terrestre, choisie de manière à coller au plus près à la surface réelle. Aujourd'hui, les altitudes sont données par rapport à ce géoïde. Le géoïde le plus précis approxime un rayon 6 370 km pour le demi-grand axe de l'ellipsoïde et 6 350 km pour le demi-petit axe une valeur d'environ.

Figure 9 : géoïde la plus précise à ce jour



Source : GéoWiki

Nonobstant, s'il n'y a qu'un géoïde reconnu comme plus précis à un instant donné, plusieurs ellipsoïdes de référence candidats plus précisément plusieurs approximations du géoïde existent.

Tableau 2 : exemples d'ellipsoïdes de référence

Système géodésique	Ellipsoïde	Demi grand axe	Demi petit axe	Inverse de l'application	Origine
NTF	Clarke	6378	6356	293,466	Paris Grade
ED 50	Hayford	6378	6356	297	Postdam Degré
RGF 93	IAG GRS	6378	6356	298,257	Greenwich Degré
W84GS	W84GS	6378	6356	298,257	Greenwich Degré

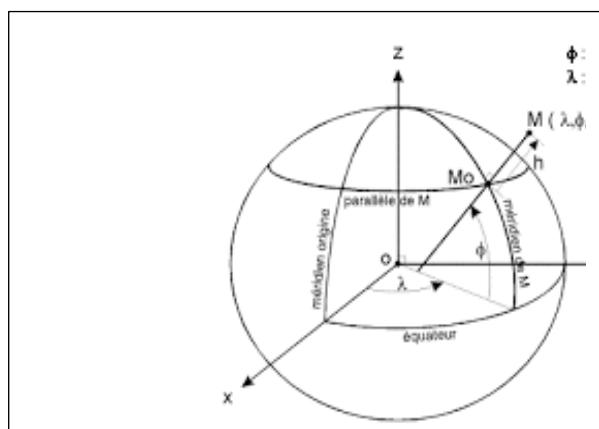
Source : élaboré par nos soins

## II.2. Les systèmes de coordonnées géographiques

Pour se localiser sur la terre, il est nécessaire d'utiliser un système de coordonnées basé sur un système de référence géodésique (datum) et d'un ellipsoïde. A chaque point est défini en des coordonnées géographiques qui ne sont rien d'autres que le **repérage** angulaire sur l'ellipsoïde mais également des coordonnées planes qui sont obtenues par projection. Les coordonnées géographiques d'un point à la surface de la terre sont représentées par trois éléments :

- la longitude  $\lambda$ : angle entre le méridien d'origine et le méridien du point M
- la latitude  $\phi$  : angle entre l'équateur et le parallèle du point M
- la hauteur ou l'altitude  $h$  : distance entre l'ellipsoïde et le point M.

En général, le méridien de Greenwich est considéré comme méridien d'origine puisqu'il est défini également comme méridien international. Cependant, l'origine peut être un méridien local propre à la géodésie d'un pays. La longitude et la latitude exprimées en degrés décimaux ou degrés minutes secondes sexdécimaux. Parfois, elles sont exprimées également à grad.



La représentation cartographique d'un territoire se fait dans le plan, ainsi il faut trouver un moyen de quitter des coordonnées géographiques vers des coordonnées planes : c'est la projection.

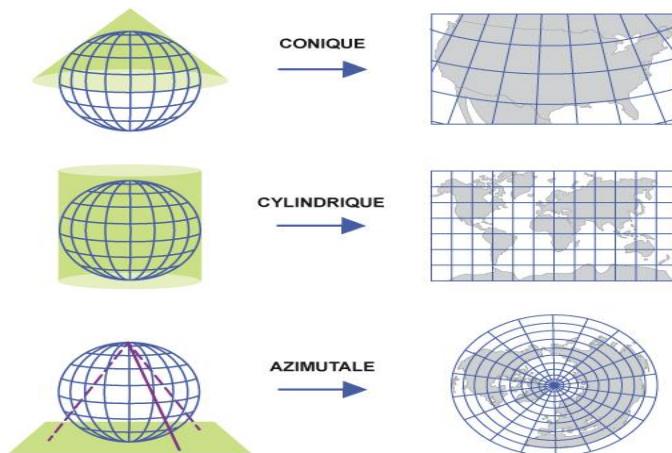
### **II.3. Les projections cartographiques**

Ces projections sont des techniques géodésiques qui ont pour objectifs la représentation de la forme de terre ou d'une portion de la terre sur une surface plane. Cependant cette transformation mène à une sorte d'arbitrage entre plusieurs restrictions car toutes les propriétés de l'espace de départ ne peuvent être conservées. En fonction de ces restrictions, il existe une classification des projections en trois types. Ce sont :

- ✓ la projection équivalente qui conserve localement les surfaces mais à avec une distorsion des surfaces ;
- ✓ la projection conforme qui conserve localement les angles, donc les formes avec une restriction qui est la distorsion des angles ;
- ✓ La projection aphylactique qui ne conserve ni les surfaces, ni les angles, mais elle peut être équidistante, c'est-à-dire conserver les distances sur les méridiens.

Pour obtenir une carte, des méthodes de la projection procèdent par la représentation de l'ellipsoïde sur une surface développable. Une surface développable permet de conserver les surfaces de départ. Elle est associée seulement aux cônes, les cylindres et le plan. Selon, la forme utilisée on trouve les projections coniques, les projections cylindriques et les projections azimutale.

**Figure 10: Types de projection**



Les autres projections différentes de ces trois sont appelés projections uniques ou individuelles. Les systèmes de coordonnées géographiques sont souvent identifiés par un code EPSG. Les deux systèmes les plus utilisés dans les applications de cartographie web et qui nous intéressent particulièrement sont EPSG: 4326 et EPSG: 3857.

EPSG: 4326 (alias WGS84, non projeté) est un système géographique de coordonnées non-projeté. On y trouve les longitudes et latitudes. Ses unités sont décimales mais ce système de coordonnées ne peut pas être affiché de manière significative sur une carte plate. Quant à EPSG :3857 (encore appelée Pseudo-Mercator, ou Web Mercator) est un système de coordonnées projeté. Il s'agit du système de coordonnées utilisé par Google Maps et à peu près toutes les autres applications de cartographie web. Il s'agit là d'une légère modification de la projection Mercator. Tout d'abord, il faut savoir que la projection Mercator est l'un des plus utilisées actuellement en navigation et généralement. Il s'agit d'une projection conforme et cylindrique. Alors que les formules de Web Mercator sont pour la forme sphérique du Mercator, les coordonnées géographiques doivent être dans le datum ellipsoïdale WGS84. Cet écart fait que la projection est légèrement non conforme ,et ces défaillances deviennent plus importantes lorsqu'on s'éloigne de l'équateur. Cet écart peut atteindre jusqu'à 40 km au sol.<sup>5</sup>. D'ailleurs, c'est pourquoi European Petroleum Survey Group, référence internationale de l'information géographique a annoncé ce qui suit :« Nous avons examiné le système de référence de coordonnées utilisé par Microsoft, Google, etc. et croyons qu'il est techniquement défectueux. Nous ne dévaluerons pas l'ensemble de données EPSG en incluant une géodésie et une cartographie aussi inappropriées ».

Souvent, les données sont stockées dans EPSG :4326 et affichées dans EPSG :3857. En outre, une API cartographique peut prendre latitude, longitude (EPSG: 4326 ) comme une entrée, mais lorsque ces coordonnées sont affichées sur une carte, ils seront montrés une carte basée sur un Mercator Web (c.-à-EPSG:3857) projection . C'est pourquoi nous ramènerons à EPSG :4326 à chaque fois que nous utiliserons les données de localisation pour des analyses quantitatives.

---

<sup>5</sup> Battersby, Sarah E.; Finn, Michael P.; Usery, E. Lynn; Yamamoto, Kristina H. (2014). Implications of Web Mercator and Its use in online Mapping.

Le système géodésique local utilisé au Maroc et en Algérie est ' Ellipsoïde de Clarke 1880 et Datum Merchich.<sup>6</sup> Les projections appliquées sont généralement conique Lambert. Mais les spécificités diffèrent d'une région à une autre. C'est pourquoi les données de localisation relatives aux découpages administratifs seront tous recodées par rapport à EPSG : 4326. En effet, La conversion d'un système de coordonnées projeté à un autre nécessite l'utilisation des équations de projection inverse.<sup>7</sup>

## ***II .4. Distance géodésique***

Comme nous l'avons mentionné plus haut, la terre est un géoïde. Elle n'a donc pas de forme mathématiquement parfaite. Par conséquent, le choix de la méthode de calcul de la distance à la surface de la terre affecte grandement la précision des résultats à obtenir. La formule Haversine est une méthode qui calcule la distance considérée de manière appropriée et précise<sup>8</sup> en faisant l'hypothèse que la terre est une sphère.

La loi d'Haversine s'écrit comme suit :

$$\text{hav}\left(\frac{d}{r}\right) = \text{hav}(\phi_2 - \phi_1) + \cos(\phi_1)\cos(\phi_2)\text{hav}(\lambda_1 - \lambda_2) \quad (1)$$

avec  $\text{hav}\left(\frac{d}{r}\right) = \sin^2\left(\frac{\theta}{2}\right)$

Avec  $r$  étant le rayon de la terre, une constante est fixée à = 6371 Km.  $\text{hav}$  et  $d$  représentent respectivement la fonction de Haversine et la distance géodésique entre les deux points sur la terre.

<sup>6</sup> Comment savoir dans quel système de référence et dans quelle projection a été réalisée une ancienne carte IGN de l'Algérie, de la Tunisie ou du Maroc ? Institut national de l'information géographique et forestière, Paris.

<sup>7</sup> Cain, Jim (9 May 2013). "Coordinate Reference Systems (Best Practices for Assignment, Manipulation and Conversion in GIS Systems)" (PDF). 2013 ESRI Petroleum GIS Conference.

<sup>8</sup> Prasetyo, D., & Hastuti, K. (2012). Application of Haversine Formula for selecting Location and Christian Church based on Mobile. Semarang: Universitas Dian Nuswantoro

Obtenir la distance  $d$  revient juste à une résolution de l'équation ( 1 ).

$$\text{hav} \left( \frac{d}{r} \right) = \text{hav}(\phi_2 - \phi_1) + \cos(\phi_1) * \cos(\phi_2) * \text{hav}(\lambda_1 - \lambda_2)$$

$$\Leftrightarrow d = r * \text{hav}^{-1}(h) \text{ avec } h = \text{hav} \left( \frac{d}{r} \right)$$

$$\Leftrightarrow d = 2 * r * \arcsin \left( \sqrt{\text{hav}(\phi_2 - \phi_1) + \cos(\phi_1) * \cos(\phi_2) * \text{hav}(\lambda_1 - \lambda_2)} \right)$$

$$\Leftrightarrow d = 2 * r * \arcsin \left( \sqrt{r * \sin^2 \left( \frac{\phi_2 - \phi_1}{2} \right) + \cos(\phi_1) * \cos(\phi_2) * \sin^2 \left( \frac{\lambda_1 - \lambda_2}{2} \right)} \right)$$

L'hypothèse de la terre comme une sphère a amené le géodésien Vincenty à mettre en place une autre distance géodésique qui part de l'hypothèse que la terre a une forme ellipsoïdale. Cette distance est appelée distance de Vincenty. Elle a l'avantage d'être plus précis que la distance Haversine en revanche son calcul est basé sur une méthode itérative et peut ne pas converger souvent pour deux points presque antipodaux<sup>9</sup>, c'est-à-dire deux points opposés par le centre de la terre. Le choix de la distance à utiliser va donc dépendre de la précision voulue et également du temps d'exécution du calcul. Les distances données par Google Maps ou OpenStreetMap sont basées sur la distance Haversine.

## Conclusion

A la sortie de ce chapitre, nous retenons que les points sont localisés sur la terre à partir d'un système géodésique. Lorsque nous avons une base de données spatiales, nous pouvons désormais savoir le type de données spatiales utilisées. Néanmoins, nous sommes avisés que la vérification du EPSG est une nécessité avant toute type d'analyse spatiale sinon nous aurions de faux calculs dans les conditions où les systèmes de coordonnées diffèrent. Dans le chapitre suivant, nous utilisons ces acquis pour formaliser la définition des zones de chalandise.

---

<sup>9</sup> Karney, C.F.F. Algorithms for geodesics. *J Geod* **87**, 43–55 (2013).

## Chapitre 2 :

---

Introduction.....	- 41 -
I. Le comportement spatial du consommateur .....	- 41 -
II. le « trading area » ou zone de chalandise.....	- 42 -
II.1. Définitions.....	- 42 -
II.2. Modèles de zone de chalandise .....	- 44 -
II.2. Index spatial .....	- 47 -
Conclusion .....	- 48 -

## Introduction

L'analyse spatiale est un type d'analyse géographique qui cherche à expliquer les modèles de comportement humain et son expression spatiale en termes de mathématiques, statistiques et de géométrie comme l'analyse du voisin le plus proche et l'analyse des bassins versants et bien d'autres. La microéconomie y trouve une place très importante. Les modèles, à travers des conditions basées sur la microéconomie telles que le principe du moindre effort des consommateurs pour acquérir les produits dont ils ont besoin ou la maximisation des bénéfices des entreprises, prédisent les modèles spatiaux qui doivent se produire. L'objectif des méthodes d'analyse spatiale est de déterminer les caractéristiques de la répartition spatiale des individus géographiques (ménages, entreprises, AG ou restaurants) ou de leurs valeurs (ventes pour une AG donnée ou la clientèle d'un restaurant).

### I. Le comportement spatial du consommateur

La demande n'est pas la même à travers l'espace. Empiriquement, on peut la mesurer en fonction du revenu, du nombre de ménages, des styles de vie. En effet, une des fonctions de demande les plus utilisées en modélisation économiques et qui présente moins de défauts d'ordre théoriques est le système AIDS (Almost Ideal Demand System , AIDS) . Pour un ménage donné, ce dernier relie le coefficient budgétaire du bien i noté  $w_i$ , au logarithme du revenu total,  $R$ . Lorsqu'on suppose l'existence de n biens au sens économique, l'on pourrait dire que la demande du SKU Hawaï de ECCBC Algeria ou Morocco s'écrit :

$$w_i = \alpha + p_i \log(R_i) + u_i \quad \text{Avec } w_i = \frac{p_i q_i}{\sum_{j=1}^n p_j q_j}$$

Avec  $p_i$  et  $q_i$  sont respectivement le prix et la quantité acquise en bien i qui est ici le SKU Hawaï, Comme tout modèle économétrique,  $u_i$  est l'aléa du modèle.

Le terme de dénominateur n'est rien d'autre que la dépense totale,

Dès lors que nous savons que les revenus sont distribués inégalement à travers le territoire et également la consommation en bien i d'un ménage dépend du style de vie mais d'autre facteurs qui sont lié à l'espace, L'hétérogénéité spatiale devient une réalité à prendre en compte dans l'étape de la modélisation. Dans cette perspective, nous avons envisagé la prise en compte des dépenses de consommations surtout en boissons et produits alcoolisées dans l'aboutissement de notre projet. Le revenu également fera partie de nos variables clés dans la suite.

Parallèlement, l'offre varie dans l'espace car les prix, les services, les produits et les magasins ne sont pas les mêmes partout. ECCBC n'est pas partout au Maroc ou en Algérie. L'offre est bien définie dans un territoire donné mais également, l'offre de produit varie d'un type de points de ventes à un autre. On offrira plus de produits single serve (boisson inférieure à 0.5 litre) comme la canette, la bouteille dans un restaurant ou Snack par rapport à un épicer, surtout si ce dernier est situé en zone résidentielle.

En économie, l'offre et la demande sont en règle générale séparées, le commerçant doit donc faire face à cette distance en étudiant le comportement spatial du consommateur, les zones de chalandise ou zone d'influence. Les techniques utilisées supposent généralement un principe d'attraction polaire : Le stocks de clients résidant dans une zone géographique doivent a priori se rendre dans un point dévente plutôt proche de leur domicile. Cependant, on note la complexité qui peut apparaître lorsqu'on regarde de près le stock de client à la loupe : La mobilité est un point à prendre à considération. On parle de méthodes de "captation des flux". L'idée est d'arriver à capter le client qui transite à proximité de l'espace commercial. C'est dans ce sens que dans la suite de notre travail nous utilisons des données de mobilités. Qu'en est -il exactement de cette zone de chalandise ?

## **II. le « trading area » ou zone de chalandise**

### ***II.1. Définitions***

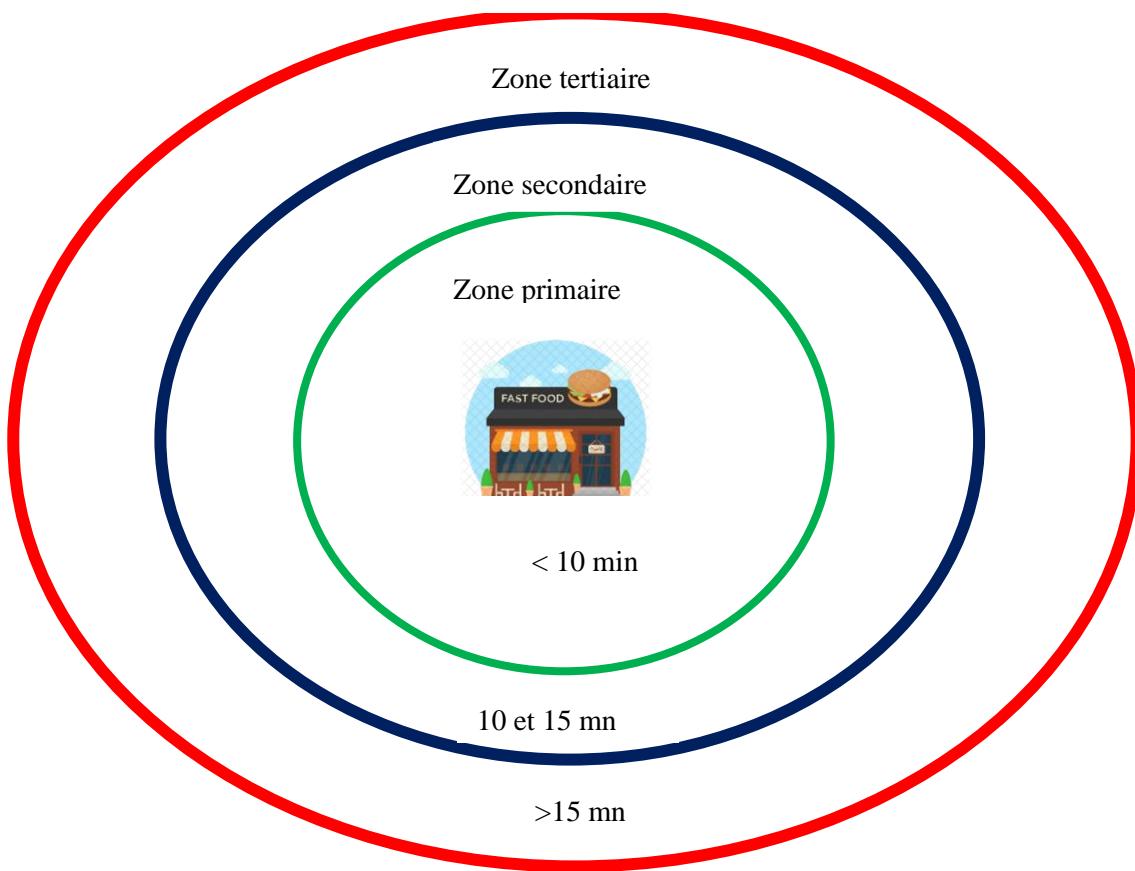
Il s'agit du territoire d'où proviennent les clients d'un point de vente (Applebaum et Cohen, 1961; Huff, 1964). Les zones commerciales des points de vente

ont longtemps été considérées comme formées de trois cercles concentriques afin d'estimer les ventes potentielles (Applebaum 1966):

- la zone principale, où se concentre environ 60 à 70% de la clientèle;
- la zone secondaire, avec entre 15 et 25% de la clientèle;
- la zone tertiaire ou marginale, avec le reste de la clientèle.

Ces zones primaires, secondaires et tertiaires ont été rapidement mesurées en termes de temps (Brunner et Masson 1968) donnant naissance aux isochrones.

**Figure 11: Zone de chalandise**



Source : réalisé par nos soins

Le temps pour définir ces zones peuvent changer selon le type de points de vente. Mais en général, ce temps est utilisé pour les petits commerces. Parfois, les zones sont décrites par distances parcourues. La forme de ces isochrones ne sont pas forcément circulaires.

## II.2. Modèles de zone de chalandise

Avec l'avènement du géomarketing, la forme réelle des zones commerciales a considérablement changé. La détermination des zones de chalandise de façon formalisé a commencé avec la loi de Reilly et les travaux de Converse.

La loi de la gravitation en géomarketing qui a été introduite par Reilly en 1931 peut être énoncée comme suit : « deux centres (situés dans deux villes différentes) attirent les achats des populations situées entre elles en proportion directe du nombre total d'habitants des villes considérées et en proportion inverse du carré de la distance qu'il faut parcourir pour s'y rendre ». La loi repose sur deux hypothèses :

- Les deux pôles (centres) sont accessibles de manière équivalente par le consommateur, l'espace est donc homogène,
- Les deux commerces ont la même efficacité.

Il a fallu attendre les travaux de Converse 1949 pour obtenir la formalisation mathématique de cette loi. Supposons  $x$  et  $y$  deux villes.

$$\text{La loi s'écrit : } \frac{V_x}{V_y} = \left( \frac{P_x}{P_y} \right) \times \left( \frac{D_y}{D_x} \right)^\beta$$

Avec  $V_x$  et  $V_y$  représentant respectivement les ventes dans la ville auprès de la population de la ville  $x$  et  $y$ . Plus généralement, ces deux termes indiquent l'attraction respective de chacune des villes  $x$  et  $y$ ;  $P_x$  and  $P_y$  sont les populations respectives de villes  $x$  et  $y$ ;  $D_x$  et  $D_y$  sont les distances respectives entre le point de rupture et les villes  $x$  et  $y$

$\beta$  indique le taux auquel l'attraction d'une ville décroît à mesure que la population de cette ville augmente. Des expériences ont montré que le coefficient de  $\beta$  était vraisemblablement égal à 2. En plus, cette valeur est couramment utilisée pour  $\beta$  car elle est recommandée depuis longtemps dans l'un des livres les plus célèbres de géomarketing qu'est Applebaum, 1968.

Quant à  $\alpha$ , il exprime le taux auquel l'attraction d'une ville s'accroît à mesure que la population de cette ville augmente. Il est fixé en général à 1.

Ainsi, afin de délimiter les frontières des aires d'influence entre deux pôles commerciaux, il suffit de chercher le point d'équilibre appelé breaking point entre les zones de desserte de deux centres urbains. Ce breaking point est le point de partage des zones de chalandises. En effet, à ce niveau le consommateur est neutre pour le choix du point de vente à visiter. Il revient à écrire que

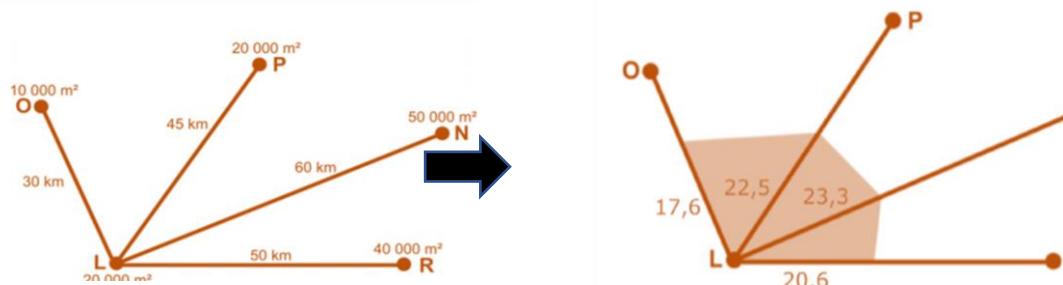
$$\frac{V_x}{V_y} = 1 \Rightarrow 1 = \frac{P_x}{P_y} \times \left( \frac{D_y}{D_x} \right)^2 \Rightarrow \frac{D_y}{D_x} = \left( \frac{P_x}{P_y} \right)^{-\frac{1}{2}}$$

En posant  $D_x = D_{XY} - D_y$  alors  $D_y = \frac{D_{xy}}{1 + \left( \frac{P_x}{P_y} \right)^{\frac{1}{2}}}$ .

Pour des magasins de la même ville, on utilise l'hypothèse selon laquelle  $\frac{P_x}{P_y}$  est équivalente au rapport des espaces commerciales de points de vente. Dans les points de ventes comme les grosseries ou snacks, on pourrait même supposer que les surfaces commerciales sont identiques.

Lorsque nous disposons des données de géolocalisations alors cette zone de chalandise est calculable. La figure ci-dessus illustre bien l'implémentation de cette formule.

Figure 12:La loi de Reilly et point de partage



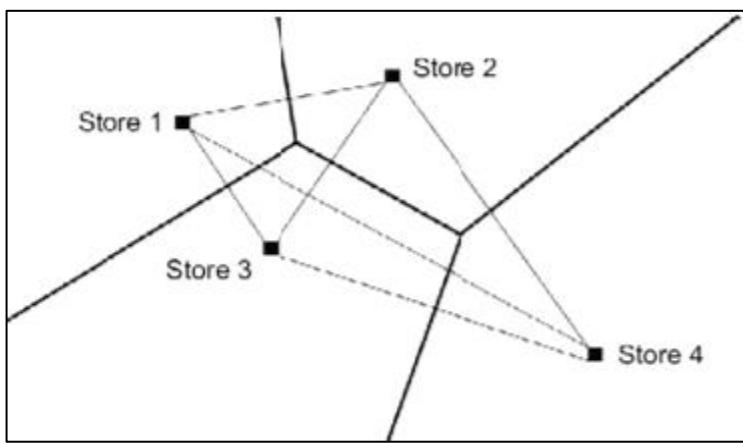
Des méthodes un peu plus sophistiquées sont apparues pour délimiter les zones commerciales. La méthode de la zone proximale est basée sur la théorie des lieux

centraux. Christaller 1993 et toujours utilisé dans les logiciels de géomarketing. Le principe repose sur la délimitation de l'aire proximale par construction de polygones dits de Thiessen (Thiessen et Alter 1911, voir figure ci-dessous) en :

- Identifiant les points de vente concurrents par le biais de segments ;
- Traçant la médiatrice de chacun de ces segments,

En définissant les polygones dits de Thiessen délimités par les médiatrices et leurs intersections, on forme la zone commerciale de chaque ensemble de points.

Figure 13: Délimitation de la zone de chalandise



**Source :** Extrait du livre Location Strategies for Retail and Service Firms

Ghosh, A., McLafferty, S.L. (1987)

On rappelle également que cette méthode de surface utilise deux hypothèses simplificatrices mais pas loin de réalité :

- les points de vente sont de taille similaire ;
- la population est bien répartie uniforme à travers le territoire.

On pourrait toujours se ramener à ces hypothèses sans les transgresser en segmentant notre territoire mais également nos points de ventes.

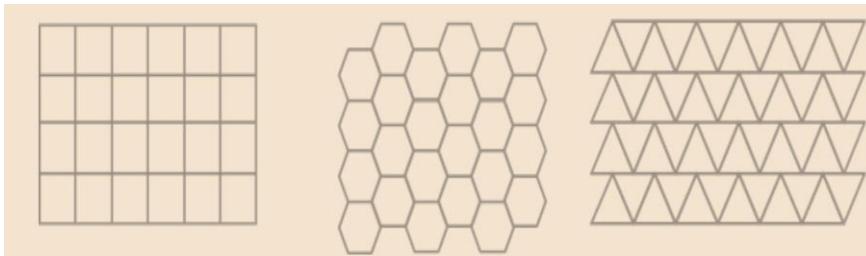
L'analyse spatiale connaît un avènement sans précédent de nos jours avec le développement du numérique. Grâce à des fichiers clients et des données de

\* géolocalisation (latitude et longitudes), il est devenu possible d'avoir une analyse spatiale assez formalisée.

## ***II.2. Index spatial***

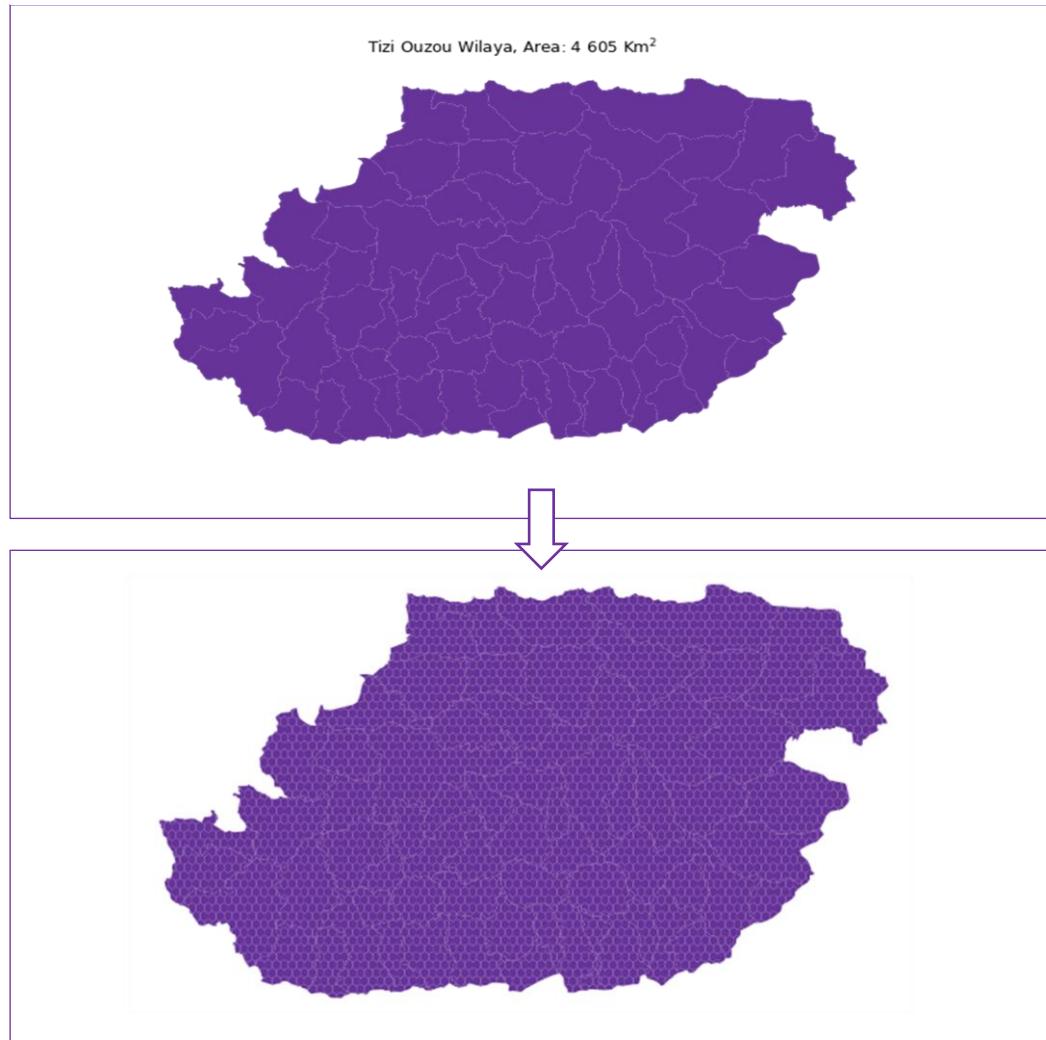
En analyse spatiale, le découpage administratif du pays n'est pas toujours efficace pour des analyses très fine. Le découpage administratif national nous permet pas de percevoir l'hétérogénéité de la population. La solution est alors de résumer une variable d'intérêt spatiale dans un territoire plus restreint. C'est dans cette lancée que nous devons partitionner les zones géographiques en cellules de grille identifiables. Ces grilles sont généralement composées de triangles équilatéraux, de carrés ou d'hexagones, car ces trois formes de polygones sont les seules qui sont des pavages du plan plus précisément elles permettent de couvrir une zone en utilisant de manière répétée une seule forme, sans espaces ni chevauchement.

Figure 14: Les types de grilles couramment utilisées



La grille carrée est la forme la plus utilisée dans l'analyse spatiale. Cependant, une grille carrée peut introduire certaines distorsions. La grille hexagonale est celle qui présente de faibles distorsions. Pour les entreprises comme Uber qui utilisent les applications de covoiturage ou également les entreprises qui s'appuient sur une cartographie précise des zones géographiques pour leurs services il est essentiel de choisir une carte de grille qui minimise les distorsions et les erreurs de quantification introduites lorsque les utilisateurs se déplacent dans une ville. Vu le succès de cette grille, Uber a rendu Open Source sa bibliothèque de cartographie hexagonale H3. Lorsqu'on indexe la wilaya de Tizi Ouzou d'Algérie on trouve la figure ci-dessous

Figure 15: Découpage de la wilaya de Tizi Ouzou en grille hexagonale



Cette figure montre une hexagonalisation de la wilaya de Tizi Ouzou avec un diamètre de 1250 mètres. Nous avons gardé ce diamètre pour la partie modélisation du projet.

## Conclusion

Plusieurs modèles de zone de chalandise existent. Chacun possède des hypothèses bien spécifiques. Nous avons présenté la détermination des points de partages. Lorsque nous avons des données textuelles, comment les utiliser pour une analyse spatiale ? La réponse à cette question sera abordée dans le chapitre suivant

## Chapitre 3 :

# Eléments d'analyse textuelle

Eléments d'analyse textuelle .....	- 49 -
Introduction.....	- 50 -
I.     Mesure de similarité de données textuelles .....	- 50 -
I.1.   La distance de Levenshtein .....	- 51 -
I.2.   Distance Damerau-Levenshtein .....	- 52 -
I.3.   Distance Hamming .....	- 53 -
I.4.   Distance Jaro .....	- 53 -
I.5    Distance Jaro-Winkler.....	- 54 -
I.6.   N-grammes .....	- 54 -
I.7.   Autres mesures de similarités.....	- 55 -
II.    La préparation de texte en Natural Language Processing .....	- 55 -
II.1.   Détection de la langue.....	- 56 -
II.2.   Suppressions des ponctuations et les caractères spéciaux .....	- 56 -
II.3.   Tekenisation .....	- 56 -
II.4.   Suppression des stopword ou mots vides.....	- 57 -
II.5.   le Stemming .....	- 57 -
II.6.   Le pliage ASCII .....	- 58 -
Conclusion .....	- 58 -

## Introduction

En analyse spatiale, l'une des techniques les plus utilisées de nos jours sont les analyses de similarité pour les données textuelles. Comment mesure-t-on la similitude entre deux chaînes de caractères ? Également comment les données textuelles sont prétraitées afin d'entrainer des modèles statistiques ? Ce paragraphe a pour but en premier lieu d'introduire les mesures de similarité et de distance des textes. En second, il récapitule les étapes de prétraitement des données textuelles.

### I. Mesure de similarité de données textuelles

Mesure de distance et mesure de similarité sont les faces d'une même pièce. Elles déterminent le degré de similarité entre deux chaînes. Lorsque nous voulons fusionner plusieurs bases de données alors que nous ne disposons pas de clés primaires ou d'identifiants sur lequel nous pouvons appliquer des jointures alors des mesures de similarité sont appliquées. Un exemple est la recherche de POI comme les écoles ou universités ou encore des sites touristiques autour des points de ventes. Il est évident que nous pouvons acquérir la base de données de points de ventes grâce à notre système de données internes. On pourrait également rentrer en possession des données des POI grâce à des sources externes. Comment alors concilier ces bases de données ? L'analyse de similarité des adresses est une approche convenable dans ce cas. Autrement dit, lorsque nous prenons un point de vente située sur le campus Madinat Al Irfane, nous pourrons alors chercher ses similarités en termes d'adresse dans la base des universités. Toutes les écoles ayant une similarité élevée avec cette adresse seront repérées.

**Figure 16: Exemple de similarité des adresses**



Restaurant Campus,  
Adresse : Madinat Al Irfane,  
Rabat Institut

Etablissement	Adresse
Ecole Supérieure de Génie Biomédical	Zénith mellinium, Batiment 6, Sidi Maarouf , Casablanca
Ecole des Sciences de l'Information – Rabat	Av Allal El Fassi, Madinat Al Irfane, Rabat-Instituts
Institut National de Statistique et d'Economie Appliquée INSEA	Av Allal El Fassi, Madinat Al Irfane, Rabat-Instituts

Dans les lignes qui suivent, nous tenteront de donner un bref aperçu de ces mesures de similarités.

Tout d'abord, pourquoi avons-nous affirmé que similarité et distance vont de pair ? En effet, Plus la distance entre deux données textuelles(chaines), mesurée par une mesure de distance, est petite, meilleur est l'accord. A cette distance, on associe une mesure de similarité. Elle est comprise entre zéro et un, l'une indiquant un accord parfait des chaînes et zéro aucun accord. Pour formaliser les différentes mesures dans cette partie, nous utiliserons les notations ci-après.

- $s_1$  est la chaîne 1 et  $s_2$  est la chaîne 2
- $n$  est la longueur de la chaîne la plus courte
- $m$  est la longueur de la chaîne la plus longue
- $i$  est la lettre à la  $i$ -ème position dans la chaîne 1
- $j$  est la lettre à la  $j$ -ème position dans la chaîne 2

Il y a une multitude de mesures de similarités et de distance dans la littérature. Nous nous contenterons des mesures qui se base sur une comparaison des caractères.

### *I.1. La distance de Levenshtein*

C'est la mesure la plus couramment utilisée. Il est défini comme le plus petit nombre d'opérations de mise à jour requises pour transformer  $s_1$  en  $s_2$ . C'est pourquoi cette méthode est considérer comme une méthode d'édition. Les éditions concernent les insertions, des suppressions et les substitutions de caractères. Dans sa forme de base, chaque opération d'édition utilisée est égalisée à un coût unitaire. La distance de Levenshtein est égale à :

$$lev_{s_1,s_2}(i,j) = \begin{cases} \max(i,j) & min(i,j)=0 \\ \min \left\{ \begin{array}{l} lev_{s_1,s_2}(i-1,j)+1 \\ lev_{s_1,s_2}(i,j-1)+1 \\ lev_{s_1,s_2}(i-1,j-1)+[s_1_i \neq s_2_j] \end{array} \right\} & \text{Sinon} \end{cases}$$

et est résolu par programmation dynamique. Pour bien illustrer cette formule, considérons les deux mots : NICHE et CHIENS. Selon la formule, la distance entre les deux chaînes est 5. En effet, les étapes sont les suivantes :

- Suppressions respectives des lettres N et I → cout=1+1+=2
- Insertions respectives des I, N et S → cout=1+1+1=3

Cette distance est comprise entre  $\max(|s1|, |s2|)$  et  $\text{abs}(|s1| - |s2|)$ . La première étant la distance maximale. Ainsi dans l'exemple des deux chaînes utilisées, le cout maximal est 6 et le cout minimal 1.

Mathématiquement, on montre que la distance de Levenshtein est une métrique réelle, car elle satisfait l'inégalité du triangle<sup>10</sup>. De plus, l'inégalité triangulaire peut être utilisée pour filtrer lors de la détermination de la meilleure correspondance d'un dictionnaire de chaînes, en sélectionnant des chaînes comme « pivots ». Pour transformer la métrique de distance de Levenshtein dans une mesure de similarité, la distance doit être divisée par la limite supérieure sur la distance de Levenshtein, qui est la longueur de la chaîne la plus longue<sup>11</sup> et retranché de 1 :

$$\text{simLev}(s1, s2) = 1 - \frac{\text{lev}_{s1, s2}}{m}$$

## I.2. Distance Damerau-Levenshtein

Cette distance est une version modifiée de la distance Levenshtein. En effet, il permet en plus des éditions de Levenshtein, des transpositions de caractères adjacents.

---

<sup>10</sup> G.R. Hjaltason and H. Samet. Index-driven similarity search in metric spaces. ACM Transactions on Database Systems, 28(4):517–580, December 2003.

<sup>11</sup> P.E. Christen. A comparison of personal name matching: Techniques and practical issues. <http://astro.temple.edu/joejupin/entitymatching/tr-cs-06-02.pdf>. Technical Report TR-CS06-02, Australian National University, 2006

### I.3. Distance Hamming

La distance Hamming entre deux chaînes se définit comme le nombre de substitutions nécessaires pour transformer une chaîne en l'autre chaîne.

#### Algorithme

```
Function hamm(s1, s2):
    compteur = 0
    Pour k in allant de 0 à la longueur de s :
        Si le k-ième caractère de s1 est différent du k-ième caractère de s2 :
            Incrémenter compteur de 1
    Finsi
    Finpour
    Retourner compteur
Fin
```

La mesure de similarité correspondante est égale à :

$$\text{simHam}(s1, s2) = 1 - \frac{\text{hamm}_{s1, s2}}{m}$$

Pour exemple, considérons JAPON – SAVON comme deux chaînes de caractère, dans ce cas, la distance de Hemming entre JAPON et SAVON est égale à 2 et la similarité est égale à 0,6

### I.4. Distance Jaro

La distance Jaro est principalement appliquée aux noms de personnes correspondants et aux adresses courtes. Ce n'est pas une véritable métrique de distance, car elle n'obéit pas à l'inégalité triangulaire. Dans cette mesure, les chaînes sont plus similaires si elles ont des caractères correspondants dans une plage spécifique, en fonction de la longueur de la chaîne la plus longue. Une pénalité est accordée si les caractères correspondants sont transposés (c'est-à-dire s'ils ne sont pas dans le même ordre).

La mesure de similarité se calcule comme suite :

$$\text{simjaro}_{c1, c2}(i, j) = \begin{cases} 0 & \text{Si } c=0 \\ \frac{1}{3} \left( \frac{c}{|s1|} + \frac{t}{|s2|} + \frac{c-t}{c} \right) & \text{Sinon} \end{cases}$$

Où  $c$  est le nombre de caractères correspondants et  $t$  le nombre de caractères correspondants qui est transposé. Deux caractères sont définis comme correspondants s'ils sont égaux et positionnés dans l'intervalle de position de 0 à  $\frac{\max(|s1|, |s2|)}{2} - 1$  l'un de l'autre .

### I.5 Distance Jaro-Winkler

La distance de Jaro-Winkler n'est rien d'autre qu'une extension de la précédente distance. En effet la similarité y associée est égale à la similarité associée à la distance Jaro, plus un score si les préfixes des chaînes sont égaux. Winkler en 1999 en partit de l'hypothèse que les erreurs ont tendance à se produire moins souvent au début des chaînes<sup>12</sup>. La nouvelle mesure de distance robuste devrait s'écrire :

$$\text{Simjaro\_Winkler} = \text{Simjaro}(s1, s2) + (l * \rho * (1 - \text{Simjaro}(s1, S2)))$$

Où  $l$  est le nombre de caractères égaux au début des chaînes avec un maximum souvent défini sur quatre caractères.  $\rho$  est un paramètre à régulariser. Par défaut, il est fixé à 0,1.

### I.6. N-grammes

On appelle n-grammes des sous-chaînes de longueur n. Ici la méthode se base sur une comparaison des sous chaines. On cherche alors à savoir si chaque N-gramme de S1 existent dans s2. La similarité est égale à :

---

<sup>12</sup> W.E. Winkler. The state of record linkage and current research problems. Statistical Research Division, U.S. Census Bureau, 1999

$$SimNgram = \frac{1}{N - n + 1} \sum_{i=1}^{N-n+1} h(i)$$

où  $h(i) = 1$  si la sous-séquence à  $n$  éléments commençant à partir de la position  $i$  dans  $s1$  apparaît dans  $s2$ ,  $h(i) = 0$  dans le cas contraire . , $N-n+1$  = nombre de sous-séquences à  $n$  éléments dans  $s1$ .

Exemple : fixons  $n=2$  pour  $s1= JAPON$  –et  $s2 = SAVON$ .

Les sous chaines s'écrivent respectivement : **JA, AP, PO, ON**, vs **SA, AV, VO, ON**.

$$SimNgram = \frac{1}{N-n+1} \sum_{i=1}^{N-n+1} h(i) = \frac{1}{4} * (0 + 0 + 0 + 1) = 0,25$$

### I.7. Autres mesures de similarités

La Similarité de Jaccard et le coefficient de chevauchement se basent sur l'intersection des caractères communes aux deux chaines  $s1$  et  $s2$ . La première rapport l'intersection à l'union des caractères dans  $s1$  et  $s2$  alors que la deuxième ne rapporte que par le minimum de la taille des  $s1$  et la taille de  $s2$ .

$$Simjaccard = \frac{|s1 \cap s2|}{|s1 \cup s2|}$$

$$SimOverlap = \frac{|s1 \cap s2|}{\min(|s1|, |s2|)}$$

On définit également une similarité basée sur la taille moyenne des deux chaines. Dans ce cas, on parle de l'indice de Sørensen-Dice, ou le coefficient de Dice :

$$SimDice = \frac{|s1 \cap s2|}{\frac{|s1 \cup s2|}{2}}$$

## II. La préparation de texte en Natural Language Processing

Avant de réaliser tout calcul de mesure de similarité, des prétraitements sont faits afin d'avoir de bonnes performances. Ces prétraitements sont bien connus dans le domaine du traitement de langage naturel ou Natural Language Processing (NLP) en

anglais. Ce domaine est l'un des domaines de recherche les plus actifs en Data science de nos jours. Il s'agit d'une combinaison d'apprentissage statistique et de la linguistique. L'étude de mesure de similarités dans un contexte de NLP est l'avant dernière étape. La dernière étape étant la modélisation proprement parlée. Comme exemple, à partir de quel score de similarité peut-on déclarer deux chaînes  $s_1$  et  $s_2$  équivalentes ? La première étape qu'est le prétraitement peut être divisée en six étapes successives étapes qui seront abordées dans les lignes qui suivent.

### *II.1. Détection de la langue*

Il est évident que la comparaison directe de deux chaînes qui proviennent de langues différentes n'a pas de sens. Il faut alors vérifier la langue utilisée avant d'appliquer les analyses. C'est ce qui explique pourquoi cette étape est considérée comme première. La détection de langue est réalisable grâce à des bibliothèques d'open source. La plus connue est celle de Google qu'est « Compact Language Detector » CLD3. Il est construit sur un modèle de réseau de neurones. Nous utiliserons cette dernière afin de faire la détection des langues des chaînes. Lorsque des textes sont de langues différentes autre que le français, nous passerons à une traduction de celles-ci. Nous allons traduire ces textes en utilisant des services du Cloud. Pour plus détail, voir le paragraphe Partie 2, Chapitre 1, II : Infrastructure du projet

### *II.2. Suppressions des ponctuations et des caractères spéciaux*

### *II.3. Tokenisation\**

Nous sommes désormais à la troisième étape. La tokenisation est le processus consistant à décomposer un texte en mots, signes de ponctuation ou chiffres numériques. Plus grossièrement, on peut la définir comme la décomposition des phrases en des mots. Dans ce sens je peux décomposer « Je suis étudiant de l'INSEA

depuis Septembre 2018" » en des mots comme ci-après.

```
Je
suis
étudiant
de
l'INSEA
depuis
Septemnre
2018
```

#### *II.4. Suppression des stopword ou mots vides*

Les mots vides sont les mots dans toute langue qui n'ajoute pas beaucoup de sens à une phrase. Ils sont appelés vides car leur absence ne change profondément le sens du texte. Ils sont en général les mots courts les plus courants, tels que « le », « est », « à », « qui » « et » « sur ».

Autres exemples de stopword en français :

'ai', 'aie', 'aient', 'aies', 'ait', 'as', 'au', 'aura', 'aurai', 'auraient', 'aurais', 'suis', 'sommes', 'est'...

#### *II.5. le Stemming*

Le stemming ou radicalisation succède l'étape de suppression de stopword. Il permet de réduire un mot dans sa forme « racine »<sup>13</sup>. Le stemming a pour objectif de regrouper de nombreuses variantes d'un mot comme un seul et même mot. Pour preuve, lorsque nous appliquons du stemming sur « Etudiantes » ou « étudiant », nous trouverons un radical unique. Les techniques de stemmers sont nombreuses dans la littérature, cependant l'une des plus connus et couramment utilisés est le Snowball Stemmer.

---

<sup>13</sup> <https://tartarus.org/martin/PorterStemmer/>, consulté le 15 mars 2021.

## ***II.6. Le pliage ASCII***

Il s'agit d'un filtre qui consiste à convertir les caractères alphabétiques, numériques et symboliques qui ne sont pas dans le bloc Unicode latin de base (127 premiers caractères ASCII) en leur équivalent ASCII, le cas échéant. Exemple : « à » devient « a ». Cette étape clôture le prétraitement textuel.

## **Conclusion**

Le traitement de données textuelles et la mesure des similarités des textes vont de pair. Jamais l'un sans l'autre. Dans la partie modélisation, nous serons amenés à chercher des similarités entre des PdV en se basant sur les noms ou les adresses. C'est pourquoi nous avons introduit ce chapitre.

Une fois que les données sont extraites et prêtées à être utilisées, la modélisation statistique intervient. Le chapitre qui va suivre résume la théorie de techniques d'apprentissage supervisé nécessaire pour comprendre ce projet.

## Chapitre 4 :

---

# Eléments d'apprentissage statistique

Introduction.....	- 59 -
I. Typologie d'apprentissage statistique .....	- 60 -
I.1. Apprentissage supervisé .....	- 60 -
I.2 Apprentissage non supervisé .....	- 60 -
I.3Apprentissage semi supervisé.....	- 61 -
II. Techniques d'apprentissage non-supervisée : Clustering.....	- 61 -
I. Critères de clustering .....	- 62 -
I .1. Le choix des variables. ....	- 62 -
I.2. Choix de la métrique .....	- 63 -
I.3. La Forme des clusters .....	- 68 -
I.4. Stabilité des clusters .....	- 70 -
II. Techniques.....	- 71 -
II.2. Clustering hiérarchique .....	- 71 -
II.2. Les méthodes basées sur les centroïdes .....	- 74 -
II.3. Cas particulier des centroïdes :k-mean .....	- 75 -
II.3.1. Principe .....	- 76 -
II.3.2. Forme des clusters du k-means.....	- 76 -
II.3.3. K-mean++.....	- 77 -
II.4. Les méthodes de densités : DBSCAN .....	- 77 -
Tableau recapitualtif .....	- 80 -
III. Techniques d'apprentissage supervisé.....	- 80 -
III.1 Généralités .....	- 80 -
III.2. Évaluation des performances des modèles .....	- 86 -
III.2.1. Les métriques régression .....	- 86 -
III.3.2. Les métriques de classification.....	- 88 -
Conclusion .....	- 91 -

## Introduction

Dans leur livre Elements of statistical learning, l'un des plus célèbres en matière de formalisation de l'apprentissage statistique, Trevor Hastie, Robert Tibshirani Jerome Friedman Stanford affirmaient: "Vast amounts of data are being generated in many fields, and the statistician's job is to make sense of it all : to extract important patterns and trends, and understand "what the data says." We call this learning from data." Ce chapitre présente les techniques d'apprentissage statistique.

### I. Typologie d'apprentissage statistique

#### *I.1. Apprentissage supervisé*

Pour chaque unité statistique  $i$  avec  $i = 1, \dots, n$ , on a une mesure des caractéristiques  $x_i$  et mesure d'une ou plusieurs variables d'intérêt  $y_i$  appelées aussi réponses associées. Les données sont dites annotées ou étiquetées par la (ou les) variable(s) d'intérêt. On souhaite ajuster un modèle qui relie la ou les variables réponses  $Y$  aux caractéristiques  $X$  (prédicteurs, features).

Selon Cornuéjols et Miclet dans Apprentissage artificiel : Deep Learning, concepts et algorithmes, ont subdivisé l'apprentissage en deux approches :

- **Approche en 2 étapes**
  1. Inférence pour estimer  $P(Y = y|X = x)$
  2. Décision - prédition de  $\hat{y}$  s'appuyant sur  $\hat{P}(Y = y|X = x)$
- **Approche directe** - prédition directe de  $\hat{y}$

#### *I.2 Apprentissage non supervisé*

Pour chaque unité statistique  $i$  avec  $i = 1, \dots, n$  on a une mesure de caractéristiques  $x_i$  mais pas de variables réponses associées. On peut souhaiter identifier des groupes d'unités statistiques qui se ressemblent au regard des variables caractéristiques. On peut aussi souhaiter identifier des groupes de variables caractéristiques qui sont liées entre elles.

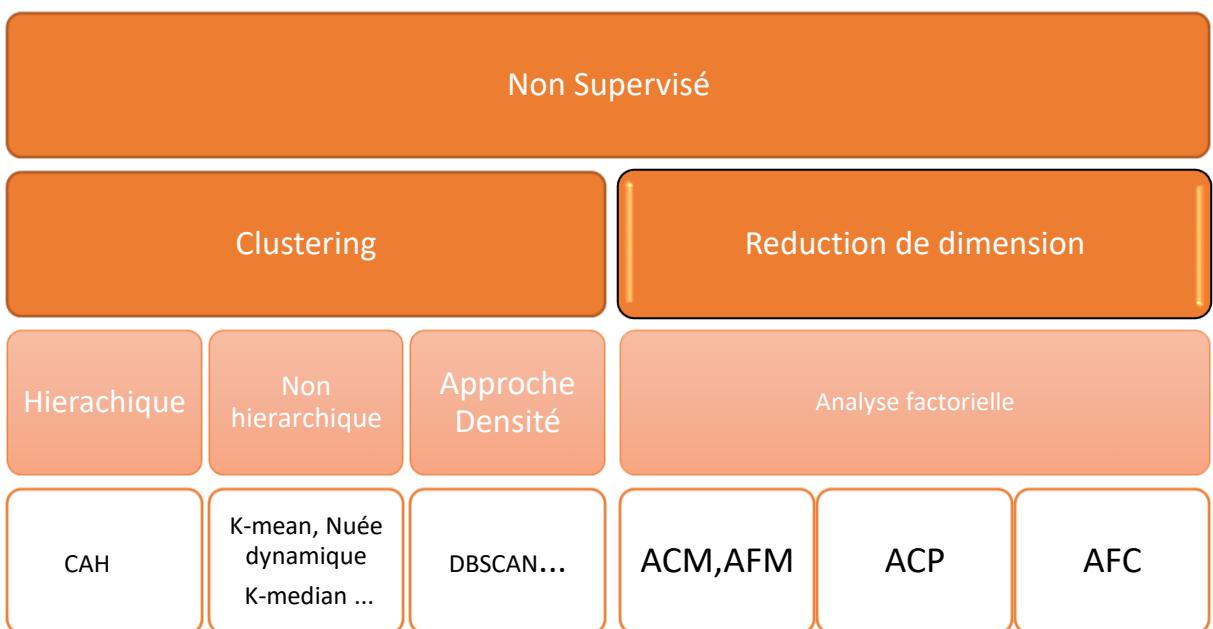
### I.3 Apprentissage semi supervisé

Pour chaque unité statistique  $i$  avec  $i = 1, \dots$ , On a une mesure de caractéristiques  $x_i$ . Certaines unités statistiques sont étiquetées et d'autres non » (Blum et al.1998). Ce type d'apprentissage peut se ramener à un type d'apprentissage supervisées.

Dans le cadre de notre projet, nous appliquerons toutes ces techniques d'apprentissage. Pour un aperçu avant les chapitres qui concernent la modélisation, on pourrait dire que notre algorithme de Matching ou de jointure de bases de données est un apprentissage semi-supervisé. Dans cette étape de notre modélisation, nous construisons nos propres étiquettes manuellement sur un sous échantillon de petite taille, si le match est vrai  $y=1$ , sinon  $y=0$ . Cela nous permettra de généraliser le modèle. Également, l'étape de segmentation dynamique des points de ventes constitue un apprentissage non-supervisé. Quant à la modélisation des fonctions de demandes par segments, il s'agira d'un apprentissage supervisé.

## II. Techniques d'apprentissage non-supervisée : Clustering

**Figure 17: Mapping des techniques d'apprentissages non supervisé**



Source : élaboré par nos soins.

Nous pouvons diviser les techniques d'apprentissages supervisée en deux grandes parties. Les techniques de Clustering et les techniques de réduction de dimension. On parle de clustering lorsque l'objectif est la représentation des liens entre individus d'une population statistique sur lesquels on a observé un certain nombre de variables afin de dégager l'existence de sous-groupes dans lesquelles se répartissent ces individus. Ainsi, ces techniques permettent d'identifier les groupes d'individus en fonctions de similitudes. Les techniques de réductions de dimension ont pour but de résumer l'information contenu dans plusieurs variables en un nombre plus petit de variables orthogonales appelés axes factoriels. Dans cette partie, nous ne présenterons pas ces techniques.

## I. Critères de clustering

Dans tout problème de classification, tout statisticien ou data scientist avéré se propose en général six questions voire plus. Ces questions lui permettront de déterminer la technique adéquate ?

- Quelle est le nombre de clusters que l'on souhaite avoir ? Existe-t-il des contraintes sur le nombre de cluster voulues ?
- Le nombre d'individus est -il élevé ?
- Quelles variables doivent être prises en compte et quelle est leur nombre ?
- Quelle métrique de distance utiliser pour mesurer la distance entre individus ?
- Quelles méthodes d'agrégations des individus utiliser ?
- Est-il nécessaire d'appliquer des transformations sur les variables ?

### *I .1. Le choix des variables.*

Le plus souvent on est amené à diviser les variables en variables actives (utilisées pour l'exercice de segmentation) et des variables descriptives qui nous permettre d'avoir des informations agrégées sur chaque cluster. En marketing ou généralement dans le secteur du retail, les ventes ou les profits sont des indicateurs sur lesquels, on décrit chaque variable. Dans le système Coca Cola, les variables descriptives (passives) sont

utilisées pour étudier d'autres aspects des points de vente (principalement internes) mais ne sont pas utilisées pour dériver activement les clusters. Ces sont par exemple, les variables relatives aux ventes ou la médaille du PdV. Lorsque le nombre de variables est énorme, des réductions de dimensions sont suggérées car on peut être confronté au problème de malédiction de la dimension. C'est Richard Bellman qui a suggéré cette appellation ésotérique en 1961, pour désigner un ensemble de problèmes qui se posent uniquement lorsque l'on analyse des données dans des espaces de grandes dimensions. Pour sélectionner les informations pertinentes, une ACP ou ACM est réalisée au préalable selon le type de variables. Par la suite, les axes factoriels renfermant le maximum d'informations sont utilisés pour exécuter l'exercice.

## *I.2. Choix de la métrique*

Ici il est question de distance et de similarité. Ce sont ces concepts qui permettent de formaliser à quel point des individus sont semblables ou à quel point un individu est proche d'un cluster donné.

### *Données quantitatives*

Soit  $x_k$  et  $x_l$  étant les caractéristiques vectorielles respectives des individus k et l. Pour des données quantitatives, la distance euclidienne est la plus souvent utilisée

$$d_M^2(k, l) = (x_k - x_l)M(x_k - x_l)'$$

Avec  $d_M^2$  le carré de la distance euclidienne si M=matrice identité

$d_M^2$  le carré de la distance Mahalobis si M représente l'inverse de la matrice de covariance-variance.

On peut également utiliser la distance Manhattan qui est le cas particulier de la distance Minkowsky de norme  $\lambda$ .

Soit  $x_{kj}$  et  $x_{lj}$  étant les caractéristiques respectives des individus l et k pour la variable j. La distance Minkowsky s'écrit :

$$d(k, l) = \sum_{j=1}^p (|x_{kj} - x_{lj}|^\lambda)^{\frac{1}{\lambda}}$$

Pour  $\lambda=1$ , on retrouve la distance Manatthan.

A partir de la distance, on peut facilement dériver une mesure de similarité comme ci-après :

$$s(k, l) = \frac{1}{1 + d(k, l)} ;$$

Plus deux individus sont distants ou éloignées alors moins ils sont similaires, et réciproquement.

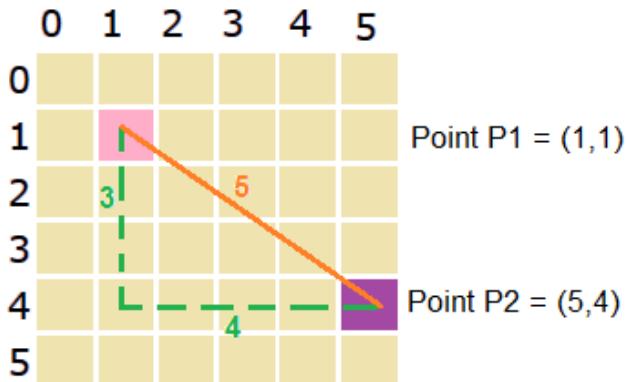
Une autre façon de mesurer la similarité est d'utiliser *la corrélation de Pearson* entre les variables :

$$\rho(k, l) = \frac{\sum_{j=1}^p (x_{kj} - \bar{x}_k)(x_{lj} - \bar{x}_l)}{\sqrt{\sum_{j=1}^p (x_{kj} - \bar{x}_k)^2} \sqrt{\sum_{j=1}^p (x_{lj} - \bar{x}_l)^2}}$$

où  $\bar{x}_k$  représente la moyenne des entrées de  $x_k$ . Lorsque les données sont centrées  $\bar{x}_k = \bar{x}_l = 0$ , la formule devient  $\rho(k, l) = \frac{\langle x_k, x_l \rangle}{\|x_k\| \|x_l\|}$ .

De près, on se rend compte qu'il s'agit en réalité de l'angle formé par les deux individus k et l. C'est de là que vient la similarité cosinus traduit littéralement de l'anglais cosine similarity.

Il y a lieu de noter qu'il n'y a pas de meilleure distance cependant, il n'y a qu'une meilleure mesure de distance pour un ensemble de données donné. Le choix de la mesure de distance va influencer le résultat du partitionnement. Donc l'objectif et la nature des données nous orientera vers la mesure de distance adéquate. Pour bien illustrer ces distances, referrons nous à la figure ci-dessous.

Figure 18: Distance euclidienne et Manhattan<sup>14</sup>

$$\text{Euclidean distance} = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

$$\text{Manhattan distance} = |5-1| + |4-1| = 7$$

Les lignes en pointillés et en couleur vert représentent la distance Manhattan. La ligne orange Euclidienne illustre la distance euclidienne.

Ainsi, lorsque deux individus sont proches sur la plupart des variables, mais plus divergents sur l'une d'entre elles, la distance euclidienne exagérera cet écart, tandis que la distance de Manhattan l'ignorera, étant plus influencée par la proximité des autres variables. La distance de Manhattan devrait donner des résultats plus robustes, tandis que la distance euclidienne est susceptible d'être influencée par les valeurs aberrantes. La même chose s'applique aux valeurs plus élevées de "  $\lambda$  " dans la formule de distance de Minkowski. Au fur et à mesure que nous augmentons la valeur de  $\lambda$ , la mesure de la distance devient plus susceptible de perdre de sa robustesse et les valeurs aberrantes dans quelques dimensions commencent à dominer la valeur de la distance.

Une pratique commune dans en analyse de données est de centrer et réduire les données. Ceci est particulièrement recommandé lorsque les variables sont mesurées à

---

<sup>14</sup> Dip Ranjan Chatterjee, Log Book — Guide to Distance Measuring Approaches for K- Means Clustering, Juillet 2019.

différentes échelles (ex : mètres, dirham, population, ...), sinon, les mesures de dissemblance obtenues seront sévèrement affectées. La normalisation rend les quatre méthodes de mesure de distance Euclidienne, Manhattan, et Corrélation de Pearson plus similaires<sup>15</sup> qu'elles ne le seraient avec des données non standardisées<sup>16</sup>. Notez que, lorsque les données sont standardisées, il existe une relation fonctionnelle entre le coefficient de corrélation de Pearson et la distance euclidienne.

**Tableau 3: Les Distance utilisées en clustering**

mesure de distance	Equation	Complexité (temps)	Avantages	Désavantages	Applications
Distance euclidienne	$d_{euc} = \left[ \sum_{i=1}^n (x_i - y_i)^2 \right]^{1/2}$	$O(n)$	Très courant, facile à calculer et fonctionne bien avec des ensembles de données ayant des clusters compacts ou isolés	Sensible aux outliers	Algorithme de K-means, Algorithme de Fuzzy c-means
Distance moyenne	$d_{ave} = \left[ \frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2 \right]^{1/2}$	$O(n)$	Meilleur que la distance euclidienne dans le traitement des outliers	Les variables contribuent indépendamment de la mesure de distance. Des valeurs redondantes pourraient dominer la similitude entre les points de données	Algorithme de K-means
Distance euclidienne ponderée	$d_{euc} = \left[ \sum_{i=1}^n w_i (x_i - y_i)^2 \right]^{1/2}$	$O(n)$	La matrice de poids permet d'augmenter l'effet des points de données plus importants que ceux moins importants	Pareil que ma distance euclidienne	Algorithme de Fuzzy c-means
Distance de Chord	$d_{chord} = \left( 2 - 2 \frac{\sum_{i=1}^n x_i y_i}{\ x\  \ y\ } \right)^{1/2}$	$O(3n)$	Peut fonctionner avec des données non normalisées	Elle n'est pas invariante par transformation linéaire	Détection de ressemblance écologique
Distance de Mahalanobis	$d_{mah} = \sqrt{(x - y) S^{-1} (x - y)^T}$	$O(3n)$	une mesure basée sur les données qui peut atténuer la distorsion de distance causée par une combinaison linéaire d'attributs	Cela peut être coûteux en termes de calcul	Algorithme de clustering hyperellipsoïdal
Distance Cosinus	$\text{Cosinus}(x, y) = \frac{\sum_{i=1}^n x_i y_i}{\ x\ _2 \ y\ _2}$	$O(3n)$	Indépendante de la taille du vecteur et invariante par rotation	Elle n'est pas invariante par transformation linéaire	Principalement utilisé dans les applications de similarité de documents
Distance de Manhattan	$d_{man} = \sum_{i=1}^n  x_i - y_i $	$O(n)$	Est commune et comme les autres distances pilotées par Minkowski, il fonctionne bien avec des ensembles de données avec des clusters compacts ou isolés	Sensible aux outliers	Algorithme de K-means
Distance Mean Character Difference	$d_{MCD} = \frac{1}{n} \sum_{i=1}^n  x_i - y_i $	$O(n)$	Donne des résultats précis à l'aide de l'algorithme K-medoids	*Faible précision pour les ensembles de données de grande dimension utilisant K-means.	Algorithmes de partitionnement et de clustering hiérarchique
Index d'association	$d_{IOA} = \frac{1}{n} \sum_{i=1}^n \left  \frac{x_i}{\sum_{j=1}^n x_j} - \frac{y_i}{\sum_{j=1}^n y_j} \right $	$O(3n)$	.	*Faible précision en utilisant les algorithmes K-means et K-medoids.	Algorithmes de partitionnement et de clustering hiérarchique.
Métrique de Canberra	$d_{canb} = \sum_{i=1}^n \frac{ x_i - y_i }{(x_i + y_i)}$	$O(n)$	Donne des résultats précis pour les ensembles de données de grande dimension à l'aide de l'algorithme K-medoids.	.	Algorithmes de partitionnement et de clustering hiérarchique.
Coefficient de Divergence	$d_{canb} = \left( \frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2 / (x_i + y_i) \right)^{1/2}$	$O(2n)$ $O(n)$	Donne des résultats précis à l'aide de l'algorithme K Means.	.	Algorithmes de partitionnement et de clustering hiérarchique.

Source : réalisé à partir des informations de l'article de Shirkhorshidi AS, Aghabozorgi<sup>17</sup>

<sup>15</sup> Pandit, S. and S. Gupta. "A Comparative Study on Distance Measuring Approaches for Clustering." International Journal of Research 2 (2011) : 29-31

<sup>16</sup> La formule de standardisation revient à retrancher de l'observation de l'individu i, la moyenne de la série :  $\frac{x_i - \bar{x}}{\sigma}$

<sup>17</sup> , Wah TY (2015) A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. PLoS ONE 10(12) : e0144059. doi :10.1371/journal.pone.0144059

Données qualitatives<sup>18</sup>

Pour des variables nominales, la distance de Khi-deux est utilisée. Supposons deux variables nominales X et Y avec J et K comme nombres respectifs de modalités.

X \ Y	1	...	j	...	J	Ensemble
.						
k				n <sub>kj</sub>		n <sub>k.</sub>
.						
l				n <sub>lj</sub>		n <sub>l.</sub>
.						
K						
Ensemble			n <sub>j</sub>			n

La distance de Khi-deux est calculée avec la formule suivante

$$d_M^2(k, l) = \sum_{j=1}^J \frac{\left( \frac{n_{kj}}{n_{k.}} - \frac{n_{lj}}{n_{l.}} \right)^2}{\frac{n_j}{n}}$$

Lorsqu'on a à faire à une table de données binaire, on peut utiliser des indices plus spécifiques :

$$\text{Indice de Jaccard : } d(k, l) = \frac{a}{a+b+c}$$

$$\text{Indice de Solak : } d(k, l) = \frac{a+d}{a+b+c+d}$$

$$\text{Indice de Pearson } d(k, l) = \frac{|ad-bc|}{[(a+b)(c+d)(a+c)(b+d)]^2}$$

Avec : a désignant le nombre de fois où  $x_{kj} = x_{lj} = 1$  ;

b désignant le nombre de fois où  $x_{kj} = 1$  &  $x_{lj} = 0$  ;

c désignant le nombre de fois où  $x_{kj} = 0$  &  $x_{lj} = 1$  ;

d désignant le nombre de fois où  $x_{kj} = x_{lj} = 0$  .

---

<sup>18</sup> Mohammed El Haj Tirari, Cours d'Analyse de données, INSEA Rabat, Année universitaire 2019-2020

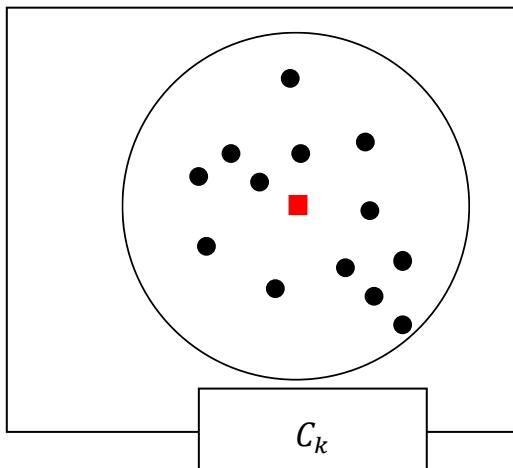
### I. 3. La Forme des clusters

Un bon clustering devrait minimiser la variance intra-cluster et maximiser la variance inter-cluster. En d'autres nous voulons que les points soient<sup>19</sup> :

- D'une part resserrée sur eux-mêmes : deux points qui sont proches doivent appartenir au même cluster ;
- Et d'autre part loin les uns des autres : deux points qui sont éloignés doivent appartenir à des clusters différents ;

Imaginons  $d$  comme étant la distance choisie. On définit alors un centroïde pour chaque cluster. Il s'agit plus précisément du barycentre noté  $u_k$  des points de ce cluster

$$u_k = \frac{1}{|c_k|} \sum_{x \in c_k} x \text{ avec } C_k \text{ le cluster k}$$



Pour un cluster donné, on peut définir l'homogénéité notée  $T_k$  (l'abréviation T vient de l'anglais tightness).

$$T_k = \frac{1}{|c_k|} \sum_{x \in c_k} d(x, u_k)$$

Cette formule signifie que l'homogénéité est la distance moyenne entre le centroïde et les différents points appartenant au cluster. Il revient alors à dire qu'un cluster resserré aura une hétérogénéité plus faible qu'un cluster de points dispersés. Pour aller plus loin, on s'intéresse à la caractérisation de l'ensemble des différents clusters.

---

<sup>19</sup>Yannis Chaouche, Chloé-Agathe Azencot, Nathalie Turck, Explorez vos données avec des algorithmes non supervisés, Ecole CentraleSupélec

Plus précisément, on introduit alors la mesure  $T$ , la moyenne des homogénéités de chaque cluster

$$T = \frac{1}{K} \sum_{k=1}^K T_k$$

Pour évaluer le degré de similitude entre deux clusters, on définit maintenant la *séparation* de deux clusters comme la distance entre leurs centroïdes  $S_{k,l} = d(u_k, u_l)$ . Là également, on caractérise le résultat de clustering à une itération donnée en évaluant la moyenne notée  $S$  des séparations de l'ensembles des pairs de clusters  $C_k$  et  $C_l$ .

$$S = \frac{2}{K(K - 1)} \sum_{k=1}^K \sum_{l=k+1}^K S_{k,l}$$

La séparation des clusters est la distance entre les centroïdes des clusters concernés.

Les techniques de clustering essaient d'optimiser ces deux critères. D'une part, à maximiser  $S_{k,l}$  ou les distances inter-clusters et d'autre part à minimiser  $T$  ou les distances intra-clusters. Ces deux critères sont regroupés pour former un indice appelé indice de Davies-Bouldin calculé pour chaque cluster que nous notons ici  $D_k$ .

$$D_k = \max_{k \neq l} \frac{T_l + T_k}{S_{k,l}}$$

Plus  $D_k$  est faible plus les clusters sont homogènes et bien séparés. Parfois, on s'intéresse à la moyenne de ces indices d'où l'indice de Davies-Bouldin

$$D = \frac{1}{K} \sum_{k=1}^K D_k$$

Pour mesurer à la fois l'homogénéité et la séparation en même temps, l'une des mesures les plus utilisées est le coefficient de silhouette. Ce coefficient permet de savoir si un point  $x$  donné appartient au « bon cluster ». Pour être plus simpliste, disons que ce coefficient permet en premier lieu de vérifier si un point donné est proche des autres points du cluster auquel il a été assigné en calculant la distance moyenne notée  $a(x)$  entre le point et les autres points appartenant au cluster

$$a(x) = \frac{1}{|c_k|-1} \sum_{u \in c_k, x_k \neq u} d(x_k, u)$$

En second lieu, il vérifie à quel degré de ce point est loin d'un autre cluster en calculant la plus petite valeur potentielle de  $a(x)$  notée  $b(x)$  s'il appartenait à un autre cluster.

$$b(x) = \min_{k \neq l} \frac{1}{|c_l|} \sum_{u \in c_l} d(x_k, u)$$

Naturellement si le point est dans le bon cluster déjà, on aurait  $a(x) \leq b(x)$ . Le coefficient de silhouette se calcule alors comme ci-après

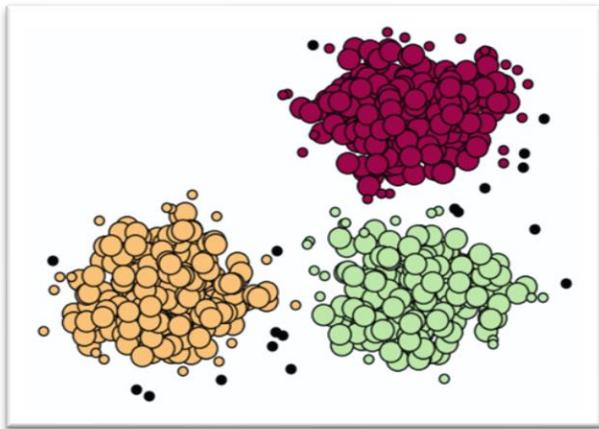
$$s(x) = \frac{b(x) - a(x)}{\max(b(x), a(x))}$$

$s(x)$  est comprise entre -1 et 1. Une valeur proche de 1 est synonyme d'appartenance au bon cluster. On évalue alors le résultat du clustering en calculant la silhouette moyenne de l'ensembles de points.

#### I.4. Stabilité des clusters

La *stabilité* est un terme récurrent dans les livres ou articles traitant des performances des algorithmes de clustering. Sommes-nous capables, d'obtenir les mêmes résultats à partir des mêmes données mais à chaque fois avec une initialisation différente ? si oui, on parle de clustering stable. Il s'agit d'un critère intéressant que permet de choisir le nombre de cluster. En effet, en faisant plusieurs itérations, l'on pourrait savoir si les résultats trouvés sont le reflet de la structure naturelle des données.

**Figure 19: structure naturelle à trois segments**



Source : tiré de la bibliothèque des exemples de sklearn et adapté par nos soins

Cette figure montre bel et bien trois clusters. Ainsi lorsqu'un algorithme donné choisit de déterminer 3 clusters, il s'ensuit que le résultat final sera le même que la figure ci-dessus. Cependant, lorsque nous choisissons 4 comme nombre de cluster alors le partitionnement que nous obtiendrons serait aléatoire. D'une itération à une autre, nous trouverons des résultats différents.

## II. Techniques

### II.2. Clustering hiérarchique

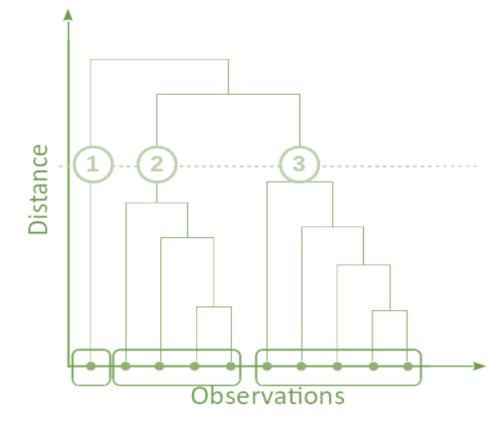
Un peu plus haut, nous avons présenté les concepts de la **séparation des clusters** et **l'homogénéité que l'on peut synthétiser grâce aux** coefficients de silhouette ou l'index de Davies-Bouldin. Il s'avère que les calculs exacts de ces indicateurs requièrent un temps de calcul colossal, C'est pourquoi des heuristiques sont adoptés. La technique de classification hiérarchique utilise en effet une approche par récurrence. Selon cette approche, on peut distinguer deux grandes familles :

- Les algorithmes ascendants, qui construisent les classes par agglomérations successives des objets deux à deux ;
- Les algorithmes descendants, qui réalisent des dichotomies

progressives de l'ensemble d'objets.

Dans les deux cas, les algorithmes aboutissent à la constitution d'un arbre appelé dendrogramme, qui rassemble des individus de plus en plus dissemblables au fur et à mesure qu'on s'approche de la racine de l'arbre.

La figure ci-dessous illustre un exemple de dendrogramme. En coupant au niveau de la ligne horizontale en pointillée, on obtient 3 clusters



Source : réalisé par nos soins

Supposons que l'on distancie d'une table de donnée avec N individus. Un clustering hiérarchique ascendante de ces individus se fait à étapes.

**Etape 1 :** Chaque individu est considéré comme un cluster. On a à cette étape autant de clusters que d'individu. L'algorithme calcule les distances inter-clusters afin de trouver les deux clusters les plus proche et les fusionner. A la fin de cette étape, il ne restera que N-1 clusters.

**Etape 2 :** L'algorithme commence avec les N-1 classes. Les distances inter-clusters sont encore calculé afin de trouver les deux clusters ayant la plus petite distance. Ces derniers sont fusionnés et l'on se retrouve avec N-2 clusters.

**Etape k :** l'algorithme commence avec les N-k+1 classes. Selon le même principe, sont calculés les distances inter-clusters afin de trouver les deux clusters les plus proches. Ces deux clusters sont fusionnés à la fin de l'étape pour donner place une seule ramenant le nombre de cluster à N-K.

**Etape N :** A Cette étape, on démarre avec deux classes pour finir par les fusionner en une seule.

Pour trouver le nombre de cluster, le dendrogramme présenté ci-haut permet de visualiser ce détail. En complément de ce dernier, on utilise généralement la perte d'inertie après chaque agrégation afin de trouver le nombre de clusters optimal ;

Mais comment se fait l'agrégation des classes ? En effet, cette agrégation est faite en calculant la distance entre les classes. Le tableau ci-après nous permet de résumer les distances utilisées.

Tableau 4: Stratégies d'agrégation

Strategies	Description	$d(C_k, C_l)$
Single linkage	Distance entre deux clusters est la distance minimale entre deux points, l'un appartenant au premier cluster et l'autre au deuxième. Sensible aux valeurs extrêmes, incapacité de différencier les classes proches	$\min_{x \in C_k, y \in C_l} d(x, y)$
Complete linkage	Distance entre deux clusters est la distance <b>maximale</b> entre deux points, l'un appartenant au premier cluster et l'autre au deuxième. Sensible aux valeurs extrêmes, incapacité de différencier les classes proches	$\max_{x \in C_k, y \in C_l} d(x, y)$
Baverage	Distance entre deux clusters est la distance <b>moyenne</b> entre toutes les paires de points deux à deux. Sensible aux valeurs extrêmes, incapacité de différencier les classes proches	$\frac{1}{ C_l } \frac{1}{ C_k } \sum_{x \in C_k} \sum_{y \in C_l} d(x, y)$
<i>centroid linkage</i>	La distance entre les deux centroïdes. Robuste et peu sensible aux valeurs extrêmes, mais incapacité de différencier les classes proches. $u_k, u_l$ étant les centroïdes	$d(u_k, u_l)$
Stratégie ward	Basée sur la décomposition de la variance. Minimise l'homogénéité et maximise la séparation. $w_{C_k}$ et $w_{C_l}$ étant les poids des cluster k et l. $g_k, g_l$ , leurs centres de gravités	$\frac{w_{C_k} w_{C_l}}{w_{C_l} + w_{C_k}} d(g_k, g_l)$

La technique de clustering hiérarchique ascendante est recommandé pour des fichiers de données contenant moins de 200 observations.

## II.2. Les méthodes basées sur les centroïdes

L'objectif de ces méthodes reste le même que pour le clustering hiérarchique, mais pour ce type de clustering, on sait à l'avance le nombre de clusters à constituer. Le principe fondamental est contenu dans la méthode des centres mobiles dont les étapes sont :

Étape 1 : déterminer  $k$  centres provisoires de clusters, qui induisent un premier clustering de l'ensemble des individus en  $k$  groupes. Ainsi, l'individu  $i$  appartient par exemple au groupe  $I^0$ , s'il est plus proche de  $C^0$  que tous les autres centres.

Étape 2 : déterminer  $k$  nouveaux centres en prenant les centres de gravité des clusters qui viennent d'être obtenus. Ces nouveaux centres induisent un nouveau clustering construit selon la même règle que pour le précédent.

Étape 3 : déterminer  $k$  nouveaux centres en prenant les centres de gravités de clusters qui viennent d'être obtenus. Ces nouveaux centres induisent un nouveau regroupement.

Étape finale : L'Algorithmme s'arrête lorsque les centres ne bougent plus.

Les variantes de cette méthode sont entre autres les k-median, k-mean, nuée dynamiques ,Isodata...

- ❖ La méthode K-means : elle fonctionne exactement comme les centres mobiles, sauf pour le calcul des centres. Un recentrage est effectué dès qu'un individu change de cluster. On n'attend plus que tous les individus soient affectés à un cluster pour en calculer les centres de gravité, ces derniers sont modifiés au fur et à mesure des réaffectations ;
- ❖ La méthode Isodata : le principe des centres mobiles est conservé, mais des contraintes vont permettre de contrôler l'élaboration du clustering. Ces contraintes servent à empêcher la formation de groupes à effectifs trop faibles ou de diamètre trop grand ;

- ❖ La méthode des nuées dynamiques : elle favorise la recherche de groupements stables. C'est une généralisation des centres mobiles dont l'idée est d'associer à chaque cluster un représentant différent de son centre de gravité. Dans la majorité des cas, on remplace le centre de gravité par un ensemble d'individus, qu'on appelle « étalons » et qui constituent un noyau. Ce dernier est censé avoir un meilleur pouvoir descriptif que des centres ponctuels. Parfois, nous disposons déjà d'un clustering donnée et nous voudrons affecter de nouveaux individus aux différents clusters existants. Nous pouvons utiliser les résultats du clustering existant, déduire les barycentres pour enfin trouver les clusters auxquels appartiennent ces individus.
- ❖ La méthode des k-median : elle est similaire à la méthode des k-means, à une différence près. En effet, elle ne va plus définir un cluster par une valeur moyenne, mais par son représentant le plus central plus précisément sa médiane. C'est donc un individu du cluster qui va représenter ce dernier. Utilisant des caractéristiques de médiane, naturellement cet algorithme aura l'avantage d'être non sensibles aux points aberrants.

Nous nous étalerons en particulier sur la méthode de K-means.

### *II.3. Cas particulier des centroïdes: K-mean*

Comme toutes les méthodes basées sur les centroïdes, le nombre de cluster est supposé connu. En pratique, on pourra ainsi se contenter d'explorer une fourchette de valeurs de K, qui sera déterminée en fonction des besoins. Dans la partie métrique, nous avons présenté l'homogénéité globale.

Étant donné K clusters envisagés, nous allons donc chercher à répartir les points  $x_1 \ x_2 \dots \ x_n$  entre les K clusters  $C_1 \ C_2 \ \dots \ C_k$  de telle sorte que la valeur de l'homogénéité globale T soit minimale

$T = \frac{1}{K} \sum_{k=1}^K T_k$  où  $T_k$  représente l'homogénéité du cluster. Comme pour le cas du CAH, des heuristiques sont adoptés. La stratégie d'agrégation de Ward, l'un des plus utilisés en CAH requiert beaucoup de temps de calcul. Dans le k-mean, On utilise l'algorithme de Lloyd.

### II.3.1. Principe

Pour initialiser l'algorithme, K points sont choisi aléatoirement comme centroïdes parmi nos observations. On associe ensuite chaque point au centroïde dont il est le plus proche, formant ainsi K clusters. On peut maintenant recalculer le centroïde de chaque cluster, et recommencer l'opération jusqu'à ce qu'à la convergence.

Il s'agit d'une stratégie gloutonne. L'algorithme converge en général très rapidement, mais peut tomber dans un minimum local. Pour cette raison, il peut être utile de le relancer plusieurs fois et d'évaluer la variance intra-cluster pour chacune de ces répétitions.

### II.3.2. Forme des clusters du K-means

Lorsque nous regardons de près les résultats d'un clustering, les points appartenant à un cluster  $C_k$  sont plus proche du centroïde  $u_k$  que de n'importe quel autre centroïde. Cela implique que l'algorithme du K-means partitionne l'espace selon une **tessellation de Voronoi**. Ainsi, les clusters obtenus sont *convexes*. On ne peut donc pas, avec l'algorithme du K-means, obtenir de cluster en forme de croissant de lune, d'anneau, etc.

On reproche à l'algorithme du K-means d'être non déterministe et que l'initialisation aléatoire des centroïdes peuvent conduire à des résultats très mauvais plus précisément les résultats sont très éloignés de la solution optimale que l'on obtiendrait

si l'on pouvait résoudre exactement notre problème de minimisation de variance intra-cluster. Pour éviter ce problème, on peut utiliser la variante **k-means++** du k-means<sup>20</sup>.

### *II.3.3. K-mean++*

K étapes préalables sont exécutées afin d'obtenir les centroïdes qui initialiseront l'algorithme de k-mean. En effet, on recherche les centroïdes initiaux qui permettent d'avoir les données les plus dispersées possibles. En effet,

Étape 1 : un premier centroïde est choisi de façon aléatoire. Puis on calcule la distance D entre ce centroïde et chacun des autres points.

Étape 2 : Un deuxième centroïde est maintenant choisi de telle sorte qu'un point  $x$  a une probabilité d'être choisi proportionnelle à  $d(x)$ , et donc d'autant plus grande que  $x$  est loin du premier centroïde.

On répète cette opération (calcul de la distance D au i-ième centroïde puis utilisation de cette distance pour sélectionner le (k+1)-ème) jusqu'à avoir K centroïdes.

Une fois les K centroïdes trouvés, on utilise alors l'Algorithme du K-mean.

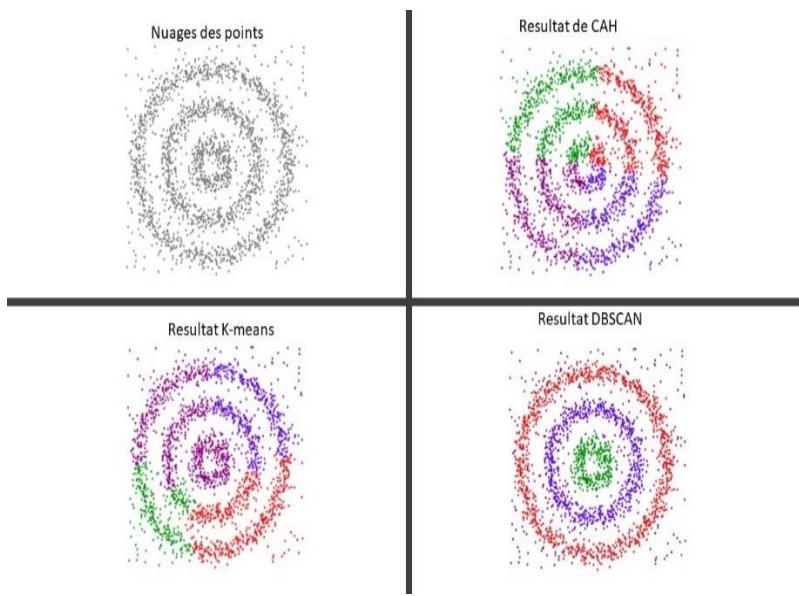
### *II.4. Les méthodes de densités : DBSCAN*

L'algorithme DBSCAN est l'abréviation de Density-Based Spatial Clustering of Applications with Noise). L'algorithme a vu le jour en 1996 mais reste aujourd'hui l'un des algorithmes de clustering les plus utilisés si bien qu'en 2014, une distinction particulière<sup>21</sup> a été décernée à ce dernier. K-Means et CAH échouent tous deux à créer des clusters de formes arbitraires. Ils ne sont pas capables de former des grappes basées sur des densités variables. C'est pourquoi nous avons besoin du clustering DBSCAN. La figure 20 nous permet de mettre en évidence la force de DBSCAN

---

<sup>20</sup> Open Classroom

<sup>21</sup> Une distinction de contribution scientifique ayant résisté à l'épreuve du temps

**Figure 20 Illustration de l'algorithme de DBSCAN par rapport aux CAH et K-means**

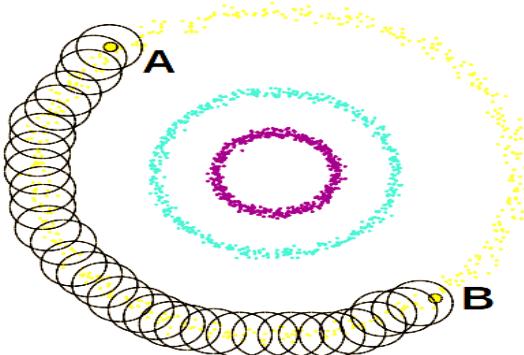
Source : élaboré par nos soins

### Concepts généraux

Pour introduire l'algorithme de DBSCAN, nous avons besoin de définir la notion d'**epsilon voisinage** et **point intérieur**. Étant donné un nombre réel positif  $\varepsilon$ , on appelle **epsilon-voisinage** d'un point  $x$  l'ensemble des points du jeu de données  $X$  dont la distance à  $x$  est inférieure à  $\varepsilon$  :  $N_\varepsilon(x) = \{u \in X / d(u, x) < \varepsilon\}$ .

Etant donné un entier naturel  $n_{min}$ , on dit que  $x$  est un **point intérieur** ou *core point* si son epsilon-voisinage contient au moins  $n_{min}$  points :  $|N_\varepsilon(x)| \geq n_{min}$

On dit maintenant que deux points  $u$  et  $x$  sont **connectés par densité** si l'on peut passer de l'un à l'autre par une suite d'epsilon-voisinage contenant chacun au moins  $n_{min}$  points. Autrement dit, il existe une suite de points *intérieurs*  $v_1, v_2 \dots v_m$  tel que  $v_1$  appartient à l' $\varepsilon$ -voisinage de  $u$ ,  $v_2$  appartient à l' $\varepsilon$ -voisinage de  $v_1$ , et ainsi de suite, jusqu'à ce que  $x$  appartienne au voisinage de  $v_m$ . On dit aussi alors que  $x$  est **atteignable par densité** depuis  $u$ . Parallèlement aux points coeurs, on définit le **point frontière**. Il s'agit d'un point accessible à partir d'un core point et mais on a  $|N_\varepsilon(x)| < n_{min}$ .

**Figure 21: Illustration de l'epsilon voisinage**

Comme montre le graphique de gauche, l'on peut connecter A et B par de petits cercles mais avec des segments sans passer par les autres clusters.

### Principe

On suppose que  $n_{min}$  et  $\varepsilon$  sont connus.

#### *Etape 1 :*

Un point de départ est choisi aléatoirement. Dans un premier temps, **epsilon-voisinage** est déterminé. S'il y a au moins  $n_{min}$  points intérieurs, le point est marqué comme point central et une formation de cluster commence. Sinon, le point est marqué comme bruit. Une fois qu'une formation de cluster commence (disons le cluster A), tous les points dans le voisinage du point initial deviennent une partie du cluster A. Si ces nouveaux points sont également des points centraux, les points qui sont dans leur voisinage sont également ajoutés à groupe A. Cependant il y a lieu de noter qu'un point, qui a une étape donnée a été définie comme bruit peut être revisité et faire partie d'un autre cluster.

#### *L'étape suivante :*

Choisir aléatoirement un point parmi les points qui n'ont pas encore été visités lors des étapes précédentes. Ensuite, la même procédure s'applique.

#### *Etape finale :*

le processus s'arrête lorsque tous les points sont visités.

En Appliquant ces étapes, l'algorithme DBSCAN est capable de trouver des régions à haute densité et de les séparer des régions à faible densité.

A la fin, on obtient des clusters qui contiennent chacun des points atteignables les uns des autres par densités et tous les points frontières de ces points intérieurs. La condition requise pour former un cluster est d'avoir au moins un point central.

### Tableau récapitulatif

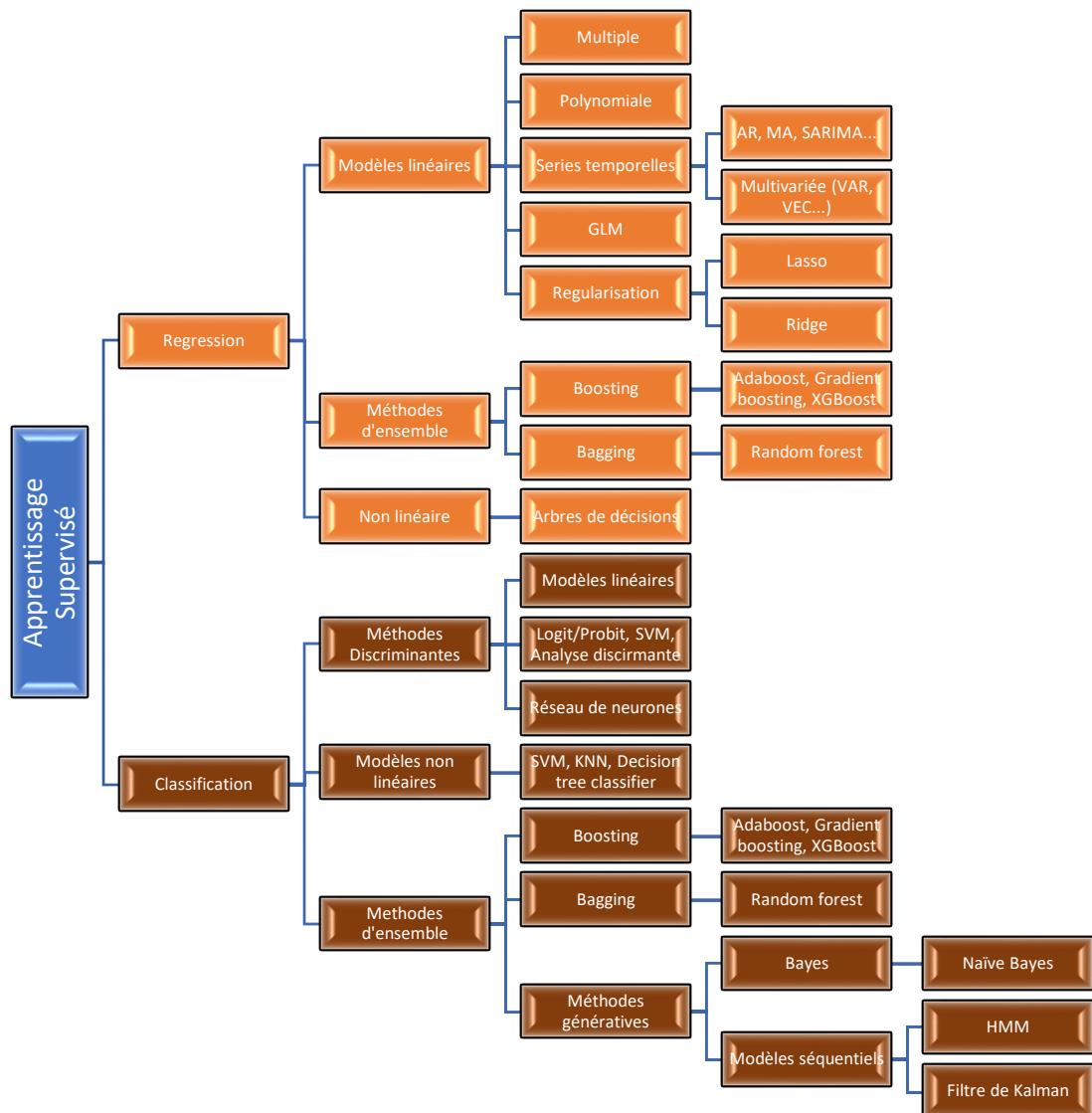
Techniques	Plus	Difficultés
K-means	Grande simplicité Rapidité	Choix du nombre de cluster
CAH	-Stabilité -Non nécessite de définir le nombre de clusters à l'avance	-Temps de calculs énormes -Choix du nombre de cluster
DBSCAN	-Pas de choix du nombre de cluster -Robuste au données aberrantes -Performant sur les formes arbitraires	-Mis en défaut en cas de densités locale -Choix des paramètres

## III. Techniques d'apprentissage supervisé

### III.1 Généralités

Dans l'apprentissage supervisé, on a deux types de modèles catégorisés par le type de variable en sortie : la régression (où la variable de sortie est continue) et la classification (où la variable de sortie prend place dans des classes).

Figure 22: Mapping des techniques d'apprentissage supervisé



Source : élaboré par nos soins.

La régression est un processus de recherche des corrélations entre les variables dépendantes et indépendantes. Elle aide à prédire les variables continues telles que la prédition des tendances du marché, la prédition des prix des logements, etc. La tâche de l'algorithme de régression est de trouver la fonction de mappage pour mapper la variable d'entrée ( $x$ ) à la variable de sortie continue ( $y$ ).

Il existe différents types de régression utilisés en statistique. Chaque type a sa propre importance sur différents scénarios, mais au fond, toutes les méthodes de régression analysent l'effet des variables indépendantes sur la variable dépendante. La structure hiérarchique donne une liste indicative des différentes méthodes possibles.

La classification est un processus de recherche d'une fonction qui aide à diviser l'ensemble de données en classes en fonction de différents paramètres. En classification, un algorithme est formé sur l'ensemble de données de formation, et, en se basant sur cette phase de formation, il classe les données en différentes classes. La tâche de l'algorithme de classification est de trouver la fonction de mappage pour mapper l'entrée ( $x$ ) à la sortie discrète ( $y$ )

Dans les problèmes de classification, il existe deux types d'apprenants :

- Lazy Learners : stockent d'abord l'ensemble de données d'apprentissage et attendent qu'ils reçoivent l'ensemble de données du test. Dans le cas du Lazy Learner, la classification est effectuée sur la base des données les plus connexes stockées dans l'ensemble de données d'apprentissage. Cela prend moins de temps à l'entraînement mais plus de temps pour les prédictions. Exemples : K plus proches voisins, raisonnement basé sur les cas
- Eager Learners : développent un modèle de classification basé sur un ensemble de données d'apprentissage avant de recevoir un ensemble de données test. Exemples : Arbres de décision, Naïve Bayes, ANN, SVM.

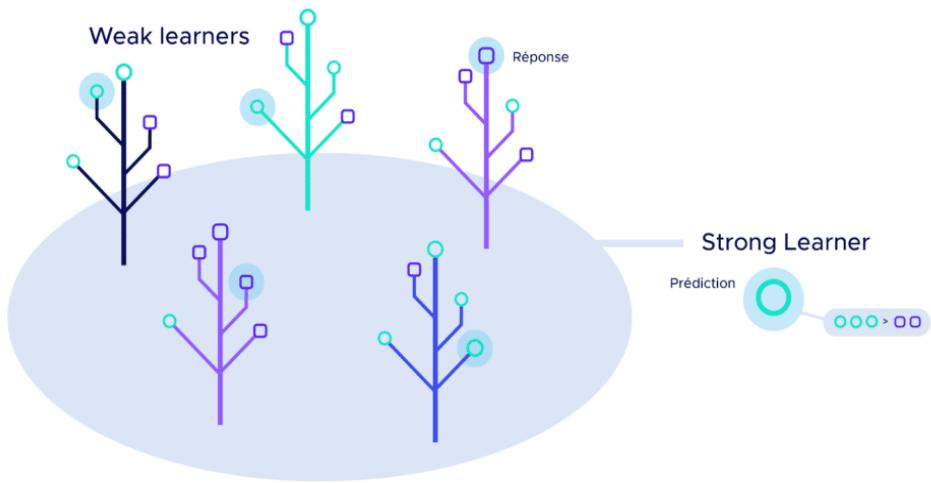
Les algorithmes de classification peuvent être divisés en deux catégories principales : modèles linéaires et modèles non linéaires. La structure hiérarchique dessinée ci-haut en donne une liste indicative de chaque catégorie.

Il existe des techniques qui sont à la fois valable pour les problèmes de classification et de régression : le bagging et le boosting

Le Bagging est une technique qui consiste à assembler un grand nombre d'algorithmes avec de faibles performances individuelles (weak learners) pour en créer un algorithme beaucoup plus efficace (strong learners).

L'idée derrière cet algorithme est que plusieurs petits algorithmes peuvent être plus performants qu'un seul grand algorithme.

**Figure 23: Weak learners et Strong learners**



Source : Extrait du blog de Alban T.<sup>22</sup>

Les weak learners peuvent être de différentes natures et avoir des performances variées, mais ils doivent être indépendants les uns des autres.

L'assemblage des « weak learners » (arbustes) en « strong learner » (forêt) se fait par règle de décision. La règle de la majorité est très utilisée dans ce cas. Pour un problème de classification, si deux weak learners prédisent un individu comme 1 et un autre weak learner prédit 0, s'il s'agit d'une règle de la majorité alors l'algorithme final attribuerait 1 à l'individu. Le Random Forest utilise ce principe.

Les algorithmes de Boosting sont également des méthodes d'ensemble mais la différence apparaît lors de la création des « weak learner » car ces derniers sont désormais dépendants. On dit que l'algorithme apprend de ses erreurs. En effet, chaque weak learner est entraîné afin minimiser les erreurs des « weak learner » qui l'ont précédé. Adaboost, Gradient Boosting et XGBOOST qui sont des exemples qui utilisent ce principe.

<sup>22</sup> Alban T. , Algorithmes de Boosting – AdaBoost, Gradient Boosting, XGBoost 19 Octobre 2020

<https://datascientest.com/algorithmes-de-boosting-adaboost-gradient-boosting-xgboost>

Nous ne nous étalerons pas sur la partie théorique des techniques supervisé. Néanmoins, aux regards des différents tests de modèles et performances constatées dans ce projet sont, les techniques d'apprentissage qu'avons retenus sont :

- Le Random forest pour le problème de classification des clusters non appariés sur TripAdvisor (Partie 2, chapitre 3, III)
- Une régression multiple avec et sans contraintes de la non-négativité et non-positivité des coefficients (Partie 2, chapitre 4, I).

## Principe du Random Forest

C'est un algorithme passe-partout, il est rapide à entraîner, car parallélisable<sup>23</sup> robuste et implémenté dans la plupart des outils utilisés. Il est beaucoup plus intuitif à comprendre que d'autres algorithmes non linéaires, comme le SVM (Support Vector Machine).

On doit les random forests à Leo Breiman, éminent statisticien américain connu pour ses travaux sur les arbres décisionnels. Lui-même avait parfaitement conscience du défaut majeur d'un arbre de décision : sa performance est trop fortement dépendante de l'échantillon de départ.

De plus, on peut s'attendre à ce que l'ajout de quelques nouvelles données dans la base d'apprentissage (ce qui est une bonne nouvelle en soit !) ne modifie pas drastiquement le modèle, qu'il le modifie de façon marginale pour l'améliorer. Ce n'est pas le cas avec un arbre de décision, dont la topologie peut totalement changer avec l'ajout de quelques observations supplémentaires. Plutôt que de lutter contre ces défauts des arbres de décisions, Breiman a eu l'idée géniale d'utiliser plusieurs arbres pour faire des... forêts d'arbres ! C'est de là que vient le terme « forest » dans random forest.

Notons que l'assemblage d'arbres de décision construits sur la base d'un tirage aléatoire parmi les observations constitue déjà un algorithme à part entière connu

---

<sup>23</sup> Le parallélisme consiste à mettre en œuvre des architectures d'ordinateur permettant de traiter des informations de manière simultanée, ainsi que les algorithmes spécialisés pour celles-ci. Ces techniques ont pour but de réaliser le plus grand nombre d'opérations en un temps le plus petit possible.

sous le nom de tree bagging. Les Random forests ajoutent au tree bagging un échantillonnage sur les variables du problème, qu'on appelle feature sampling.

On retiendra que: ***Random forest = tree bagging + feature sampling***

### ***Principe de la régression multiple avec contraintes***

En économétrie, il est connu que le meilleur modèle n'est pas toujours le vrai modèle. Un modèle de régression peut respecter toutes les hypothèses stochastiques, avoir un pouvoir prédictif énorme mais en termes de pouvoir explicatif, ce dernier peut échouer. Lorsque qu'on constraint un modèle, c'est parce que la théorie nous donne des informations à priori. Généralement les modèles des régressions avec contraintes sur les coefficients sont des modèles qu'on réestime après une première estimation de modèle de régressions sans contraintes. On montre théoriquement que pour un modèle constraint, la variance des coefficients est plus faible par rapport à un modèle non constraint. Par conséquent, nos intervalles de confiances deviennent assez étroits. En marketing, ces modèles de régressions avec constraint sur les coefficients est souvent utilisées en présence de cannibalisation. Lorsque nous voulons avoir un modèle explicatif, une régression avec contrainte est souvent recommandée. On rappelle que parmi les données mise à dispositions, se trouve la matrice de cannibalisations des SKUs de l'entreprise.

### **III.2. Évaluation des performances des modèles**

#### *III.2.1. Les métriques régression*

Pour un problème d'apprentissage supervisé, plusieurs métriques peuvent être calculés. La notion de erreurs de prédiction est également liée aux sur-apprentissage ou sur-ajustement overfitting en anglais) et au sous-apprentissage. Le Surajustement signifie que le modèle va (trop) bien s'ajuster sur les données d'apprentissage et il aura du mal à s'adapter à de nouveaux ensembles de données (individus).<sup>24</sup> On montre mathématiquement que selon la complexité du modèle, il existe un compromis entre biais et variance. Ainsi, pour ne pas tomber dans des situations d'overfitting ou underfitting, l'échantillon est divisé en deux : un échantillon d'apprentissage et un échantillon de validation. Il est alors recommandé d'effectuer les mesures de performances sur l'échantillon test. En statistique, les critères d'évaluations de problèmes de régressions sont nombreux. Généralement, en fonction du problème, on privilégiera un critère par rapport à l'autre.

Considérons la valeur réelle d'une série à prédire noté  $y_i$ , et sa valeur prédictive par le modèle noté  $\hat{y}_i$ ; On appelle prévision naïve la moyenne de la série des  $y_i$  notée  $\bar{y}$ . Les métriques les plus connues sont formalisées dans les lignes qui suivent.

- L'erreur moyenne absolue (MAE, Mean Absolute Error)

$$MAE = \frac{1}{n} \sum_{i=p}^n |y_i - \hat{y}_i|$$

- La racine carrée de la moyenne du carré des erreurs (RMSE, Root Mean Squared Error) :

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=p}^n (y_i - \hat{y}_i)^2}$$

---

■<sup>24</sup> <sup>24</sup> Mohammed El Haj Tirari, Apprentissage statistique, INSEA Rabat, Année universitaire 2020-2021

- Erreur de pourcentage absolu moyen (MAPE, Mean Absolute Percentage Error)

:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- Le coefficient de détermination :

$$R^2 = 1 - \frac{\sum_{i=p}^n (y_i - \hat{y}_i)^2}{\sum_{i=p}^n (y_i - \bar{y}_i)^2}$$

- Le critère d'information d'Akaike et de Schwartz:

$$AIC = -2 \log(L) + 2k$$

$$BIC = -2 \log(L) + \ln(n) * k$$

Avec L étant la vraisemblance du modèle et k le nombre de paramètres à estimer.

Le *MAPE* a l'avantage d'être une mesure relative et par conséquent ne dépend pas de l'échelle de données. Cependant, il est sensible aux transformations des données qui ont été appliquées. Les critères MAE et RMSE sont réputés pour pénaliser les grandes erreurs. Vu que ce sont des mesures absolues, ils permettent de connaître l'ordre de grandeurs des erreurs et facilite les interprétations. Le *R*<sup>2</sup> mesure l'ajustement du modèle pour les modèles linéaires. Il s'agit du critère de performance générale. Un *R*<sup>2</sup> élevée est synonyme d'un bon ajustement linéaire du modèle. Bien que ce dernier ne soit pas très conseillé dans l'étude de certains types de données (séries temporelles) car parfois trompeur. Pour une comparaison de modèles linéaires multivariés, on utilisera plutôt Le *R*<sup>2</sup> ajusté car ne favorisant pas forcément les modèles avec plus de variables. D'ailleurs le *R*<sup>2</sup> ou le *R*<sup>2</sup> ajusté perdent de sens théorique pour des modèles dont on ignore la constante. Sans oublier les hypothèses de normalité des erreurs. Une alternative est l'utilisation des critères d'informations. Meilleur est un modèle, plus petit est l'AIC. Il en est de même pour le BIC. A la différence, cette dernière pénalise plus fortement les modèles trop complexes. Lorsqu'on désire un modèle très parcimonieux,

le BIC sera privilégié. Dans la littérature, il existe plusieurs autres critères d'informations.

Quel indicateur choisir ? En général, On ne peut dire que l'un est supérieur à l'autre. Cependant, pour des problématiques bien définies, on privilégiera quelques un. D'ailleurs, pour plus de robustesse, les spécialistes combinent les indicateurs afin d'évaluer leurs modèles.

### ***III.3.2. Les métriques de classification***

Les métriques qui permettent d'évaluer sont nombreuses. Les plus connus sont celles qui permettent de mesurer les prédictions pour les classifieurs binaires.

#### **a) Matrice de confusion et ses métriques**

Il s'agit en réalité d'un tableau croisé réalisé pour résumer les performances du modèle. Il est calculé également sur l'échantillon de validation. Ce tableau de contingence, est obtenue en comparant les données ajustées par le modèle (**prédictes**) avec des données de référence (**réelles**) **contenue dans l'échantillon de test**. Même si nous la présentons dans un cas binaire, la matrice de confusion peut être tracée pour une classification multiclasse (classes supérieures à 2).

On suppose qu'on est dans un cas de classification binaire. Le tableau de contingence aura en lignes les données de références et en colonne les colonnes données classées. Il se présente comme ci-après :

Prédites Réelles	Positifs	Négatifs
Positifs	TP	FN
Négatifs	FP	TN

Avec :

TP représentant les vrais positifs c'est-à-dire le nombre d'observations correctement prédictes dans la classe 1 ;

TN représentant les vrais Négatifs, donc le nombre d'observations correctement prédictes dans la classe 2 ;

FP désigne les Faux positifs, plus précisément nombre d'observations prédictes dans la classe 1 alors qu'elles font partie de la classe 2 ;

FN désigne les Faux Négatifs

A partir de la matrice de confusion, on calcule des métriques comme l'exactitude, la précision, le recall ou le score F1.

$$\text{Accuracy(Exactitude)} = \frac{TP + TN}{TP + FP + FN + FP}$$

C'est le rapport entre le nombre de prédictions correctes et le nombre total d'échantillons d'entrée. Il n'est pas recommandé lorsqu'il y a un déséquilibre des classes dans l'échantillon.

$$\text{Précision} = \frac{TP}{TP + FP}$$

La précision est une bonne mesure pour déterminer quand les coûts des faux positifs sont élevés

$$\text{Recall} = \frac{TP}{TP + FN}$$

**Le recall** est une métrique utilisée lorsque qu'il y a un coût élevé associé aux faux négatifs.

$$F1 = 2 * \frac{\text{Précision} * \text{Recall}}{\text{Précision} + \text{Recall}}$$

Lorsqu'on cherche le juste milieu entre le Recall et la précision, le score F1 est une solution. En réalité, parfois, la différence entre le score F1 et la précision est que parfois un grand nombre de vrais négatifs peut largement contribuer à la précision. Dans les activités commerciales plus comme FMCG, les faux négatifs et les faux positifs ont généralement des coûts commerciaux, si le score F1 pourrait être une meilleure mesure à utiliser si nous devons rechercher un équilibre entre la précision et le Recall.

### b) L'aire sous la courbe ROC : AUC

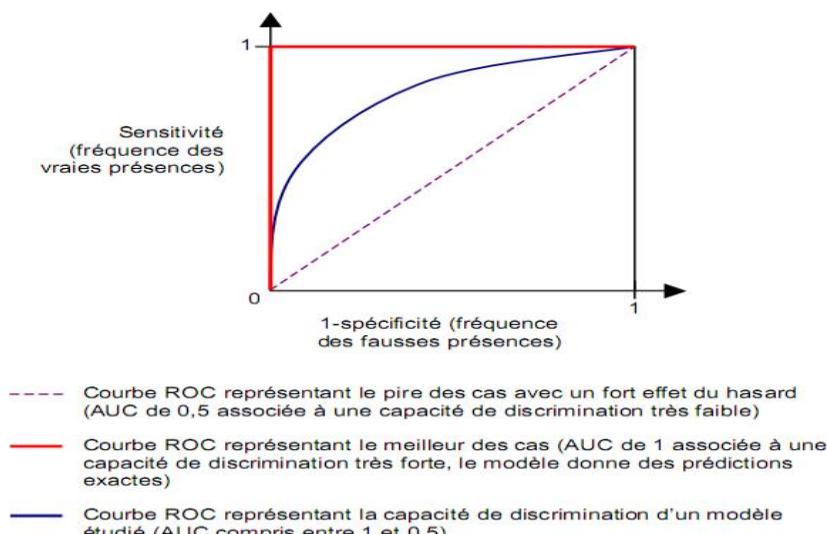
L'aire sous la courbe ROC est un indicateur synthétique de classificateur. Mais que désigne la courbe ROC ? On désigne respectivement par sensibilité et spécificité, le taux de vrai positif et le taux de vrai négatif.

*Taux de vrai positif (TPR) : sensibilité*

*Taux de faux positifs (FPR): 1-Specificité*

L'idée de la courbe ROC est de faire varier le seuil<sup>25</sup> de 1 à 0 et, pour chaque cas, calculer le TVP et le TFP. Ces résultats sont représentés sous forme de graphique avec en absence 1-spécificité et en ordonnées la sensibilité.

Figure 24:courbe ROC



Plus la courbe se rapproche de l'angle supérieur gauche, plus le classifieur est considérant performant. Il permet également de déterminer le seuil qui maximise le taux de vrais positifs et minimise par la même occasion le taux de faux positifs. Il suffit de résoudre le programme contraint :

$$\operatorname{argmin} (0 - FPR)^2 + (1 - TPR)^2 \text{ Avec } 0 \leq FPR \leq 1 \text{ et } 0 \leq TPR \leq 1$$

<sup>25</sup> Il s'agit d'un seuil à partir duquel, on prédit un cas positif ou pas.

L'aire sous la courbe indique la probabilité pour que la fonction SCORE place un positif devant un négatif (dans le meilleur des cas AUC = 1) .

L'aire sous la courbe ROC est utilisé le plus souvent dans des problèmes de classification binaire parce que :

- Il Indépendant des matrices de coûts de mauvaise affectation.
- Il est opérationnel même dans le cas des distributions très déséquilibrées
- Il permet la Comparaison plusieurs modèles quelle que soit la matrice de coût
- Les résultats restent valables même si l'échantillon test n'est pas représentatif
- Un indicateur synthétique facilement interprétable

En général, la courbe ROC est utilisé pour les modèles de classification qui retournent des valeurs réelles.

## Conclusion

Ce projet est intéressant dans la mesure que nous appliquerons à la fois des techniques d'apprentissage statistique supervisées mais également non supervisés. Ce chapitre clos la première partie du livre qui concerne la partie théorique. Nous commencerons la réalisation du projet dans les pages prochaines.

# Deuxième partie

## Modélisation

## Chapitre 1 :

---

# Données et infrastructures et schéma de modélisation

Introduction .....	- 94 -
I.      Données utilisées .....	- 94 -
I.1.    Données internes.....	- 94 -
I.1.1.    Les données de ventes .....	- 95 -
I.1.2.    Les données d'indicateur d'exécution de livraison (RED).....	- 95 -
I.1.3.    Les données sur notre force de vente.....	- 95 -
I.1.4.    Les données sur les réfrigérateurs .....	- 95 -
I.1.5.    Les données relatives au temps .....	- 95 -
I.1.6.    Les données relatives au calendrier .....	- 96 -
I.1.7.    Matrice de cannibalisation .....	- 96 -
I.1.    Données externes.....	- 96 -
I.1.1.    Les données d'Experian .....	- 97 -
I.1.2.    Les données d'Unacast.....	- 98 -
I.1.3.    Les données de WorldPop .....	- 99 -
I.1.4.    OpenStreetMap et de Pitney Bowes .....	- 100 -
I.1.5.    Les données scrapées de TripAdvisor .....	- 103 -
I.1.6.    Les données de Flickr.....	- 107 -
II.    Infrastructures du projet .....	- 108 -
II.1.    Cloud et architectures .....	- 108 -
II.2    Azure Blob storage .....	- 110 -
II.3.    Python .....	- 110 -
II .4.    Databricks, Apache Spark et Azure Databricks .....	- 111 -
II.5.    Azure Translator.....	- 113 -
III. Schéma de la modélisation du projet .....	- 114 -
Conclusion.....	- 115 -

## Introduction

Dans ce chapitre, nous allons d'abord présenter les données utilisées pour ce travail tout en essayant d'être brève pour des soucis de confidentialités puis parler des outils utilisés pour mener à bien la modélisation.

### I. Données utilisées

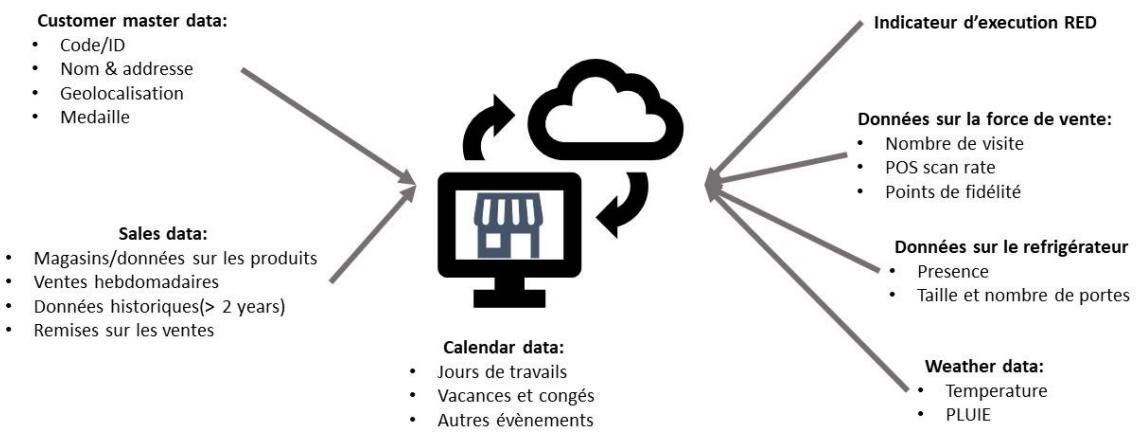
#### I.1. Données internes

Nous avons utilisé plusieurs dimensions de données afin d'avoir une bonne connaissance du client. On rappelle que nos clients sont les points de ventes. Le lien entre les différentes bases de données étant le code client. En effet, le temps et le calendrier mis à part, pour chaque dimension, la table de données a comme un attribut code client.



### Données internes

Plusieurs sources de données internes du système Coca Cola seront exploitées.



### *1.1.1. Les données de ventes*

Nous disposons des données par point de ventes grâce au stockage qui est réalisé directement dans le cloud via les données recueillies par les applications relatives à nos services. Plusieurs niveaux d'agrégations sont possibles. On pourrait agréger les données à l'échelle d'un cluster ou d'une région ou encore des données de ventes mensuelles. Également, ces données de ventes concernent chaque SKU. Une unité de gestion des stocks est un identifiant unique pour un article vendu par un détaillant. Les SKU permettent de différencier les produits les uns des autres. Il serait difficile de suivre les ventes et les stocks sans les classer par marque, emballage, type, gout, taille ou tout autre trait d'identification.

### *1.1.2. Les données d'indicateur d'exécution de livraison (RED)*

Disponible à l'échelle des points de ventes, elles permettent de mesurer la bonne exécution du réfrigérateur et des produits vendus. Les détails de calculs ne seront pas mentionnés dans ce rapport pour des questions de confidentialités.

### *1.1.3. Les données sur notre force de vente*

Des visites sont fréquemment faites auprès de chaque PdV afin d'enregistrer les commandes et scanner les réfrigérateurs. Un indicateur est calculé sur ces critères.

### *1.1.4. Les données sur les réfrigérateurs*

Ces données permettent de savoir à quel point de ventes, des réfrigérateurs de l'entreprise ont été placées. S'ils existent, des données sur leurs caractéristiques (le nombre de porte, la taille ...) sont récupérées.

### *1.1.5. Les données relatives au temps*

Nul n'ignore que le temps impacte significativement les ventes dans le secteur des biens de grande consommation (FMCG). Ces données concernent les prévisions météorologiques et également les données historiques des températures et des pluies.

### I.1.6. Les données relatives au calendrier

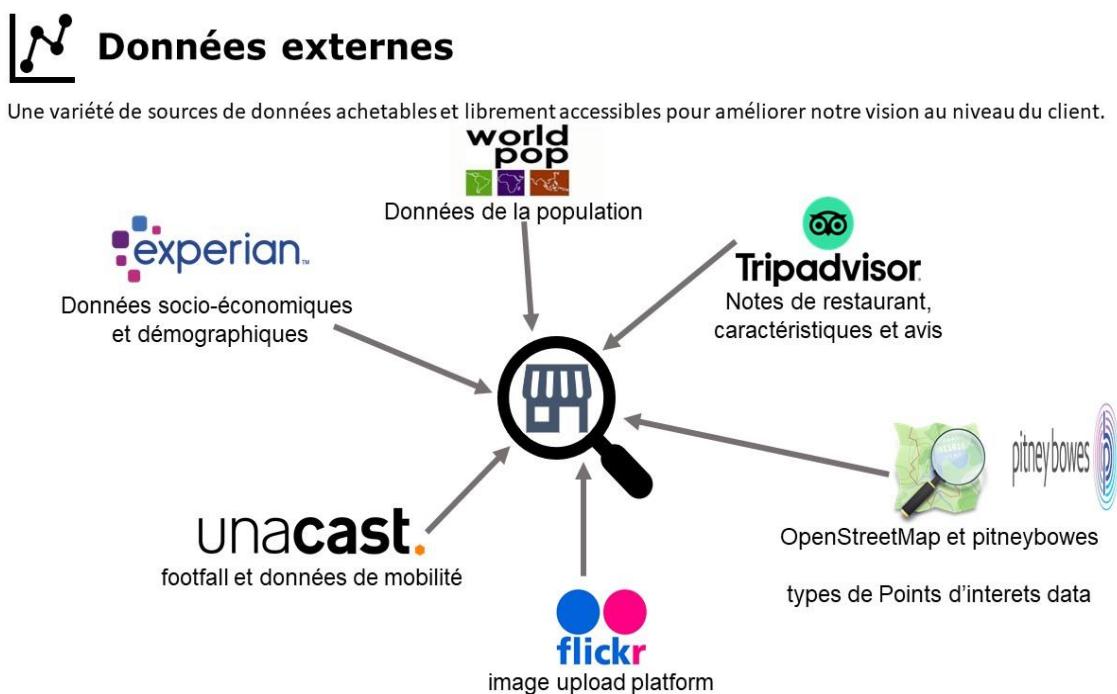
Elles concernent les jours de travaux, les jours de ramadan, les vacances... Elles peuvent servir de variables de contrôles. Par exemple, La consommation des ménages connaît des pics pendant les événements extraordinaires.

### I.1.7. Matrice de cannibalisation

On parle de cannibalisation lorsque les ventes d'un nouveau produit d'une entreprise sont en partie expliquées à une baisse des ventes d'un autre produit plus ou moins substituable de la même entreprise. Pour les boissons d'ECCBC, entre une marque de produits à une autre marque il y a risque de cannibalisation selon leur degré de ressemblance en termes de saveur ou de format. Nous utiliserons les données de la matrice de cannibalisations dans la construction des fonctions de demandes ;

## I.1. Données externes

Afin de mieux enrichir nos sources de données, nous utilisons des données économiques, démographiques, de mobilités mais également des données sur les POI.



Dans l'économie en évolution rapide d'aujourd'hui, être en mesure d'accéder et d'analyser des données spatiales de haute qualité est fondamental pour les compagnies utilisant la géolocalisation . Qu'il s'agisse de consolidation de sites dans le commerce de détail, ou d'étude de marché, il est essentiel de disposer des données spatiales adaptées afin de créer des modèles efficaces et recommandations pertinentes. Des entreprises spécialisées (Experian, Unacast, Pitney Bowes) et de projets collaboratifs (OpenStreetMap, WorldPop) mettent en places de données à l'échelles granulaires.

### *1.1.1. Les données d'Experian*

Experian fournit des attributs sociodémographiques clés à travers des grilles enrichies qui peuvent aller jusqu'à 250 mètres \* 250 mètres couvrant le globe et dans notre cas le Maroc et l'Algérie, offrant un accès immédiat à des informations précises pour mener des analyses poussées. Experian fournit des données sociodémographiques et des données relatives aux dépenses de consommation grâce à des données de recensement, des cartes de crédits et banques. Ci-dessous se trouvent les attributs les plus important des données socio-économiques et démographiques disponibles

Experian Sociodemographics	Experian Consumer Spending (en anglais)
Moyenne annuelle de la population : nombre total	Dépenses en alimentation et boissons non alcoolisées (€)
Ménages : nombre total	Dépenses pour les boissons alcoolisées et le tabac (€)
Population : hommes	Dépenses en vêtements et chaussures (€)
Population : femmes	Dépenses de logement (€)
Population par âge : 0 - 14 ans, total	Dépenses en biens et services ménagers (€)
Population par âge : 0 - 14 ans, hommes	Dépenses en biens de santé et services médicaux (€)
Population par âge : 0 - 14 ans, femmes	Dépenses de transport (€)
Population par âge : 15 - 29 ans, total	Dépenses de communication (€)
Population par âge : 15 - 29 ans, hommes	Dépenses pour les loisirs et la récréation (€)
Population par âge : 15 - 29 ans, femmes	Dépenses d'éducation (€)
Population par âge : 30 - 44 ans, total	Dépenses d'hôtellerie et de restauration (€)
Population par âge : 30 - 44 ans, hommes	Dépenses en biens et services divers (€)
Population par âge : 30 - 44 ans, femmes	
Population par âge : 45 - 59 ans, total	
Population par âge : 45 - 59 ans, hommes	
Population par âge : 45 - 59 ans, femmes	
Population par âge : 60 ans et plus, total	
Population par âge : 60 ans et plus, hommes	
Population par âge : 60 ans et plus, femmes	
Pouvoir d'achat : millions d'euros	
Pouvoir d'achat : Euro par habitant	
Pouvoir d'achat : indice (pays eq.100)	

Toutes ces données sont anonymisées. Ces données sont individuelles et nous disposons de la géométrie des individus à l'échelle des grilles.

### *1.1.2. Les données d'Unacast*

Unacast est une société de données sur la mobilité humaine qui s'engage à comprendre comment les gens se déplacent sur la planète. Les professionnels de l'immobilier commercial sophistiqués et axés sur les données, les détaillants, les chercheurs, et les Data scientists utilisent Unacast pour réaliser des études plus granulaires. Grace aux images satellitaires et aux applications mobiles nous pouvons disposer par le biais d'UNACAST des données de mobilités anonymisées avec leur géolocalisation à l'échelle de grille carrée. L'avantage de ces données est qu'elles sont d'une granularité très fine au niveau des dates. Des agrégations à l'échelle journalière ou mensuelle sont possibles

<b>mobility data</b>
Année
Mois
Jour
géometry
Résidents
travailleurs
Autres
Habitants quittant la localité et retournant après

### I .1.3. Les données de WorldPop

La production des ensembles de données spatiales par WorldPop suit principalement les méthodologies décrites dans Stevens et al (2015)<sup>26</sup>, Alegana et al (2015)<sup>27</sup>, Deville et al (2014)<sup>28</sup>, Gaughan et al (2013)

Les méthodes utilisées sont conçues dans l'optique d'un accès ouvert et d'une application opérationnelle, en utilisant des méthodes transparentes, entièrement documentées et évaluées par des pairs pour produire des cartes facilement actualisables, accompagnées de métadonnées et de mesures d'incertitude. Le projet WorldPop a été lancé en octobre 2013 pour combiner les projets de cartographie démographique AfriPop, AsiaPop et AmeriPop. Il vise à fournir des archives en libre accès des données démographiques spatiales pour l'Amérique centrale et du Sud, l'Afrique et l'Asie afin de soutenir les applications de développement, de réponse aux catastrophes et de santé.

WorldPop fournit des données à haute résolution, ouvertes et contemporaines sur les distributions de la population humaine, permettant une mesure précise des distributions, compositions, caractéristiques, croissance et dynamiques de la population locale, à l'échelle nationale et régionale. Elles sont stockées sous format raster. La granularité peut aller jusqu'à des grilles carrées 100m\*100\*m.

Les jeux de données de WorldPop incluent des estimations du nombre de personnes résidant dans chaque cellule de grilles, et leurs structures âge/sexe pour chaque pays à revenu faible ou intermédiaire. Son approche intègre des données de recensement, d'enquête, de satellite et de SIG dans un cadre flexible d'apprentissage

---

<sup>26</sup> Forrest R. Stevens, Andrea E. Gaughan, Catherine Linard, Andrew J. Tatem, désagrégation des données du recensement pour la cartographie des populations à l'aide de forêts aléatoires à l'aide de données de télédétection et de données auxiliaires février 17, 2015 <https://doi.org/10.1371/journal.pone.010704>

<sup>27</sup> Alegana V. A., Atkinson P.M., Pezzulo C., Sorichetta A., Weiss D., Bird T., Erbach-Schoenberg E. et Tatem A. J. 2015 Fine resolution mapping of population age-structures for health and development applications J. R. Soc. Interface.12201500732015007

<sup>28</sup> Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R. Stevens, Andrea E. Gaughan, Vincent D. Blondel, Andrew J. Tatem, Proceedings of the National Academy of Sciences Dynamic population mapping using mobile phone data Nov 2014, 111 (45) 15888-15893; DOI : 10.1073/pnas.1408439111

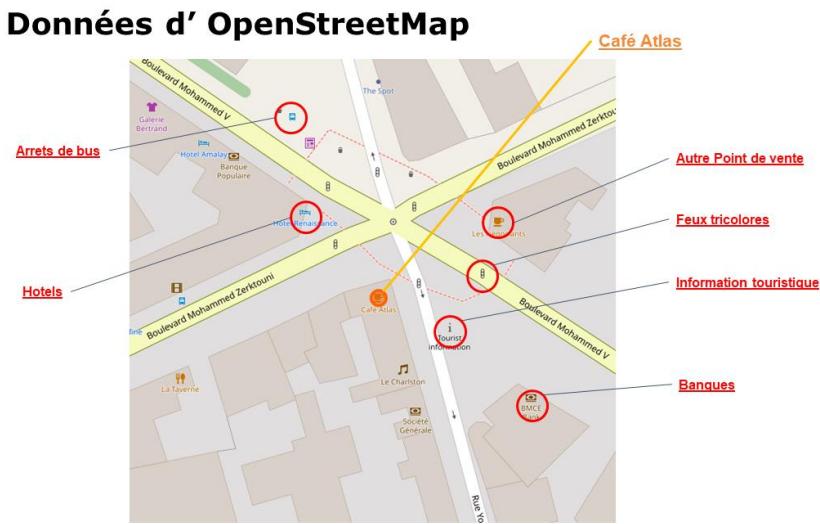
automatique. Ainsi, des cartes à haute résolution des chiffres et des densités de population pour la période 2000-2020 sont produites, accompagnées de métadonnées et d'articles universitaires évalués par des pairs sur les méthodes. Lorsque les données de recensement sont obsolètes ou peu fiables, des approches d'estimation de la population par satellite et par enquête sont mises en œuvre en collaboration avec les bureaux statistiques nationaux. La collection de données WorldPop Global comprend des surfaces de population pour les populations totales (à la fois ajustées pour correspondre aux estimations nationales des Nations Unies et non ajustées) ainsi que des ventilations par classes d'âge et de sexe, à des pas de temps annuels entre 2000 et 2020, avec une résolution spatiale de 3 secondes d'arc (environ 100 m à l'équateur). L'expérience a montré que ces données sont très précises. Elles sont disponibles en format raster. Une non-maitrise des SIG rend l'utilisation de ces données plus compliquées. C'est pourquoi des sociétés comme Experian utilisent les données de WorldPop afin de produire des données plus accessibles aux non spécialistes mais également aux spécialistes qui préfèrent ne pas perdre du temps dans l'extraction de la bonne base de données.

#### *1.1.4. OpenStreetMap et de Pitney Bowes*

OpenStreetMap (OSM) est bâti par une communauté de cartographes bénévoles qui contribuent et maintiennent les données des points d'intérêts géographiques, plus précisément des routes, sentiers, cafés, stations ferroviaires et bien plus encore, partout dans le monde. OpenStreetMap encourage et met en avant la connaissance locale du terrain. Les contributeurs utilisent l'imagerie aérienne, les récepteurs GPS et les cartes classiques du terrain pour vérifier qu'OSM est exact et à jour. Parmi les contributeurs, se trouvent des cartographes, des professionnels du SIG, des ingénieurs qui font fonctionner les serveurs d'OSM, des humanitaires cartographiant les zones dévastées par une catastrophe et beaucoup d'autres.

OpenStreetMap est en données ouvertes. Nous utilisons alors les données de ce projet afin de repérer les points d'intérêts et extraire leur localisation et attributs.

**Figure 25 : Données d'OpenStreetMap**



L'ensemble des données sont disponibles à partir du serveur de Geofabrik, un centre créé pour rendre les données d'OSM plus attrayantes pour des utilisations commerciales.

« Les fichiers OSM sont des fichiers au format XML contenant la description d'une carte avec trois éléments notables (les nœuds, les chemins et les relations). Les nœuds décrivent une position fixe dans l'espace, comme une ville. Les chemins décrivent un segment ou un polygone, comme le contour d'un pays. Tandis que les relations sont simplement des groupes d'éléments. Chaque élément est décrit par différents attributs XML et des tags permettant d'associer des informations à l'élément sous forme de clef-valeur ».<sup>29</sup> L'exemple suivant décrit par exemple la ville de Rabat sous forme d'un nœud avec sa latitude et sa longitude comme attributs ainsi qu'une liste de tags.

Pour utiliser ces données, nous devons alors parser les fichiers OSM disponible en open source sur le site de geofabrik<sup>30</sup> vers des fichiers Json ou Excel. Par parser, on s'intéressera aux géométries des points d'intérêts, leur nom, leur adresse et bien d'autres attributs. Bien que ce travail informatique derrière soit fastidieux, il en vaut la peine car

---

<sup>29</sup> Sacha schutz, Explorer des données cartographiques avec osmium, <https://dridk.me/osmium-tools.html>

<sup>30</sup> <http://download.geofabrik.de/>

les données recueillies sont presque complètes. Parfois, lorsque le besoin est spécifique et l'on ne dispose pas des compétences nécessaires pour utiliser les données d'OSM, Pitney Bowes (PB) est une entreprise qui vend des données de point d'intérêt géographiques. Cette dernière ne donne pas une liste exhaustive des points d'intérêt d'une localité mais les plus importants.

Groupe de point d'intérêt de Pitney Bowes
Groupe de point d'intérêt
Agriculture, Chasse et Pêche
Industrie minière
Construction
Manufactures
Transport and biens communs
Commerce grossiste
Commerce de détail
Finance et assurance
Service
Administration publique
Tourisme

Quant aux données d'OSM, vu la multitude des points d'intérêts, nous avons choisi de récupérer les attributs suivants recensés dans la liste ci-après.

### Liste des points d'intérêt choisis dans OSM

comptable	commercial	fondation	centre commercial	snack	commerce de détail
centre de jeux	centre communautaire	aliments surgelés	marina	football	aviron
agence de publicité	entreprise	carburant	marché	centre sportif	rugby league
alcool	confiserie	jeux gaéliques	golf miniature	salle de sport	rugby union
salle de jeux	commodité	jeux d'argent	motel	stade	sandwich
aquarium	palais de justice	jardinerie	motocross	gare	école
architecte	coworking	général	moteur	papeterie	fruits de mer
centre artistique	cricket	géodésie	entreprise de déménagement	vendeur ambulant	services
association	croquet	golf	multi	strip-club	tir
agent de cautionnement	crémerie	gouvernement	musée	supermarché	snack
boulangerie	curling	primeur	marchand de journaux	géomètre	football
banque	cyclisme	maison d'hôtes	journal	natation	centre sportif
bar	danse	guide	ONG	piscine	salle de sport
basket	fléchettes	salle de sport	boîte de nuit	baby-foot	stade
station balnéaire	épicerie fine	hackerspace	notaire	impôts	gare
boissons	grand magasin	alimentation saine	bureau	conseiller fiscal	papeterie
biergarten	courses de chiens	courses de chevaux	bio	télécommunication	vendeur ambulant
billard	pharmacie	équitation	pâtisserie	tennis	strip-club
boxe	établissement d'enseignement	hôtel	pharmacie	terminal	supermarché
gare routière	agence de placement	auberge	pizza	théâtre	géomètre
arrêt de bus	fournisseur d'énergie	hôtel	planétarium	parc à thème	natation
café	ingénieur	glace	police	mairie	piscine
casino	équitation	hockey sur glace	parti politique	gare	baby-foot
association caritative	jeu d'évasion	patinoire	détective privé	université	impôts
pharmacie	ferme	assurance	gestion immobilière	visa	conseiller fiscal
poulet	restauration rapide	cybercafé	pub	eau	télécommunication
cinéma	gare maritime	informatique	bains publics	parc aquatique	tennis
escalade	financier	karting	quango	water-polo	terminal
aventure d'escalade	poisson	kiosque	raquette	service des eaux	théâtre
combat de coqs	centre de remise en forme	école de langues	religion	oui	parc à thème
café	aire de restauration	avocat	recherche	zoo .	mairie
collège	visa	bibliothèque	restaurant	université	gare
	eau	logistique			

### I.1.5. Les données scrapées de TripAdvisor

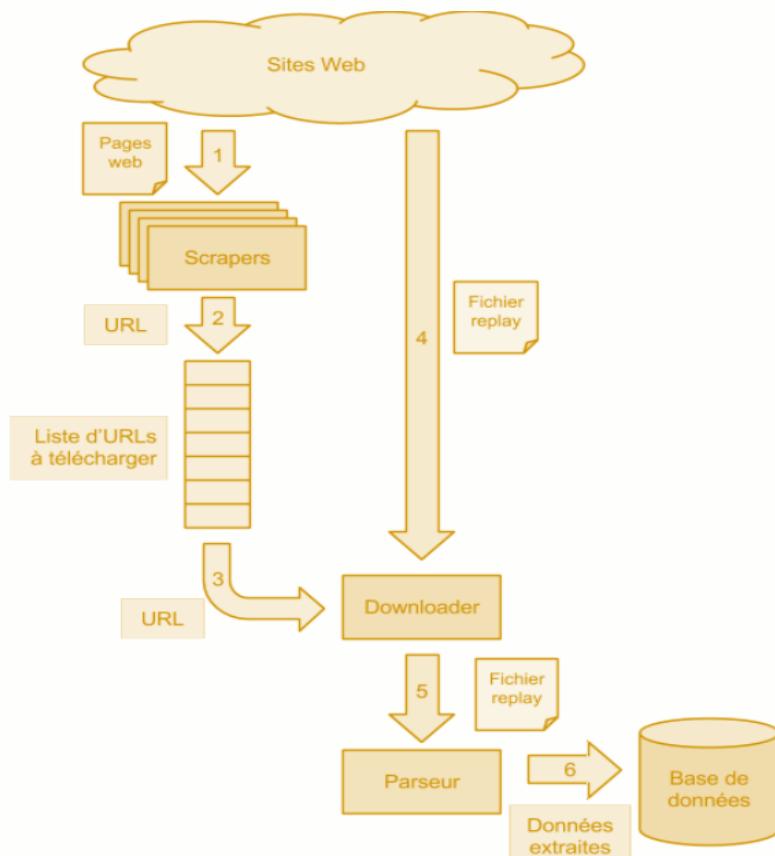
TTripAdvisor (TA) est un site web américain qui offre des avis et des conseils touristiques émanant de consommateurs sur des hôtels, restaurants, villes et régions, lieux de loisirs, etc., à l'international. Il s'agit également du plus grand site de voyage au monde. Pour des touristes qui viennent au Maroc ou également des personnes cherchant un restaurant spécialisé, TripAdvisor est une solution pratique en matière d'avis ou une recommandation.

**Figure 26: Données de TripAdvisor**



Lorsque nous nous rendons sur le site, nous pouvons voir la figure de la page précédente. Les points commentés avec les flèches nous intéressent dans notre projet. Les notes des voyageurs ou les commentaires peuvent nous permettre de dériver un score appelé touristness(une façon de capter les fréquentations touristiques). Le nombre d'étoiles pourraient être un proxy de premiumness(. C'est dans ce sens que nous voudrions recueillir ces informations. La seule solution qui s'offre à nous est le Scraping car collecter ces données manuellement est presque impossible. Le Web Scraping est une technique permettant d'extraire les données des pages Web, mais de manière automatisée. Un script de scraping web peut charger et extraire les données de plusieurs pages en fonction des besoins. Bien que nous ayons utilisé le scraping, il ne s'agit pas d'un objectif clés du projet, donc la présentation technique n'a pas été introduite. Néanmoins, l'on peut résumer cette technique à travers la figure ci-après.

**Figure 27: Etapes de scraping**



Il se fait en 6 étapes dont la plus importante reste l'étape du Parseur. Les trois premières consistent à se diriger vers l'ensemble des adresses URL à scraper. Généralement, il suffit de créer un algorithme qui utilisera la première page comme

input et se diriger vers les pages suivantes automatiquement jusqu'à parcourir tous les URLs voulus. L'étape commence à une vérification du code source de la page car le Parser est basé sur les balises. Grâce aux parseurs, nous récupérons les données qui nous intéressent. Pour parser le code HTML, l'outil que nous avons utilisé est Scrapy.

Scrapy est un cadre d'exploration du Web qui fournit un outil complet pour le scraping. Dans Scrapy, nous créons des araignées qui sont des classes python qui définissent comment un ou plusieurs sites particuliers seront scrapés. BeautifulSoup et sélenium sont d'autre outils mais que nous n'avons pas utilisés car notre projet est évolutif et est amené à être à grande échelle.

Nous avons ici scraper des données de restaurants, cafés et d'hôtels. Les catégories trouvées sur le site sont définies comme le montre la figure

Restaurant et vie nocturne	Hôtels
✓ Bouchées rapides	✓ Clubs de danse et discothèques,
✓ Dessert	✓ Clubs de <u>comédie</u> ,
✓ Café et thé	✓ Bars à <u>vins</u>
✓ Boulangeries	✓ Magasins de café
✓ Bars et pubs	✓ Clubs et bars",
✓ Dîner <u>avec</u> un chef local	✓ Clubs et bars de jazz",
✓ <u>Marché des aliments spécialisés</u>	✓ Bars <u>karaoké</u>
	✓ Piano Bars"
	✓ Hôtels d'affaires
	✓ Hôtels de plage
	✓ Hôtels <u>romantiques</u>
	✓ Hôtels de <u>luxe</u>
	✓ Familiale"
	✓ SPA Resorts,
	✓ Hôtels Casino,
	✓ <u>Hôtel</u> <u>voile</u>

Les features collectées après scraping sont au nombre de 26

Liste des features obtenus
id
statut
nom
adresse
géolocalisation
heures de travail
établissement
types d'établissement
catégories de restaurant
prix (min et max)
notes moyennes
header_line
critiques et avis
thèmes de reviews
Nombres de reviews
notation par catégorie
ratio des notations
commande en ligne
menu sur la table
site
wifi
email
phone_number
Détails
certificat
classement

	url	tched	id	status	name	address	location	opening_hours	establishment_type	jurisdiction	feature	price_range	ranking	avg_rating	header_line	reviews	view_topview	view_courses	category_of_restaurant
0	<a href="https://www.2021-c240512">https://www.2021-c240512</a>	200	Hotel d'Or 13 Cour de la Revolution, Al	[{"lat": 36.89935, "lon": 7.76158}]									4		["title": "A Reasonable"]	5		Excellent	
1	<a href="https://www.2021-c99d4674">https://www.2021-c99d4674</a>	200	Hotel Dar 14 Rue ben ouhiba, Annaba	[{"lat": 36.88319, "lon": 7.75682}]									3		["title": "MAUVAIS H3"]			Excellent	
2	<a href="https://www.2021-c5c4858a">https://www.2021-c5c4858a</a>	200	Hotel Ned Route de la Rocade Mansoura, Tlemcen Algeria										3		["title": "Contacto", "text": "15"]			Excellent	
3	<a href="https://www.2021-c6bdff44">https://www.2021-c6bdff44</a>	200	Hotel La R Cite des Palmiers Aokas, Bel	[{"lat": 36.63277, "lon": 5.246185}]									4,5		["title": "Treated like 27"]			Excellent	
4	<a href="https://www.2021-c5e739c4">https://www.2021-c5e739c4</a>	200	Sophotel RN 9 Route de Setif Ireyane, Bejaia 06000 Algeria																Excellent
5	<a href="https://www.2021-c5b1a05a">https://www.2021-c5b1a05a</a>	200	La Bravou 09 Rue des Freres Akouta, Bejaia 06000 Algeria																Excellent
6	<a href="https://www.2021-c0774f72">https://www.2021-c0774f72</a>	200	Hotel Braf 25 Rue de la Liberte Bejaia, Al	[{"lat": 36.753033, "lon": 5.072349}]									4		["title": "Very Good"]	38		Excellent	
7	<a href="https://www.2021-c546dbb6">https://www.2021-c546dbb6</a>	200	Hotel Hor Tighremt Plage, Bejaia Alger	[{"lat": 36.75328, "lon": 5.08633}]									4,5		["title": "Hala", "text": "3"]			Excellent	
8	<a href="https://www.2021-c7f05656">https://www.2021-c7f05656</a>	200	Auberge D La Zone D'Activite, Larbaa, Al	[{"lat": 36.75105, "lon": 5.07978}]									4		["title": "Auberge des 2"]			Excellent	
9	<a href="https://www.2021-c9ff68d1">https://www.2021-c9ff68d1</a>	200	Residence Plage de Boulimat, Bejaia 06000 Algeria	[{"lat": 36.75956, "lon": 5.08923}]															Excellent
10	<a href="https://www.2021-c237f742">https://www.2021-c237f742</a>	200	Hotel Roy Route De L'Universite Ihada, Al	[{"lat": 36.74526, "lon": 5.041719}]									2				11		Excellent
12	<a href="https://www.2021-c8485de5">https://www.2021-c8485de5</a>	200	Hotel Zida Cite Mohamed Hamdi Laarouci, Tlemcen	[{"lat": 36.200512, "lon": 5.412854}]									2,5		["title": "A reasonable"]	19		Excellent	
13	<a href="https://www.2021-cd7e454">https://www.2021-cd7e454</a>	200	Hotel les 12 Road Khebdli Ali, Tlemcen	[{"lat": 34.879974, "lon": -1.302466}]									2		["title": "Do not try.", "text": "52"]			Excellent	
14	<a href="https://www.2021-cd7e454">https://www.2021-cd7e454</a>	200	Residence 9 Boulevard 7irou Youcef, Tlemcen	[{"lat": 36.816814, "lon": 5.777984}]									3,5		["title": "I like it", "text": "18"]			Excellent	

R	S	I	U	V	W	X	Y	Z
1	view_topview_courses_by_cate		ratio_of_ratings	order_online	book_a_table	details	certificate	outlet_rankings
2	Reasoanb15	{'location':	('Excellent': 2, 'Very Good': 0, 'Average': 3, 'Poor': 0, 'Terrible': 0)	FAUX	FAUX	{'about': None, 'property_amenities': '# # of 4 B&Bs / Inns in Anna		
3	AUVAIS_H3	{'location':	('Excellent': 1, 'Very Good': 0, 'Average': 1, 'Poor': 0, 'Terrible': 1)	FAUX	FAUX	{'about': None, 'null': None} ] # # of 10 hotels in Annaba]		
4	ontacto_115	{'location':	('Excellent': 5, 'Very Good': 2, 'Average': 3, 'Poor': 1, 'Terrible': 4)	FAUX	FAUX	{'about': None, 'null': None} ] # # of 9 B&Bs / Inns in Tem		
5	eated like_27	{'location':	('Excellent': 17, 'Very Good': 10, 'Average': 0, 'Poor': 0, 'Terrible': 1)}	FAUX	FAUX	{'about': None, 'good_to_know': ['fre # # of 7 B&Bs / Inns in Bejaï		
6		}		FAUX	FAUX	{'about': None, 'good_to_know': None, 'null': None} ]		
7		}		FAUX	FAUX	{'about': None, 'null': None} ]		
8	ery Good_C38	{'location':	('Excellent': 8, 'Very Good': 21, 'Average': 6, 'Poor': 1, 'Terrible': 2)	FAUX	FAUX	{'about': None, 'property_amenities': '# # of 7 B&Bs / Inns in Bejaï		
9	ala', 'text': 3	{'location':	('Excellent': 2, 'Very Good': 0, 'Average': 1, 'Poor': 0, 'Terrible': 0)	FAUX	FAUX	{'about': None, 'null': None} ] # # of 7 B&Bs / Inns in Bejaï		
10	uberge des_2	{'service':	('Excellent': 1, 'Very Good': 0, 'Average': 1, 'Poor': 0, 'Terrible': 0)	FAUX	FAUX	{'about': [Located in Toudja(Algeria)], '# # of 7 B&Bs / Inns in Bejaï		
11		}		FAUX	FAUX	{'about': None, 'property_amenities': 'free parking', 'free high spe		
12	11	{'location':	('Excellent': 0, 'Very Good': 1, 'Average': 3, 'Poor': 2, 'Terrible': 5)	FAUX	FAUX	{'about': None, 'property_amenities': '# # of 7 B&Bs / Inns in Bejaï		

Les colonnes **Location** ou **ratio\_of\_rating** sont respectivement les features de géolocalisation et classement . Comme on peut le remarqué, ces features ne sont pas utilisable pour de la modélisation.Pour la plupart des features, ces problèmes existent. Donc une autre étape était nécessaire : **feature engineering**

Pour contourner ce problème, nous avons procédé à d'autres manipulations.

## Features engineering

- Transformer les données catégorielles : One-hot encoding
- Nettoyage
- Transformer les horaires de travail

```
1 derive_weekly_work_hours({'Sun': ["11:00 - 15:00", "16:00 - 21:00"]})  
{'Sun': 9.0, 'Mon': 0, 'Tue': 0, 'Wed': 0, 'Thu': 0, 'Fri': 0, 'Sat': 0}
```

- Générer le ratio de notes

ratio_of_ratings	ratio_of_ratings_excellent	ratio_of_ratings_very_good	ratio_of_ratings_average	ratio_of_ratings_poor	ratio_of_ratings_terrible
('Excellent': 2, 'Very Good': 0, 'Average': 3,...)	0.400000	0.000000	0.600000	0.000000	0.000000
('Excellent': 1, 'Very Good': 0, 'Average': 1,...)	0.333333	0.000000	0.333333	0.000000	0.333333
('Excellent': 5, 'Very Good': 2, 'Average': 3,...)	0.333333	0.133333	0.200000	0.066667	0.266667
('Excellent': 17, 'Very Good': 10,...)					

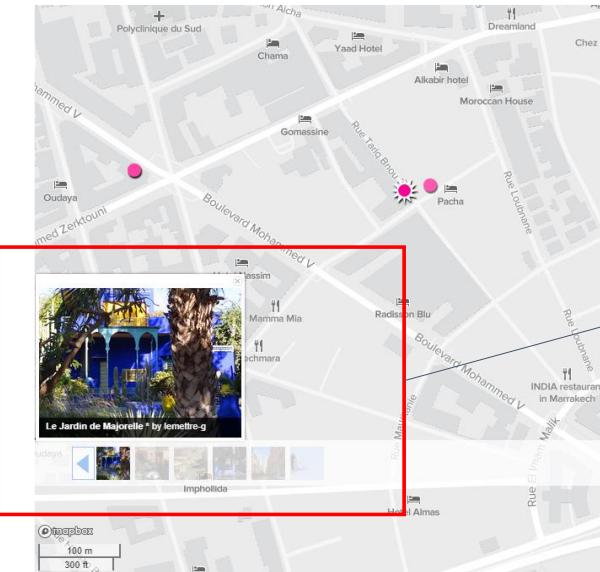
Après ces transformations, nous avons recueilli les données. Ci-dessous se trouvent l'ensemble des informations sur les données collectées.

**Tableau 5: Couverture des données de TripAdvisor**

	Algeria			Morocco		
	Restaurant	Hôtel	Base de données finale	Restaurant	Hôtel	Base de données finale
<b>Lignes</b>	821	416	1237	5553	3829	6790
<b>Features</b>	220	269	461	281	373	544

### 1.1.6. Les données de Flickr

Adaptée au tourisme, Flickr est une plateforme de partage de photos. Des avis, des données de géolocalisation y sont disponibles. Ces données permettent également d'analyser le comportement spatial et les schémas de déplacement des touristes. Nous pourrions comme le montre la figure retrouver les sites touristiques. Lorsque nous calculons la densité des points d'intérêts de Flickr, nous avons un proxy du tourisme dans la localité.

**Figure 28: Données de Flickr**

Nous pouvons obtenir le nombre de photos à proximité d'une sortie, (par exemple 250m).

## II. Infrastructures du projet

Dans cette partie, nous présenterons l'architecture qui tient nos modèles. Ils ont été hébergés dans le cloud

### II.1. Cloud et architectures

Le cloud computing est la fourniture de services informatiques (notamment des serveurs, du stockage, des bases de données, la gestion réseau, des logiciels, des outils d'analyse, l'intelligence artificielle) via Internet (le cloud) dans le but d'offrir une innovation plus rapide, des ressources flexibles et des économies d'échelle. En règle générale, vous payez uniquement les services cloud que vous utilisez (réduisant ainsi vos coûts d'exploitation), gérez votre infrastructure plus efficacement et adaptez l'échelle des services en fonction des besoins de votre entreprise »<sup>31</sup>. Pour notre cas,

---

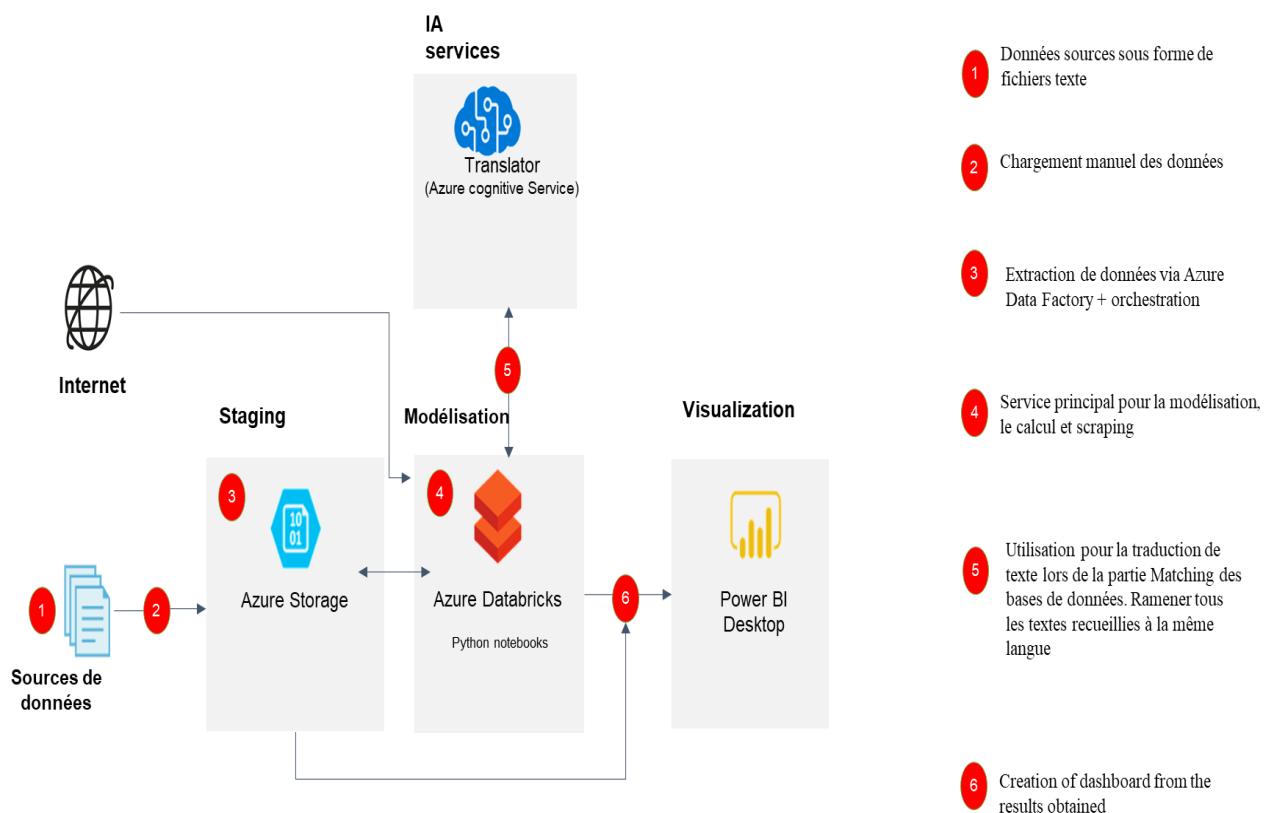
31 <https://azure.microsoft.com/fr-fr/overview/what-is-cloud-computing/#uses>, consulté le 11 juin 2021

l'utilisation du cloud computing consiste aux services de stockage de données, de sauvegarde et de récupération des données mais à faciliter également nos exercices de modélisation

Non seulement les coûts sont réduits car le cloud computing élimine la nécessité d'investir dans du matériel et des logiciels, et de configurer et de gérer des centres de données mais également la productivité des utilisateurs du cloud augmente. D'énormes ressources de calcul peuvent donc être mises en œuvre en quelques minutes et en quelques clics, offrant ainsi aux entreprises un haut niveau de flexibilité et les dégageant de la pression liée à la planification de la capacité. Quant à la sécurité, de nombreux fournisseurs de cloud offrent un vaste éventail de stratégies, technologies et contrôles qui renforcent globalement votre situation de sécurité, contribuant ainsi à protéger les données, les applications contre les menaces et attaques extérieures.

Pour mener à bien notre projet, nous utiliserons l'architecture présentée dans le graphique ci-dessous. Pour chaque étape, des services du cloud adaptés seront exécutés.

**Figure 29: Architecture du projet**



Cette figure montre que nos données doivent être stockés via Azure Storage.

## ***II.2 Azure Blob Storage***

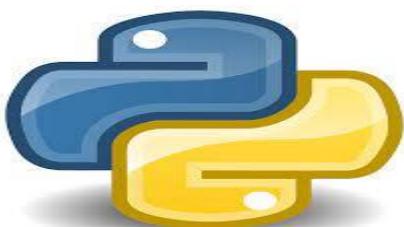
Azure Blob Storage est un service permettant de stocker de grandes quantités de données d'objet non structurées, telles que du texte ou des données binaires. Vous pouvez utiliser le stockage d'objets blob pour exposer des données publiquement au monde entier ou pour stocker des données d'application en privé. Les utilisations courantes du stockage d'objets blob sont les suivantes :

- Diffusion d'images ou de documents directement dans un navigateur
  - Stockage de fichiers pour un accès distribué
  - Streaming vidéo et audio
  - Stockage des données pour la sauvegarde et la restauration, la reprise après sinistre et l'archivage
  - Stockage de données à des fins d'analyse par un service local ou hébergé par Azure
- La dernière est celle qui nous intéresse. Toutes nos données internes et externes seront y stockés.

## ***II.3. Python***

Python est un langage de programmation, dont la première version est sortie en 1991. Créé par Guido van Rossum, il a voyagé du Macintosh de son créateur, qui travaillait à cette époque au Centrum voor Wiskunde en Informatica aux Pays-Bas, jusqu'à se voir associer une organisation à but non lucratif particulièrement dévouée, la Python Software Foundation, créée en 2001.

**Figure 30: Logo Python**



Python un langage puissant, à la fois facile à apprendre et riche en possibilités. Il est, en outre, très facile d'étendre les fonctionnalités existantes. Ainsi, il existe ce qu'on appelle des bibliothèques qui aident le développeur à travailler sur des projets particuliers. Plusieurs bibliothèques peuvent ainsi être installées pour, par exemple, développer l'analyse spatiale avec géopandas.

## ***II .4. Databricks, Apache Spark et Azure Databricks***

Databricks a été conçu par les créateurs d'Apache Spark, Delta Lake et MLflow. Plus de 2 000 entreprises internationales y compris le système Coca Cola utilisent la plateforme Databricks sur un cycle de vie de Big Data et de Machine Learning. Il s'agit de la version non-opens source d'Apache Spark. Mais qu'est-ce qu'Apache Spark ? Il s'agit d'un moteur de traitement unifié capable d'analyser du Big Data en utilisant SQL, le Machine Learning, le traitement des graphes ou l'analyse de flux en temps réel. Plus rapide et plus générale, Spark permet d'exécuter des programmes jusqu'à 100 fois plus vite en mémoire ou 10 fois plus vite sur disque que Hadoop.



- Le moteur Spark est central.
- L'API<sup>32</sup> DataFrame fournit une abstraction au-dessus des RDD<sup>33</sup> tout en améliorant entre 5 et 20 fois les performances par rapport aux RDD traditionnels avec Catalyst Optimizer.

---

<sup>32</sup> API : Application programming interface : interface permettant d'accéder à des services tiers (ici Spark)

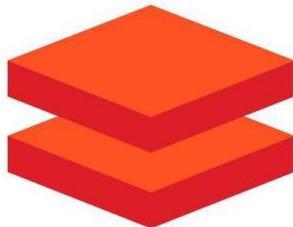
<sup>33</sup> Un RDD (Resilient Distributed Dataset) : C'est un jeu de données qui se parcourt comme une collection. Il est distribué, car il sera vraisemblablement partitionné (découpé en partitions), et chacune des partitions traitées sur un nœud du cluster. Il est résilient, car il sera peut-être partiellement relu en cas de problème (perte d'un noeud par exemple)

- Spark ML fournit des algorithmes Machine Learning de haute qualité et ajustés pour le traitement du Big Data.
- L'API de traitement Graph nous offre une API facile d'accès pour modéliser les relations de paires entre des personnes, des objets ou des nœuds dans un réseau.
- Les API Streaming nous donnent une tolérance de panne de bout en bout, avec une sémantique « exactement une fois », ainsi que la possibilité d'une latence inférieure à la milliseconde.

Tout fonctionne conjointement sans interruption.

Azure Databricks a été proposé par Microsoft.

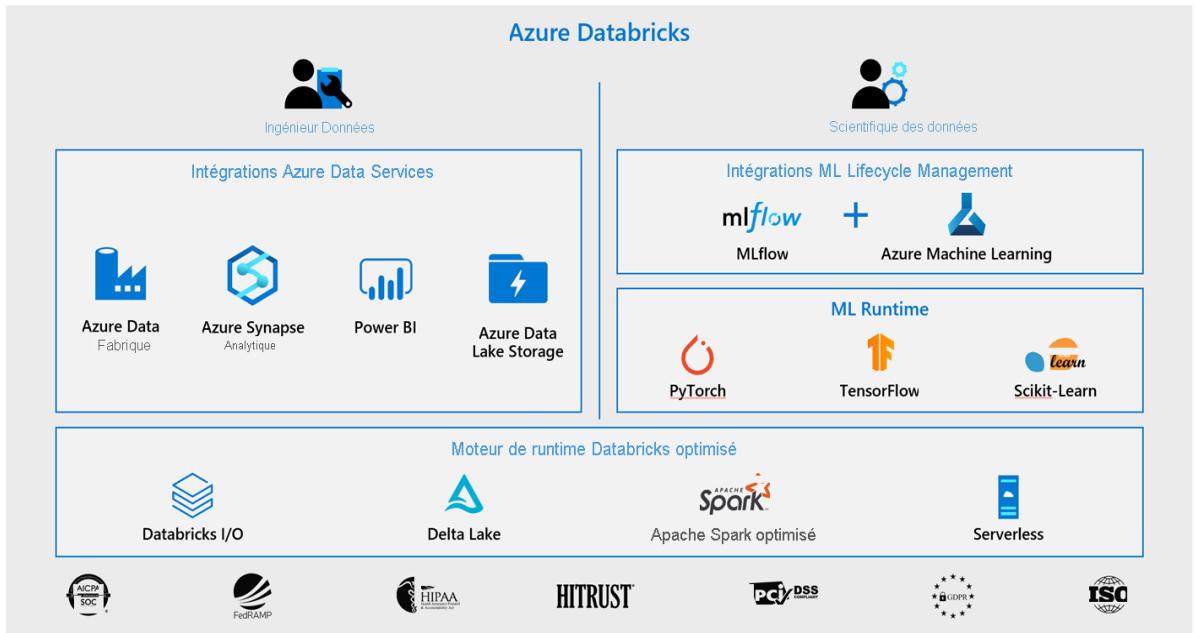
**Figure 31: Logo Databricks**



Il permet le travail collaboratif ainsi que le travail dans plusieurs langages comme Python, Spark, R et SQL. Travailler sur Azure Databricks offre les avantages du cloud computing : traitement et stockage des données évolutifs, à moindre coût et à la demande.

En combinant la puissance de Databricks, une plateforme Apache Spark de bout en bout, managée et optimisée pour le cloud, à la sécurité et à la dimension « entreprise » de la plateforme Azure de Microsoft, Azure Databricks simplifie l'exécution des charges de travail Spark à grande échelle.

**Figure 32: Architecture d'Azure Databricks**



En tant que moteur de calcul, Azure Databricks se trouve au centre de la plateforme logicielle et permet de lier les autres services d'Azure comme Azure Storage Blob, Azure Translator.

En plus de pouvoir s'exécuter dans de nombreux environnements, Apache Spark rend la plateforme encore plus accessible en prenant en charge plusieurs langages :

- Scala : langage principal d'Apache Spark
- Python : plus communément appelé PySpark
- R: SparkR (R sur Spark)
- Java

Comme mentionné aux débuts de ce livre, les notebook d'Azure Databricks seront alors exécutés en Python ou en PySpark lorsque nous solliciterons des calculs qui nécessiteront plus de ressources.

## II.5. Azure Translator

Translator est un service de traduction automatique basé sur le cloud et fait partie de la famille d'API cognitives Azure Cognitive Services utilisées pour créer des applications intelligentes. Translator est facile à intégrer dans vos applications, sites

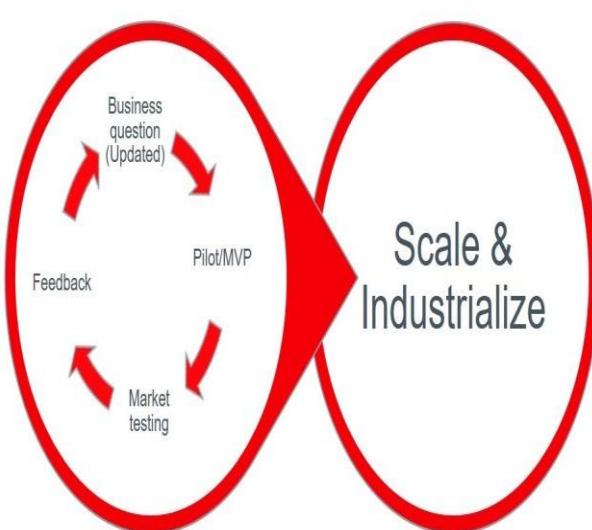
web, outils et solutions. Il vous permet d'ajouter des expériences utilisateur multilingues dans 90 langues et dialectes et peut être utilisé pour la traduction de texte avec n'importe quel système d'exploitation.

La mise en place de ces infrastructures est la partie cachée de l'iceberg, lorsqu'on assimile notre projet à un iceberg. C'est la partie non visible du projet mais indispensable. Sans ces ressources, développer nos modèles sur nos propres machines en locales seraient très compliquées.

### III. Schéma de la modélisation du projet

La segmentation dynamique vue par le système Coca Cola est une discipline relativement jeune qui utilise une MVP approche.

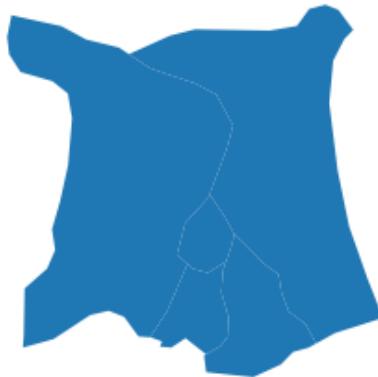
**Figure 33: Schéma de segmentation par pilote**



Ce projet est assimilable à la première journée d'un voyage qui durera plusieurs jours. Vu les territoires d'ECCBC, le choix de la zone pilote a été fait en amont. Les territoires concernés par ces projets sont Maroc et l'Algérie. Ainsi, les choix de pilotes ont été réalisés en prenant en compte leur représentativité.

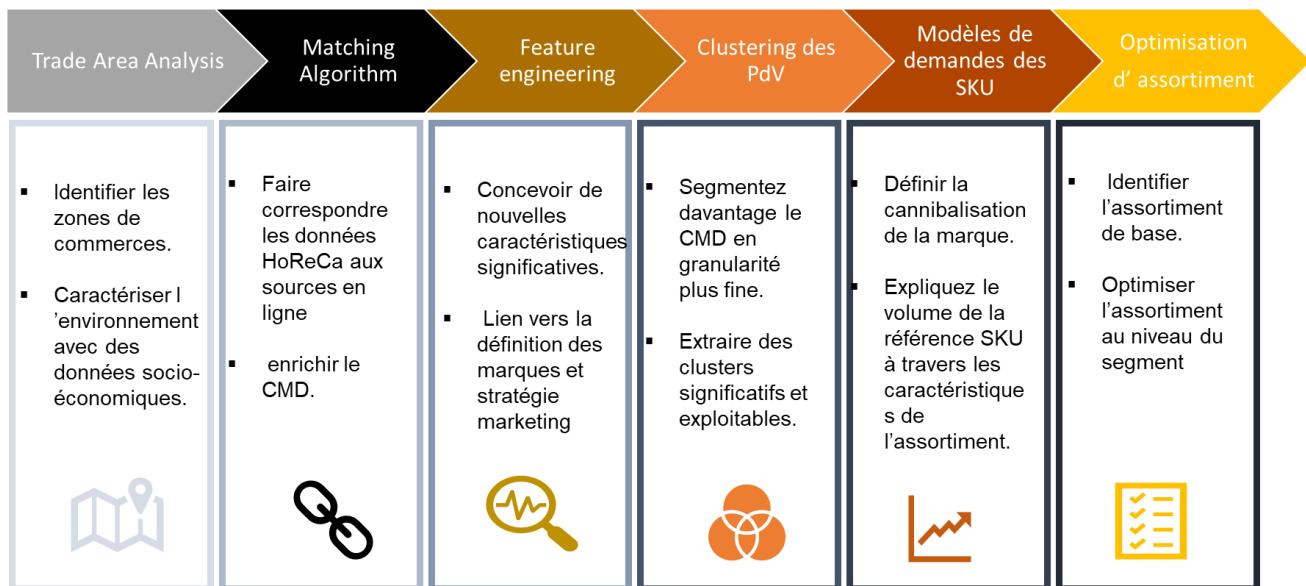
La ville de Marrakech et la Wilaya de Tizi Ouzou ont été respectivement désigné comme pilotes pour le Maroc et l'Algérie. Tout au long de la modélisation, nous nous intéresseront au cas de Marrakech. Pour s'en convaincre, nous avons comparé les proportions des AG/DL, SNACK et HORECA dans tout le territoire d'ECCBC Maroc par rapport à la ville de Marrakech. Il en est de même pour les moyennes de volumes de boissons vendues. Les chiffres ne sont pas présentés afin de rester confidentiel.

**Figure 34: Carte de la ville de Marrakech**



Pour structurer cette partie modélisation, les étapes ont été structurées comme le montre le schéma évolutif.

**Figure 35: Etapes de modélisation du projet**



## Conclusion

Tout au long de ce chapitre nous avons étayé les données qui rentrent dans le cadre de ce projet en un premier temps. Dans un second temps, nous avons donné un aperçu de l'infrastructure utilisée pour réaliser le projet. Nous espérons également que le schéma de modélisation vous a permis de mieux comprendre la destination de notre voyage de modélisation. Le chapitre suivant débute avec la première étape du modèle.

## Chapitre 2 :

---

# Trade area analysis et Matching

Trade area analysis et Matching .....	- 116 -
Introduction .....	- 117 -
I. Trade Area Analysis.....	- 117 -
I.1. Préparation des composants de l'analyse de spatiale .....	- 118 -
I.2. Détermination de la zone de chalandise.....	- 119 -
I.1.1    Calcul des scores d'urbanicité .....	- 119 -
I.1.2.    Feature engineering du score de densité : urbanicité.....	- 120 -
I.1.3    Calcul de distance des zones de chalandise .....	- 120 -
I.1.4.    Profil de la zone chalandise .....	- 122 -
II.    Matching.....	- 123 -
II.1.    Principe du matching.....	- 123 -
II.2.    Construction de l'échantillon de matching manuel .....	- 123 -
II.3.    Similarité des noms et des adresses.....	- 124 -
II.4.    Similarité des distances .....	- 126 -
II.5.    Seuil optimal de similarité et courbe ROC .....	- 126 -
II.5.    Algorithme de matching.....	- 127 -
III. En route vers le clustering .....	- 129 -
Conclusion.....	- 129 -

## Introduction

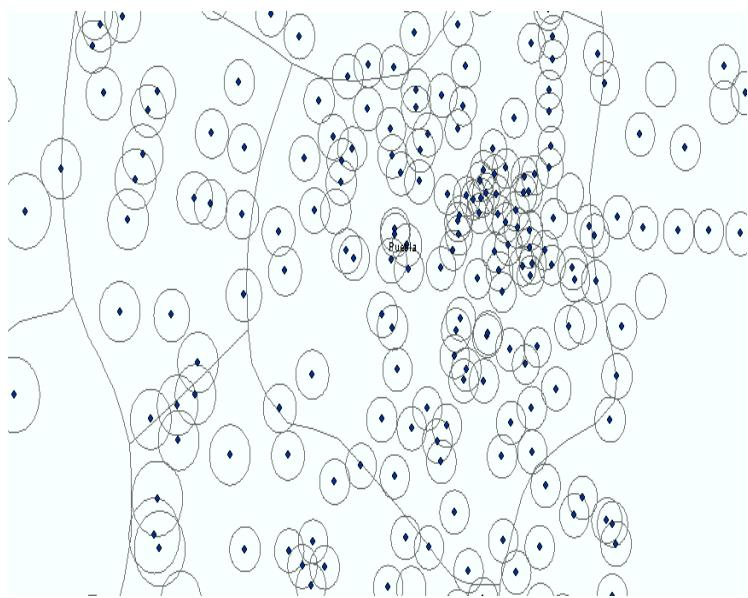
Dans ce chapitre nous nous intéressons à la réalisation de deux premiers points du schéma évolutifs à savoir le Trade area analysis et le Matching. Tous ces points ont tous un seul but : enrichir la base de données.

### I. Trade Area Analysis

Ce type d'analyse, lorsqu'elle est précise nous permettra de découvrir le bon profil démographique et socio-économique des environs du point de vente.

Supposons que le graphique ci-dessous comme la distribution spatiale des points de ventes, les cercles délimitant leurs zones de chalandises respectives.

**Figure 36: Distribution de PdV**



**Business inputs :** Les points de vente situés dans des zones plus denses ont une zone de chalandise plus petite en raison d'une concurrence plus sévère.

Ce business input à droite de la figure nous servira de point de repère lorsque nous modéliserons cette partie. Nos résultats devraient être en concordance avec ladite phrase. Nous définissons le score de densité comme la mesure de l'urbanité ou de la ruralité. Logiquement, il est donc proportionnel aux rayons des zones de chalandise.

Le but de ce paragraphe est alors d'expliquer comment modéliser cette zone de chalandise introduite dans la partie théorique, chapitre 3.

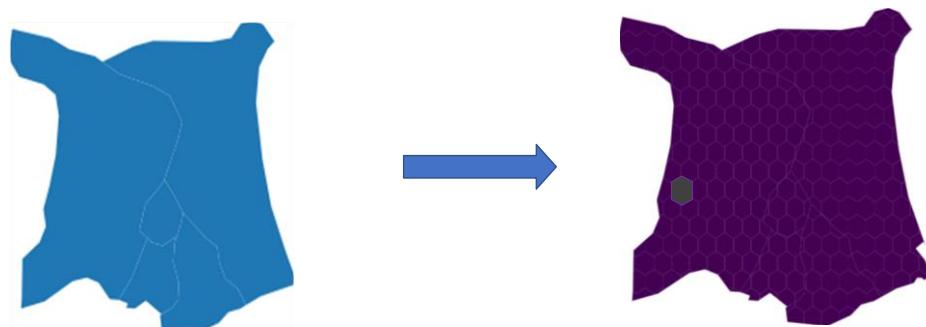
### **I.1. Préparation des composants de l'analyse de spatiale**

Les étapes sont deux.

- **Préparation de la grille hexagonale sur le territoire**

- Former des hexagones de diamètre 1250 m à travers la Région de Marrakech. La grille hexagonale a été choisie pour mener le projet pour les mêmes raisons évoquées dans la partie théorique ;
- Superposition de la grille construite sur la ville de Marrakech ;

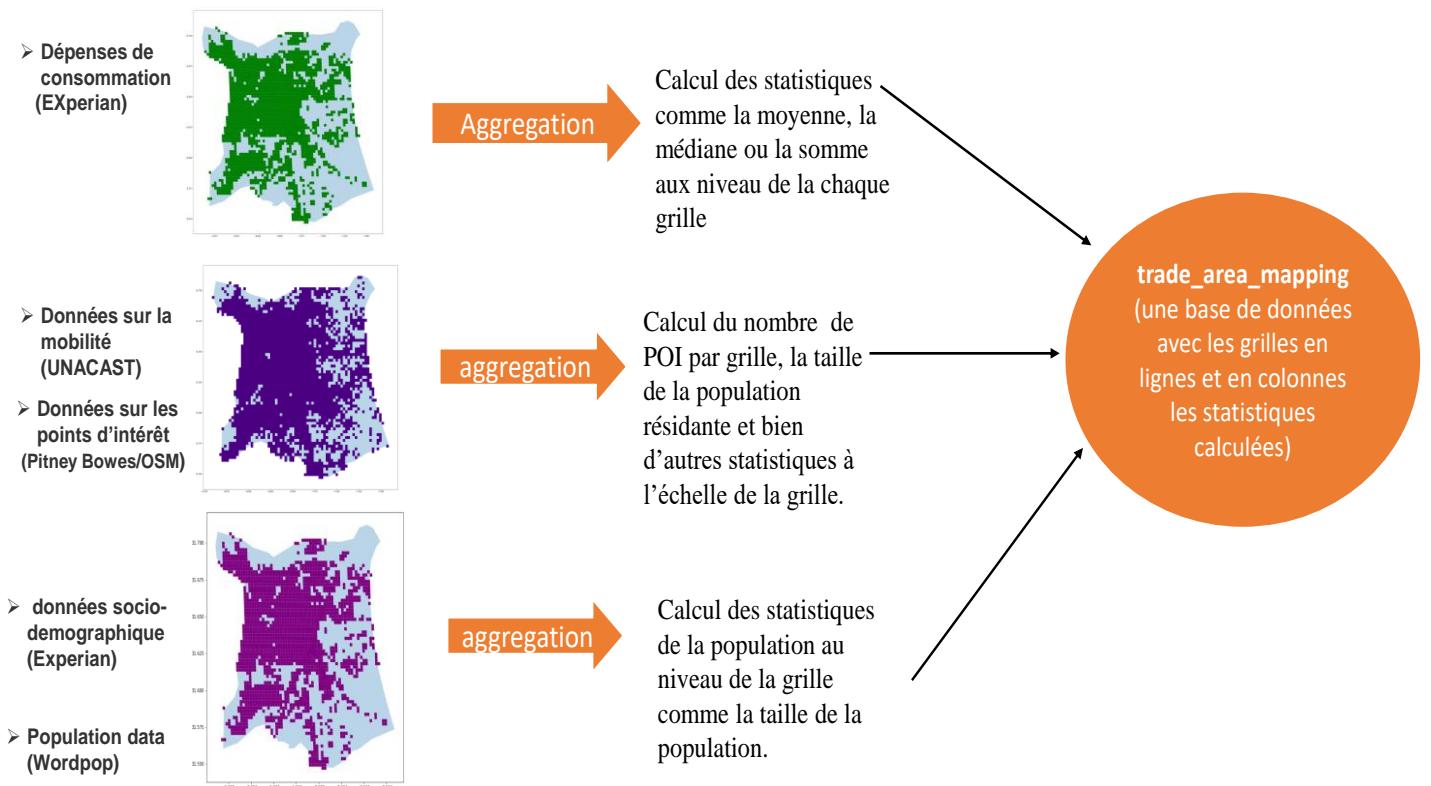
**Figure 37: Hexagonalisation du territoire de Marrakech**



- **Préparation des agrégations hexagonales sur les données externes**

Dans l'ensemble de nos données externes, nous disposons des géolocalisations de l'individu statistique en question. Par exemple, Les données de population de WorldPop ont comme géolocalisation les coordonnées de la grille carré 100m\*100m. Les autres sources de données ont des grilles qui leur sont spécifiques.

Tout d'abord nous identifions les individus statistiques par les géolocalisations. Ensuite, nous réalisons des jointures spatiales afin de se ramener à notre découpage. A ce niveau, les individus statistiques sont identifiés par la géolocalisation de la grille hexagonale à laquelle elle appartient. Des résumés statistiques des données comme la moyenne, la médiane ou l'effectif sont alors disponibles. Enfin, nous mettons en œuvre une jointure par la clé primaire des grilles de toutes les bases de données. On retrouve ainsi une base de données finale avec les hexagones(territoires) comme lignes et les résumés statistiques comme colonnes. Nous pouvons schématiser cette étape à travers ce graphique ci-après.

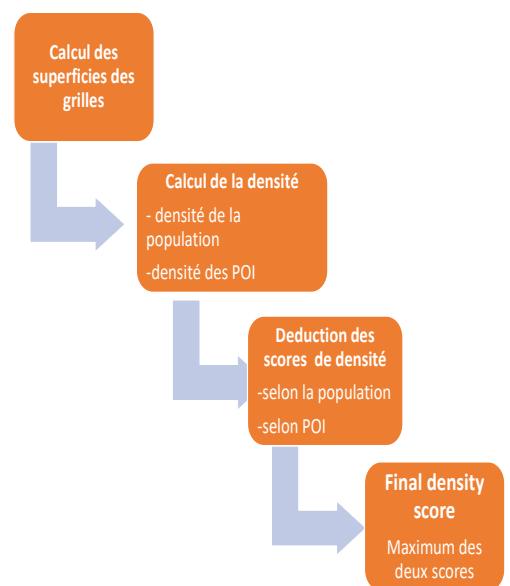


## I.2. Détermination de la zone de chalandise

La détermination de la zone de chalandise s'est faite successivement par un calcul de score d'urbanicité, un features engineering, la détermination du point de partage et des opérations de jointures de tables.

### I.1.1 Calcul des scores d'urbanicité

Pour mesurer la ruralité, nous avons deux manières de la définir. La première est la plus classique. Nous allons raisonner dans ce cas précis par agglomération de la population. La deuxième manière est de raisonner par agglomération des POI. Les centres urbains ont naturellement plus de POI qu'un territoire rural. La table de données construite nous permet de calculer la superficie, compter la population et les POI par grille.



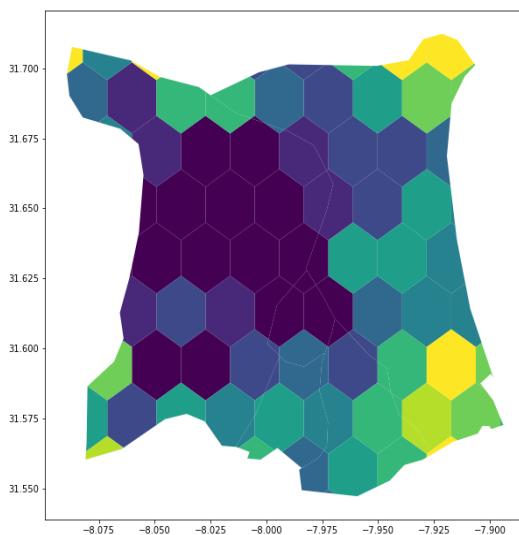
$$\text{Score de densité} = \frac{\text{densité} - \min(\text{densité})}{\max(\text{densité}) - \min(\text{densité})}$$

Le score final retenu est le maximum des deux scores

### *1.1.2. Feature engineering du score de densité : urbanicité*

Nous avons essayé de regrouper ce score de densité en des groupes de telles sortes à minimiser la variance intragroupe des scores. En utilisant une technique de binning, le regroupement optimal trouvé est la formation de 10 classes d'urbanicité. Chaque groupe, la somme de la population est d'environ 10% de la population totale. Cette approche nous permet ainsi de dériver sur une variable importante de ce projet. Ruralité étant juste l'opposé du score urbanicité.

**Figure 38: Ruralité de la Ville de Marrakech**



La figure illustre la ville de Marrakech selon notre variable urbanicité. La base de données obtenue à cette étape est utilisée comme input de l'étape de la détermination de la zone de chalandise. Notons la base de données *urba\_set*.

### *1.1.3. Calcul de distance des zones de chalandise*

Nous nous referons ici à La loi de Reilly évoqué dans la partie théorique.

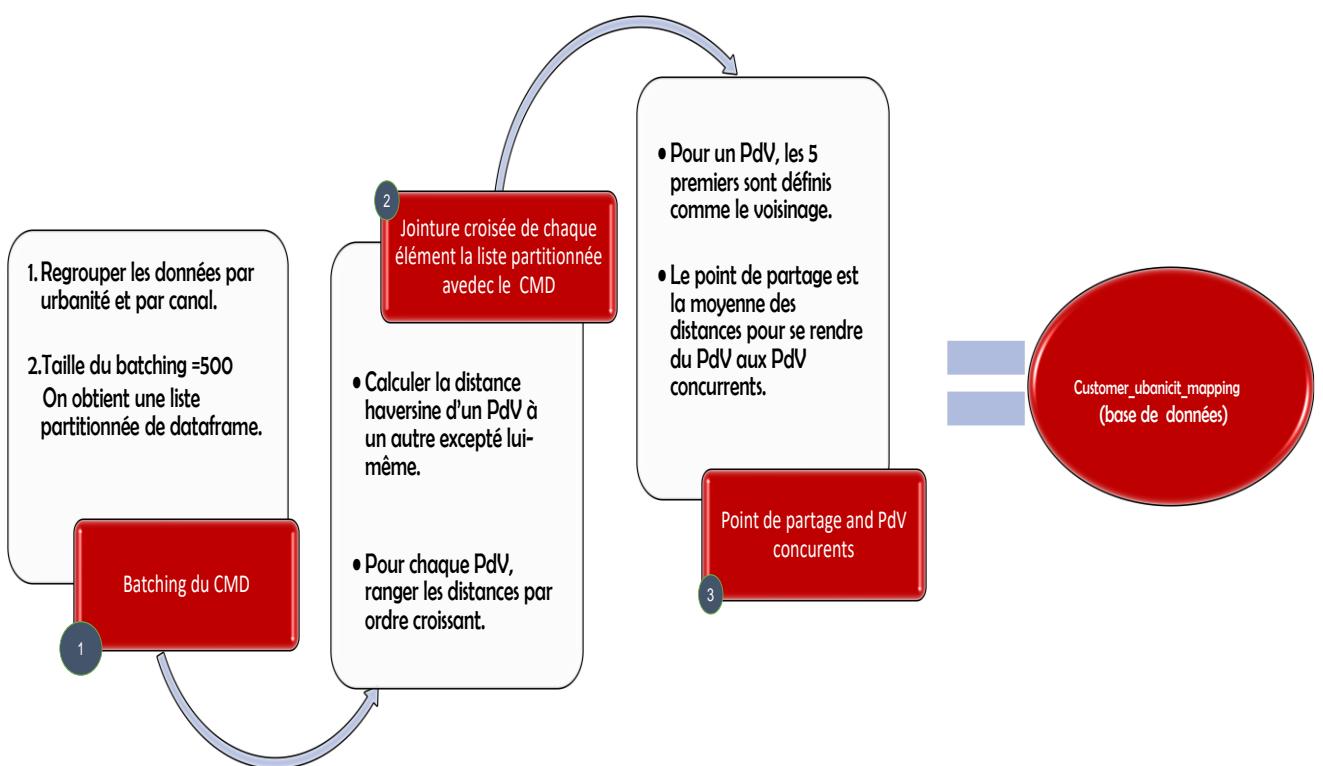
L'idée revient à trouver le point de partage entre les zones de chalandises des PdV environnants. Une fois ce point trouvé, il suffit de calculer la distance du PdV

au point de partage. Cette distance est le rayon de la zone de chalandise si nous supposons des zones de chalandises en forme de cercle.

Nous avons fait l'hypothèse qu'au sein de chaque canal (HORECA, Snack, grosseries), les points de ventes ont les mêmes superficies. Cette hypothèse même simplificatrice n'est pas irréelle. Également, pour chaque PdV d'un canal donné et d'une fixée donné nous avons choisi comme points de ventes concurrents, les cinq plus proches PdV du même canal et urbanité

### Principe de l'algorithme utilisé pour le calcul du point de séparation

**Inputs:** CMD et urba\_set



nous nous éloignons des zones urbaines, les données confirment un agrandissement des zones des chalandises.

**Tableau 6: Vérifications des rayons de zones de chalandise**

CHANNEL_CUSTOM	urbanicity	trade_area_radius_km	nr_outlets
AG	1	0.070735	1629
AG	2	0.124161	1891
AG	3	0.119195	1994
AG	4	0.412396	864
AG	5	0.337388	1204
AG	6	0.296497	2337
AG	7	0.501805	1593
AG	8	0.553787	2043
AG	9	0.749432	1817
AG	10	0.830071	3699
DL	1	0.316118	165
DL	2	1.838201	162
DL	3	0.735177	234
DL	4	2.684028	79
DL	5	2.193807	134
DL	6	1.138967	386
...	...	...	...

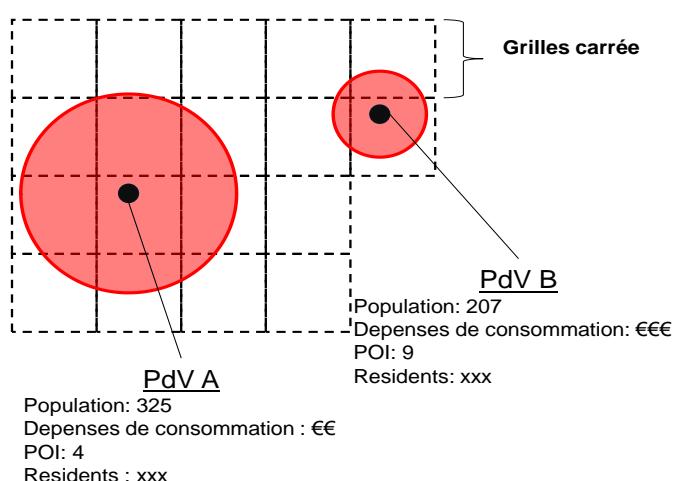
La colonne nr\_outlets désigne le nombre de PdV dans le canal (CHANNEL\_CUSTOM) et urbanité (urbanicity) considéré.

#### 1.1.4. Profil de la zone chalandise

Des opérations de jointure des bases de données obtenues dans les deux dernières étapes et du CMD nous permet de mettre fin à cette analyse de zone de chalandise :

- Jointure à partir des clés primaires des grilles (urba\_set et trade\_area\_mapping) ;
- Jointure spatiale des clients avec le résultat obtenu précédemment.

Cette analyse nous a permis d'obtenir les informations mentionnées comme ci-dessous :



## II. Matching

Pour les données concernant l'environnement économique, la présence de grille dans les données de départ à faciliter les jointures spatiales. Les données de TripAdvisor n'ont pas ces grilles. D'ailleurs, les informations qu'elles renferment sont propres aux PdV. Il nous faut alors trouver un moyen pour concilier notre CMD et les données de TripAdvisor. C'est dans que s'inscrit l'objet de ce grand paragraphe.

### *II.1. Principe du Matching*

Il s'agit d'enrichir les données de base des clients (CMD) en recherchant pour chaque PdV appartenant au canal HORECA, le PdV correspondante dans TripAdvisor. Chaque PdV HORECA du CMD est comparé à tous les PdV de TripAdvisor dans un rayon de 2 km. La comparaison se fait sur la base de 3 paramètres :

- Nom
- Adresse
- Distance (basée sur la géolocalisation - coordonnées GPS)

Ces 3 paramètres n'ont pas la même importance. L'importance relative utilisée est : 75% Nom, 15% Adresse, 10% Distance.

Sur cette base, un score global de similarité est calculé, et les entrées TripAdvisor sont classées en fonction de ce score. Le PdV présentant la plus grande similarité est le candidat potentiel.

Dans certains cas, même le candidat potentiel n'est pas un bon candidat, c'est pourquoi nous avons défini un seuil de similarité au-delà duquel le candidat est considéré comme correspondant.

### *II.2. Construction de l'échantillon de Matching manuel*

Cet exercice revient tout d'abord à transformer le problème en une forme d'apprentissage supervisé. Pour ce faire, un échantillon représentatif de taille 400 (échantillonnage stratifié selon les médailles (or, argent, bronze ...)) des hôtels,

restaurants café de notre CMD a été choisi afin de respecter les mêmes proportions dans la population des HORECA. Nous ne pouvons pas fournir les proportions dans ce rapport pour des raisons de confidentialités.

Cet échantillon de 400 individus a alors été utilisé pour apprécier les données de TripAdvisor (5457 individus) sur la base des similarités. Lorsque nous calculons notre algorithme trouve le point de vente le plus similaire, nous vérifions manuellement cette correspondance. Une variable est créée ainsi à cette occasion. Si la correspondance est bonne, nous assignons 1 sinon 0. Cette vérification manuelle est faite sur les 400 HoReCa choisis. Mais comment avons calculé les similarités ?

### *II.3. Similarité des noms et des adresses*

Il s'agit de contenu textuel, nous allons donc utiliser la théorie présentée dans le chapitre 2, éléments d'analyse textuelle. L'idée est de créer des fonctions qui permettent de mesurer la similarité entre deux chaînes de caractère et de les vectoriser sur l'ensemble des lignes d'une base de données donnée.

Un algorithme de prétraitement a d'abord été réalisé suivant les mêmes étapes de nltk énoncées dans la partie théorique. Les textes traités sont alors utilisés pour dériver la similarité.

Dans la page ci suivante se trouve l'algorithme que nous avons utilisé.

**fonction compound\_similarity(s1: str, s2: str) -> float:**

"""\p>Cette fonction calcule le score composé mesurant la similarité

entre deux chaînes. Le score est basé sur les 7 mesures suivantes :

- **Damerau-Levenshtein** - modifier la distance qui prend également en compte les transpositions.

- **Jaro-Winkler** - similarité basée sur des lettres communes ajustées pour tenir compte de la probabilité plus élevée que deux chaines soit les mêmes si leurs débuts se correspondent

- **n-gram** - Cette similarité est basée sur les nombres de n-grammes (séquence de sous-chaînes de longueur  $n$ ) qui correspondent. Il a été empiriquement sélectionné que la longueur des n-grammes dans ce cas est défini sur  $N=2$ .

- **Jaccard** - comme n-grammes sans tenir compte de la cardinalité (longueur) de la n-grammes. En effet, cela donne un score de similarité de n-gramme pour  $N=1$ .

- **Sorensen-Dice** - Logique similaire à jaccard mais avec de légers ajustements.

- **Overlap** - mesure le « chevauchement » entre deux chaînes en fonction du nombre de caractères en eux.

- **Ratcliff-Obershelp** - prend en compte la longueur des sous-chaînes entièrement correspondantes mais aussi le nombre de caractères correspondants à partir de sous-chaînes qui ne correspondent pas complètement.

**Inputs:**

s1 {str} -- La première chaîne.

s2 {str} -- La deuxième chaîne.

**Output:**

float -- Moyenne des scores de similarité provenant des 7 algorithmes. 0 signifie non similaire du tout et 1 signifie que les deux cordes correspondent parfaitement. Si l'une des deux chaînes est vide, la similarité sera traitée comme 0.

"""

***Si s1 est vide:***

s1 = ""

***Si s2 est vide:***

s2 = ""

***Si s1 == "" et s2 == "":***

**return 0.**

scores = [ damerau\_levenshtein.normalized\_similarity(s1, s2),

jaro\_winkler.normalized\_similarity(s1, s2), #toutes les scores sont récupérés dans une liste

sorensen\_dice.normalized\_similarity(s1, s2),

jaccard.normalized\_similarity(s1, s2),

overlap.normalized\_similarity(s1, s2),

ratcliff\_oberhelp.normalized\_similarity(s1, s2),

NGram.compare(s1, s2, N=2) ]

**return moyenne(scores)**

#la fonction retourne la moyenne de ses scores de la liste

Lorsque nous vectorisons cette fonction pour calculer par exemple la similarité de la chaîne *café rosalia* avec des chaînes d'une table de données, on trouve

**Tableau 7 : Exemple de similarité de texte**

s1	s2	Similarité des noms
café rosalia	touhfa	0.277437641723356
café rosalia	société générale	0.45254256673429605
café rosalia	café rosalia	1
café rosalia	carretera farhana - b° constitución	0.3545605930972844
café rosalia	farès	0.38447891977303744

L'algorithme marche bien sur les noms. De la même façon on peut calculer la similarité des adresses.

## II.4. Similarité des distances

Soit un PdV du CMD et un autre PdV sur TripAdvisor dont nous disposons les géolocalisations.

$$\text{Similarité de distance entre les deux PdV} = \frac{\text{distance entre les PdV} - \min(\text{distance})}{\max(\text{distance}) - \min(\text{distance})}$$

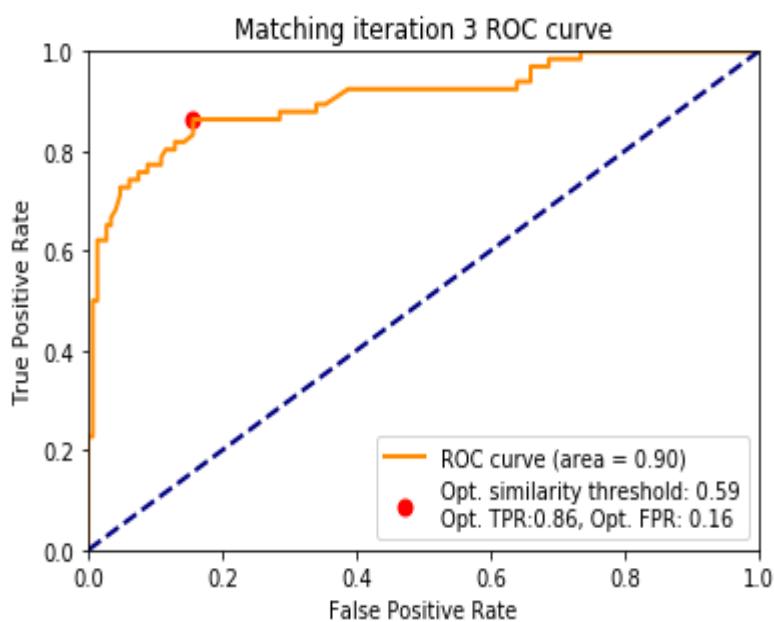
Avec  $\min(\text{distance})$  et  $\max(\text{distance})$  respectivement la distance minimale et la distance maximale entre le PdV du CMD et toutes les autres PdV de TripAdvisor.

## II.5. Seuil optimal de similarité et courbe ROC

L'apprentissage est réalisé sur l'échantillon de Matching manuel.

La similarité optimale est sélectionnée de manière à maximiser le taux vrai positif (TPR / sensibilité) et le vrai taux négatif (FPR / spécificité)

**Figure 39: Courbe ROC et sélection du seuil optimal**



Au regard de ce graphique, nous pouvons dire que vraisemblablement, lorsque la similarité globale (moyenne pondérée des mesures de similarités des noms, des adresses et des distances) d'un PdV du CMD et un PdV De TripAdvisor dépassera 0.59 nous pouvons considérer que l'algorithme trouvé sa bonne correspondance sur TripAdvisor.

LA performance du modèle est assez importante car l'aire sous la courbe ROC dépasse 0,9. En plus, malgré que notre échantillon ait deux classes déséquilibrées, nous arrivons à avoir jusqu'à un taux de vrai positif de l'ordre 0.86.30 de l'échantillon a eu une correspondance sur TripAdvisor.

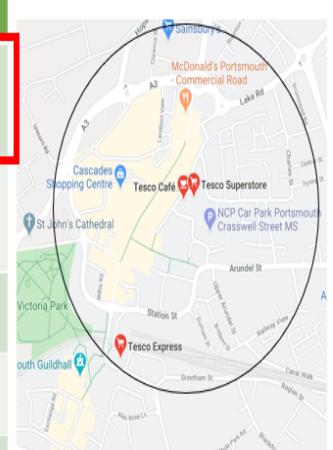
Les performances de cet apprentissage supervisé, nous permet alors d'utiliser notre algorithme pour trouver les bonnes correspondances des PdV sur TripAdvisor.

### ***II.5. Algorithme de Matching***

Une fois que nous ayons trouvé les bons appariements, la jointure des deux tables devient triviale. Comme résultat de cet algorithme, La ligne en rouge est choisie par l'algorithme car le score de similarité dépasse 0,59, seuil trouvé à partir la courbe ROC.

**Tableau 8: Exemple de l'algorithme de Matching**

CMD_ID	CMD_Name	CMD_Address	EXT_Name	EXT_Address	Name_Score	Address_score	Score
1	tesco	fratton road	tesco superstore	fratton road	0.75	1	0.85
1	tesco	fratton road	tesco express	pompey place	0.8	0.2	0.56
1	tesco	fratton road	sainsburys	cook road	0.23	0.22	0.226
1	tesco	fratton road	mcdonalds	fleet street	0.25	0.25	0.25
1	tesco	fratton road	cascades	bond street	0.27	0.27	0.27



A la page suivante se trouve comment les inputs de l'algorithme et son fonctionnement. Le code entier de ce programme est disponible en annexe.

Cette fonction effectue une correspondance basée sur la similarité de nom et d'adresse avec un filtrage de géolocalisation facultatif (si les latitudes et longitudes sont disponibles) entre deux tables de données. La table de données gauche (première) est traitée comme la table primaire même si les résultats produits sont dans le sens où chaque élément de la table de données droite (deuxième) source est apparié à juste un élément de la table de gauche. Dans le cas où un élément de la table de données droite est la meilleure correspondance pour plusieurs éléments de la table de données, l'égalité est rompue en choisissant l'élément de gauche qui a le score de similarité le plus élevé. La boucle correspondante continue jusqu'à ce qu'il n'y ait plus d'autre des correspondances vraisemblablement correctes pourraient être faits.

Fonction *match\_join ()*,

#### inputs :

**l\_sdf** Première table de données à mettre en correspondance.

**l\_id** : Nom de la colonne de clé primaire de la première source.

**l\_lon** : Nom de la colonne de longitude de la première source.

**l\_lat** : Nom de la colonne de latitude de la première source.

**l\_name** : Nom de la colonne nom de l'élément de la première source.

**l\_addr** : Nom de la colonne d'adresse de l'élément de la première source.

**r\_sdf** : La deuxième source à appairier.

**r\_id** : Nom de la colonne de clé primaire de la deuxième source.

**r\_lon** : Nom de la colonne de longitude de la deuxième source.

**r\_lat** : Nom de la colonne de latitude de la deuxième source.

**r\_name** : Nom de la colonne nom de l'élément de la deuxième source.

**r\_addr** : Nom de la colonne d'adresse d'élément de la deuxième source.

**distance\_threshold\_m**: Si la géolocalisation est disponible, le rayon en mètres dans lequel les éléments de la deuxième source sera mis en correspondance avec les éléments de la première source. La valeur par défaut est 0.

**similarity\_threshold** (: Le seuil de similarité au-delà duquel une paire appariée est acceptée et

en dessous de laquelle la correspondance est rejetée. La valeur par défaut est 0,59.

**name\_similarity\_weight** : Poids accordé à la similarité du nom. La valeur par défaut est 0,75.

**address\_similarity\_weight** : Le poids accordé à la similarité d'adresse. Defaults to 0 ,15.

**dist\_similarity\_weight** : Le poids donné à la similarité de distance. Defaults to 0,1.

**no\_geolocation** : Indique si le filtrage de géolocalisation est à effectuer. Par défaut, Non.

#### Output :

`pyspark.sql.DataFrame` : un DataFrame Spark avec les correspondances 1-1, y compris les colonnes d'entrée et certains des résultats numériques tels que les différents scores de similarité.

### III. En route vers le clustering

En guise de rappel de notre modélisation, nous avons maintenant les points de ventes avec les profils socio-économiques et d'autres informations externes de leur zone de chalandises. Pour le groupe HORECA, cette étape a permis d'enrichir les informations des points de ventes (PdV) que nous avons déjà.

Désormais, nous disposons dans notre table de données finale, les caractéristiques de services des hôtels, des restaurants et des cafés. Dans la suite de ce projet, nous ferons un focus sur la partie HORECA. C'est pourquoi nous présenterons dans ce qui suit la segmentation clientèle du canal HORECA. La couverture des HORECA est 30% sur TripAdvisor. Comme nous nous y attendons, tous les HORECA ne s'y trouvent pas. Nous avons décidé de réaliser tout d'abord d'implémenter un clustering robuste sur ces 30% présents sur TripAdvisor. Par la suite, nous utiliserons des variables pertinentes pour entraîner un modèle d'apprentissage supervisé des 30% des HORECA vers les clusters afin de prédire les 70% restants vers les clusters trouvés. D'ailleurs, si notre modèle de clustering est optimal le modèle d'apprentissage supervisé que nous mettrons en place devraient prédire les bons clusters pour chaque PdV.

### Conclusion

A la sortie de ce chapitre, les éléments de réponses commencent à s'éclaircir. La base de données est désormais très enrichie. Pour le chapitre suivant, cette base de données a été le principal outil. Juste un peu de feature engineering, nous pouvons retirer des variables clés à l'analyse.

## Chapitre 3 :

---

# Clustering

Introduction.....	- 131 -
I. Exécution de l'algorithme pour les restaurants et café.....	- 131 -
I.1 Transformations et distance .....	- 131 -
I.2. Variables de segmentation.....	- 131 -
I.3. Nombre de cluster retenus et performances .....	- 134 -
I.4 . Interprétation des clusters.....	- 136 -
I.5. Structure hiérarchique des clusters .....	- 141 -
II. Exécution de l'algorithme pour les Hotels.....	- 141 -
III. Classifications des PdV à l'aide du Random Forest.....	- 142 -
Conclusion .....	- 142 -

## Introduction

Ce chapitre présente les résultats obtenus du clustering. La méthode finalement retenue a été la technique du K-means. Une initialisation K-means++ a été également introduite afin de s'assurer de la robustesse du partitionnement réalisé. Pour plus de concision, nous avons décidé de ne pas présenter les résultats de l'algorithme retenu. Cependant, nous avons testé un algorithme CAH afin de comparer les résultats. Les individus ici sont les HORECA. Quel type de transformation et de distance avons-nous utilisé ? Combien de clusters ont été retenus ? Quelle sont les interprétations des clusters ? Qu'en-t-il de la structure hiérarchique du clustering des PdV ? Tant de questions pertinentes auxquelles nous apporterons des réponses dans les lignes qui suivent. Selon des directives orientées business, une segmentation est réalisée sur la partie restaurant et cafés et une autre sur la partie hôtel.

### I. Exécution de l'algorithme pour les restaurants et café

#### *I.1 Transformations et distance*

Les données utilisées pour le clustering ont été centrées et réduites. Une distance euclidienne a été finalement retenue selon les mêmes raisons annoncées dans la partie théorique. D'ailleurs, un test avec la distance Manhattan nous conduit aux mêmes résultats.

#### *I.2. Variables de segmentation*

Comme précisé dans la partie théorique, toutes les variables n'ont pas été utilisées pour l'exercice de segmentation. Étant donné que seul un sous-ensemble de tous les points de vente sont couverts par TripAdvisor, les variables actives doivent être une sélection bien équilibrée des features de TripAdvisor et d'autres variables basées sur les données sociodémographiques, POI et mobilité. Une bonne différenciation sur les autres

variables est nécessaire pour une cartographie réussie des points de vente sans données TripAdvisor avec les segments.

Le tableau 8 présente la liste des variables utilisées pour segmenter les restaurants et cafés. Il faut noter que des variables ont été obtenus par feature engineering ou par transformations. Nous avons décidé de laisser volontairement le nom des variables en anglais pour faciliter la compréhension et pour des soucis de précisions. Le nom des variables est formulé de telle sorte qu'on puisque comprendre son sens. Etant donné les centaines de features dont nous disposons, nous ne pouvons tous les présenter. L'approche de les nommer en anglais et garder les mots clés des variables a donc été retenue.

Par exemple La variable *Population Age 30-59 Prop.* a été en en réalité obtenue par agrégation des variables *Population Age 30-34 Prop*, *Population Age 35-39 Prop* ... jusqu'à *Population Age 54-59 Prop*. Également, les variables *Blue Collar POI Prop* et *White Collar POI Prop* sont également un regroupement des activités similaires. Ces variables nous permettront ainsi de mesurer les POI des activités et lieu de travail. Ce regroupement des activités est d'ailleurs connu en marketing. Nous mettrons la liste des activités en annexe. *Prop* dans le nom d'une variable veut dire proportions.

**Tableau 9: Variables actives**

Variables actives	Source
Rurality	Calcul
Population Age 30-59 Prop.	Experian
Disposable Income per Cap.	Experian
Residents Staying Peak Season Prop.	Unacast
Residents Staying Night time Prop.	Unacast
Touristiness Score	Flickr
Touristic POI Prop.	Pitney Bowes
Transportation POI Prop.	Pitney Bowes
Blue Collar POI Prop.	Pitney Bowes
White Collar POI Prop.	Pitney Bowes
Education & Sports POI Prop.	Pitney Bowes
TripAdvisor Top Ranking	TripAdvisor
TripAdvisor Premium Features Flag	TripAdvisor
TripAdvisor Average Rating	TripAdvisor
TripAdvisor Price Range	TripAdvisor
TripAdvisor Meals Types	TripAdvisor
TripAdvisor Alcohol Presence Flag	TripAdvisor

51 features ont été sélectionnées comme passives (explicatives) afin de mieux expliquer nos clusters. Les plus significatives dans l'explication des clusters des restaurants sont recensées dans la figure ci-après. Les variables commençant S sont des variables extraites directement de notre table de données enrichies. Celles débutant par – sont des variables obtenues par features engineering.

**Tableau 10: Variables inactives**

Variables inactives	Source
S: Dis. Income per Capita	Experian
S: Touristic POI Prop	Pitney Bowes
S: Transportation POI Prop	Pitney Bowes
S: TA Top Ranked	TripAdvisor
S: TA Premium Features	TripAdvisor
S: TA Avg. Rating	TripAdvisor
S: TA Price Range	TripAdvisor
S: TA Meals: Breakfast	TripAdvisor
S: TA Meals: Dinner,	TripAdvisor
S: TA Serves Alcohol	TripAdvisor
-TA Features: Entertainment	TripAdvisor
-TA Cuisines: French	TripAdvisor
-TA Cuisines: Asian	TripAdvisor
-TA Cuisines: Moroccan & Middle Eastern	TripAdvisor
-TA Cuisines: International	TripAdvisor
-TA Cuisines: Healthy	TripAdvisor
-TA Meals: Lunch	TripAdvisor

Le tableau ci-dessous une liste de quelques features qui n'ont pas été retenus finalement dans l'explications de nos clusters car possédant une faible variabilité. Des variables comme le nombre de ménages (S: carto\_household\_sum) dans la zone de chalandise du PdV, la présence des boissons en emballage plastique (S : UNIT\_CASES\_PACKAGE\_PET ) ou des cannettes (S : UNIT\_CASES\_PACKAGE\_Retornable\_Glass) ou encore le type des boissons multiserve (S: UNIT\_CASES\_SERVE\_TYPE\_Multiserve). C'est-à-dire les boissons dont les volumes sont strictement supérieurs à 0.5

**Tableau 11 : liste des variables inactives avec une variabilité interclasses négligeable**

S: UNIT_CASES_PACKAGE_Returnable_Glass	CMD
-TA_touristiness	TripAdvisor
S: carto_staying_weekdays	Unacast
S: carto_staying_off_peak	Unacast
carto_staying_peak	Unacast
-TA_features_serves_alcohol	TripAdvisor
S: carto_staying_weekends	Unacast
-TA_radius	Calculé
S: carto_age_60pl_prop	Experian
-TA_cuisines_contemporary_and_fusion	TripAdvisor
S: carto_staying_night	Unacast
-TA_meals_after_hours	TripAdvisor
S: UNIT_CASES_PACKAGE_PET	CMD
S: carto_household_sum	Experian
S: carto_beverage_spend_per_capita	Experian
S: carto_staying_day	Unacast
S: md_horeca_prop	CMD
-flickr_density	Flickr
-TA_cuisines_african	TripAdvisor
-TA_rating_food	TripAdvisor
-TA_rating_service	TripAdvisor
-TA_rating_value	TripAdvisor
-TA_meals_lunch	TripAdvisor
-TA_dietary_restrictions_halal_kosher	TripAdvisor
S: UNIT_CASES_SERVE_TYPE_Multiserve	CMD
-TA_dietary_restrictions_vegan_vegetarian_glutenfree	TripAdvisor
S : LOYALTY	CMD
S : medal	CMD

### I.3. Nombre de cluster retenus et performances

Nous fixons un nombre de cluster pouvant aller jusqu'à 20 en utilisant la méthode du coude. Nous pouvons alors regarder comment varie l'homogénéité (nous l'avons T dans la partie théorique) au fur et à mesure que nous varions le nombre de clusters. Pour trouver le point optimal, l'on peut se référer à la méthode de coude. Visuellement, nous dirons que le nombre de cluster optimale est situé entre 7 et 9 ;

**Figure 40 : Choix du nombre de cluster**

Vu la subjectivité de cette méthode, nous serons alors plus méthodiques en utilisant nos connaissances mathématiques. Il s'avère que le point qui indique l'équilibre entre homogénéité et la séparation des clusters est le point de la courbe qui est le plus éloigné d'une ligne tracée entre les points  $P_0$ (nombre de cluster minimale) et  $P_1$ (nombre de cluster maximale). Nous chercherons alors le couple  $(x, y)$  qui maximise cette distance. La formule de la distance est :

$$d(P_0, P_1, (x, y)) = \frac{(y_1 - y_0)x - (x_1 - x_0)y + x_1y_0 - x_0y_1}{\sqrt{(y_1 - y_0)^2 + (x_1 - x_0)^2}}$$

Où le couple  $(x, y)$  représente les coordonnées de tout point dont nous pourrions vouloir calculer la distance à la ligne. Regardons encore une fois notre intrigue de coude.

*En appliquant cette méthode, le nombre de cluster optimale est 8.*

Pour aller plus loin, nous affichons ci-dessous l'inertie (somme des carrés des distances au centroïde le plus proche pour tous les points de l'ensemble de données d'entraînement), le coefficient de silhouette et la taille de chaque cluster.

**Tableau 12: Evolution du nombre de classes et performances**

K	Silhouette	Inertie	Taille des clusters
2	<b>0,475</b>	<b>2 344,6</b>	[164, 12]
3	<b>0,174</b>	<b>2 114,5</b>	[60, 41, 75]
4	<b>0,239</b>	<b>1 747,7</b>	[61, 64, 15, 36]
5	<b>0,244</b>	<b>1 636,2</b>	[48, 56, 28, 13, 31]
6	<b>0,178</b>	<b>1 548,2</b>	[28, 38, 26, 12, 49, 23]
7	<b>0,237</b>	<b>1 451,6</b>	[32, 23, 47, 12, 30, 6, 26]
8	<b>0,198</b>	<b>1 452,5</b>	[28, 24, 43, 7, 23, 5, 21, 25]
9	<b>0,231</b>	<b>1 281,6</b>	[19, 30, 27, 12, 4, 19, 11, 21, 33]

Au regard de ce tableau, on peut dire que vraisemblablement tout va bien. Que l'on se penche sur le coefficient de silhouette globale des clusters ou de l'inertie, nous sommes satisfaits.

#### 1.4. Interprétation des clusters.

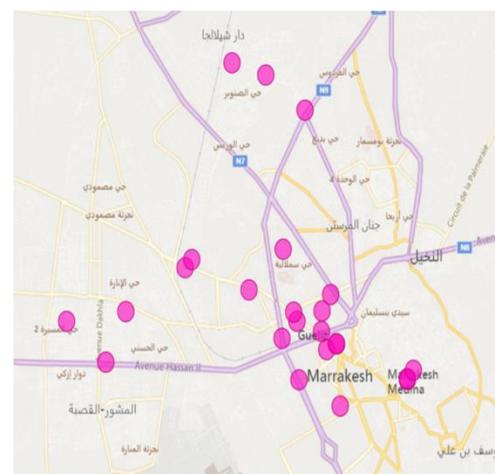
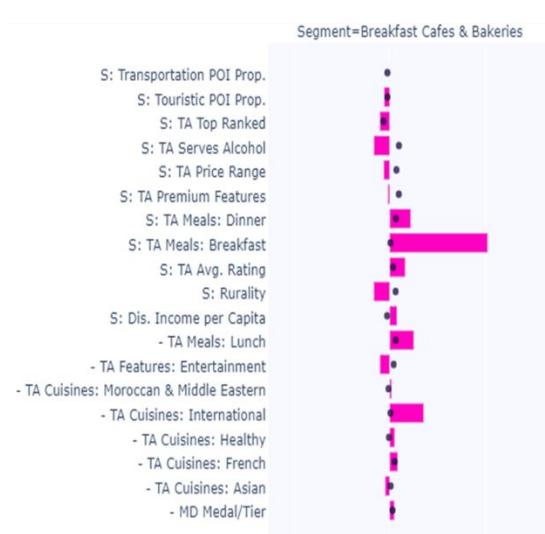
Cet exercice est difficile à réaliser manuellement ou sans programmer des algorithmes qui nous aidera à faciliter les interprétations au regard la multitude d'exécutions de clustering que nous avons à faire. Pour interpréter, nous visualisons les statistiques des clusters pour les variables dites inactives. Celles dont la variabilité interclasses sont négligeables ne sont plus considérées. Nous utilisons pour chaque variable :

- L'ampleur et la direction de l'écart par rapport à la moyenne globale (centrée à 0) (représenter sous forme de barre dans la visualisation)
  - Les moyennes entre clusters. Sur le graphique, ils seront repérés par un point.
- Nous avons gardé le nom des clusters en anglais.

#### Cluster 1 : Breakfast Cafés & Bakeries

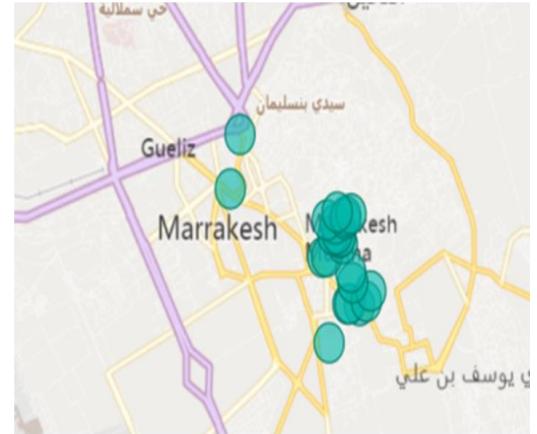
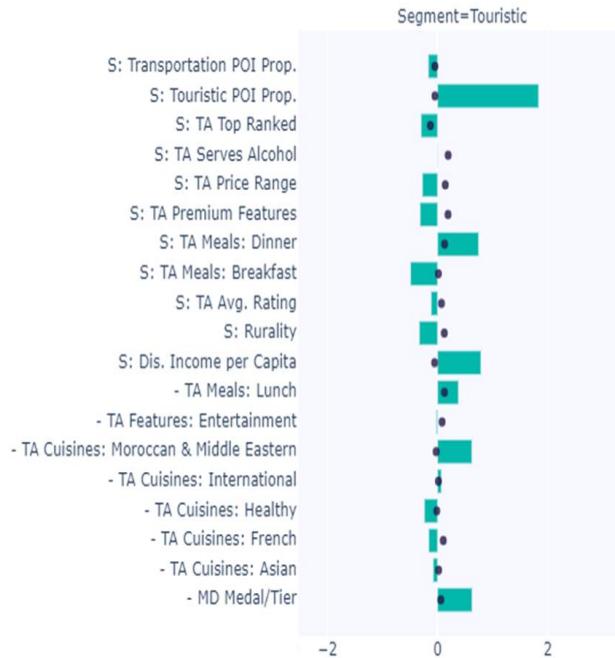
Par Bakeries, nous désignons boulangeries. Ce cluster comprend des PdV ouverts pendant l'heure du petit déjeuner tels que les cafés et les boulangeries. Ces cafés ont tendance à rester ouverts toute la journée.

#### Segment: Breakfast Cafés & Bakeries



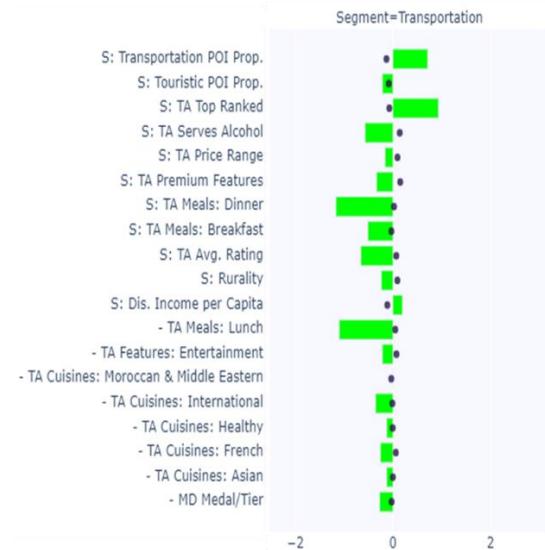
## Cluster 2 : Touristic

PdV sont situés dans le centre-ville, dans une zone très touristique  
Le revenu disponible par habitant est plus élevé.



## Cluster 3 : Transportation

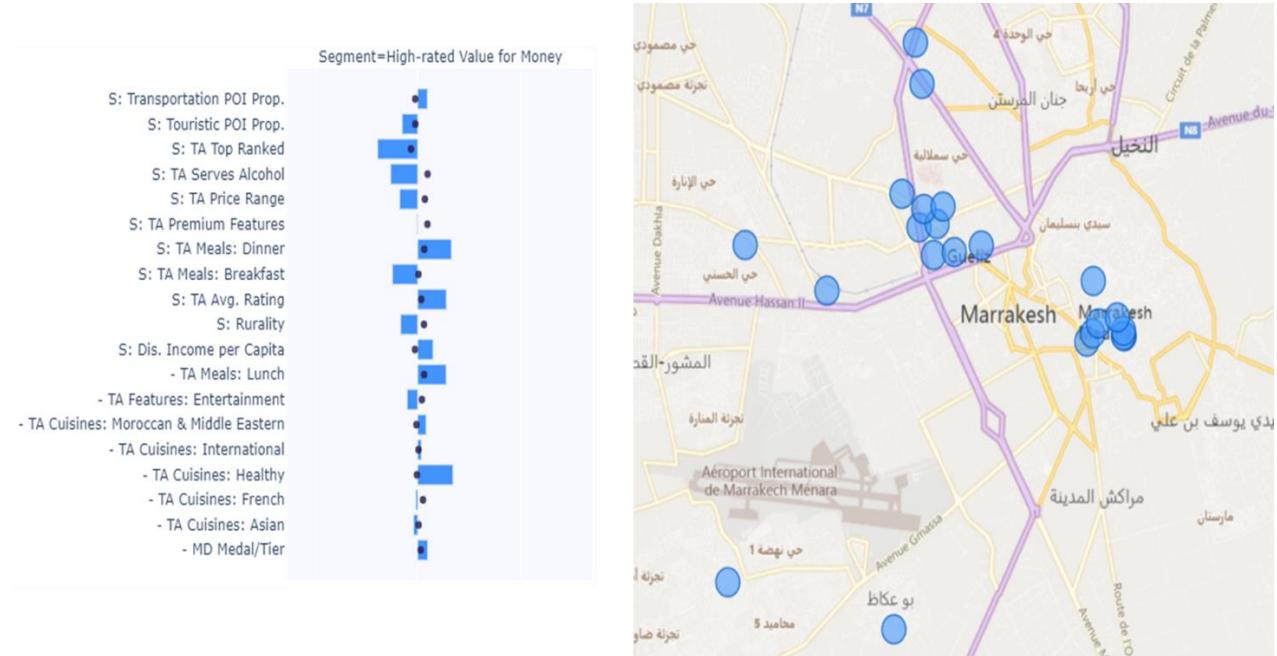
Endroits situés à proximité des POI de transport et des POI col bleu et blanc  
Principalement des pages sur TripAdvisor à faible reviews avec un nombre relativement faible d'avis et de faibles notes.



#### **Cluster 4: High-rated value for money**

Ce sont des PdV marquées par :

- Top Rang sur Trip Advisor,
  - Gamme de prix inférieure,
  - Présence accrue de points de vente servant des repas sains, options végétariennes et végétaliennes



## Cluster 5 : Premium & Entertainment

- Établissements haut de gamme
- Servant de l'alcool et offrant des divertissements (comme de la musique en direct)
- Fourchette de prix plus élevée  
De la cuisine de type française est souvent servie
- Top classement sur TripAdvisor



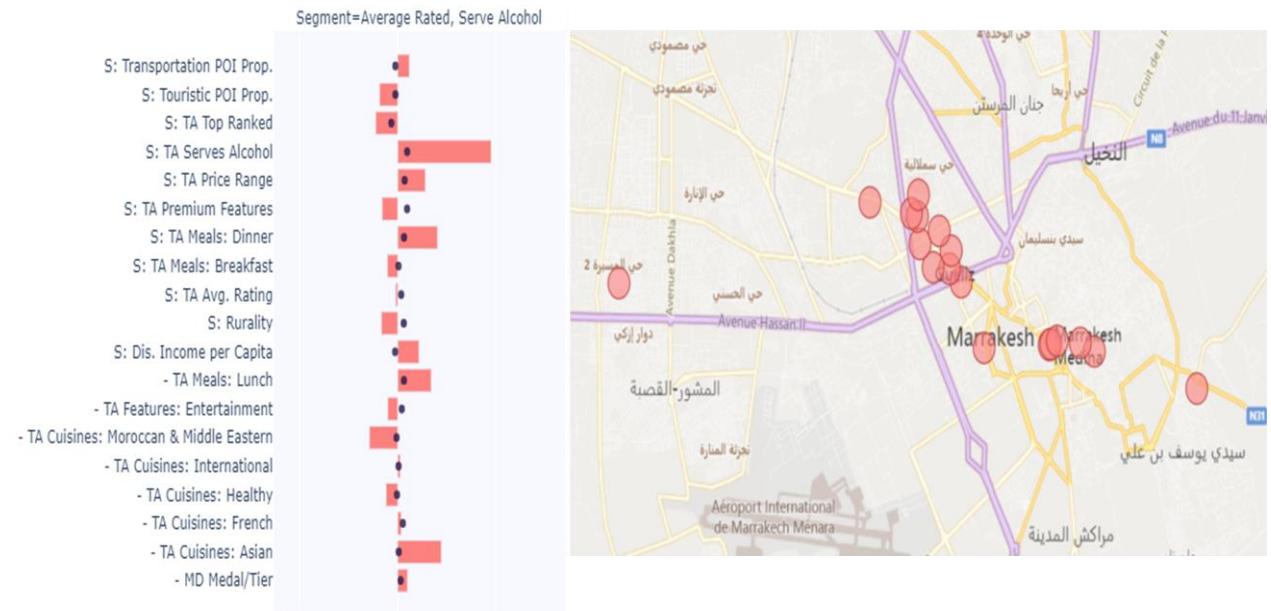
## Cluster 6 : Semi-Rural

- Restaurants et cafés situés dans les zones rurales, dans des endroits où le revenu disponible par habitant est plus faible et où les dépenses en aliments et boissons sont plus faibles
- Population biaisée vers les groupes d'âge plus jeunes



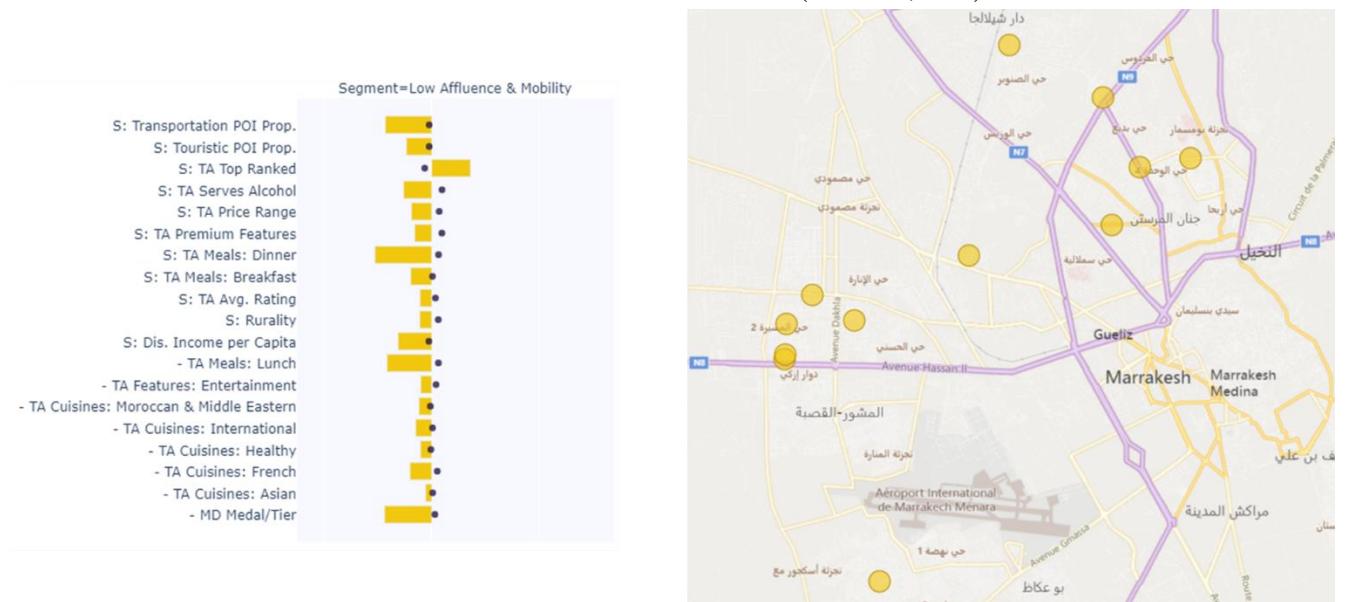
## Cluster 7: Average Rated, Serve Alcohol

- PdV servant de l'alcool
- Note moyenne sur TripAdvisor
- Inclinaison vers la cuisine asiatique



## Cluster 8: Low Affluence & Mobility

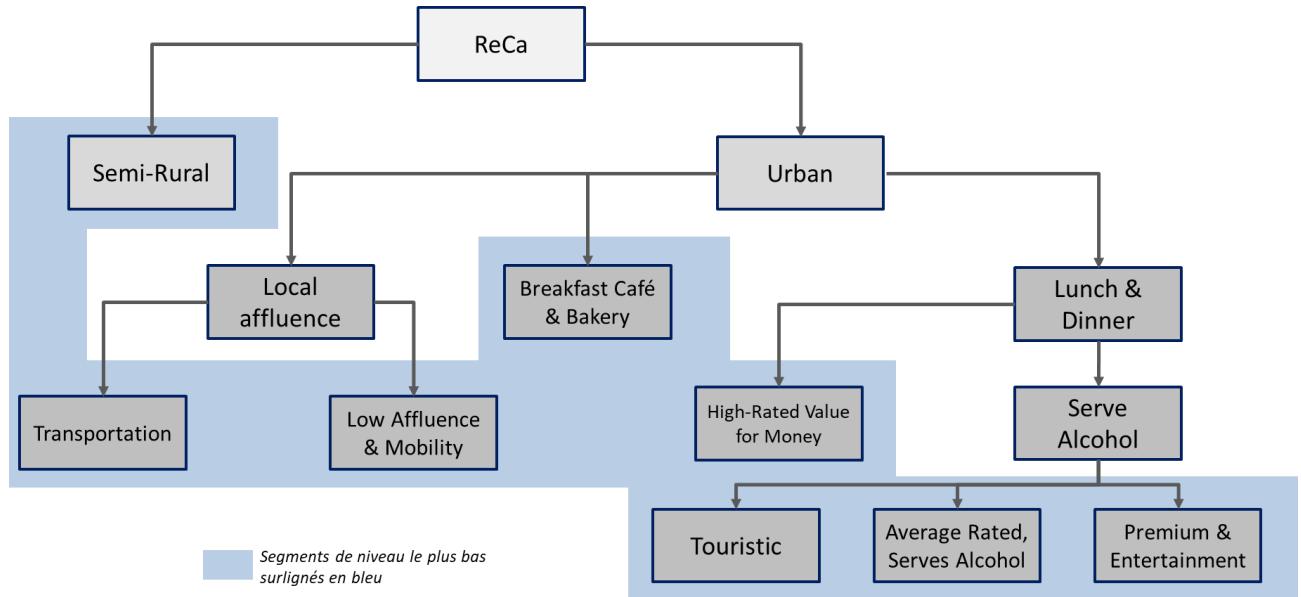
- Restaurants et cafés situés dans des zones de faible affluence, sans de nombreux POI de transport.
- Le classement de Trip Advisor est inférieur, ainsi que la gamme de prix
- Inclinaison vers les niveaux des médailles basses (Bronze, Tin)



### I.5. Structure hiérarchique des clusters

Réalisée sur la base des expériences acquises en essayant méthodes CAH et K-mean.

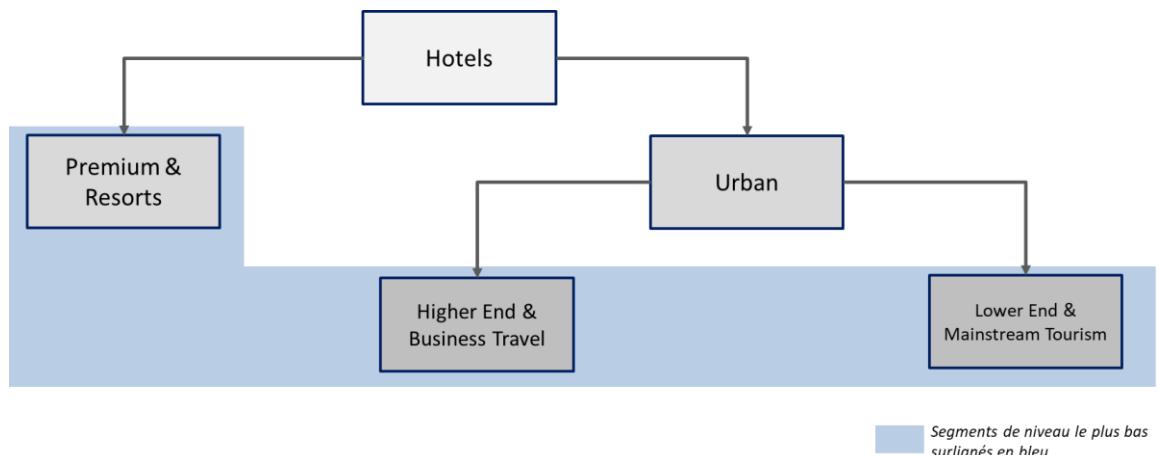
**Figure 41: Structure hiérarchique des Restaurant et Café**



### II. Exécution de l'algorithme pour les Hotels

Les mêmes méthodes ont été appliquées. Trois clusters ont été trouvés pour répartition optimal. Ainsi le résultat des interprétations des classes est présenté à travers la structure hiérarchique ci-après.

**Figure 42: Structure hiérarchique des Hotels**



### III. Classifications des PdV à l'aide du Random Forest

Pour les hôtels trois segments ont été créés et pour les Restaurants-Café 8 segments. Nous expliquons dans les lignes qui suivent la démarche effectuée pour les Restaurants-Café. La démarche étant la similaire.

Notre Baseline est l'ensemble des résultats obtenus à partir du clustering. Dans la base de données, chaque client a un numéro de cluster. Les données sont ensuite séparées en données d'apprentissage et en données de test. Pour les hôtels par exemple, parmi 176 PdV, 80% des PdV ont été tiré aléatoirement mais stratifié par la taille des clusters.

Toutes les variables passives du clustering ont été utilisées pour cet exercice.

La première étape fut l'optimisation des paramètres de tuning comme le nombre d'arbre à retenir, le nombre maximum d'entités prises en compte pour diviser un nœud, le nombre maximum de niveaux dans chaque arbre de décision par validation croisée de Monte Carlo.

Notre modèle est par la suite entraîné à partir de ces paramètres optimisés. Comme performance, l'accuracy ou exactitude est l'ordre de 92% et F1 90%. Cette performance était largement suffisant. Ces performances sont en réalité le reflet de la bonne division de nos variables pour la segmentation.

Tout est désormais bien dans le meilleur des mondes. Le modèle supervisé est à cet effet utilisé sans crainte afin d'assigner l'ensembles des PdV qui n'étaient pas sur TripAdvisor vers les clusters que nous avons formés dans l'exercice du clustering.

## Conclusion

A la sortie de chapitre, chaque PdV du canal HORECA de la ville de Marrakech est désormais assigné à un cluster donné.

En entreprise lorsqu'on segmente la clientèle, c'est pour des raisons bien déterminées. L'une de ses raisons est la modélisation de la demande des clusters.

## Chapitre 4 :

---

# Modélisation de la demande et optimisation des assortiments.

Introduction .....	- 144 -
I. Modèles de demandes.....	- 144 -
I.1. Variables et forme des modèles de demandes. ....	- 144 -
I.2 Aperçu des modèles de demandes .....	- 145 -
II. Optimisation des assortiments .....	- 148 -
Conclusion.....	- 150 -

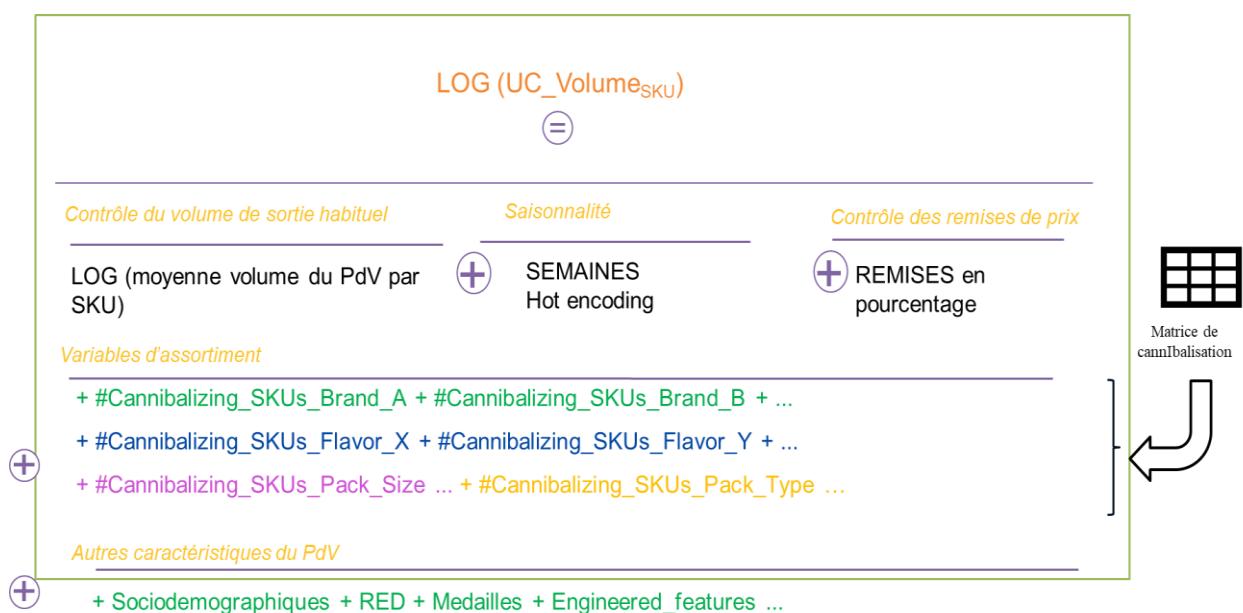
## Introduction

Le but de ce chapitre est la présentation de la méthodologie adoptée pour la modélisation de la demande au sein de chaque cluster. Nous rappelons que lorsque nous parlons de demande nous parlons de demande de boissons auprès des PdV. Les modèles de demandes qui sont construits sont des modèles pour chaque type de produit. Ce chapitre donne également le schéma de l'optimisation des assortiments.

### I. Modèles de demandes

#### I.1. Variables et forme des modèles de demandes.

Pour chaque segment, nous découvrons comment le volume d'une référence SKU particulière dépend de l'assortiment actuel (produits d'ECCBC disponible dans le Pdv), des caractéristiques et des facteurs externes (par exemple, météo, événements spéciaux). En effet, nous modélisons la performance des SKUs (capturée par les ventes observées représentées) à partir d'un modèle de demande à effet fixe (contrôle du volume de sortie habituel). Un input important au modèle est une matrice signalant la cannibalisation entre les marques et les saveurs. Pour simplifier la compréhension, ci-dessous le modèle par SKU pour un cluster donné s'écrit comme suit :



Avant l'étape de l'implémentation de la régression, nous avons créé des fonctions afin de nous aider à réaliser automatiquement certaines analyses exploratoires. Ainsi, des fonctions étaient chargées de :

- Eliminer les variables dont la variance est égale ou inférieure au seuil 0 .001
- Supprimer les variables fortement corrélées s'il ne vienne pas de la matrice de cannibalisation. Seuil fixé à 0 .8
- Supprimer les variables avec des VIF supérieur à 10

En appliquant ces étapes, les variables socio-économiques ont été retirées du modèle. Cela n'est pas étonnant nous sommes dans des clusters assez homogènes.

Cependant, il y a lieu de rappeler que le travail que nous réalisons actuellement est dynamique. De nouveaux SKUs peuvent arriver une année donnée. Comment seront -ils prises en compte ? Déjà, pour l'année 2019, des nouveaux SKUs sont sur le marché. Pour ces cas, nous n'avons pas assez d'informations les concernant ni comment se fait la cannibalisation.

Dans les programmes développés, nous tenons compte de la nouveauté du SKU par une variable appelée limited coverage. Lorsque le nombre d'observation pour un SKU donné est inférieur à un certain seuil, il est déclaré comme nouveau ou à couverture limitée.

Les SKUs à couverture limitée ne sont pas modélisés directement. Nous estimons leur performance par ajustement. En effet, l'idée est de chercher les SKUs similaires du SKU à couverture limitée en utilisant la distance cosinus par rapport aux caractéristiques de l'article (la marque, le goût, l'emballage, taille ...). Les trois plus similaires sont récupérés. Lorsque nous aurons besoin des estimations de volumes du modèle du SKU à couverture limitée, nous prenons la moyenne des estimations des 3 SKUs similaires pondérées par la distance cosinus de similarités.

## *I.2. Aperçu des modèles de demandes*

Pour des raisons liées au business, deux types de modèles à savoir à un modèle de régressions constraint et un modèle non constraint ont été estimées. Dans le modèle avec contraintes, des contraintes de non-positivité des coefficients ont été imposées

aux variables de cannibalisation alors que des contraintes de non-négativité sont assignées aux variables de médaille, le nombre de porte du frigo ou les données RED. Cela nous permet d'améliorer les écarts types des effets estimés. Les intervalles de confiance deviennent par définition plus étroits.

L'algorithme de modèles de demandes/ SKU par cluster est programmé comme suit :

*Sélection des données hebdomadaires de janvier 2019 jusqu'à 2021*

*Récupérer la liste des SKUs disponibles dans le canal*

*Itérer par cluster*

*Sélection des SKUs qui ne sont pas à couverture limitée*

*Itérer par SKU*

*Appliquer la sélection var à partir des variables selon le principe expliqué*

*Fractionnement de l'échantillon d'entraînement (80%\*) et validation*

*Exécuter le Modèle sans contraintes*

- Génération automatique d'un rapport qualité du modèle ( $R^2$ , MAPE, MAE)
- Rapport sur tous les coefs du modèle SKU -> distribution --> rechercher les limites inférieure et supérieure des contraintes, faire attention aux coefs négatifs
- Génération automatique des résultats tests des hypothèses du modèle

*Exécuter le modèle avec contraintes*

- Génération automatique d'un rapport qualité du modèle ( $R^2$ , MAE+MAPE)
- Rapport sur tous les coefs du modèle SKU -> distribution --> rechercher les limites inférieure et supérieure des contraintes, vérifier l'amélioration
- Génération automatique des résultats des tests des hypothèses du modèle

*Fin sku*

*Fin Cluster*

*Fin*

Même si nous ne pouvons pas afficher les coefficients des modèles obtenus dans ce rapport, nous pouvons quand même présenter quelques résultats concernant leurs

performances. Les modèles de non contraints que nous avons réalisés ont respecté les critères de viabilités d'un modèle de régression. En effet Les hypothèses de normalité des erreurs ont été validées par des tests de Shapiro Wilks<sup>34</sup>. En parti, nous devons la validité du test de normalité à la transformation logarithme appliqué à notre variable dépendante mais également à notre variable indépendante qui représente l'effet fixe. Les hypothèses d'homoscédasticité et d'autocorrélation des erreurs sont réalisées respectivement grâce aux tests de White<sup>35</sup> et de Durbin-Watson<sup>36</sup>. Dans les deux cas, nous avons observé pour l'ensemble des SKUs modélisés des p value supérieurs à 0.05. Théoriquement, ces performances étaient prévisibles car nos modèles sont exécutés ont sein de clusters homogènes. Les erreurs ont tendance à être généralement non corrélées pour un cluster homogène isolé.

Ci-dessous se trouvent les performances des modèles contraints du Sku 26 ET 27. Le mean absolute error (MAE) et MAPE (mean absolute percentage Error) ont été calculé après retransformation. En effet, nous avons utilisés des logarithmes dans la modélisation des volumes, donc une retransformation appliquant la fonction exponentielle était une nécessité afin de bien voir l'ampleur du MAPE et MAE

Données d'entraînement			Données de test					
R2 ajusté	MAE	MAPE	R2 ajusté	MAE	MAPE	channel	cluster	sku
0.926694	0.916594	0.134976	0.871369	1.058682	0.13681	RECA	Premium and entertainment	26
0.934678	0.994904	0.105678	0.940045	0.958682	0.1045681	RECA	Premium and entertainment	27

Que ce soit le R2 ou MAPE, nous pouvons juger que les deux modèles présentés sont performants. Le plus grand MAPE sur les données de test pour l'ensemble des régressions exécuté est de 14%.

<sup>34</sup> Le test de Shapiro Wilk a pour comme hypothèse H0 : ‘les erreurs suivent une loi normale’ vs H1 : ‘les erreurs ne sont pas distribuées normalement’

<sup>35</sup> Le test de White a pour hypothèse H0 : Homoscédasticité (les résidus ont tous la même variance) vs H1 : Hétéroscédisticité

<sup>36</sup> Durbin-Watson a pour hypothèses ‘H0 : Non-autocorrélation des erreurs’

Les modèles contraints également réalisés respectent les hypothèses de la régression. Les écart types des coefficients estimés se sont rétrécis comme le montre la théorie.

## II. Optimisation des assortiments

Actuellement les services d'ECCBC sont faites par canal. Au niveau des HORECA il existe un seul assortiment. Vu la richesse d'informations que nous avons apportés dans ce projet, le chemin de votre voyage prendra fin lorsque nous utiliserons nos résultats obtenus pour proposer de nouveaux assortiments.

**Figure 43: Exemple d'assortiment du réfrigérateur**



Cette fois-ci, chaque cluster aura un assortiment qui lui est précis. L'idée c'est de proposer l'ensemble des produits selon le cluster à mettre dans le réfrigérateur d'un PdV afin de maximiser le volume vendu de l'entreprise. On rappelle ici qu'on peut également changer le paramètre volume à profits.

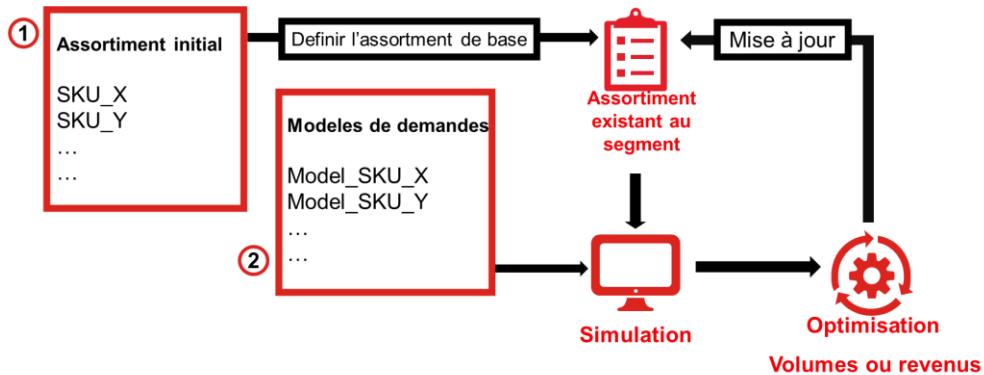
L'assortiment optimal pour un segment sera calculé en 2 étapes :

Etape 1 : Identification de l'assortiment de base. Les SKUs avec des performances et une présence solide de longue date dans les points de vente.

Etape 2 : Simuler et optimiser. A l'aide des modèles de demande estimés, on prédit tout d'abord le volume de tous les SKUs et par la suite, on sélectionne le SKU qui

maximise la valeur générée. Répétez l'opération jusqu'à ce que la liste des références SKUs soit épuisée et qu'un classement soit créé.

**Figure 44: Algorithme d'optimisation de l'assortiment du réfrigérateur**



### Algorithme

Pour cet exercice, nous rappelons que nous travaillons sur des données de janvier 2019 à fin décembre 2019. L'optimisation que nous allons mettre en œuvre est dynamique. En 2019, de nouveaux SKUs ont été mis sur le marché. L'algorithme repose sur une évaluation des rangs.

Pour un canal donné

Pour un cluster donné

Pour un sku donné

### Evaluation des rangs

[Rang 1] :

*Le point de départ est un magasin vide fictif. Sélectionner la référence SKU la plus performante en l'absence de cannibalisation (réflétée dans les variables SIM). Dans ce cas, les variables de cannibalisations sont tous manquantes.*

[Rang 2] :

*Le point de départ est le magasin fictif avec le SKU déjà ajoutées dans l'étapes précédente. Sélectionner la référence SKU la plus performante en tirant parti de la cannibalisation réestimée. Le nombre de variable de cannibalisation passe à 1.*

[Rang j] :

*Le point de départ est le magasin fictif avec les SKU déjà ajoutées dans les étapes précédentes (itérations de classement). Sélectionner la référence SKU la plus performante en tirant parti de la cannibalisation réestimée - compte tenu des références SKU disponibles dans le PdV fictif.*

L'algorithme est répété jusqu'à  $j = \text{nombres de SKUs}$  dont la couverture n'est pas limitée.

#### **Calculs communs à réaliser dans les étapes d'évaluations de rang**

- *Construire l'ensemble de données 2019 pour noter / prédire, caractéristiques disponibles*
- *Utiliser les modèles de demande pour prédire les volumes par sku selon le scénario actuel du PdV fictif*
- *Sélectionnez la référence SKU la plus performante - en termes de volume NSR maximal, pour remplir la position de classement respective*

#### **Traitement spécial pour les nouveaux SKUs non modélisé par manque d'observation (couverture limitée/nouvelles) :**

- *Trouver des références sku similaires (en fonction de la distance cosinus par rapport aux caractéristiques de l'article (la marque, le goût, l'emballage, taille ...))*
- *Utiliser les ventes observées en 2019 pour chaque SKU similaire pour calibrer les données de ventes moyennes de 2019 disponibles pour le sku à couverture limitée.*
- *Utilisez les modèles de demande estimés pour chaque sku similaire afin d'obtenir des prédictions pour le SKU à couverture nouvelle/limitée.*
- *Prendre une moyenne pondérée des prédictions des SKU similaires considérées - en utilisant les distances de similarité obtenues comme pondérations*

**Fin sku**

**Fin cluster**

Comparez les rangs optimisés aux rangs réels (observés sur les ventes réelles de 2019)

**Fin canal**

**Fin**

Cet algorithme a été déjà validé au sein du département pour implémentation.

Vu le délai accordé à ce stage, nous n'avons pas pu finir la programmation de cet algorithme et présenter un résultat.

## **Conclusion**

Ce chapitre a demandé assez d'attention en programmation car les SKUs à modéliser étaient nombreux et il fallait trouver une manière automatisée de construire

les modèles. Néanmoins nous avons pu construire des modèles de demandes contraintes et non contraintes. Chacun de ces types de modèles aura un rôle bien défini à jouer au sein du département de data science et business intelligence de l'ECCBC. Déjà les modèles constraints sont déjà mis en application pour la recommandation des assortiments des réfrigérateurs afin de maximiser les ventes ou les volumes.

# Conclusion générale

---

Pour une bonne compréhension de ce rapport, nous avons séparé sa rédaction en deux grandes parties. La première concerne l'ensemble de présentations théoriques et revue de la littérature nécessaire pour mieux comprendre la seconde. Dans cette dernière, corps principal de la modélisation, nous avons poursuivi trois principaux objectifs. Le premier étant d'enrichir la base de données internes avec des informations spatiales pour la ville de Marrakech. En utilisant les techniques d'analyse spatiale et un algorithme de Matching, nous y sommes arrivés. Notre deuxième objectif a été atteint avec la découverte de 8 segments pour les PdV de types Restaurant-Café et trois (03) segments pour les hôtels.

Notre segmentation a révélé qu'elle a l'avantage d'être robuste et optimale car CAH et K-mean donnent les mêmes résultats. Le choix de la distance également n'altère pas la formation des mêmes classes. Par la suite, nous avons utilisés les travaux précédents afin de réaliser des modèles de demandes au sein de chaque cluster et pour chaque SKU. Les modèles estimés ne souffrent d'aucune violation des hypothèses stochastiques. Une application de ces modèles de demande est l'optimisation des assortiments des réfrigérateurs au sein des clusters afin de maximiser les volumes vendus.

Néanmoins, même si les résultats de ce projet sont satisfaisants, nous pensons que des améliorations sont possibles. Dans la détermination des zones de chalandises, nous avons supposé que les PdV appartenant à la même ruralité et au même canal aient des superficies identiques. Ceci est discutable. Également, il serait plus scientifique de bien vouloir confirmer les rayons de zones de chalandises obtenu via le point de partage selon la loi Reilly avec une autre méthode comme la méthode de la zone proximale basée sur la théorie des lieux centraux.

Quant aux modèles de demandes, pour un futur proche, évidemment après l'achèvement de l'implémentation de notre algorithme d'optimisation de l'assortiment des réfrigérateurs, nous pensons à implémenter des modèles de demandes basés sur une

approche de données de panel. Cette approche est connue chez les économètres pour sa richesse en termes d'interprétation comme l'étude de la convergence des consommations par cluster. Non seulement les modèles de panels permettent de nous affranchir des problèmes de biais en cas d'omissions de variables non observables (très fréquent en marketing d'ailleurs) mais ils permettent d'avoir des estimations plus précises. Au lieu de modéliser par cluster un SKU donnée, une modélisation panel nous permettra de le faire en une seule étape, occasionnant des degrés de libertés de plus, ayant pour conséquence une précision des estimations.

# Annexes

---

## Annexe : Classification POI des activités

### Blue-Collar (Col-bleu)

Trade division	Detailed Group
DIVISION B. - MINING	COAL MINING METAL MINING MINING AND QUARRYING OF NONMETALLIC MINERALS, EXCEPT FUELS OIL AND GAS EXTRACTION
DIVISION C. - CONSTRUCTION	CONSTRUCTION - GENERAL CONTRACTORS AND OPERATIVE BUILDERS CONSTRUCTION - SPECIAL TRADE CONTRACTORS HEAVY CONSTRUCTION, EXCEPT BUILDING CONSTRUCTION, CONTRACTOR
DIVISION D. - MANUFACTURING	APPAREL, FINISHED PRODUCTS FROM FABRICS AND SIMILAR MATERIALS CHEMICALS AND ALLIED PRODUCTS ELECTRONIC AND OTHER ELECTRICAL EQUIPMENT, AND COMPONENTS FABRICATED METAL PRODUCTS FOOD AND KINDRED PRODUCTS FURNITURE AND FIXTURES INDUSTRIAL AND COMMERCIAL MACHINERY, AND COMPUTER EQUIPMENT LEATHER AND LEATHER PRODUCTS LUMBER AND WOOD PRODUCTS, EXCEPT FURNITURE MEASURING, PHOTOGRAPHIC, MEDICAL, AND OPTICAL GOODS, AND CLOCKS MISCELLANEOUS MANUFACTURING INDUSTRIES PAPER AND ALLIED PRODUCTS PETROLEUM REFINING AND RELATED INDUSTRIES PRIMARY METAL INDUSTRIES PRINTING, PUBLISHING AND ALLIED INDUSTRIES RUBBER AND MISCELLANEOUS PLASTIC PRODUCTS STONE, CLAY, GLASS, AND CONCRETE PRODUCTS TEXTILE MILL PRODUCTS TOBACCO PRODUCTS TRANSPORTATION EQUIPMENT
DIVISION I. - SERVICES	AUTOMOTIVE REPAIR, SERVICES AND PARKING MISCELLANEOUS REPAIR SERVICES

### White-Collar (Col blanc)

Trade division	Detailed Group
DIVISION H. - FINANCE, INSURANCE, AND REAL ESTATE	DEPOSITORY INSTITUTIONS HOLDING AND OTHER INVESTMENT OFFICES INSURANCE AGENTS, BROKERS AND SERVICE INSURANCE CARRIERS NONDEPOSITORY CREDIT INSTITUTIONS REAL ESTATE SECURITY AND COMMODITY BROKERS, DEALERS, EXCHANGES AND SERVICES
DIVISION I. - SERVICES	BUSINESS SERVICES ENGINEERING, ACCOUNTING, RESEARCH, AND MANAGEMENT SERVICES LEGAL SERVICES MOTION PICTURES

# Bibliographie

---

## Livres

- Trevor Hastie, Robert Tibshirani, Jerome Friedman (2008) The Elements of Statistical learning Data Mining, Inference, and Prediction Springer Series in Statistics.
- Gérard Cliquet, Jerome Baray (2020), Location- based Marketing , Géomarketing and Géolocation, Informations systems, web and pervasive computing series
- Jason Brownlee, Ensemble Learning Algorithms With Python Make Better Predictions with Bagging, Boosting, and Stacking (2021)
- Insee (2018) - Eurostat Manuel d'analyse spatiale, Théorie et mise en œuvre pratique avec R, *Insee Méthodes* n° 131- octobre 2018

## Cours

- Mohammed El Haj TIRARI, Apprentissage statistique, INSEA, Rabat, Année universitaire 2020-2021
- Mohammed El Haj TIRARI, Nouvelles techniques d'échantillonnage, INSEA Rabat,
- Année universitaire 2020-2021
- Serge LHOMME, Introduction à l'analyse spatiale, Université Paris-Est Créteil,
- Année universitaire 2020-2021
- Serge LHOMME, Introduction à la géomatique, Université Paris-Est Créteil 2020-2021
- Mohammed El Haj Tirari, Analyse de données, INSEA, Rabat,
- Année universitaire 2019-2020
- Brigitte Gelein, Apprentissage supervisé, ENSAI France, Année universitaire 2019-2020
- Touhami ABDELKHALEK, Econométrie 1, INSEA, Rabat, Année universitaire 2019-2020.
- Yannis Chaouche, Chloé-Agathe Azencot, Nathalie Turck, Explorez vos données avec des algorithmes non supervisés, Ecole CentraleSupélec/OpenClassroom

## Articles scientifiques et autres

Wah TY (2015) A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. PLoS ONE 10(12): e0144059. doi:10.1371/journal.pone.0144059, <https://doi.org/10.1371/journal.pone.0144059>.

Shirkhorshidi AS, Aghabozorgi S, Wah TY (2015) A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. PLoS ONE 10(12): e0144059. doi:10.1371/journal.pone.0144059.

Alegana V. A., Atkinson P.M., Pezzulo C., Sorichetta A., Weiss D., Bird T., Erbach-Schoenberg E. et Tatem A. J. (2015) Fine resolution mapping of population age-structures for health and development applications J. R. Soc. Interface. 12(2015) 00732015007.

Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R. Stevens, Andrea E. Gaughan, Vincent D. Blondel, Andrew J. Tatem, (2014) Proceedings of the National Academy of Sciences Dynamic population mapping using mobile phone data, 111 (45) 15888-15893; DOI : 10.1073/pnas.1408439111.

Cain, Jim (2013). "Coordinate Reference Systems (Best Practices for Assignment, Manipulation and Conversion in GIS Systems)" (PDF). 2013 ESRI Petroleum GIS Conference.

Bas Ranzijn (2013), A Geocoding Algorithm Based On A Comparative Study Of Address Matching Techniques, Master Thesis, Operations Research and Quantitative Logistics Erasmus Universiteit Rotterdam

Prasetyo, D., & Hastuti, K. (2012). Application of Haversine Formula for selecting Location and Christian Church based on Mobile. Semarang: Universitas Dian Nuswantoro

Pandit, S. and S. Gupta. (2011) “A Comparative Study on Distance Measuring Approaches for Clustering.” International Journal of Research.

Alastair Aitchison (2011) The Google Maps / Bing Maps Spherical Mercator Projection.

Forrest R. Stevens, Andrea E. Gaughan, Catherine Linard, Andrew J. Tatem (2006), désagrégation des données du recensement pour la cartographie des populations à l'aide de forêts aléatoires à l'aide de données de télédétection et de données auxiliaires

P.E. Christen(2006). A comparison of personal name matching: Techniques and practical issues. <http://astro.temple.edu/~joejupin/entitymatching/tr-cs-06-02.pdf>. Technical Report TR-CS06-02, Australian National University.

G.R. Hjaltason and H. Samet(2003). Index-driven similarity search in metric spaces. ACM Transactions on Database Systems.

W. Cohen, P. Ravikumar, and S. Fienberg(2003) A Comparison of String Metrics for Matching Names and Records *KDD Workshop on Data Cleaning and Object Consolidation*.

W.E. Winkler(1999). The state of record linkage and current research problems. Statistical Research Division, U.S. Census Bureau.

CRESSIE, Noel A.C. (1993). « Statistics for spatial data: Wiley series in probability and statistics ». Wiley-Interscience, New York.

OXYGEN(Coca Cola System) , SEGMENTED EXECUTION 2.0 , Case Study FLAWLESS SEGMENTED EXECUTION and Case Study POGO MICRO SEGMENTATION

