Université des Sciences et de la Technologie
Houari Boumediene Faculté Informatique

TP DATA MINING

Exploitation des données et Extraction des règles d'associations

---

Étudiants :
HOUACINE Maya
OUCHAR Manel
SII G02

Table des matières

Introduction Générale **6**

1 Analyse et prétraitement des données **7**

1.1 Données statiques : Dataset 1 7
1.1.1 Description du dataset 7
1.1.2 Analyse des caractéristiques des attributs 8
1.1.2.1 Calcul des mesures de tendance centrale et déduction Des symétries 8
1.1.2.2 Boîtes à moustache et données aberrantes. 9
1.1.2.3 Histogrammes et distribution des données 11
1.1.2.4 Diagrammes de dispersion des données 12
1.1.3 Prétraitement 15
1.1.3.1 Traitement des valeurs manquantes et aberrantes : 15
1.1.3.1.a Choix de la méthode de remplacement des valeurs manquantes. 15
1.1.3.1.b Choix de la méthode de traitement des valeurs aberrantes 16
1.1.3.2 Réduction des données horizontales / verticales. 16
1.1.3.3 Normalisation des données : 17
1.1.3.3.a Méthode Min-Max 17
1.1.3.3.b Méthode z-score. 17

1.2 Données temporelles : Dataset 2 18
1.2.1 Description du dataset 18
1.2.2 Prétraitement 19
1.2.2.1 Traitement des Valeurs Manquantes 19
1.2.2.2 Traitement des Données Aberrantes 19
1.2.3 Visualisation 20
1.2.3.1 La distribution du nombre total des cas confirmés et tests positifs par zones 20
1.2.3.1.a Graphe 20
1.2.3.1.b Analyse 20

1.2.3.2 L'évolution des tests COVID-19, tests positifs et le nombre de cas confirmés au fil du temps pour la zone 94085 21
1.2.3.2.a Graphes 21
1.2.3.2.b Analyse 22
1.2.3.3 Distribution des cas COVID-19 positifs par zone et par année 22
1.2.3.3.a Graphe 22
1.2.3.3.b Analyse 22
1.2.3.4 Le rapport entre la population et le nombre de tests effectués 23
1.2.3.4.a Graphe 23
1.2.3.4.b Analyse 23
1.2.3.5 Les 5 zones les plus fortement impactées par le coronavirus 24
1.2.3.5.a Graphe 24
1.2.3.5.b Analyse 24
1.2.3.6 Le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant une periode choisie 25
1.2.3.6.a Graphes 25
1.2.3.6.b Analyse 26

2 Extraction de motifs fréquents, règles d'associations et corrélations 27

2.1 Dataset3 27
2.1.1 Description du dataset 27
2.1.2 Discrétisation de l'attribut Température 28
2.1.2.1 En classes d'effectifs égaux (Equal Frequency) 28
2.1.2.2 En classes d'amplitudes égales (Equal Width) 28
2.1.3 Extraction des motifs fréquents : Apriori 29
2.1.4 Extraction des règles d'associations sous plusieurs mesures 32
2.1.4.1 Confiance 32
2.1.4.2 Lift 33
2.1.4.3 Cosine 33
2.1.4.4 Jaccard 33
2.1.4.5 Kulczynski 33
2.1.5 Résultats de l'expérimentation des différentes valeurs de MinSupp et MinConf 34

3 Conclusion et perspectives 37

Table des figures

1.1 Boxplots des 14 attributs du dataset1 avec échelle logarithmique 10
1.2 Histogramme de l'attribut Mn (Manganèse) avant remplacement des valeurs abberantes 11
1.3 Histogramme de l'attribut Mn (Manganèse) après remplacement des valeurs abberantes 11
1.4 Histogramme de l'attribut EC 11
1.5 Histogramme de l'attribut Cu 11
1.6 Histogramme de l'attribut N 11
1.7 Histogramme de l'attribut S 11
1.8 Matrice de corrélation des attributs du Dataset 1. 12
1.9 Scatter plot des deux attributs N et Fertility. 13
1.10 Scatter plot des deux attributs OC et OM avant traitement des valeurs aberrantes. 13
1.11 Scatter plot des deux attributs OC et OM après traitement des valeurs aberrantes. 13
1.12 Scatter plot des deux attributs K et Zn avant traitement des valeurs aberrantes. 14
1.13 Scatter plot des deux attributs K et Zn après traitement des valeurs aberrantes. 14
1.14 Scatter plots des deux attributs B et Mn, Cu et pH, avant et après traitement des valeurs aberrantes. 15
1.15 Graphe de distribution du nombre total des cas confirmés et tests positifs par zones. 20
1.16 Graphe de l'évolution hebdomadaire du nombre de cas, tests positifs et tests effectués dans la zone 94086 en août 2021 21
1.17 Graphe de l'évolution hebdomadaire du nombre de cas et tests positifs dans la zone 94086 en août 2021 21
1.18 Graphe de l'évolution mensuelle du nombre de cas, tests positifs et tests effectués dans la zone 94086 en 2021 21
1.19 Graphe de l'évolution mensuelle du nombre de cas et tests positifs dans la zone 94086 en 2021 21

1.20 Graphe de l'évolution annuelle du nombre de cas, tests positifs et tests effectués dans la zone 94086 21
1.21 Graphe de l'évolution annuelle du nombre de cas, tests positifs dans la zone 94086 21
1.22 Graphe de distribution des cas COVID-19 positifs par zone et par année. . 22
1.23 Graphe représentant le rapport entre la population et le nombre de tests effectués. 23
1.24 Graphe de distribution des cas COVID-19 positifs par zone et par année. . 24
1.25 Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 22. 25
1.26 Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 26. 25
1.27 Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 30. 25
1.28 Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 35. 25
1.29 Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 40. 25
1.30 Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 45. 25
1.31 Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 50. 25
1.32 Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 56. 25
1.33 Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 62. 25
1.34 Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 67. 26

2.1 Graphe 3D du nombre de règles fréquentes générées par rapport au support minimum et à la confiance minimale. 34
2.2 Graphe de l'évolution du nombre de motifs fréquents selon supmin. 35
2.3 Graphe de l'évolution du temps d'exécution selon supmin. 35
2.4 Graphe l'évolution de l'espace mémoire alloué à l'algorithme Apriori et les règles d'associations selon le support minimum et la confiance minimale. . 36

Liste des tableaux

1.1 Description des attributs du Dataset1 8
1.2 Tendances centrales, quartiles et ecart type des attributs du Dataset1 . . . 9
1.3 Tableau descriptif des attributs du Dataset 2. 18

2.1 Tableau descriptif des attributs du Dataset3 27

3.2 Tableau descriptif des attributs du dataset1 39
3.3 Tableau de l'évolution du temps d'exécution de l'algorithme Apriori selon le support minimum. 59
3.4 Tableau de l'évolution du nombre de règles générées de l'algorithme Apriori selon la confiance minimale et le support minimum. 68
3.5 Tableau de l'évolution de l'éspace mémoire alloué à l'algorithme Apriori selon la confiance minimale et le support minimum. 70

Introduction générale

Au vu de la croissante exponentielle de quantitées de data générés ces dernieres années, nous sommes submergés par les données, mais reste affamés de connaissance, ce qui rend le Datamining est d'autant plus important de nos jours. Dans cette première partie de notre projet nous allons d'abord netoyer et prétraités nos données, car avec des données incompletes, redondantes et bruité aucun bon resultats ne peut etre tiré des modeles. Ensuite, nous nous concentrons sur l'extraction de règles d'associations et la découverte de motifs fréquents.

Dans le premier chapitre, nous explorons deux types de datasets : le dataset statique (dataset1) portant sur les propriétés du sol, et le dataset temporel (dataset2) représentant l'évolution du nombre de cas de COVID-19 par code postal. Cette exploration comprend l'analyse et pré-traitement des données, et est essentielle pour garantir la fiabilité des résultats dans la prochaine phase du projet.

Le deuxième chapitre se consacre à l'extraction de motifs fréquents et de règles d'association dans un dataset spécifique (dataset3), Ce processus, impliquant la discrétisation des données et l'application de l'algorithme Apriori, vise à fournir des informations cruciales sur les associations existantes entre les differents attribut liés au climat, au sol, à la végétation, et à l'utilisation d'engrais... pour des prises de décision judicieuses.

La section suivante détaille les résultats expérimentaux, en variant les valeurs du support minimum ainisi que la confiance minimale afin de deduire leurs effets sur differentes mesures. En parallèle, le développement d'une interface utilisateur conviviale offre une expérience interactive pour explorer les résultats.

En conclusion, nous analysant ,pretraitant et appliquant un nouvel algorithme "Appriori" sur de reel dataset de domaines variés tels que l'agriculture, et la santé publique.

Chapitre 1

Analyse et prétraitement des données

1.1 Données statiques : Dataset 1

1.1.1 Description du dataset

Le dataset que nous explorons dans cette première partie du projet, comprend **14** caractéristiques concernant le sol, notamment des éléments nutritifs tels que l'azote et le potassium, des indicateurs de santé comme le pH et la matière organique, ainsi que des métaux essentiels. Le Tableau1.1 ci-dessous présente une brève description de chacune d'entre elles.

Ces données fournissent une vue globale sur les propriétés de **884** sols (instances), ainsi que leurs fertilités respectives comprises entre 0 et 2, offrant une base pour des analyses approfondies en agriculture et en gestion environnementale.

Attribut	**Description**	**Type**
N	Azote	entier
P	Phosphore	reel
K	Potassium	entier
pH	Potentiel Hydrogène	reel
EC	Conductivité Électrique	reel
OC	Matière Organique	reel
S	Soufre	reel
Zn	Zinc	reel
Fe	Fer	reel
Cu	Cuivre	reel
Mn	Manganèse	reel
B	Bore	reel
OM	Matière Organique	reel
Fertility	Fertilité du Sol F	{0,1,2}

TABLE 1.1 – Description des attributs du Dataset1

1.1.2 Analyse des caractéristiques des attributs

1.1.2.1 Calcul des mesures de tendance centrale et déduction Des symétries

En comparant entre les valeurs de moyenne, mode et médiane de chaque attribut, nous pouvons déduire la nature de leurs distributions.

PH, EC, Zn et OC se distinguent par des distributions symetriques/légérement asymétrique, car leurs mode = médiane = moyenne. Tandis que les attributs K, Mn et B ont une Moyenne > Médiane >Mode ce qui correspond à une asymétrie positive. Ils indiquent que les valeurs de ses caractéristiques sont généralement faibles dans les sols, mais cela peut aussi être dû à la présence de valeurs aberrantes. Quant au reste des attributs, ils sont soit asymétriques simples soit suivent une distribution asymétriques négative comme

Fertility

En ce qui concerne les écart-types, elles varient entre 0.14 et 129.03 indiquant que certains attributs dénotent d'une grande variabilité dans les données alors que d'autres sont denses dans leurs distributions.

À noter que tous les attributs sont unimodale à l'exception d'EC et S qui sont bimodale

Nom	**Mean**	**Mode**	**Min**	**Q1**	**Q2**	**Q3**	**Max**	**Ecart type**
N	246.997	[207.0]	6.0	201.0	257.0	307.0	383.0	77.315
P	14.555	[8.3]	2.9	6.8	8.1	10.7	125.0	21.918
K	501.338	[444.0]	11.0	412.0	475.0	581.0	1560.0	129.031
Ph	7.511	[7.5]	0.9	7.35	7.5	7.63	11.15	0.4643
EC	0.5439	[0.53, 0.62]	0.1	0.43	0.55	0.64	0.95	0.1412
OC	0.617	[0.88]	0.1	0.38	0.59	0.78	24.0	0.84064
S	7.545	[4.22, 5.13]	0.64	4.7	6.64	8.75	31.0	4.415
Zn	0.468	[0.28]	0.07	0.28	0.36	0.47	42.0	1.887
Fe	4.126	[6.32]	0.21	2.05	3.56	6.31	44.0	3.10
Cu	0.952	[1.25]	0.09	0.63	0.93	1.25	3.02	0.52
Mn	8.6536	[7.54]	0.11	6.21	8.34	11.44	31.0	4.298
B	0.593	[0.34]	0.06	0.27	0.41	0.61	2.82	0.5744
OM	1.0637	[1.51]	0.17	0.65	1.0148	1.34	41.28	1.445
fertility	0.592	[1.0]	0.0	0.0	1.0	1.0	2.0	0.578

TABLE 1.2 – Tendances centrales, quartiles et ecart type des attributs du Dataset1

1.1.2.2 Boîtes à moustache et données aberrantes.

la figure suivante englobante les boxplots des 14 attributs de notre dataset sur une

échelle logarithmique afin de les comparer entre eux met en évidence les faits suivants :

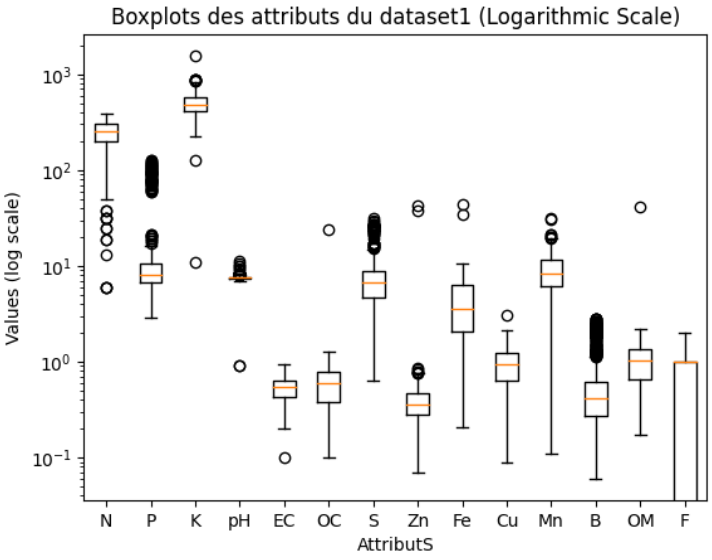

FIGURE 1.1 – Boxplots des 14 attributs du dataset1 avec échelle logarithmique

- Les caractéristiques de nos sols appartiennent à des échelles distinctes allant de 10∧(-1) à 10∧3
- Nous pouvons visuellement constater les différentes distributions des attributs et confirmer qu'elles concordent avec les observations et deductions du tableau precedant
- Pour ce qui est des valeurs aberrantes, nous remarquons que le phosphore, le Bore, le Souffre et le PH sont sensible aux Outliers de par leurs nombres conséquents. À l'inverse OC, EC, le Cuivre et la matière organique n'en contiennent que très peu.

1.1.2.3 Histogrammes et distribution des données

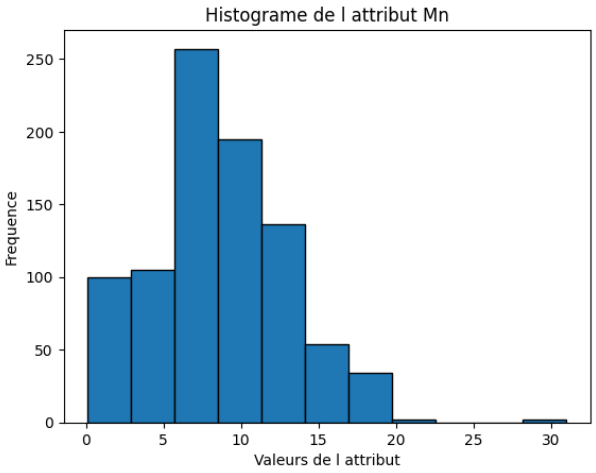

FIGURE 1.2 – Histogramme de l'attribut Mn (Manganèse) avant remplacement des valeurs abberantes

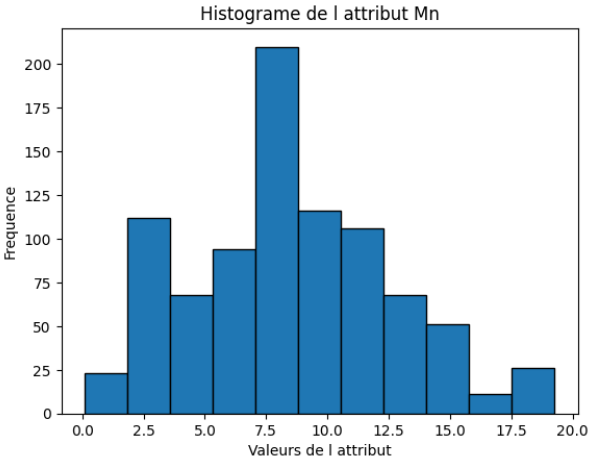

FIGURE 1.3 – Histogramme de l'attribut Mn (Manganèse) après remplacement des valeurs abberantes

L'observation initiale de l'histogramme de fréquence du manganèse (Mn) montre l'existance de valeurs abberantes ,causant une classe vide et affectant la représentation globale de la distribution (Figure 1.2). Afin d'assurer une interprétation plus fiable, nous avons traité ces valeurs aberrantes et généré un nouvel histogramme dépourvu de ces anomalies (Figure 1.3).

Ce deuxième histogramme montre une distribution legerement asymétrique positive, car la majorité des échantillons ont des concentrations modérées en manganèse. Ceci met en évidence la tendance des sols, à ne pas accumuler beaucoup de manganèse. Cette petite inclination peut être due entre autres à la nature géologique du sol qui limite l'absorption du manganèse.

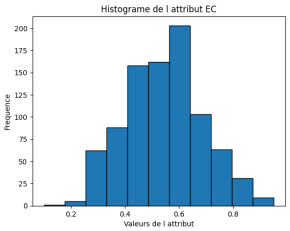

FIGURE 1.4 – Histogramme de l'attribut EC

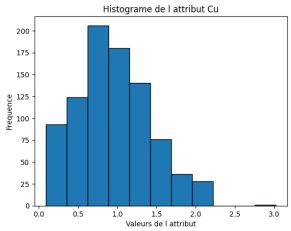

FIGURE 1.5 – Histogramme de l'attribut Cu

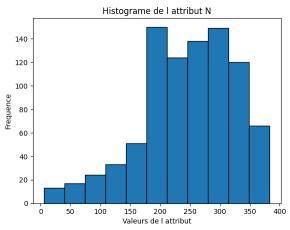

FIGURE 1.6 – Histogramme de l'attribut N

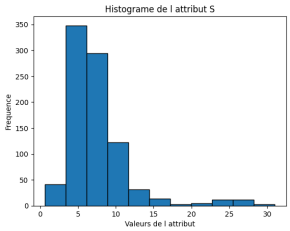

FIGURE 1.7 – Histogramme de l'attribut S .

L'histogramme de la conductivité électrique (EC) des sols, dévoile une distribution symétrique, car la plupart des sols ont des concentrations moyennes. Les fréquences les

plus élevées se situent entre 0,2 et 0,6, ce qui indique qu'il y a généralement une tendance vers des niveaux modérés comme le confirme sa valeur faible d'ecart type.

L'observation des l'histogrammes représentant la fréquence du soufre (S) et "Cuivre" (Cu) sur les sols, révèle une distribution asymétrique. Cette asymétrie est positive car la majorité des données sont concentrées à gauche Les raisons de cette tendance peuvent être multiples, allant des propriétés géologiques du sol aux pratiques agricoles spécifiques et meme l'activité humaine et d'autres facteurs.

Tandis que l'histogramme indiquant la fréquence de l'Azote (N) sur les sols, révèle une distribution asymétrique négative etant donnée que ses valeurs on tendance à etre grandes et effectivement ce qui etait moins facile à deduire à partir du tableau des tendances centrales.

1.1.2.4 Diagrammes de dispersion des données

Les diagrammes de dispersions nous fournissent un aperçu sur les corrélations existantes entre les attributs ce qui est primordial, que se soit lors du prétraitement dans la réduction de dimensions, la découverte de nouvelles relations implicite entre attributs et l'amélioration des modèles utilisé sur la dataset par la suite

Étant donné qu'il y a 91 possibles scatter plots, nous avons calculé la matrice de corrélation afin de choisir judicieusement nos graphes.

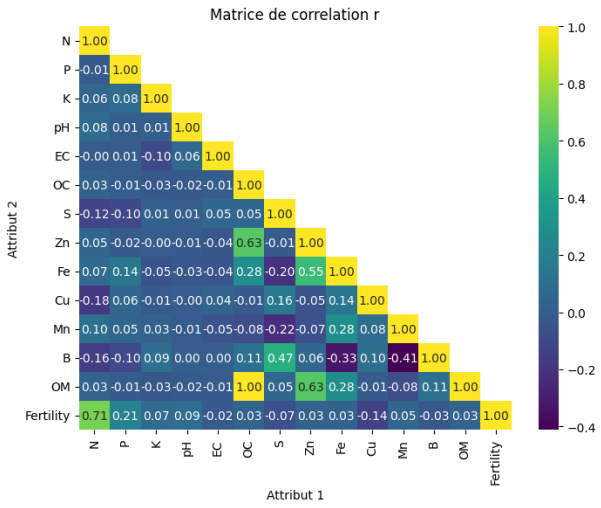

FIGURE 1.8 – Matrice de corrélation des attributs du Dataset 1.

Nous pouvons observer à partir de cette matrice que la plupart des attributs sont indépendants, et que ceux corrélés négativement relativement fortement sont presque inexistant. Cependant, il y a plusieurs propriétés du sol qui sont corrélées positivement avec EC et OM atteignant une valeur de 1.0 (Selon Pearson corrélation).

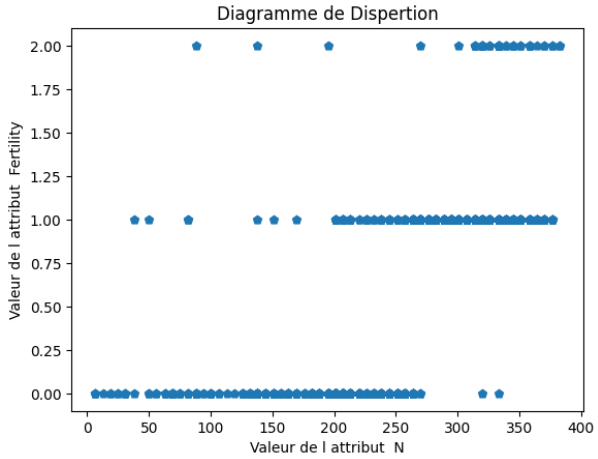

FIGURE 1.9 – Scatter plot des deux attributs N et Fertility.

Une autre paire correlée positivement est la paire Fertility et N, comme l'attribut Fertility est de type categorique les points s'accumulent a 0,1 et 2 seulement. Donc le Nitrogene est un bon indicateur de la fertilité du sol.

Pour ce qui est des graphes concerant la paire d'attribut OM et OC apres avoir remplacer les valeurs abberantes par regression lineaire, nous constatons qu'elle correspond à la fonction x=y indiquant une claire correlation positive entre les 2 (r=1). Ceci s'explique par le fait que le carbone organique constitue 58% de la matiere organique.

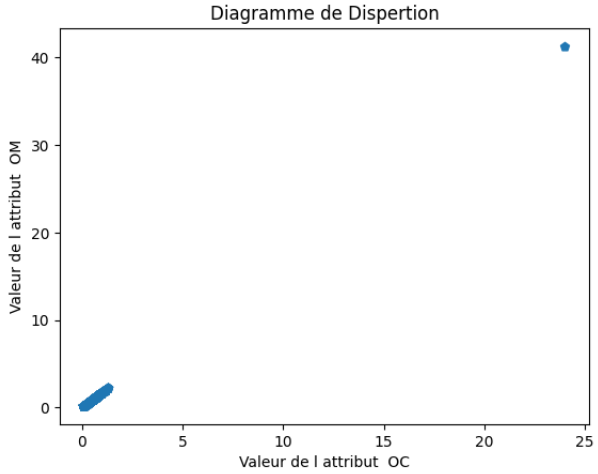

FIGURE 1.10 – Scatter plot des deux attributs OC et OM avant traitement des valeurs aberrantes.

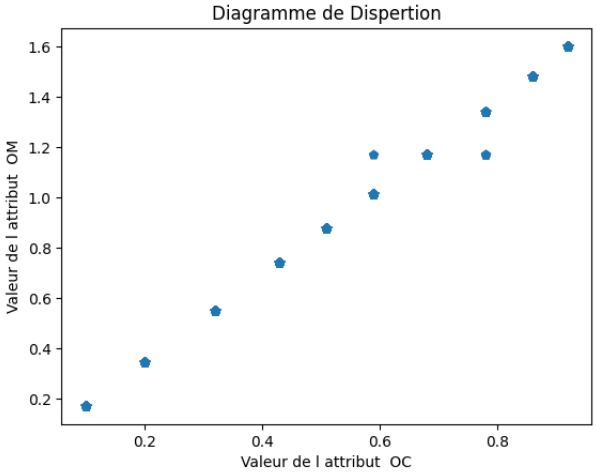

FIGURE 1.11 – Scatter plot des deux attributs OC et OM après traitement des valeurs aberrantes.

Malgré le remplacement des valeurs aberrantes, ainsi que l'obtension de r<0 pour la paire (B,Mn) nous ne voyons pas de corrélations négative calire entre les 2 caractéristiques, cela est attribuable au fait que r=-0.4 est plus proche de 0 (aucune corrélation) que de -1

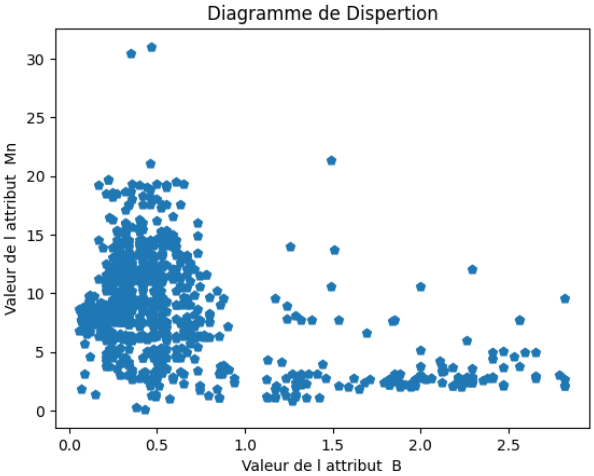

FIGURE 1.12 – Scatter plot des deux attributs K et Zn avant traitement des valeurs aberrantes.

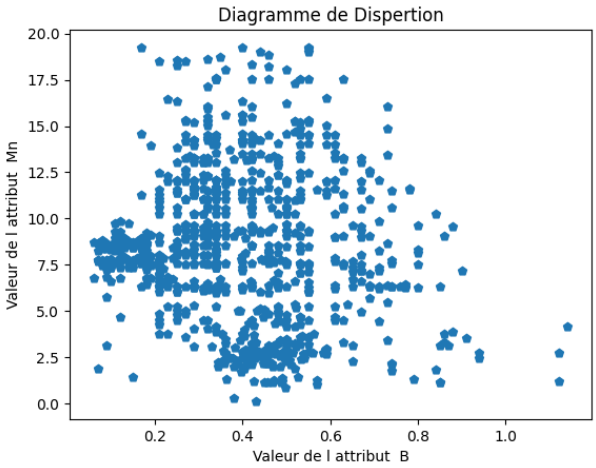

FIGURE 1.13 – Scatter plot des deux attributs K et Zn après traitement des valeurs aberrantes.

Pour ce qui est des 2 paires (Cu,Ph) et (K,Zn) leurs points sont dispersées de manière aléatoire prouvant l 'indépendance entre ces propriétés.

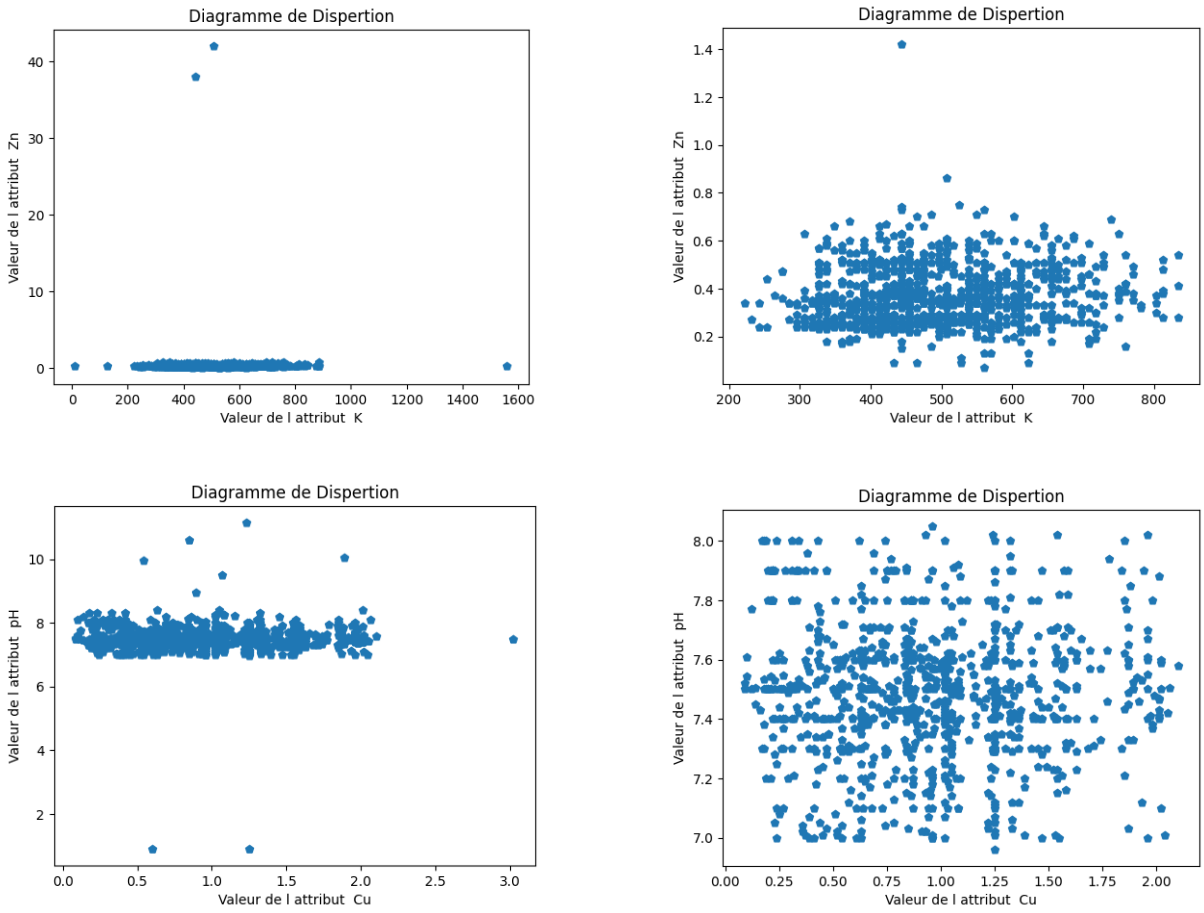

FIGURE 1.14 – Scatter plots des deux attributs B et Mn, Cu et pH, avant et après traitement des valeurs aberrantes.

1.1.3 Prétraitement

1.1.3.1 Traitement des valeurs manquantes et aberrantes :

1.1.3.1.a Choix de la méthode de remplacement des valeurs manquantes.

Après avoir analysé nos données, il est temps de nettoyer notre dataset. Pour les valeurs manquantes, nous en avons détecté 4 (Qui sont ” ” , ” ?“ , ” ?” , ”NA”) en utilisant une expression régulière. Cela en retournant les valeurs qui ne respectent pas le pattern correspondant à des valeurs numériques (étant donné que tous nos attributs sont de ce type)

Pour le remplacement 2 techniques fut implémenter

1. **Remplacement par Mode** : Cette méthode consiste à remplacer chaque valeur manquante par la valeur la plus fréquemment observée de sa colonne. Nous avons identifié le mode en comptabilisant la fréquence de chaque valeur non manquante dans la colonne respective. En cas de plusieurs modes, nous avons choisi aléatoirement l'une de ces valeurs.
2. **Remplacement par Moyenne** : Adaptée aux données numériques, cette approche remplace les valeurs manquantes par la moyenne des valeurs non manquantes de la même classe (valeur de Fertility) assurant un remplacement contextuel des valeurs manquantes. Cette opération préserve la cohérence des données.

1.1.3.1.b Choix de la méthode de traitement des valeurs aberrantes

Les valeurs aberrantes quant à elles sont détectées si elles verifient la condition suivante :

$$\text{si } x > Q3 + 1.5 \times IQR \text{ ou } x < Q1 - 1.5 \times IQR]$$

L'IQR est la différence entre le troisième quartile (Q3) et le premier quartile (Q1) des données.

$$IQR = Q3 - Q1$$

— **Remplacement par regression lineaire** : Nous avons créer un modèle de régression linéaire qui nous permet d'estimer la valeur aberrante trouvée en utilisant les autres variables du groupe. Cette méthode permet un remplacement plus contextuel en tenant compte des relations entre la variable cible et les autres variables.
— **Remplacement par discretisation par frequence** : Apres avoir etablit les intervalles des classes de l'attribut contenant des valeurs abberantes, Nous avons choisis de remplacer la valeur aberrante trouvée par la mediane de l'intervalle ou il se trouve. Cette méthode de discrétisation est insensible au outliers contrairement à Equal Width (Plus de details dans la Discretisation à la partie 3).

1.1.3.2 Réduction des données horizontales / verticales.

Réduire les dimensions de notre dataset a une influence considérablement sur les performances des modèles d'apprentissage automatique de plusieurs manières : en réduisant le temps d'exécution et la mémoire consommée, améliorant les résultats en prévenant

l'overfitting et en réduisant les redondances.

Pour ce qui est de la réduction horizontale, nous avons éliminé les lignes redondantes ainsi, 2 instances ont été supprimées. Concernant la réduction verticale, nous avons calculé la corrélation de Pearson ente chaque paire d'attributs et parmi celles dont la valeur depassent le seuil (paramètre empirique à expérimenter en partie 2), l'une des deux est retirée.

1.1.3.3 Normalisation des données :

La dernière étape de notre prétraitement consiste à normaliser nos données. Elle implique de ramener à une échelle commune tous les attributs de notre dataset afin de s'assurer qu'ils contribuent de manière équitable lors de l'exécution des algorithmes de clustering. Sans cela, certaines domineront plus que d'autres simplement en raison de la disparité entre ses valeurs et ceux des autres attributs.

Nous avons appliqué les 2 méthodes de normalisation suivantes :

1.1.3.3.a Méthode Min-Max

Nous permet de normaliser les données au sein d'un intervalle dont les bornes sont choisies en entrée, voici sa formule :

$$Valeur_{(i,new)} = \frac{Valeur_{(i,old)} - Valeur_{(min,old)}}{Valeur_{(max,old)} - Valeur_{(min,old)}} \left(Valeur_{(max,new)} - Valeur_{(min,new)}\right) + Valeur_{(min,new)} \quad (1.1)$$

1.1.3.3.b Méthode z-score.

Elle quantifie à quel point chaque valeur est éloignée de la moyenne de l'attribut courant en termes d'unités d'écart-type comme le montre la formule suivante

$$Valeur_{(i,new)} = \frac{Valeur_{(i,old)} - Valeur_{(mean,old)}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} \left|Valeur_{(i,old)} - Valeur_{(mean,old)}\right|^2}} \quad (1.2)$$

1.2 Données temporelles : Dataset 2

1.2.1 Description du dataset

Le deuxième dataset de cette partie "**dataset2**" fournit des informations détaillées sur la propagation du COVID-19 de 2019 à 2023 dans plusieurs régions aux Etats Unis, identifiées par un code postal. Ce dataset comprend 11 colonnes, chacune décrivant un attribut spécifique de la situation épidémiologique au fil du temps. Le tableau 2 représente une description de chacun d'entre eux.

Attribut	**Description**	**Type**
ZCTA	Identifiant unique d'une zone géographique	Caractère
Time Period	Période temporelle pendant laquelle les données ont été enregistrées	Entier
Population	Nombre de personnes résidant dans la zone géographique considérée	Entier
Start Date	Dates de début de chaque période temporelle, indiquant le début des enregistrements	Date
End Date	Dates de fin de chaque période temporelle, indiquant la fin des enregistrements	Date
Case Count	Nombre total de cas de COVID-19 enregistrés au cours de la période spécifiée	Entier
Test Count	Nombre total de tests de dépistage effectués pendant la période considérée	Entier
Positive Tests	Nombre de tests qui ont donné un résultat positif pour le COVID-19	Entier
Case Rate	Taux de cas	Réel
Test Rate	Taux de tests	Réel
Positivity Rate	Taux de positivité	Réel

TABLE 1.3 – Tableau descriptif des attributs du Dataset 2.

Chaque ligne du dataset représente une période temporelle spécifique et fournit des informations détaillées sur le nombre de cas, les tests effectués, et d'autres métriques

associées à la propagation du COVID-19 dans une zone géographique spécifique.

1.2.2 Prétraitement

Afin d'assurer la qualité et la fiabilité de nos analyses ultérieures, nous avons effectué deux traitements sur ce dataset : le traitement des valeurs manquantes et le traitement des données aberrantes.

1.2.2.1 Traitement des Valeurs Manquantes

Nous avons utilisé un processus méthodique pour trouver les valeurs manquantes dans le Dataset 2. Tout d'abord, compte tenu de la nature des données temporelles du dataset, nous avons identifié les valeurs manquantes et sélectionné les méthodes de remplacement appropriées. Nous avons garder les 2 methodes de la partie 1 c'est à dire remplacement par mode et par moyenne de meme classe à part qu'ici la classe represente la period pour prendre en compte la temporalité dans nos données

Ensuite pour les attributs de type date un traitement speciale a lieu. D'abord nous rendons les formats des dates uniforme ensuite pour les dates sans année, nous avons remarquer une relation entre l'attribut time period et date debut ce qui nous permet de retrouver l'année manquante en mettant l'année de date debut des instances ayant le meme period

1.2.2.2 Traitement des Données Aberrantes

Ici les methodes definis en partie 1 : par discretisation et par regression lineaire fonctionnent tres bien dans notre dataset temporelle, cependant, il est plus judicieux d'utiliser la regression lineaire dans ce cas la car discretiser des attributs tel que positif test et case count.. nous fera perdre en pertinence d'information

Pour ce qui est des attributs de type dates, aucun remplacement n'a lieu car les dates sont valides et comprises entre 2019 et 2023 comme cité à l'enoncé.

1.2.3 Visualisation

1.2.3.1 La distribution du nombre total des cas confirmés et tests positifs par zones

1.2.3.1.a Graphe

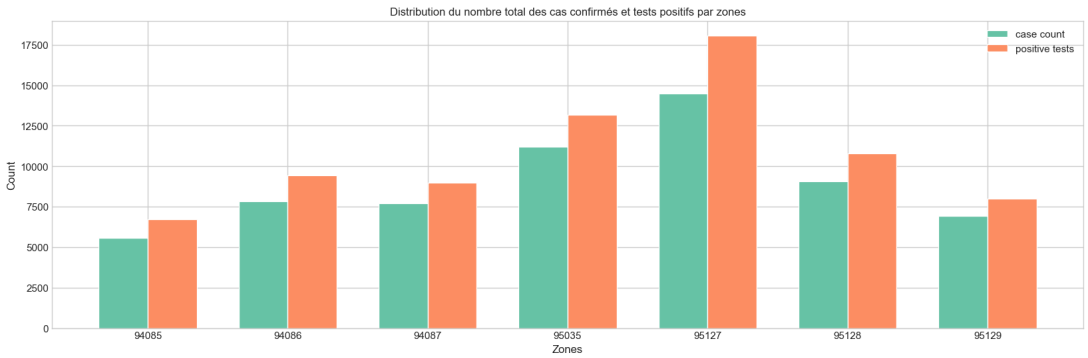

FIGURE 1.15 – Graphe de distribution du nombre total des cas confirmés et tests positifs par zones.

1.2.3.1.b Analyse

En analysant le graphe de la Figure 1.17, nous avons remarqué des variations significatives entre les différentes zones dans le dataset, en termes de cas confirmés et de tests positifs. La zone 95127 a le nombre de cas confirmés et de tests positifs le plus élevé, suivie de près par la zone 95035. Les chiffres des zones 94085, 94086 et 95128 sont intermédiaires, tandis que les chiffres des zones 94087 et 95129 sont plus faibles. Ceci peut être expliqué par plusieurs facteurs tels que la densité de population, l'accès aux tests, et les mesures de santé publique dans cette zone.
Nous avons observés aussi une corrélation positive entre les cas confirmés et les tests positifs pour toutes les zones.

1.2.3.2 L'évolution des tests COVID-19, tests positifs et le nombre de cas confirmés au fil du temps pour la zone 94085

1.2.3.2.a Graphes

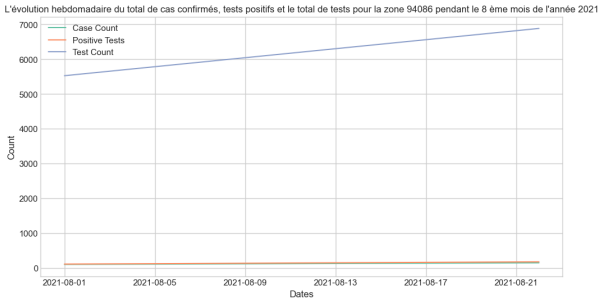

FIGURE 1.16 – Graphe de l'évolution hebdomadaire du nombre de cas, tests positifs et tests effectués dans la zone 94086 en août 2021

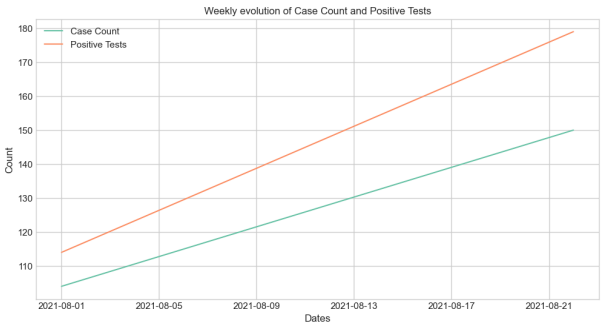

FIGURE 1.17 – Graphe de l'évolution hebdomadaire du nombre de cas et tests positifs dans la zone 94086 en août 2021

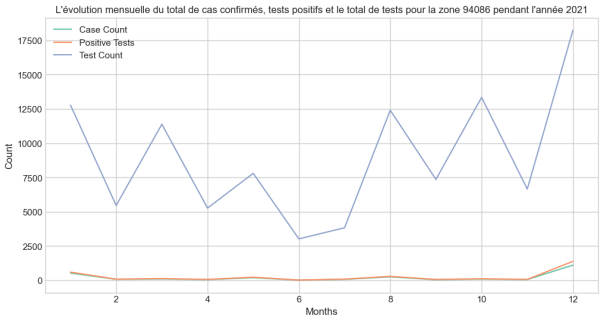

FIGURE 1.18 – Graphe de l'évolution mensuelle du nombre de cas, tests positifs et tests effectués dans la zone 94086 en 2021

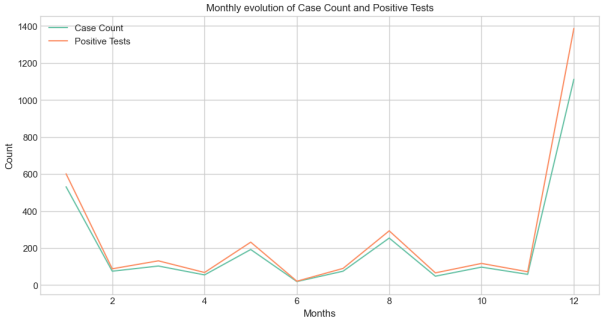

FIGURE 1.19 – Graphe de l'évolution mensuelle du nombre de cas et tests positifs dans la zone 94086 en 2021

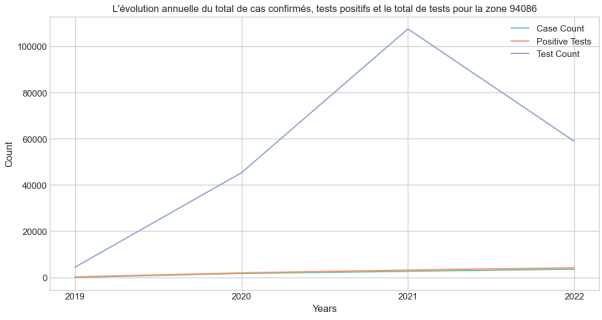

FIGURE 1.20 – Graphe de l'évolution annuelle du nombre de cas, tests positifs et tests effectués dans la zone 94086

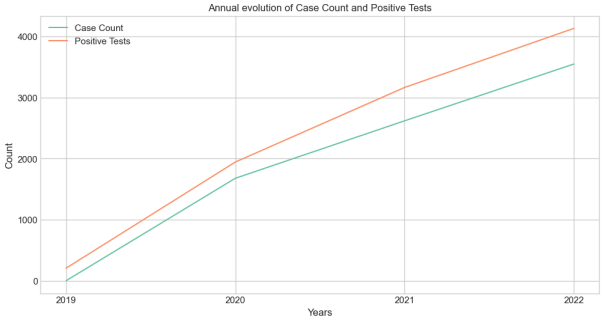

FIGURE 1.21 – Graphe de l'évolution annuelle du nombre de cas, tests positifs dans la zone 94086

1.2.3.2.b Analyse

Les résultats des différentes zones notamment la zone 94086 nous ont indiquer une croissance respective du nombre de cas confirmés, nombre de tests positifs et nombre de tests effectués. Cette augmentation montre une évolution de la situation épidémiologique dans la zone 94086 au cours de ces semaines.

Nous avons remarqué aussi que le nombre total de tests atteint des valeurs relativement élevées pour les 3 périodes. En comparaison, les attributs de cas confirmés et de tests positifs ont des valeurs beaucoup plus basses. Cette disparité indique une grande proportion de tests négatifs.

Nous avons également observé que les valeurs de tests positifs sont généralement légèrement plus élevées que celles des cas confirmés pour chaque période. Ceci suggère la possibilité de faux positifs dans les tests.

1.2.3.3 Distribution des cas COVID-19 positifs par zone et par année

1.2.3.3.a Graphe

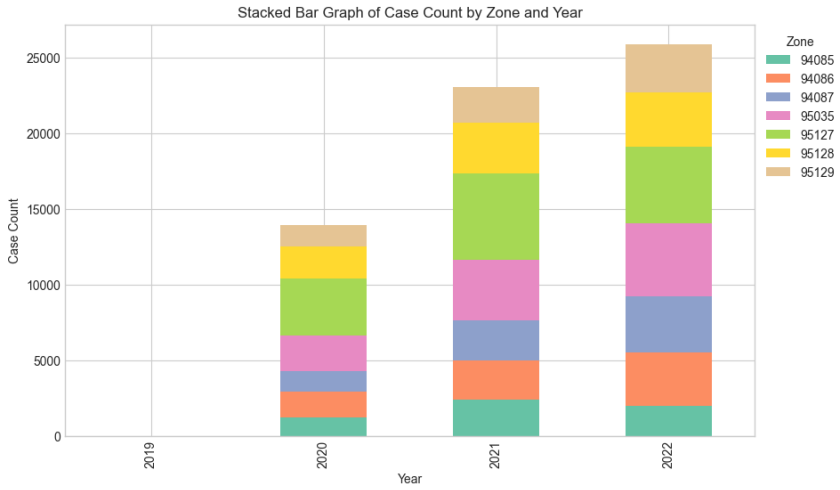

FIGURE 1.22 – Graphe de distribution des cas COVID-19 positifs par zone et par année.

1.2.3.3.b Analyse

Plusieurs tendances distinctes se dégagent en examinant la répartition des cas de COVID-19 par zone. Nous distingons aussi des zones de concentration récurrents qui sont 95035, 95127 et 95128 qui peuvent être expliqués par la population élevée de ses zones. Les

zones 94085 et 95127 ont connu une forte croissance en 2020 et 2021, mais elles ont connu une légère baisse en 2022. Cependant, les autres zones sont en constante augmentation au fil du temps, ce qui révèle des différences importantes dans l'impact de la pandémie en fonction des zones géographiques.

Une analyse par année, montre que le nombre de cas a augmenté considérablement en 2020, atteignant son pic en 2022 ce qui souligne la persistance du virus au fil des années. En

En résumé, l'impact de la pandémie de COVID-19 varie considérablement d'une région à l'autre et montre des variations notables chaque année. Les facteurs géographiques, socio-économiques et les stratégies de gestion de la pandémie spécifiques à chaque région pourraient influencer ces tendances divergentes.

1.2.3.4 Le rapport entre la population et le nombre de tests effectués

1.2.3.4.a Graphe

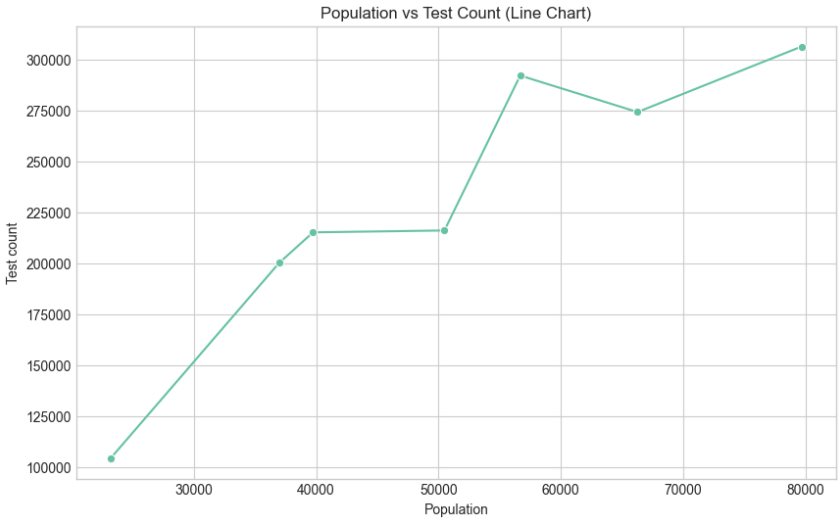

FIGURE 1.23 – Graphe représentant le rapport entre la population et le nombre de tests effectués.

1.2.3.4.b Analyse

L'analyse du graphique montrant le rapport entre la population et le nombre de tests effectués indique une augmentation rapide du nombre de tests par rapport à la population, suivie d'une tendance à la stagnation ou à une croissance moins prononcée dans les zones

densément peuplées. Les tests disponibles, les politiques de santé publique et certaines caractéristiques démographiques peuvent expliquer cette variation.

1.2.3.5 Les 5 zones les plus fortement impactées par le coronavirus

1.2.3.5.a Graphe

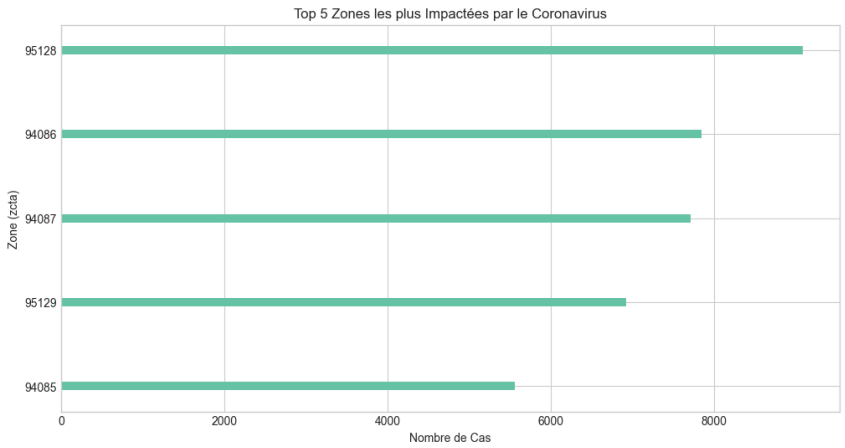

FIGURE 1.24 – Graphe de distribution des cas COVID-19 positifs par zone et par année.

1.2.3.5.b Analyse

D'après les résultats obtenus du graphe, les 5 zones les plus touchées par le coronavirus sont les suivantes, classées de la plus touchée à la moins touchée :

1. **95128** avec 9082 cas confirmés.
2. **94086** avec 7844 cas confirmés.
3. **94087** avec 7710 cas confirmés.
4. **95129** avec 6917 cas confirmés.
5. **94085** avec 5555 cas confirmés.

1.2.3.6 Le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant une periode choisie

1.2.3.6.a Graphes

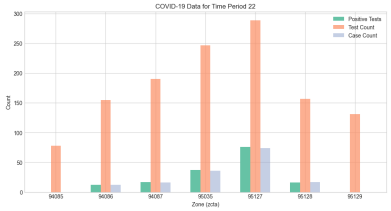

FIGURE 1.25 – Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 22.

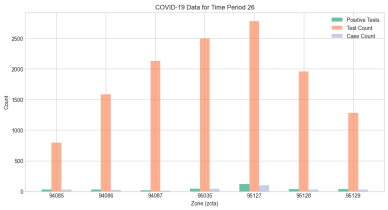

FIGURE 1.26 – Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 26.

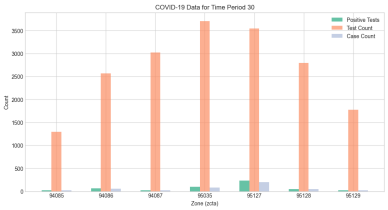

FIGURE 1.27 – Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 30.

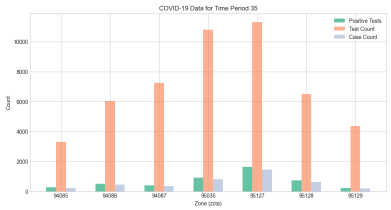

FIGURE 1.28 – Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 35.

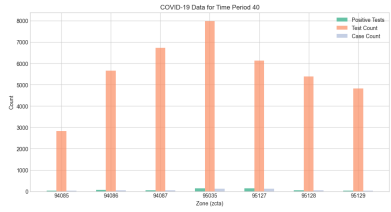

FIGURE 1.29 – Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 40.

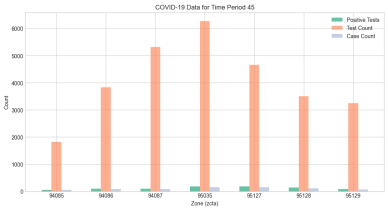

FIGURE 1.30 – Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 45.

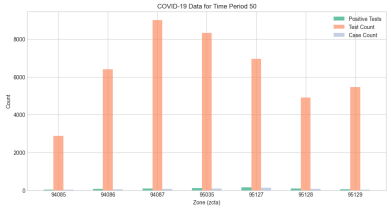

FIGURE 1.31 – Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 50.

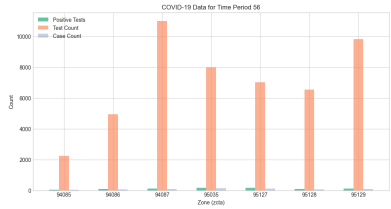

FIGURE 1.32 – Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 56.

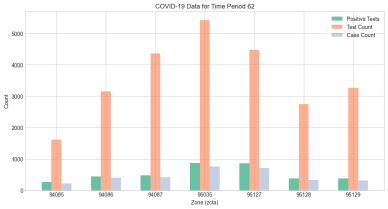

FIGURE 1.33 – Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 62.

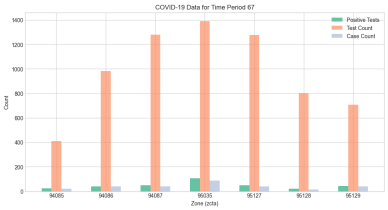

FIGURE 1.34 – Graphe représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour chaque zone pendant la période 67.

1.2.3.6.b Analyse

D'après l'analyse des 10 graphes ci-dessus représentant le rapport entre les cas confirmés, les tests effectués et les tests positifs pour différentes périodes dans chaque zone, nous observons que toutes les zones semblent suivre une trajectoire similaire avec une augmentation ou une baisse du nombre de tests effectués, des tests positifs et des cas confirmés selon la période choisie.

Nous remarquons aussi une corrélation positive entre le nombre de tests effectués, le nombre de cas confirmés et nombre tests positifs. Ceci peut etre expliqué par le fait que le nombre de tests effectués englobe à la fois les cas confirmés et les tests positifs, établissant ainsi un lien entre ces trois attributs.

Notons aussi que certaines zones ont des ratios de tests positifs, tests effectués et tests positifs plus élevés par rapport aux autres. Cela est dû à divers facteurs notamment la population.

Chapitre 2

Extraction de motifs fréquents, règles d'associations et corrélations

2.1 Dataset3

2.1.1 Description du dataset

Dans cette troisieme partie du projet, les données concernent également le sol avec cette fois seulement 295 instances et des caracteristiques relatifs au climat, la vegetation .. expliquées avec plus details dans le tableau 2.1 suivant :

Attribut	**Description**	**Instervalles des valeurs**
Température	Température en degrees celsius	[20.05-29.87]
Humidity	Humidité en	[80.12-99.98]
Rainfall	Precipitation	[131.09-298.56]
Soil	Type de sol	{alluvial, clay loam, sandy, laterite, silty clay, coastal, Clayey}
Crop	Type de vegetation	{Coconut, rice}
Fertilizer	Engrais utilisé	{Good NPK, MOP, Urea, DAP}

TABLE 2.1 – Tableau descriptif des attributs du Dataset3

Les informations qui seront tiré de ce dataset nottament les associations deduites sont cruciale dans la gestion des ressources environnementales et agricoles.

2.1.2 Discrétisation de l'attribut Température

Afin d'utiliser l'attribut "Température" dans ce qui suit, il est nécessaire de la discrétiser, nous l 'avons fait en 2 manières :

2.1.2.1 En classes d'effectifs égaux (Equal Frequency)

Ici, chaque classe a la meme frequence mais avec des largeur differentes. On détermine la fréquence en divisant le nombre de valeurs de l'attribut par le nombre de classes souhaité.

2.1.2.2 En classes d'amplitudes égales (Equal Width)

Dans cette approche, les classes ont toute la même étendue (intervalles de même taille) que nous déterminons à partir du nombre de classes k choisis par l'utilisateur $\frac{Valeur_{\max} - Valeur_{\min}}{k}$.

Critère	**Equal Width (Largeur égale)**	**Equal Frequency (Fréquence égale)**
Avantages	Simple à implementer et à interpréter, Preserve la distribution des données	Assure une distribution égale des valeurs dans chaque intervalle, insensible au valeurs abberantes
Inconvénients	Peut générer des classes vide surtout si il y a des outliers, Peut ne pas refléter la densité des données	Déforme la distribution des données, Nécessite des calculs plus complexes pour déterminer les intervalles

Comparaison entre les 2 methodes de discrétisation Le choix entre Equal width et EqualFrequency dépend du cas d'utilisation et de l'objectif de la discrétisation. Comme nous l'avons vu dans la partie "Remplacement de valeurs aberrantes" , la méthode de classe de fréquences égales est plus adapté car elle est insensible au Outliers

2.1.3 Extraction des motifs fréquents : Apriori

En Datamining, l'extraction des motifs fréquents est une tache courante visant à trouver des associations récurrentes (patterns) dans des ensembles de données volumineux. Le but étants d'y découvrir des informations pertinentes et implicites, ce qui est en soit le but ultime du domaine.

Avant d'appliquer cette méthode, nous devons d'abord fixer la valeur le support minimum (paramètre empirique) et définir la table de transaction.

L'état initial d'un dataset n'est souvent pas adapté à la structure d'une table de transanctions. Dans ce cas, il est nécessaire de la transformer.

Dans le cadre de notre dataset, une transaction représente les instances car il s'agit de sols et chacun est unique avec ses propres caracteristiques donc rien à modifier de ce côté-là, tandis que les items correspondent aux différentes valeurs des attributs qui sont de type catégorique. Lors de la discretisation de l'attribut Temperature nous avons choisis k=3 sans risque de perte d'information car les valeurs ont un ecartype tres faible (les donneées varie entre 20 et 29 degres) et suit une distribution symetrique.

Maintenant que nos prérequis sont satisfait, nous allons expliquer l'algorithme permettant d'effectuer notre tâche à savoir l 'algorithme Apriori

La 1ère opération est celle de la génération de candidats. Pour k=1 il suffit de retourner simplement la liste de tous les items sans répétitions, sinon elle retourne les combinaisons d'items de taille k à partir des items des motifs fréquents de taille k-1. Cependant à partir de k=3, il faut effectuer le Pruning, qui consiste à filtrer ces combinaisons en ne conservant que celles qui contiennent des sous-combinaisons frequentes. Autrement dit en vérifiant si toute les sous-combinaisons se trouvent dans la liste des motifs fréquents en entrée.

Algorithm 1 Generateur Ck

```
1: procedure GenerateurCk(k, L, itemliste)
2:     if k == 1 then
3:         return itemListe
4:     else
5:         C ← liste_vide()
6:         if longueur(L) == 0 then
7:             return C
8:         end if
9:         liste_items_uniques ← ensemble des éléments uniques de tous les tuples des L.cles()
10:        if longueur(liste_items_uniques) < k then
11:            return C
12:        end if
13:        combinations_list ← combinations(liste_items_uniques, k)
14:        if k == 2 then
15:            return combinations_list
16:        end if                                                    ▷ Le pruning
17:        for combi in combinations_list do
18:            existe ← vrai
19:            sous_combinations_list ← combinations(combi, k − 1)
20:            for sous_combi in sous_combinations_list do
21:                if sous_combi not in L.cles() then
22:                    existe ← faux
23:                    break
24:                end if
25:            end for
26:            if existe == vrai then
27:                C.ajouter(combi)
28:            end if
29:        end for
30:        return C
31:    end if
32: end procedure
```

Après avoir généré les candidats de taille k, nous calculons leurs supports respectifs en parcourant la table de transaction une seule fois transaction par transaction. Dès que l'une d entre elles contient un sous-ensemble d'item correspondant à un candidat, nous incrémentons le compteur de ce dernier.

A la fin nous divisant ces valeurs (nombre de transactions contenant notre itemset) par le nombre totale de transactions dans notre table.

Algorithm 2 Calculateur Support

```
1: procedure CalculateurSupport(C, dataset)
2:     dico ← {val : 0 for val in C}
3:     for ligne in dataset do
4:         combinations_list ← combinations(ligne, len(dico.keys()[0]))
5:         for val in combinations_list do
6:             if val in dico then
7:                 dico[val] ← dico[val] + 1
8:             end if
9:         end for
10:    end for
11:    dico ← {key : val/len(dataset) for key, val in dico.items()}
12:    return dico
13: end procedure
```

Enfin, la dernière étape de génération des motifs fréquents de taille k est de filtrer les candidats selon leurs supports comme présenté dans le pseudo-code suivant :

Algorithm 3 Generateur Lk

```
1: procedure GenerateurLk(C, supp_min)
2:     l ← {}
3:     for each item c in C do
4:         if c.valeur() ≥ supp_min then
5:             l.ajouter(c)
6:         end if
7:     end for
8:     return l
9: end procedure
```

À présent, nous pouvons nous attaquer à l'algorithme principal d'Appriori. Comme nous pouvons le constater dans le pseudo-code suivant, il consiste à générer des candidats de taille k, calculer leurs supports respectifs et ne garder que ceux supérieurs au Supmin. Ce processus est répété de manière itérative en incrémentant à chaque fois la taille k et ceci jusqu'à qu'il soit impossible d'en générer de nouveaux.

Algorithm 4 Apriori Algorithm

```
1: procedure APRIORI(min_supp, dataset)
2:     L ← liste_vide()
3:     k ← 1
4:     C ← generateur_Ck(k, None, dataset)
5:     while C non vide do
6:         S ← calculateur_support(C, dataset)
7:         l ← generateur_Lk(S, min_supp)
8:         if l non vide then
9:             L.ajouter(l)
10:        end if
11:        k ← k + 1
12:        C ← generateur_Ck(k, l, None)
13:    end while
14:    return L
15: end procedure
```

2.1.4 Extraction des règles d'associations sous plusieurs mesures

Les motifs fréquents étant identifiés, nous allons maintenant trouver les corrélations existantes au sein de chacun en extrayant les règles d'association fréquentes.

Une règle d'association s'écrit sous la forme d'une implication logique entre deux itemset $A \Rightarrow B$,sémantiquement cela veut dire,nous associons à la présence de l'itemset à l'antecedant de la regle, la présence de l'itemset à la conclusion de la regle

Ainsi, nous gênerons toutes les règles possibles pour chaque itemset fréquent (en combinant les sous-motifs dans l'antecedant et conséquent) cela nous donne 2∧(k-2) règles pour un motif de taille k

Ensuite nous calculons la valeur de mesure de corrélation de chaque règle et ne gardons que celles dépassant un seuil minimal.

Dans le cadre de notre projet, nous avons choisi les 4 mesures suivantes :

2.1.4.1 Confiance

Indique la probabilité conditionnelle de la présence du conséquent sachant la présence de l'antécédant dans une règle d'association.Mais peu causer des ambiguités dans les

resultats obetnus (2 regles contradictoires avec une haute confiance)

$$\text{Confiance}(A \Rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)} \tag{2.1}$$

2.1.4.2 Lift

tente d'équilibrer le support et la confiance. Le lift mesure l'augmentation de la probabilité du conséquent sachant l'antécédent par rapport à la probabilité sans l'antécédent. A eviter quand il y a beaucoup de transaction nulles

$$\text{Lift}(A \Rightarrow B) = \frac{\text{Confiance}(A \Rightarrow B)}{\text{Support}(B)} \tag{2.2}$$

2.1.4.3 Cosine

Dans le contexte des règles d'association, il est utilisé pour évaluer à quel point deux ensembles d'items co-occurrent fréquemment.

$$\text{Cosine}(A, B) = \frac{\text{Support}(A \cup B)}{\sqrt{\text{Support}(A) \times \text{Support}(B)}} \tag{2.3}$$

2.1.4.4 Jaccard

Évalue la similarité entre deux ensembles en mesurant le rapport du nombre d'éléments communs sur le nombre total d'éléments distincts.

$$\text{Jaccard}(A, B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A) + \text{Support}(B) - \text{Support}(A \cup B)} \tag{2.4}$$

2.1.4.5 Kulczynski

Mesure la similitude entre deux itemset en considérant à la fois les éléments communs et ceux exclusifs des ensembles.

$$\text{Kulczynski}(A, B) = \frac{1}{2}\left(\frac{\text{Support}(A \cup B)}{\text{Support}(A)} + \frac{\text{Support}(A \cup B)}{\text{Support}(B)}\right) \tag{2.5}$$

Le choix de la mesure de corrélation à utiliser dépend fortement de l'objectif de l'analyse, le type d'analyse, chaque mesure capture différents aspects de la relation entre

les 2 ensembles de notre règle et a ses propres avantages et desavantages.

2.1.5 Résultats de l'expérimentation des différentes valeurs de MinSupp et MinConf

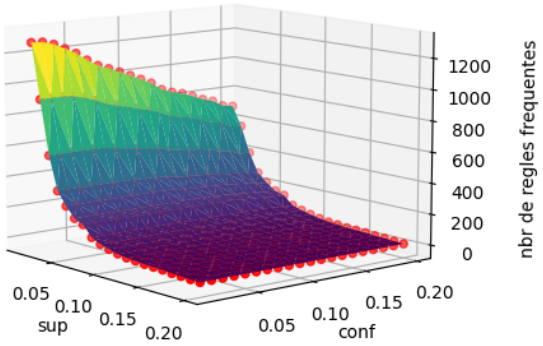

FIGURE 2.1 – Graphe 3D du nombre de règles fréquentes générées par rapport au support minimum et à la confiance minimale.

Le graphe 3D suivant décrivant le nombre de règles générées en fonction du confmin et supmin suivant révèle 2 choses.

En premier lieu, les valeurs de spumin influencent le nombre de règles fréquentes de manière telle que leur diminution entraîne une augmentation du nombre de règles.

Alors que les valeurs de confmin n'ont d'impact sur le nombre de règles fréquentes que lorsque supmin est très petit, dans ce scénario spécifique, une réduction de confmin conduit à une croissance proportionnelle du nombre de règles.

Cela provient du fait qu'un supmin relativement élevé engendrera peu de motifs fréquents, c'est pourquoi peu importe la valeur de confmin dans ce cas le nombre de règles fréquentes sera minime.

La raison pour laquelle les valeurs de supmin et confmin sont de l 'ordre de 0.1%-0.01% est dû à la taille conséquente de notre table de transaction qui ajoute de la diversité dans les motifs et diminue la probabilité qu'un d'eux se trouvent dans un nombre considérable de transactions.

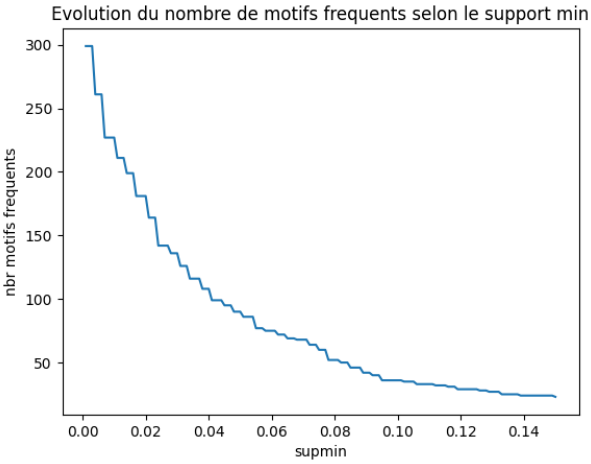

FIGURE 2.2 – Graphe de l'évolution du nombre de motifs fréquents selon supmin.

Ce graphe montre que plus la valeur du support minimum augmente, plus le nombre de motifs générés décroît, indiquant ainsi une claire corrélation négative entre les deux. Cela est dû au fait qu'une grande valeur de supmin filtrera plus sévèrement les motifs initiaux laissant passer un nombre plus petit.

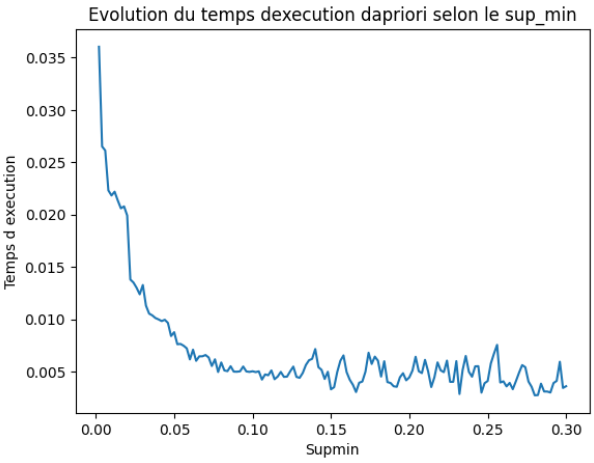

FIGURE 2.3 – Graphe de l'évolution du temps d'exécution selon supmin.

Nous constatons à partir de cette courbe que le temps d'exécution de l'algorithme Appriori est inversement proportionnel à la valeur du support minimum, chutant de 0.035 s pour un Supmin de 0.001 à moins de 0.005 s où il se stabilise au delà 0.1 de supmin. Nous expliquons cela par le fait qu'avec une valeur de Supmin relativement grande, plus il y aura moins de motifs fréquents à traiter (comme vu dans le graphe précédant).

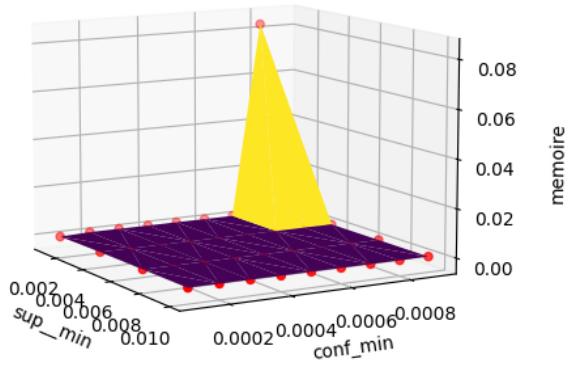

FIGURE 2.4 – Graphe l'évolution de l'espace mémoire alloué à l'algorithme Apriori et les règles d'associations selon le support minimum et la confiance minimale.

Les résultats relatifs aux ressources spatiales allouées durant l'exécution de l'algorithme Appriori sont en accord avec les résultats précédents.La consommation de mémoire est très basse globalement, à l'exception du cas ou les valeurs du support minimum et confiance minimale sont très faible, car ainsi, le nombre de motifs et de règles retenus deviennent énormes conduisant à une explosion combinatoire de l'espace utilisé pour stocker ces derniers.

Chapitre 3

Conclusion et perspectives

L'analyse et le prétraitement de dataset de différent types (statique et temporel), l'extraction des motifs et regles d'associations frequentes ainsi que les resultats des experimentations sur l'algorithme appriori nous a permis de resortir avec ces conclusions :

- L'analyse visuelle des attributs d'un dataset s'avère utile pour connaître la distribution des attributs et la corrélation entre les données
- Dans un dataset les valeurs manquantes se manifeste sous plusieurs formes vide, signe ou lettre au lieu de chiffre... Cela doit être pris en compte lors de leur détection.
- le choix de méthode de remplacement de valeurs manquantes et aberrantes doit être pertinent : remplacer par smoothing après avoir discrétisé par largeurs égales ne remplace pas les outiller, car Equal Width est sensible aux valeurs aberrantes.
- L'analyse et le prétraitement d'un dataset de type temporelle se différencie d'un dataset statique à cause du besoin d'adaptation par rapport à la dimension temporelle (graphes et méthodes de remplacement des valeurs manquantes et aberrantes spécifiques)
- Il est important de bien définir à quoi correspond une transaction et un item (et de discrétiser un attribut si besoin) afin de bien adapter notre dataset à la structure d'une table de transaction, avant l'exécution de l'algorithme Appriori.
- Le support minimum et confiance minimale sont des paramètres empiriques à fixer selon nos objectifs et après avoir effectué des expérimentations. Ils sont inversement proportionnels au nombre de motifs, règles fréquente ainsi que les ressources temporelle et spatiales déversées.
- Le choix de mesure de corrélations durant l'extraction des règles d'association fréquente est important, même si la confiance et la plus connue, elle n'est pas toujours le meilleur choix causant des ambiguïtés dans certains cas.

Annexe

Support minimum	**Temps d'exécution (s)**
0.3	0.0036019086837768555
0.298	0.0034363269805908203
0.296	0.005937981605529785
0.294	0.004112625122070312
0.292	0.0039030075073242187
0.29	0.0030003786087036133
0.288	0.003092670440673828
0.286	0.0030974388122558595
0.284	0.003845524787902832
0.282	0.002753591537475586
0.27999999999999997	0.0027333498001098633
0.27799999999999997	0.0035440921783447266
0.27599999999999997	0.004053235054016113
0.27399999999999997	0.0054159402847290036
0.27199999999999996	0.005620694160461426
0.26999999999999996	0.004890847206115723
0.26799999999999996	0.0040825605392456055
0.26599999999999996	0.0033171415328979493
0.26399999999999996	0.003920125961303711
0.26199999999999996	0.003582763671875
0.25999999999999995	0.004069089889526367
0.25799999999999995	0.003955316543579101
0.25599999999999995	0.007555413246154785
0.25399999999999995	0.006713080406188965
0.25199999999999995	0.005774641036987304
0.24999999999999994	0.004099321365356445
0.24799999999999994	0.0038920164108276365

Support min	**Confiance min**	**Nombre règles**
0.01	0.01	1270.0
0.01	0.02	1265.0
0.01	0.03	1243.0
0.01	0.04	1207.0
0.01	0.05	1169.0
0.01	0.060	1117.0
0.01	0.069	1073.0
0.01	0.08	1044.0
0.01	0.09	1014.0
0.01	0.099	982.0
0.01	0.11	948.0
0.01	0.12	914.0
0.01	0.13	862.0
0.01	0.14	835.0
0.01	0.150	817.0
0.01	0.16	790.0
0.01	0.17	760.0
0.01	0.180	738.0
0.01	0.19	698.0
0.01	0.2	685.0

0.02	0.01	914.0
0.02	0.02	914.0
0.02	0.03	914.0
0.02	0.04	908.0
0.02	0.05	889.0
0.02	0.060	863.0
0.02	0.0699	829.0
0.02	0.08	809.0
0.02	0.09	792.0
0.02	0.0999	771.0
0.02	0.11	753.0
0.02	0.12	728.0
0.02	0.13	696.0
0.02	0.14	678.0
0.02	0.150	664.0
0.02	0.16	642.0
0.02	0.17	615.0
0.02	0.180	603.0
0.02	0.19	566.0
0.02	0.2	559.0
0.03	0.01	564.0

0.03	0.02	564.0
0.03	0.03	564.0
0.03	0.04	564.0
0.03	0.05	564.0
0.03	0.060	561.0
0.03	0.0699	550.0
0.03	0.08	539.0
0.03	0.09	528.0
0.03	0.099	519.0
0.03	0.11	508.0
0.03	0.12	488.0
0.03	0.13	478.0
0.03	0.14	465.0
0.03	0.150	461.0
0.03	0.16	449.0
0.03	0.17	426.0
0.03	0.180	419.0
0.03	0.19	390.0
0.03	0.2	387.0
0.04	0.01	348.0
0.04	0.02	348.0

0.04	0.03	348.0
0.04	0.04	348.0
0.04	0.05	348.0
0.04	0.060	348.0
0.04	0.0699	348.0
0.04	0.08	344.0
0.04	0.09	341.0
0.04	0.0999	336.0
0.04	0.11	333.0
0.04	0.12	319.0
0.04	0.13	314.0
0.04	0.14	307.0
0.04	0.150	304.0
0.04	0.16	298.0
0.04	0.17	286.0
0.04	0.180	281.0
0.04	0.19	264.0
0.04	0.2	261.0
0.05	0.01	264.0
0.05	0.02	264.0
0.05	0.03	264.0

0.05	0.04	264.0
0.05	0.05	264.0
0.05	0.060	264.0
0.05	0.0699	264.0
0.05	0.08	264.0
0.05	0.09	264.0
0.05	0.099	262.0
0.05	0.11	262.0
0.05	0.12	251.0
0.05	0.13	248.0
0.05	0.14	245.0
0.05	0.150	242.0
0.05	0.16	236.0
0.05	0.17	229.0
0.05	0.180	224.0
0.05	0.19	210.0
0.05	0.2	207.0
0.060000000000000005	0.01	198.0
0.060000000000000005	0.02	198.0
0.060000000000000005	0.03	198.0
0.060000000000000005	0.04	198.0

0.060000000000000005	0.05	198.0
0.060000000000000005	0.060000000000000005	198.0
0.060000000000000005	0.06999999999999999	198.0
0.060000000000000005	0.08	198.0
0.060000000000000005	0.09	198.0
0.060000000000000005	0.09999999999999999	198.0
0.060000000000000005	0.11	198.0
0.060	0.12	198.0
0.060	0.13	196.0
0.060	0.14	193.0
0.0605	0.150	192.0
0.060	0.16	186.0
0.060	0.17	181.0
0.060	0.180	176.0
0.060	0.19	166.0
0.060	0.2	163.0
0.0699	0.01	172.0
0.0699	0.02	172.0
0.0699	0.03	172.0
0.0699	0.04	172.0
0.0699	0.05	172.0

0.0699	0.060	172.0
0.0699	0.0699	172.0
0.0699	0.08	172.0
0.0699	0.09	172.0
0.0699	0.0999	172.0
0.0699	0.11	172.0
0.0699	0.12	172.0
0.0699	0.13	172.0
0.0699	0.14	170.0
0.0699	0.150	169.0
0.0699	0.16	163.0
0.0699	0.17	158.0
0.0699	0.180	155.0
0.0699	0.19	148.0
0.0699	0.2	146.0
0.08	0.01	104.0
0.08	0.02	104.0
0.08	0.03	104.0
0.08	0.04	104.0
0.08	0.05	104.0
0.08	0.060000000000000005	104.0

0.08	0.06999999999999999	104.0
0.08	0.08	104.0
0.08	0.09	104.0
0.08	0.09999999999999999	104.0
0.08	0.11	104.0
0.08	0.12	104.0
0.08	0.13	104.0
0.08	0.14	104.0
0.08	0.15000000000000002	104.0
0.08	0.16	104.0
0.08	0.17	102.0
0.08	0.18000000000000002	100.0
0.08	0.19	93.0
0.08	0.2	91.0
0.09	0.01	76.0
0.09	0.02	76.0
0.09	0.03	76.0
0.09	0.04	76.0
0.09	0.05	76.0
0.09	0.060000000000000005	76.0
0.09	0.06999999999999999	76.0

0.09	0.08	76.0
0.09	0.09	76.0
0.09	0.09999999999999999	76.0
0.09	0.11	76.0
0.09	0.12	76.0
0.09	0.13	76.0
0.09	0.14	76.0
0.09	0.15000000000000002	76.0
0.09	0.16	76.0
0.09	0.17	76.0
0.09	0.18000000000000002	76.0
0.09	0.19	74.0
0.09	0.2	72.0
0.09999999999999999	0.01	58.0
0.09999999999999999	0.02	58.0
0.09999999999999999	0.03	58.0
0.09999999999999999	0.04	58.0
0.09999999999999999	0.05	58.0
0.09999999999999999	0.060000000000000005	58.0
0.09999999999999999	0.06999999999999999	58.0
0.09999999999999999	0.08	58.0

0.09999999999999999	0.09	58.0
0.09999999999999999	0.09999999999999999	58.0
0.09999999999999999	0.11	58.0
0.09999999999999999	0.12	58.0
0.09999999999999999	0.13	58.0
0.09999999999999999	0.14	58.0
0.09999999999999999	0.15000000000000002	58.0
0.09999999999999999	0.16	58.0
0.09999999999999999	0.17	58.0
0.09999999999999999	0.18000000000000002	58.0
0.09999999999999999	0.19	58.0
0.09999999999999999	0.2	57.0
0.11	0.01	44.0
0.11	0.02	44.0
0.11	0.03	44.0
0.11	0.04	44.0
0.11	0.05	44.0
0.11	0.060000000000000005	44.0
0.11	0.06999999999999999	44.0
0.11	0.08	44.0
0.11	0.09	44.0

0.11	0.09999999999999999	44.0
0.11	0.11	44.0
0.11	0.12	44.0
0.11	0.13	44.0
0.11	0.14	44.0
0.11	0.15000000000000002	44.0
0.11	0.16	44.0
0.11	0.17	44.0
0.11	0.18000000000000002	44.0
0.11	0.19	44.0
0.11	0.2	44.0
0.12	0.01	34.0
0.12	0.02	34.0
0.12	0.03	34.0
0.12	0.04	34.0
0.12	0.05	34.0
0.12	0.060000000000000005	34.0
0.12	0.06999999999999999	34.0
0.12	0.08	34.0
0.12	0.09	34.0
0.12	0.09999999999999999	34.0

0.12	0.11	34.0
0.12	0.12	34.0
0.12	0.13	34.0
0.12	0.14	34.0
0.12	0.15000000000000002	34.0
0.12	0.16	34.0
0.12	0.17	34.0
0.12	0.18000000000000002	34.0
0.12	0.19	34.0
0.12	0.2	34.0
0.13	0.01	30.0
0.13	0.02	30.0
0.13	0.03	30.0
0.13	0.04	30.0
0.13	0.05	30.0
0.13	0.060000000000000005	30.0
0.13	0.06999999999999999	30.0
0.13	0.08	30.0
0.13	0.09	30.0
0.13	0.09999999999999999	30.0
0.13	0.11	30.0

0.13	0.12	30.0
0.13	0.13	30.0
0.13	0.14	30.0
0.13	0.15000000000000002	30.0
0.13	0.16	30.0
0.13	0.17	30.0
0.13	0.18000000000000002	30.0
0.13	0.19	30.0
0.13	0.2	30.0
0.14	0.01	24.0
0.14	0.02	24.0
0.14	0.03	24.0
0.14	0.04	24.0
0.14	0.05	24.0
0.14	0.060000000000000005	24.0
0.14	0.06999999999999999	24.0
0.14	0.08	24.0
0.14	0.09	24.0
0.14	0.09999999999999999	24.0
0.14	0.11	24.0
0.14	0.12	24.0

0.14	0.13	24.0
0.14	0.14	24.0
0.14	0.15000000000000002	24.0
0.14	0.16	24.0
0.14	0.17	24.0
0.14	0.18000000000000002	24.0
0.14	0.19	24.0
0.14	0.2	24.0
0.15000000000000002	0.01	22.0
0.15000000000000002	0.02	22.0
0.15000000000000002	0.03	22.0
0.15000000000000002	0.04	22.0
0.15000000000000002	0.05	22.0
0.15000000000000002	0.060000000000000005	22.0
0.15000000000000002	0.06999999999999999	22.0
0.15000000000000002	0.08	22.0
0.15000000000000002	0.09	22.0
0.15000000000000002	0.09999999999999999	22.0
0.15000000000000002	0.11	22.0
0.15000000000000002	0.12	22.0
0.15000000000000002	0.13	22.0

0.15000000000000002	0.14	22.0
0.15000000000000002	0.15000000000000002	22.0
0.15000000000000002	0.16	22.0
0.15000000000000002	0.17	22.0
0.15000000000000002	0.18000000000000002	22.0
0.15000000000000002	0.19	22.0
0.15000000000000002	0.2	22.0
0.16	0.01	20.0
0.16	0.02	20.0
0.16	0.03	20.0
0.16	0.04	20.0
0.16	0.05	20.0
0.16	0.060000000000000005	20.0
0.16	0.06999999999999999	20.0
0.16	0.08	20.0
0.16	0.09	20.0
0.16	0.09999999999999999	20.0
0.16	0.11	20.0
0.16	0.12	20.0
0.16	0.13	20.0
0.16	0.14	20.0

0.16	0.15000000000000002	20.0
0.16	0.16	20.0
0.16	0.17	20.0
0.16	0.18000000000000002	20.0
0.16	0.19	20.0
0.16	0.2	20.0
0.17	0.01	14.0
0.17	0.02	14.0
0.17	0.03	14.0
0.17	0.04	14.0
0.17	0.05	14.0
0.17	0.060000000000000005	14.0
0.17	0.06999999999999999	14.0
0.17	0.08	14.0
0.17	0.09	14.0
0.17	0.09999999999999999	14.0
0.17	0.11	14.0
0.17	0.12	14.0
0.17	0.13	14.0
0.17	0.14	14.0
0.17	0.15000000000000002	14.0

0.17	0.16	14.0
0.17	0.17	14.0
0.17	0.18000000000000002	14.0
0.17	0.19	14.0
0.17	0.2	14.0
0.18000000000000002	0.01	12.0
0.18000000000000002	0.02	12.0
0.18000000000000002	0.03	12.0
0.18000000000000002	0.04	12.0
0.18000000000000002	0.05	12.0
0.18000000000000002	0.060000000000000005	12.0
0.18000000000000002	0.06999999999999999	12.0
0.18000000000000002	0.08	12.0
0.18000000000000002	0.09	12.0
0.18000000000000002	0.09999999999999999	12.0
0.18000000000000002	0.11	12.0
0.18000000000000002	0.12	12.0
0.18000000000000002	0.13	12.0
0.18000000000000002	0.14	12.0
0.18000000000000002	0.15000000000000002	12.0
0.18000000000000002	0.16	12.0

0.18000000000000002	0.17	12.0
0.18000000000000002	0.18000000000000002	12.0
0.18000000000000002	0.19	12.0
0.18000000000000002	0.2	12.0
0.19	0.01	8.0
0.19	0.02	8.0
0.19	0.03	8.0
0.19	0.04	8.0
0.19	0.05	8.0
0.19	0.060000000000000005	8.0
0.19	0.06999999999999999	8.0
0.19	0.08	8.0
0.19	0.09	8.0
0.19	0.09999999999999999	8.0
0.19	0.11	8.0
0.19	0.12	8.0
0.19	0.13	8.0
0.19	0.14	8.0
0.19	0.150	8.0
0.19	0.16	8.0
0.19	0.17	8.0

0.19	0.18000000000000002	8.0
0.19	0.19	8.0
0.19	0.2	8.0
0.2	0.01	6.0
0.2	0.02	6.0
0.2	0.03	6.0
0.2	0.04	6.0
0.2	0.05	6.0
0.2	0.060000000000000005	6.0
0.2	0.06999999999999999	6.0
0.2	0.08	6.0
0.2	0.09	6.0
0.2	0.09999999999999999	6.0
0.2	0.11	6.0
0.2	0.12	6.0
0.2	0.13	6.0
0.2	0.14	6.0
0.2	0.150	6.0
0.2	0.16	6.0
0.2	0.17	6.0
0.2	0.180	6.0

0.2	0.19	6.0
0.2	0.2	6.0

TABLE 3.3 – Tableau de l'évolution du temps d'exécution de l'algorithme Apriori selon le support minimum.

sup min	nbr motifs
0.2	11.0
0.199	11.0
0.198	11.0
0.197	11.0
0.196	13.0
0.195	13.0
0.194	13.0
0.193	13.0
0.192	13.0
0.191	13.0
0.19	13.0
0.189	13.0
0.188	13.0
0.187	13.0
0.186	13.0

0.185	13.0
0.184	13.0
0.183	16.0
0.182	16.0
0.181	16.0
0.18	16.0
0.179	17.0
0.178	17.0
0.177	17.0
0.176	17.0
0.175	17.0
0.174	17.0
0.173	17.0
0.172	17.0
0.17099999999999999	17.0
0.16999999999999998	17.0
0.16899999999999998	19.0
0.16799999999999998	19.0
0.16699999999999998	19.0
0.16599999999999998	19.0
0.16499999999999998	19.0

0.16399999999999998	19.0
0.16299999999999998	19.0
0.16199999999999998	22.0
0.16099999999999998	22.0
0.15999999999999998	22.0
0.15899999999999997	23.0
0.15799999999999997	23.0
0.15699999999999997	23.0
0.15599999999999997	23.0
0.15499999999999997	23.0
0.15399999999999997	23.0
0.15299999999999997	23.0
0.15199999999999997	23.0
0.15099999999999997	23.0
0.14999999999999997	23.0
0.14899999999999997	24.0
0.14799999999999996	24.0
0.14699999999999996	24.0
0.14599999999999996	24.0
0.14499999999999996	24.0
0.14399999999999996	24.0

0.14299999999999996	24.0
0.14199999999999996	24.0
0.14099999999999996	24.0
0.13999999999999996	24.0
0.13899999999999996	24.0
0.13799999999999996	25.0
0.13699999999999996	25.0
0.13599999999999995	25.0
0.13499999999999995	25.0
0.13399999999999995	25.0
0.13299999999999995	25.0
0.13199999999999995	27.0
0.13099999999999995	27.0
0.12999999999999995	27.0
0.12899999999999995	27.0
0.12799999999999995	28.0
0.12699999999999995	28.0
0.12599999999999995	28.0
0.12499999999999994	29.0
0.12399999999999994	29.0
0.12299999999999994	29.0

0.12199999999999994	29.0
0.12099999999999994	29.0
0.11999999999999994	29.0
0.11899999999999994	29.0
0.11799999999999994	31.0
0.11699999999999994	31.0
0.11599999999999994	31.0
0.11499999999999994	32.0
0.11399999999999993	32.0
0.11299999999999993	32.0
0.11199999999999993	32.0
0.11099999999999993	33.0
0.10999999999999993	33.0
0.10899999999999993	33.0
0.10799999999999993	33.0
0.10699999999999993	33.0
0.10599999999999993	33.0
0.10499999999999993	35.0
0.10399999999999993	35.0
0.10299999999999992	35.0
0.10199999999999992	35.0

0.10099999999999992	36.0
0.09999999999999992	36.0
0.09899999999999992	36.0
0.09799999999999992	36.0
0.09699999999999992	36.0
0.09599999999999992	36.0
0.09499999999999992	36.0
0.09399999999999992	40.0
0.09299999999999992	40.0
0.09199999999999992	40.0
0.09099999999999991	42.0
0.08999999999999991	42.0
0.08899999999999991	42.0
0.08799999999999991	46.0
0.08699999999999991	46.0
0.08599999999999991	46.0
0.08499999999999991	46.0
0.08399999999999991	50.0
0.08299999999999991	50.0
0.0819999999999999	50.0
0.0809999999999999	52.0

0.0799999999999999	52.0
0.0789999999999999	52.0
0.0779999999999999	52.0
0.0769999999999999	60.0
0.0759999999999999	60.0
0.0749999999999999	60.0
0.0739999999999999	64.0
0.0729999999999999	64.0
0.0719999999999999	64.0
0.0709999999999999	68.0
0.0699999999999999	68.0
0.0689999999999999	68.0
0.0679999999999999	68.0
0.0669999999999989	69.0
0.0659999999999989	69.0
0.0649999999999989	69.0
0.0639999999999989	72.0
0.0629999999999989	72.0
0.0619999999999989	72.0
0.0609999999999989	75.0
0.0599999999999989	75.0

0.058999999999999886	75.0
0.057999999999999885	75.0
0.056999999999999884	77.0
0.05599999999999988	77.0
0.05499999999999988	77.0
0.05399999999999988	86.0
0.05299999999999988	86.0
0.05199999999999988	86.0
0.05099999999999988	86.0
0.04999999999999988	90.0
0.04899999999999988	90.0
0.047999999999999876	90.0
0.046999999999999875	95.0
0.045999999999999874	95.0
0.04499999999999987	95.0
0.04399999999999987	99.0
0.04299999999999987	99.0
0.04199999999999987	99.0
0.04099999999999987	99.0
0.03999999999999987	108.0
0.03899999999999987	108.0

0.03799999999999987	108.0
0.03699999999999866	116.0
0.03599999999999865	116.0
0.03499999999999865	116.0
0.03399999999999864	116.0
0.03299999999999986	126.0
0.03199999999999986	126.0
0.03099999999999986	126.0
0.02999999999999986	136.0
0.02899999999999986	136.0
0.02799999999999986	136.0
0.02699999999999857	142.0
0.02599999999999857	142.0
0.02499999999999856	142.0
0.02399999999999855	142.0
0.02299999999999854	164.0
0.02199999999999853	164.0
0.02099999999999852	164.0
0.01999999999999985	181.0
0.01899999999999985	181.0
0.01799999999999985	181.0

0.01699999999999985	181.0
0.01599999999999848	199.0
0.01499999999999847	199.0
0.01399999999999846	199.0
0.01299999999999845	211.0
0.01199999999999844	211.0
0.01099999999999843	211.0
0.00999999999999842	227.0
0.00899999999999841	227.0
0.00799999999999984	227.0
0.00699999999999984	227.0
0.00599999999999839	261.0
0.00499999999999838	261.0
0.00399999999999837	261.0
0.00299999999999836	299.0
0.0019999999999998352	299.0
0.0009999999999998344	299.0

TABLE 3.4 – Tableau de l'évolution du nombre de règles générées de l'algorithme Apriori selon la confiance minimale et le support minimum.

Support min	**Confiance min**	**Mémoire**

0.001	0.0001	-0.000387923
0.001	0.0002	7.71986166
0.001	0.0003000	6.175889328
0.001	0.0004	2.8949481
0.001	0.0005	2.894948122530
0.001	0.000600	5.7898962
0.001	0.0007000	5.7898962450
0.001	0.0008	0.08220
0.001	0.0009000	6.175889328
0.004	0.0001	7.719861660
0.004	0.0002	0.0
0.004	0.000300	0.0
0.004	0.0004	0.0
0.004	0.0005	0.0
0.004	0.0006000	0.0
0.004	0.00070000	0.0
0.004	0.0008	0.0
0.004	0.0009000	0.0
0.007	0.0001	0.0
0.007	0.0002	0.0
0.007	0.000300000	0.0

0.007	0.0004	0.0
0.007	0.0005	0.0
0.007	0.00060000	0.0
0.007	0.00070	0.0
0.007	0.0008	0.0
0.007	0.00090000	0.0
0.010000	0.0001	0.0
0.0100000	0.0002	0.0
0.0100000	0.0003000	0.0
0.010000	0.0004	0.0
0.0100000	0.0005	0.0
0.010000	0.00060000	0.0
0.0100000	0.000700	0.0
0.0100000	0.0008	0.0
0.0100000	0.000900	0.0

TABLE 3.5 – Tableau de l'évolution de l'éspace mémoire alloué à l'algorithme Apriori selon la confiance minimale et le support minimum.