République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université 8 Mai 1945 Guelma

Faculté de Science et de Technologie

Département de Génie Civil et Hydraulique

Laboratoire de Génie Civil et Hydraulique LGCH

THÈSE

EN VUE DE L'OBTENTION DU DIPLOME DE

DOCTORAT EN 3ème CYCLE

Domaine : Sciences et Technologie Filière : Hydraulique

Spécialité : Hydraulique

Présentée par

BELLOUM FARID

Intitulée

Apprentissage artificiel pour la détection des fuites et gestion des réseaux d'alimentation en eau potable

Soutenue le : 04/07/2023 Devant le Jury composé de :

Nom et Prénom	Grade	Université	
Mr MAOUI Ammar	Professeur	Univ. 8 Mai 1945 Guelma	Président
Mr HOUICHI Larbi	Professeur	Univ.Batna2 Batna	Encadreur
Mr KHEROUF Mazouz	MCA	Univ. 8 Mai 1945 Guelma	Co-encadreur
Mr MANSOURI Rachid	Professeur	Univ. 8 Mai 1945 Guelma	Examinateur
Mr ZEGHADNIA Lotfi	Professeur	Univ. Mohamed Cherif Messaadia Souk Ahras	Examinateur

Année Universitaire : 2022/2023

بسم الله الرحمن الرحيم

﴿ أَفَرَأَيْتُمُ الْمَاءَ الَّذِي تَشْرَبُونَ ﴾[الواقعة: 68]

﴿أَأَنتُمْ أَنزَلْتُمُوهُ مِنَ الْمُزْنِ أَمْ نَحْنُ الْمُنزِلُونَ﴾ [الواقعة: 69]

﴿ لَوْ نَشَاءُ جَعَلْنَاهُ أُجَاجًا فَلَوْلَا تَشْكُرُونَ ﴾[الواقعة: 70]

And have you seen the water that you drink?

Is it you who brought it down from the clouds, or is it We who bring it down?

If We willed, We could make it bitter, so why are you not grateful?

Voyez-vous donc l'eau que vous buvez

Est-ce vous qui l'avez fait descendre du nuage ? Ou [en] sommes Nous le descendeur

Si Nous voulions, Nous la rendrions salée. Pourquoi n'êtes-vous donc pas reconnaissants ?

Remerciements

À l'issue de cette thèse, je tiens à remercier tous ceux qui m'ont soutenu dans ce travail et dans la réalisation de cette thèse.

Je suis très reconnaissant à l'université de Guelma 1945 de m'avoir donné l'opportunité de poursuivre mes études doctorales ainsi qu'au personnel du laboratoire LGCH pour leur formation et leur dévouement à ce projet de recherche.

Je remercie également Monsieur HOUICHI LARBI et Monsieur KHROUF MAZOUZ d'avoir accepté de m'accompagner en tant que directeurs de thèse pendant ces années et de m'accorder carte blanche dans le cadre de ma thèse.

Je dois exprimer ma vive gratitude à Madame TEBBI FATIMA ZOHRA, sans elle, le projet de recherche n'aurait pas pu être réalisé. Je tiens à la remercier pour son aide dans la recherche et la rédaction de ma thèse, pour son appui indéfectible, pour son aide scientifique, pour sa gentillesse, et du temps qu'elle a mis à lire et exprimer ses opinions.

Je tiens également à exprimer l'honneur qui m'a été accordé par les membres de mon jury de soutenance; le Pr. MAOUI AMMAR, Pr. RACHID MANSOURI et Pr. ZEGHADNIA LOTFI qui ont accepté de juger ce modeste travail et de me faire l'honneur de leur présence.

Enfin, je tiens à remercier ma famille de m'avoir soutenu dans ce doctorat, en particulier Aridj, Zakaria et le petit Yacine pour leur soutien émotionnel dans l'accomplissement de ma thèse de doctorat.

Table des matières

Liste des figures

Liste des tableaux

ملخص

تعد المياه من الموارد النادرة التي يجب إدارتها بكفاءة، ومن ركائز تحسين هذه الكفاءة الحد من تسرب المياه وزيادة أداء شبكات الإمداد بمياه الشرب. تم تصميم شبكات مياه الشرب لتلبي بشكل مرضٍ احتياجات المشتركين من المياه، من وجهة نظر كمية ونوعية. عادة لا تكون هذه الشبكات نتيجة لعملية تصميم واحدة، فهي نتيجة سنوات من الفوضى التي تستجيب للطلبات المتزايدة باستمرار .يعد الحصول على معلومات حول الشبكات الحقيقية مهمة معقدة للغاية؛ لأن شبكات التوزيع يمكن أن تتكون من آلاف العقد الاستهلاكية المتصلة بآلاف الأنابيب والخزانات لتغذيتها .تمثل التسريبات في شبكات إمدادات مياه الشرب (AEP) تهديدًا ليس فقط للمورد، ولكن أيضًا للصحة العامة والبيئة. في هذه الأطروحة، نقترح الجمع بين أدوات محاكاة شبكات AEP وطرق التدريب الذكاء الاصطناعي من أجل تحسين إدارة الشبكات المذكورة أعلاه بهدف جعل التقسيم في المقام الأول وتحديد موقع المحتملة في المرتبة الثانية، مما يؤدي إلى تقليل وقت البحث وتقليل منطقة التفتيش في النظام، وبالتالي تسهيل اكتشاف وتحديد مواقع العرض الشاذة والتحكم فيها (الانقطاعات، والتسريبات، وانخفاض الضغط، وما إلى ذلك) على الشبكات المعنية. مبدأ التقسيم هو تقسيم الشبكة إلى قطاعات هيدرومترية تسمى(DMA District Metered Area)، والتي يتم التحكم في مداخلها ومخارجها؛ دون قطع الاتصال عن بقية الشبكة سواء ماديًا أو هيدروليكيًا. في معظم الحالات التي يصبح فيها المشروع القطاعي ضروريًا، لا يتم اتباع العملية بشكل عام بطريقة علمية وتقنية؛ على العكس من ذلك، فهي تستند عمومًا إلى التجربة والخطأ. في الشبكات الصغيرة، هذه المصالحة ليست مشكلة كبيرة. تكمن المشكلة في تحديد القطاعات في الشبكات الكبيرة بشكل صحيح، نظرًا للكم الكبير من المعلومات المرتبطة بها، سيكون من المستحيل تحقيق ذلك. يتطلب تنفيذ مثل هذه العملية من هذا النوع الاستعانة بأدوات التعلم الاصطناعي من أجل إنشاء إجراء كمبيوتر للحصول على خطة شبكة مقسمة بالإضافة إلى موقع التسرب. نقدم منهجيات تستند إلى نظرية الرسوم البيانية والشبكات من أجل إنشاء تقسيم لشبكة إمدادات المياه (AEP). سيتم التعامل مع هذه المهام من خلال مراعاة الخصائص الهندسية للشبكة بالإضافة إلى المصفوفة المجاورة وLaplacian. أولاً، استكشاف طريقتين للتقسيم، «Fast Greedy»و «Random Walk»، اللتان تستخدمان بشكل شائع لإنشاء مناطق العد (DMA) في شبكات إمدادات المياه. بعد ذلك، نستكشف التجميع الطيفي لحل مشكلة تقسيم شبكة إمدادات المياه، باستخدام نهج رياضي متقدم. تتمثل مساهمتنا في إجراء دراسة مقارنة بين خوارزميات التعلم المختلفة غير الخاضعة للإشراف لتحديد خوارزميات التقسيم المثلى لنظام إمداد المياه. يتم تطبيق الطرق على ثلاثة أنظمة لتوزيع المياه EXNET وC-TOWN وEL MA OUED. أخيرًا، تحديد التسربات في شبكة إمدادات المياه، باستخدام تقنيات المحاكاة الهيدروليكية وSVM (التصنيف والانحدار). تُستخدم شبكة AEP لتشغيل العديد من "سيناريوهات التسرب"، مع تغيير موقع التسرب وشدته، وإنشاء مجموعة بيانات تحتوي على الاختلافات في الضغط والتدفق بسبب التسرب.

Résumé

L'eau est une ressource rare qui doit être gérée avec efficacité et un des piliers pour améliorer cette efficacité est la diminution des pertes d'eau et l'augmentation de la performance de ces réseaux d'alimentation en eau potable. Les réseaux d'eau potable sont conçus pour répondre de façon satisfaisante aux besoins en eau des abonnés, tant du point de vue quantitatif que qualitatif. Ces réseaux ne sont habituellement pas le résultat d'un seul processus de conception, ils sont le résultat de plusieurs années d'anarchie répondant à des demandes toujours croissantes. L'acquisition d'informations sur des réseaux réels constitue une tâche très complexe ; parce que les réseaux de distribution peuvent être constitués de milliers de nœuds de consommation connectés par des milliers de conduites et de réservoirs pour les alimenter. Les fuites dans les réseaux d'Alimentation en Eau Potable (AEP) représentent une menace non seulement pour la ressource, mais aussi pour la santé publique et l'environnement. Dans cette thèse, on propose la combinaison des outils de simulation des réseaux d'AEP et les méthodes d'apprentissage Artificiel afin d'optimiser la gestion desdits réseaux dans un objectif de faire la sectorisation en premier lieu et de pré localiser les éventuelles en deuxième lieu, ce qui mène à réduire le temps de recherche et à la réduction de la zone d'inspection du système, facilitant ainsi la détection, la localisation et le contrôle des anomalies d'alimentation (ruptures, fuites, chute de pression, etc.) sur les réseaux en question. Le principe de la sectorisation est de diviser le réseau en secteurs hydrométriques appelé DMA (District Metered Area), dont les entrées et les sorties sont contrôlées ; sans être déconnecté du reste du réseau ni physiquement ni hydrauliquement. Dans la plupart des cas où un projet de sectorisation est devenu nécessaire, le procédé n'est généralement pas suivi d'une manière scientifique et technique ; au contraire, celle-ci est généralement fondée sur des essais et des erreurs. Dans les petits réseaux, un tel rapprochement ne constitue pas un problème majeur. Le problème est de bien définir les secteurs dans les grands réseaux, étant donné la grande quantité d'informations qui leur sont associées, il serait impossible d'y parvenir. L'exécution d'un tel processus de cette nature nécessite l'aide d'outils d'intelligence artificielle afin d'établir une procédure informatique pour obtenir un plan de réseau sectorisé ainsi que l'emplacement de la fuite. Nous présentons des méthodologies basées sur la théorie des graphes et des réseaux afin d'établir une sectorisation du réseau d'alimentation en

eau (AEP). Ces tâches seront abordées en prenant en compte les caractéristiques géométriques du réseau ainsi que la matrice adjacente et laplacienne. Premièrement, une exploration de deux méthodes de partitionnement, Fast Greedy et Random Walk, qui sont couramment utilisés pour établir les zones de comptage (DMA) dans les réseaux d'alimentation en eau. Ensuite, nous explorons le groupement spectral pour résoudre le problème du partitionnement des réseaux d'alimentation en eau, avec une approche mathématique avancée. Notre contribution est de mener, une étude comparative entre les divers algorithmes d'apprentissage non supervisés pour identifier les algorithmes de partitionnement optimaux pour le système d'alimentation en eau. Les méthodes sont appliquées à trois systèmes de distribution d'eau EXNET, C-TOWN et OUED EL MA. Enfin, l'identification des fuites dans un réseau d'alimentation en eau, à l'aide de techniques de simulation hydraulique et de SVM (classification et régression). Le réseau d'AEP est utilisé pour exécuter plusieurs « scénarios de fuite », en faisant varier l'emplacement et la gravité de la fuite, et pour créer un jeu de données contenant les variations de pression et de débit dues à la fuite.

Abstract

Water is a scarce resource that must be managed efficiently, and one of the pillars for improving this efficiency is to reduce water losses and increase the performance of these drinking water supply systems. Drinking water systems are designed to satisfactorily meet the water needs of customers, both in terms of quantity and quality. These systems are usually not the result of a single design process, but rather the result of many years of anarchy in response to ever-increasing demands. Acquiring information about real networks is a very complex task; because distribution networks can consist of thousands of consumer nodes connected by thousands of pipes and reservoirs to feed them. Leakages in drinking water supply networks represent a threat not only to the resource, but also to public health and the environment. In this thesis, we propose the combination of simulation tools for drinking water networks and artificial learning methods to optimize the management of these networks with the objective of doing the sectorization in the first place and pre-localizing the possible ones in the second place, which leads to reduce the time of research and the reduction of the inspection area of the system, thus facilitating the detection, localization and control of the supply anomalies (breaks, leaks, pressure drops, etc.) in the networks in question. The principle of sectorization is to divide the network into hydrometric sectors called DMA (District Metered Area), whose inlets and outlets are controlled; without being disconnected from the rest of the network either physically or hydraulically. In most cases where a sectorization project has become necessary, the process is not usually followed in a scientific and technical manner; instead, it is usually based on trial and error. In small networks, such approximation is not a major problem. The problem is to properly define the sectors in large networks, given the large amount of information associated with them, it would be impossible to achieve this. Performing such a process of this nature requires the help of artificial intelligence tools to establish a computational procedure to obtain a sectorized network map as well as the location of the leak. We present methodologies based on graph and network theory to establish a sectorization of the water supply network. These will be discussed taking into account the geometric characteristics of the network as well as the adjacent and Laplacian matrix. First, an exploration of two partitioning methods, Fast Greedy and Random

Walk, which are commonly used to establish metering areas (DMA) in water supply networks. Then, we explore spectral clustering to solve the problem of partitioning water supply networks, with an advanced mathematical approach. Our contribution is to conduct, a comparative study between various unsupervised learning algorithms to identify the optimal partitioning algorithms for the water supply system. The methods are applied to three water distribution systems EXNET, C-TOWN and OUED EL MA. Finally, the identification of leaks in a water supply network, using hydraulic simulation and SVM techniques (classification and regression). The water supply network is used to run several "leak scenarios", varying the location and severity of the leak, and to create a dataset containing the pressure and flow variations due to the leak.

Introduction générale

L'eau potable est une ressource indispensable à tout processus lié à la vie. Il est évident que l'eau est intarissable du point de vue qu'il représente les 3/4 de notre planète, alors qu'en réalité, le taux d'eau douce adapté aux humains est très faible et a une forte tendance à diminuer compte tenu de l'accroissement de la population et de la pollution. L'eau est la matière la plus importante pour l'existence de l'humanité, et sa disponibilité est l'une des clés de la répartition des êtres vivants à la surface de la terre. C'est donc un capital à mobiliser, valoriser, gérer et préserver qui semble être le grand défi du 21ème siècle pour tous les pays du monde. Il s'agit d'un produit de base pour les activités domestiques et urbaines ainsi que pour l'agriculture. Sa disponibilité est entièrement liée au bien-être et à la prospérité de toute société. Faire en sorte que la quantité et la qualité de l'eau soient suffisantes est l'une des questions les plus importantes de l'histoire humaine. D'où l'importance d'une gestion adéquate des réseaux d'alimentation en eau potable qui : conventuellement, peut-être définit comme l'infrastructure permettant de transporter la ressource en question depuis les sources jusqu'aux consommateurs. Les réseaux d'alimentation d'eau potable appartiennent, de la même manière que les autres réseaux techniques, à un milieu urbain et péri-urbain au sein duquel ils agissent et interagissent avec d'autres réseaux. Le fonctionnement de ces systèmes prend en considération de nombreux indices et paramètres allant de la production (ressources) au stockage (réservoirs) en passant par la distribution aux consommateurs (réseau); La gestion doit affronter les contraintes d'une mission de service public et donc d'une gestion de réseau technique ; qui a comme objectif principal la livraison aux consommateurs d'un produit (eau) conforme aux normes de qualité, en quantité suffisante (besoin de l'abonné) avec continuité de service sans défaillance(Herrera Fernández, 2011). Pour ce faire, il est nécessaire de connaître précisément le réseau et ces infrastructures, son fonctionnement hydraulique et de procéder à une maintenance régulière et continue du réseau. D'un point de vue technique, dans l'hypothèse d'une gestion administrative adéquate, les problèmes d'un réseau d'AEP peuvent se résumer en quatre aspects généraux : les

fuites et les eaux non comptabilisées ; l'intégrité physique du réseau ; la qualité de l'eau distribuée, ainsi que la fiabilité et la qualité de la base de données sur le réseau de distribution d'eau(Campbell, 2013). En ce qui concerne le premier aspect, le contrôle des pertes est préoccupant depuis la construction des premiers réseaux d'AEP. Même dans la Rome antique, des actions étaient déjà entreprises pour réduire le volume des pertes d'eau. Dans les pays en développement, les fuites peuvent constituer jusqu'à 50 % de l'eau injectée dans le réseau. Autrement dit, dans un réseau d'AEP doté d'un tel indice de perte, 2 m^3 d'eau doivent être produits pour qu'un m^3 atteigne les utilisateurs. On estime les pertes annuelles d'eau dans ces pays à 26,7 milliards de mètres cube, ou 5,9 milliards de dollars américains. La perte d'eau n'est pas l'unique conséquence des fuites, car elle crée aussi des problèmes sociaux et environnementaux. De plus, les fuites qui ne sont pas réparées sont susceptibles de se développer et de laisser des contaminants environnementaux et des agents pathogènes s'infiltrer dans le réseau d'alimentation en eau. Ceci réduit considérablement la qualité de l'eau fournie et peut influer sur la vie des humains et d'autres espèces vivantes. Les réseaux d'AEP sans perte sont considérés comme utopiques, à la fois sur le plan technique et économique; D'autre part, des progrès considérables ont été accomplis dans la connaissance et le développement de l'équipement et des techniques, ce qui a conduit à des améliorations dans la lutte contre les fuites. Ces techniques comprennent la sectorisation du réseau d'AEP, qui est perçue comme une stratégie que les gestionnaires des systèmes d'eau peuvent suivre pour améliorer l'efficacité. L'un des principaux avantages de son application est la surveillance du bilan hydrique, ce qui facilite la détection et l'exploration de toute anomalie dans le réseau d'AEP due à la réduction de l'aire d'inspection. En mesurant les quantités d'eau entrante et en sortant d'un secteur, on peut analyser la performance de ce sous-système indépendamment du comportement de l'ensemble du système. La conception des secteurs a été essentiellement basée sur des méthodes d'essai et d'erreur pour vérifier les résultats étape par étapes à l'aide de simulations hydrauliques pour valider la performance correspondante(Grayman et al., 2001). Cette procédure n'a aucune base rationnelle. Car il y a beaucoup de choix de sectorisation des réseaux, même dans les petits réseaux. Il est très difficile de choisir la meilleure sectorisation par tâtonnements. Ces derniers temps, les chercheurs ont commencé à étudier le sujet pour créer automatiquement des secteurs. Certaines procédures basées sur la théorie des graphes et des réseaux ont démontré que les solutions optimales pour le nombre, la forme et la taille peuvent être identifiées automatiquement. Pour sectoriser un réseau, des débitmètres et des vannes doivent être installés aux points stratégiques de ce réseau, tout en conservant l'efficacité du

réseau et en réduisant les coûts économiques. Devant cette difficulté, et étant donné les techniques numériques et l'évolution récente de la puissance des ordinateurs utilisés par les programmes de simulation, il est possible de réorienter l'analyse, les chercheurs ont commencé à étudier la question afin de concevoir automatiquement des secteurs. Ils font appel à l'intelligence artificielle pour partitionner les réseaux. L'objectif était de réduire le plus possible les investissements et d'améliorer l'efficacité de la gestion du réseau. Le but de l'intelligence artificielle est de développer des systèmes dotés de capacités intellectuelles semblables à celles de l'homme, capables de prendre des décisions de façon automatique dans l'environnement perçu. Dans ce large domaine, il y a un domaine de recherche relativement important qui est l'apprentissage automatique Les algorithmes d'apprentissage automatique sont particulièrement efficaces dans la représentation et l'analyse de situations complexes ou une grande quantité d'informations importantes sont disponibles. Les réseaux d'alimentation en d'eau potable (AEP) sont des objets complexes constitués de nombreux éléments avec de nombreux paramètres à partir desquels nous pouvons agir, leur modélisation pour l'optimisation est donc complexes. Les réseaux d'AEP produisent également une grande quantité de données sur leur fonctionnement (notamment les débits, pressions, diamètres, indicateurs de qualité...), ce qui permet aux opérateurs d'avoir une image précise de leur rendement. En raison de la complexité et de l'abondance des données, les opérateurs ont commencé à s'intéresser aux techniques de machine learning pour optimiser le fonctionnement des réseaux. Il y a differents modes d'apprentissage en fonction des donnees disponibles pour former l'intelligence artificielle et de la reponse desiree et des utilisations prevues (notamment l'apprentissage supervise, l'apprentissage non supervise, l'apprentissage semi-supervisé). Les algorithmes les plus frequemment utilises sont les algorithmes de classification, les algorithmes de regression lineaire, les réseaux de neurones, les arbres de decision et les machines a vecteurs supports. Un réseau d'alimentation d'eau peut être modélisé sous forme de graphe G (V, E) avec l'ensemble des nœuds V représentant les sources et les nœuds consommateurs, et l'ensemble des arêtes E les conduites de raccordement. Il existe plusieurs méthodes pour stocker un graphe (réseau) sur un ordinateur. Les matrices figurent parmi les structures de données utilisées en informatique pour représenter les graphes. Partitionner un graphe consiste à identifier des sous-graphes denses et différenciés, c'est-à-dire à grouper automatiquement les éléments (nœuds). Le partitionnement est le terme utilisé pour décrire les méthodes de partage de graphes. Cette tâche est appelée, en fonction du domaine, classification non supervisée, classification automatique ou clustering. Son objectif est de diviser un graphe en plusieurs sous-

graphes selon les critères de proximité. Afin d'obtenir un bon partitionnement, il faut tenir compte de deux critères : il faut optimiser la proximité des éléments du même sous-graphe et minimiser la proximité des différents sous-graphes. Les groupes criés sont appelés clusters.. Il existe deux principales familles de méthodes de clustring : les méthodes hiérarchiques et les méthodes de partition. Les méthodes hiérarchiques sont basées sur des mesures de similitude entre les sommets. Ils construisent des clusters en distribuant récursivement des instances descendantes ou ascendantes. Le résultat de ce type d'algorithmes est un arbre de cluster appelé dendrogramme, qui montre la façon dont les clusters sont organisés. Pour les méthodes de partition, il en résulte une partition de l'espace objet, autrement dit, chaque objet est associé à un cluster unique. De nombreuses approches sont proposées, nous nous concentrerons sur ceux qui ont attiré le plus l'attention de la communauté scientifique comme Fast Greedy, Random Walk(Liu et al., 2018) et les méthodes de partitionnement spectral(Di Nardo et al., 2018). Pour les méthodes de partition, le clustering des nœuds de graphe est effectuée par l'algorithme populaire K-MEANS qui est continuellement utilisé. Afin de sélectionner les bons algorithmes, nous testerons les différents algorithmes existants (PAM, CLARA, HIERARCHICAL et DIANA) et effectuerons une analyse comparative. Nous avons sélectionné quelques indicateurs de qualité couramment utilisés tels que la modularité, l'indice interne et l'indice de stabilité afin de déterminer les algorithmes dominants. Les réseaux EXNET, C-TOWN et OUED EL MA font l'objet de comparaison. Le choix de ces réseaux a été effectué en fonction de leur nature et de leur dimension. La plupart du temps, les fuites ne sont détectées que lorsque celles-ci deviennent visibles et qu'elles ont été réparées. Ceci aboutit généralement à une énorme perte d'eau. Par conséquent, il est important d'élaborer des méthodes de détection rapide des fuites. Les méthodes courantes de détection des fuites sont fondées sur le bruit généré par la fuite. L'un des inconvénients de ces méthodes est l'interférence provenant de sources externes. Il est possible d'utiliser des techniques non-acoustiques pour mieux détecter et localiser les fuites dans les conduites. Les réseaux neuronaux et la méthode SVM (les machines à vecteurs de support) sont des outils particulièrement appropriés pour aider les spécialistes de la maintenance à détecter et à classer les défaillances des réseaux. À partir d'un ensemble de capteurs de pression et de débit qui surveillent un réseau de canalisations, on a utilisé le SVM pour interpréter les données obtenues, pour obtenir des informations relatives à la localisation des fuites dans le réseau.

Le présent mémoire est structuré de la façon suivante :

Le premier chapitre met l'accent sur les problemes de gestion des reseaux d'eau potable, ainsi que sur les modalites des sectorisations classiques et celles fondees sur l'intelligence artificielle. Suivi d'une presentation sur l'apprentissage automatique utilise dans ce memoire, y compris la methode svm et la classification.

Le deuxieme chapitre commence par les definitions fondamentales concernant la théorie des graphes, la représentation matricielle des graphes et la classification. Ensuite l'application et la critique du travail de sectorisation de (Liu et al., 2018) ainsi que le travail (Di Nardo et al., 2018).

Le troisième chapitre est consacré à notre contribution qui consiste en une analyse comparative des méthodes de clustering appliquées aux réseaux d'eau potable. Par les différents algorithmes existants (PAM, CLARA, HIERARCHICAL et DIANA) en fonction de certains indicateurs de qualité afin de déterminer les algorithmes dominants qui ont été appliqués aux réseaux EXNET, C-TOWN et OUED EL MA.

Le quatrième chapitre décrit la mise en œuvre de la méthode SVM (classification et régression) pour les réseaux d'alimentation en eau potable. À partir des données provenant de plusieurs capteurs de pression et de débit installés sur le réseau qui contrôle le système en permanence afin de détecter l'emplacement des fuites dans le réseau de canalisations.

CHAPITRE I :

ANALYSE DU CONCEPT DE

SECTORISATION ET DE SES FINALITÉS

I ANALYSE DU CONCEPT DE SECTORISATION ET DE SES FINALITÉS

I.1 Introduction

Assurer une qualité et une quantité suffisantes de l'eau a été l'une des questions les plus importantes dans l'histoire de l'humanité. Les plus anciennes civilisations se sont installées à proximité des sources d'eau. Avec l'accroissement de la population, les demandes des usagers sont devenues difficiles à satisfaire. Les gens ont commencé à acheminer l'eau d'autres endroits vers leurs communautés. Par exemple, les Romains construisaient des canaux qui alimentaient leurs communautés à partir de sources lointaines. Aujourd'hui, un système d'alimentation en eau potable consiste en une infrastructure qui collecte, traite, stocke et distribue l'eau entre les sources d'eau et les consommateurs. Le réseau d'eau potable peut être qualifié de complexe pour un certain nombre de raisons, ils sont souvent très grands; ils sont enterrés et donc difficilement accessibles pour la surveillance et la maintenance; ils sont fortement bouclés, leur modélisation comprend des équations exigeant des méthodes de résolution numériques compliqué les pertes d'eau sont souvent importantes par rapport aux autres réseaux civils, certaines de ses caractéristiques sont particulières et rendent leur gestion difficile, avec de nombreux problèmes d'exploitation (comme les pertes d'eau et d'énergie)(Nardo et al., 2016; Slimani & Drif, 2016).Le contrôle et la gestion de l'ensemble des réseaux posent des difficultés. Bien que ce soit le cas dans plusieurs pays, des progrès importants ont été accomplis pour combler ces lacunes. Dans le cadre de la gestion des ressources en eau, des efforts majeurs ont été consentis pour développer les réseaux d'eau potable. En raison de la diminution des ressources en eau et de la croissance démographique rapide, des méthodes novatrices sont nécessaires pour gérer les réseaux d'eau potable. Pour améliorer la gestion et mieux identifier les pertes d'eau, les réseaux peuvent être physiquement divisés en secteurs, en insérant des dispositifs hydrauliques sur des conduites appropriées. La sectorisation est une disposition qui facilite le contrôle des difficultés. Le fait d'avoir un réseau sectorisé permet non seulement d'appliquer des techniques particulières de contrôle des fuites, mais aussi de mettre en place différents modèles de gestion. Le contrôle du bilan hydrique contribue à détecter les anomalies dans l'alimentation en eau et à réduire la zone d'inspection. En mesurant les quantités d'eau entrante et en sortant d'un secteur (Zevnik,

2018), on peut analyser la performance de ce sous-système indépendamment du comportement global du système (Diao, 2013). Il s'agit de la division du réseau d'alimentation en zones isolées (secteurs de l'eau), délimitées par des vannes limitrophes, en configurations fixes ou dynamiques (Wright, 2015). C'est une technique de conception et de contrôle ingénieuse qui permet d'améliorer les performances des réseaux de distribution d'eau (Campbell, 2016). La conception des secteurs reposait principalement sur des méthodes d'essai et d'erreur permettant de vérifier les résultats étape par étapes à l'aide de simulations hydrauliques pour valider les performances correspondantes (Grayman, 2001). En d'autres termes, sur la base de cartes réseau, les secteurs sont sélectionnés de manière à réduire au minimum le nombre de vannes à installer puis tester les schémas de fermeture et d'accès des secteurs (Campbell, 2013). Une telle procédure n'a pas de fondement rationnel. Parce que les choix de sectorisation de réseaux sont très nombreux, même dans les petits réseaux (Di Nardo, 2011). Il est très difficile de déterminer laquelle est la meilleure sectorisation par essais et erreurs. Récemment, des chercheurs ont entrepris d'étudier le sujet afin de créer automatiquement des secteurs. Certaines procédures fondées sur la théorie des graphes et des réseaux ont montré qu'il est possible d'identifier automatiquement les solutions optimales en matière de nombres, de forme et de taille telle que ceux proposés par (A. N. Di Nardo, MicheleSantonastaso, Giovanni FTzatchkov, Velitchko GAlcocer-Yamanaka, Victor H, 2014; Tzatchkov, 2008). La sectorisation d'un réseau d'alimentation en eau potable (AEP) est un problème du groupement d'éléments (lien, nœuds et utilisateurs). Depuis quelque temps il y a eu des algorithmes et des outils mathématiques dans les graphes et la théorie des réseaux complexes afin de mieux analyser le comportement et l'évolution des systèmes complexes. Ce regroupement peut être envisagé sur la base des techniques de regroupement des graphes ; en partant du principe que les réseaux peuvent être représentés sous forme de graphes. Un graphe est une structure couramment utilisée en informatique et en mathématiques; il exprime la relation (au moyen d'arêtes ou de liens) entre un ensemble d'éléments ou de nœuds. Il est important de noter la similarité entre le concept de graphe et le concept de réseau d'AEP, qui permet de le représenter sous forme de graphe. Diverses méthodes ont été proposées dans la littérature pour trouver une adaptation idéale de la sectorisation, elles sont essentiellement réparties en deux phases: a) regroupement, pour définir la forme et la taille des sous-ensembles du réseau, à partir d'algorithmes de théorie des graphes, approche spectrale, et b) la division physique

du réseau par le choix des conduites pour l'insertion des débitmètres ou des vannes, à partir d'ité-ratifs ou génétiques afin de définir la disposition optimale qui minimise l'investissement économique et la détérioration des performances hydrauliques (Doiron, 2016).

I.2 Problèmes de gestion des réseaux d'alimentation en eau potable

Dans le but d'améliorer la gestion des systèmes d'alimentation en eau potable (AEP), de nombreux gestionnaires ont choisi comme option de créer des secteurs, avec des entrées et sorties d'eau contrôlées (Di Nardo, 2017). Traditionnellement, la sectorisation est conçue à partir de propositions empiriques. Récemment, les chercheurs ont commencé à se pencher sur le sujet afin de créer automatiquement des secteurs.

I.3 Les réseaux de distribution d'eau

Un réseau de distribution a pour but de fournir de l'eau au consommateur en quantité, qualité et pression appropriées. Le système de distribution d'eau potable est utilisé pour décrire toutes les installations utilisées pour acheminer l'eau depuis la source jusqu'au point d'utilisation.

I.3.1 *Prescriptions applicables aux réseaux de distribution d'eau*

La conception d'un réseau AEP fait l'objet d'erreurs découlant principalement:

- d'erreurs de conception.
- de pose.
- d’exploitation.
- d'hypothèses erronées.
- données statistiques inadéquates.
- d'erreurs de saisie de calcul.
- modifications inappropriées des traces pendant la pose.

Elles entraînent souvent des ruptures fréquentes de conduites, des pertes d'eau traitée et des coûts élevés de réparation et d'entretien pendant et après les réparations. Un autre effet secondaire est le siphonnage de l'eau souillée et contaminée qui a un impact négatif sur la qualité de l'eau fournie aux consommateurs.

I.3.2 *Conditions requises pour un réseau de distribution efficace*

La qualité de l'eau ne doit pas être altérée dans les conduites. Il doit pouvoir fournir de l'eau à tous les endroits prévus et à une hauteur de pression suffisante. Il doit être en mesure de fournir la quantité d'eau nécessaire durant les opérations de lutte contre l'incendie. La disposition devrait être telle qu'aucun consommateur ne soit privé d'eau au cours de la réparation d'une section quelconque du réseau. Il est préférable que toutes les canalisations de distribution soient à un mètre au-dessus des canalisations d'égout.Il doit être assez étanche pour minimiser les pertes dues aux fuites.

I.3.3 *Les fuites d'eau dans les réseaux*

Il n'existe aucun réseau de D'AEP dans le monde sans quelques fuites et ruptures de tuyaux. Quelques-unes de ces fuites viennent souvent de joints enterrés, vannes, débitmètres, tuyaux vieux et fragiles qui sont physiquement non détectables jusqu'à ce que l'eau commence à faire surface ou à la formation de cavités. Les pertes d'eau dans un réseau sont calculées en fonction du bilan hydrique entre les volumes d'eau injectés dans le réseau de distribution et le volume d'eau facturé aux clients (Tableau I-1).

Table I-1 Répartition des consommations (Coursol Tellier 2015).

Volume d'eau entrant	Consommation autorisée	Consommation autorisée et facturée	Volume d'eau consommé et mesuré incluant le volume d'eau exporté dans un autre système (ex. : consommation résidentielle)
			Volume d'eau consommé et non mesuré
		Consommation autorisée et non facturée	Volume d'eau mesuré et non facturé (ex. : infrastructures municipales)
			Volume d'eau non mesuré et non facturé (ex. : nettoyage du réseau)
	Pertes	Pertes apparentes	Consommation non autorisée
			Imprécision des débitmètres
			Erreurs de manipulation des données
		Pertes réelles	Fuites provenant des conduites d'amenée
			Fuites provenant des conduites du RDEP
			Débordement des réservoirs
			Fuites des connexions de service

I.4 **Solutions aux problèmes de distribution d'eau**

La surveillance et la gestion de tous les réseaux constituent un problème. Pour mieux gérer et localiser les pertes d'eau, les réseaux de distribution d'eau peuvent être physiquement divisés en zones. Le suivi du bilan hydrique permet de détecter tout dysfonctionnement de l'approvisionnement en eau. Les zones de mesure sont des techniques de conception et de contrôle intelligentes

conçues pour améliorer les performances des réseaux de distribution d'eau (Han & Liu, 2017). La sectorisation d'un réseau d'eau est une étape primordiale vers un réseau intelligent (Hajebi, 2014). La sectorisation des réseaux de distribution doit tenir compte de trois grands critères : le secteur doit répondre aux questions de conception et de lutte contre les incendies, la quantité d'eau doit être mesurée pratiquement et économiquement et que la qualité de l'eau soit assurée (Falcini, 2007; Sturm, 2005). Les secteurs ont été conçus principalement en fonction de méthodes d'essai et d'erreur. Un tel procédé n'est pas rationnel. En effet, les choix de sectorisation des réseaux sont très nombreux, notamment dans les petits réseaux (Di Nardo, 2011). IL est très difficile de déterminer quelle est la meilleure sectorisation par tâtonnements. Les chercheurs se sont engagés à étudier le thème pour créer automatiquement des secteurs. Les procédures fondées sur la théorie des graphes et des réseaux ont démontré qu'il est possible d'identifier automatiquement les solutions optimales comme celles proposées par (A. N. Di Nardo, MicheleSantonastaso, Giovanni FTzatchkov, Velitchko GAlcocer-Yamanaka, Victor H, 2014; Tzatchkov, 2008).

I.4.1 ***Sectorisation***

Actuellement; la plupart des problèmes liés aux réseaux d'alimentation en eau potable sont reliés aux réseaux existants; soit parce qu'ils ont été mal conçus à l'origine, soit parce qu'ils ont été conçus en fonction de critères optimaux. En vue d'améliorer la gestion des réseaux; et comme une nécessité technique et économique beaucoup de dirigeants ont choisi comme option la création de secteurs de distribution mesurée, aussi appelé DMA (District Metered Area) avec entrées et sorties d'eau contrôlées. La sectorisation consiste à créer des secteurs autonomes mais non indépendants au sein d'un réseau de distribution; en d'autres termes il divise ou partitionne le réseau en un grand nombre de petits réseaux, afin de faciliter leur gestion. La sectorisation est un outil permettant de diagnostiquer l'état et le fonctionnement du réseau en un temps donné. Il est beaucoup plus facile de contrôler les flux entrants dans chaque secteur la pression, la demande, la consommation, et les pertes d'eau résultant d'une fuite et d'une utilisation non autorisée.

I.4.1.A *Définition du secteur de distribution contrôlée*

La sectorisation consiste à créer des zones d'approvisionnements autonomes, mais non indépendant à l'intérieur d'un réseau de distribution ; en d'autres termes, c'est la division ou le partage du réseau en plusieurs petits réseaux, dans le but de faciliter sa gestion. Ainsi, on peut mieux contrôler le volume d'eau consommé par un secteur, et une connaissance du comportement du réseau.(pression,

demande, consommation, pertes d'eau, que ce soit des fuites ou des utilisations non autorisées). Le secteur hydrométrique est un secteur distinct du réseau de distribution, qu'il soit constitué naturellement ou imposé. L'isolation des zones entre elles est effectuée correctement pour mesurer et contrôler le débit, Pression, afin d'offrir une qualité de service identique à tous les utilisateurs du réseau.

I.4.1.B *Les bases d'une sectorisation de réseaux*

La sectorisation consiste à diviser un réseau d'eau en plusieurs sous-réseaux de comportement homogène afin de mesurer les paramètres essentiels (débit et pression) de manière permanente ou temporaire. L'isolement des sous-réseaux entre eux est atteint par la fermeture des vannes, ou par l'établissement de points de mesure (débitmètre/compteur) aux frontières des sous-réseaux. Les nouveaux systèmes sont entièrement gérés à distance et permettent la récupération et le stockage continus des mesures par tranches horaires ou intra horaires. Le partitionnement d'un réseau en secteurs est l'étape la plus importante d'une sectorisation : il s'agit de trouver une taille de secteur assez petite pour donner des informations précises, mais aussi, assez large pour limiter le nombre de secteurs nécessaires pour couvrir un réseau donné.

I.4.1.C *Les buts de la sectorisation d'un réseau d'eau*

Les buts de la sectorisation d'un réseau d'eau sont :

- détecter les défaillances et trouver la cause pour agir aussi rapidement que possible. Ceci est rendu possible grâce au calcul des volumes entrants et sortants.
- Réaliser des économies réelles, puisque l'eau perdue et non redistribuée au client final n'est pas facturée.
- Le défi est aussi environnemental et durable, dans un contexte où les ressources hydriques deviennent de plus en plus rares. Par conséquent, les pertes ne sont plus permises et il faut utiliser les outils et les procédures appropriés pour les minimiser.

I.4.1.D *Avantages et inconvénients de la sectorisation*

En voici quelques-uns :

- améliorer la gestion des réseaux de distribution ainsi que leur rendement hydraulique.
- Il permet d'effectuer périodiquement des bilans hydriques.

- Facilite l'évaluation du débit de chaque secteur et, par conséquent, du niveau de fuite.
- La surface d'inspection pour détecter et localiser les anomalies est réduite.- Favorise une meilleure gestion des zones isolées.

Parmi les inconvénients, on peut citer:

- la garantie d'approvisionnement est réduite en comparaison avec les réseaux entièrement connectés; Parce Que' cas de panne éventuelle des points d'approvisionnement du secteur serait laissé sans services pour lesquels il est recommandé qu'il y ait d'autres connexions d'alimentation; qui en conditions normales, sont restées fermes.
- La durée de séjour de l'eau dans le réseau augmente, ce qui nuit à la qualité de l'eau. Cela est dû au fait que le trajet que l'eau doit parcourir pour atteindre l'utilisateur final augmente, puisque les réseaux adoptent une typologie plus ramifiée.
- La sectorisation d'un réseau nécessite un investissement initial important puisqu'il ne s'agit pas seulement d'installer des vannes et des débitmètres en certains points du système, mais il est parfois nécessaire de renforcer certaines parties des tuyaux pour assurer l'approvisionnement.

I.5 **Méthodologie de sectorisations classiques des réseaux d'alimentation d'eau potable**

Pour mieux gérer et mieux localiser les fuites d'eau, les réseaux de distribution d'eau peuvent être divisés en secteurs, en installant des dispositifs hydrauliques sur les tuyaux appropriés et en simplifiant ainsi la surveillance du bilan hydrique (Di Nardo, 2017). Traditionnellement, la sectorisation est conçue sur la base de suggestions empiriques (telles que le nombre maximal d'abonnés ou la longueur totale des conduites par secteur). Il convient de tenir compte du fait que la mauvaise sectorisation peut engendrer des problèmes d'approvisionnement et de qualité. En combinaison avec des procédures d'essais et d'erreurs où les tuyaux à fermer sont sélectionnés et un modèle de simulation du réseau de distribution exécuté à plusieurs reprises pour trouver une solution fiable en ce qui concerne la pression et les dépenses. Si une solution possible est trouvée, on ne connaît pas sa qualité par rapport aux autres solutions possibles. Cette procédure ne repose sur aucun fondement rationnel, parce que de très nombreux choix de sectorisation de réseaux, même sur de petits réseaux (Di Nardo, 2011). Cependant, on peut difficilement appliquer ces procédures aux grands réseaux.

La sectorisation peut se résumer en trois étapes :

- la conception: c'est l'étape de base et celle qui prend la majorité du temps dans le processus de sectorisation d'un réseau.
- Préparation en fonction du plan : Il s'agit d'une étape importante dans le but de connaître la structure et le fonctionnement du réseau, collecter auprès des autorités compétentes les données techniques du réseau et passer ensuite au découpage du réseau en secteurs.
- Opérations sur le terrain : Vérification et développement (l'établissement du plan définitif pour chaque secteur)

I.6 **Méthode de sectorisation basée sur l'intelligence artificielle**

Les réseaux de distribution d'eau sont utilisés au quotidien, qu'ils soient domestiques ou industriels. Elles sont généralement de grandes dimensions et nécessitent des améliorations constantes en matière de gestion des fuites d'eau. Ces fuites d'eau sont fréquemment liées à la détérioration des canalisations. D'autre part, il serait beaucoup trop coûteux de rénover l'ensemble des infrastructures. Des procédures efficaces sont nécessaires pour assurer un contrôle et une gestion optimale des réseaux de distribution d'eau (Di Nardo, 2018). Parmi les procédures, on retrouve la sectorisation qui à deux objectifs principaux le premier a pour but de faciliter le calcul du bilan hydrique du réseau en surveillant les débits de nuits de chaque secteur et en assurant la gestion de la pression, le second est que le comptage sectoriel de l'eau peut se faire de différentes manières et a différents niveaux de partage (Di Nardo, 2010). Le partitionnement des réseaux nécessite l'installation de débitmètres et de vannes à des endroits stratégiques du réseau tout en maintenant la performance du réseau et en réduisant les coûts économiques (Khoa Bui, 2020). Face à une telle complexité, les techniques numériques et la récente croissance exponentielle de la puissance des ordinateurs utilisés par des programmes de simulation qui nous permettent de réorienter l'analyse, depuis la conception et la gestion traditionnelles des réseaux d'approvisionnement en eau, les chercheurs ont commencé à étudier le sujet afin de créer automatiquement des secteurs (Herrera Fernández, 2011). La recherche axée sur l'utilisation de techniques d'intelligence artificielle dans le but de faciliter l'exploitation des données et la prise de décision par les gestionnaires de systèmes d'approvisionnement en eau. Aussi bien que de réduire les coûts dans l'installation de vannes et débitmètres.

Beaucoup de méthodes de partage sont disponibles pour diviser les réseaux de distribution d'eau en secteurs. certaines méthodes fondées sur la connectivité et la topologie du réseau également par le biais de différents algorithmes tels que la théorie des graphes, les algorithmes à modularité et les algorithmes spectraux ont montré qu'il est possible d'identifier automatiquement les solutions optimales du point de vue du nombre, de la forme et de la taille, comme suggéré par (A. N. Di Nardo, MicheleSantonastaso, Giovanni FTzatchkov, Velitchko GAlcocer-Yamanaka, Victor H, 2014; Tzatchkov, 2008).On peut recourir à l'intelligence artificielle pour segmenter les réseaux. Il s'agissait de minimiser les investissements et de parvenir à de meilleurs niveaux d'efficacité dans la gestion des réseaux.

I.6.1 ***L'apprentissage artificiel (Machine-Learning)***

I.6.1.A *L'apprentissage automatique (Artificiel):*

L'évolution technologique de ces dernières années a permis aux scientifiques de développer et d'améliorer des méthodes dans différents domaines. L'évolution des ordinateurs en particulier et la capacité d'intégration des composants extraordinaires atteints aujourd'hui ont permis une grande vitesse de calcul et une grande capacité de mémoire. Parmi ces méthodes, se trouve l'apprentissage automatique (ARTIFICIEL) qui n'est rien d'autre que la rencontre des statistiques avec la puissance de calcul disponible aujourd'hui (mémoires, processeurs, cartes graphiques). L'apprentissage ARTIFICIEL est un sous-ensemble de l'intelligence artificielle; il s'agit simplement d'une technique pour réaliser l'intelligence artificielle. L'apprentissage automatique est la capacité d'un système d'acquérir et d'intégrer des connaissances de manière autonome. Ce concept contient toute méthode selon laquelle un modèle de réalité peut être construit à partir de données; soit en améliorant un modèle partiellement ou moins général, soit en créant entièrement le modèle (Cornuéjols, 2011). Ce sous-domaine de l'intelligence artificielle s'intéresse à attribuer aux machines la capacité de s'améliorer à l'accomplissement d'une tâche, en interagissant avec leur environnement (Burkov, 2020). On se base sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'acquérir et d'apprendre de façon autonome des connaissances à partir des données, c'est-à-dire que ce programme a la capacité d'apprendre sans que cette modification ne soit explicitement programmée (Samuel 1959). L'apprentissage automatique repose sur deux piliers : les données, qui sont les exemples qu'apprendra l'algorithme, et l'algorithme d'apprentissage, qui est la procédure utilisée pour créer un modèle. L'instruction est l'exécution d'un algorithme d'apprentissage dans un

ensemble de données. Ces deux piliers revêtent une importance égale. D'une part, aucun algorithme d'apprentissage ne sera en mesure de construire un bon modèle avec des données non pertinentes. Par ailleurs, un modèle appris au moyen d'un algorithme inapproprié sur des données pertinentes ne peut pas être de bonne qualité (Gosalia, 2019) . Cette branche de la science a beaucoup d'applications pratiques, en particulier la reconnaissance des formes, dans le domaine de la santé, moteurs de recherche, classification, détection de modèles, prédiction, détection de cas particuliers, exploration des connaissances. L'apprentissage automatique peut être divisé en trois catégories générales : supervisé ou non-supervisé et Semi-Supervisé. L'apprentissage automatique utilise différents algorithmes pour prendre des décisions, prévoir les résultats, grouper les résultats et détecter les anomalies. La définition de l'apprentissage automatique selon Wikipédia (septembre 2020) est : « L'apprentissage automatique (en anglais machine learning, littéralement « apprentissage machine ») ou apprentissage statistique est un champ d'études de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'« apprendre » à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes.

I.6.2 ***Methodes d'apprentissage automatique***

Les algorithmes d'apprentissage peuvent être classés en fonction du mode d'apprentissage qu'ils emploient. On distingue généralement au moins trois types d'apprentissage machine :

- l'apprentissage supervisé.
- l'apprentissage non supervisé.
- l'apprentissage semi-supervisé.

I.6.2.A *L'apprentissage supervisé (supervised learning)*

L'apprentissage supervisé est une technique de modélisation prédictive qui consiste à enseigner à un algorithme à associer une entrée à un résultat tiré d'exemples connus. La première étape est la détermination d'un modèle de données étiqueté. Ensuite, de nouvelles données doivent être identifiées (Cornuéjols, 2011).Dans l'apprentissage supervisé, les classes auxquelles appartiennent les données sont connues à l'avance. Ces connaissances (données, classe d'appartenance) seront utilisées pour élaborer un modèle afin de classer les prochaines données. L'apprentissage supervisé

peut-être divisé en deux sous-catégories modèles de régression ou de classification la régression consiste à estimer une réponse, tandis que la classification désigne l'appartenance à un groupe (Gagan Gupta, 2017).

I.6.2.B *L'apprentissage non supervisé (unsupervised learning)*

Comme son nom l'indique, l'apprentissage en l'absence d'un superviseur. L'apprentissage non supervisé fait partie des « statistiques descriptives » (Côme, 2009). Pour ce type d'apprentissage, aucune variable cible ne figure dans la base de données d'apprentissage (comme nous l'avons vu lors de l'apprentissage supervisé). Son but est de trouver une structure cohérente au sein d'un ensemble de données qui peut faciliter l'interprétation, l'analyse et la représentation. L'entrée contient uniquement un ensemble de données. L'algorithme doit découvrir la structure masquée de ces données. Pour extraire des classes ou des groupes ayant des fonctionnalités communes le nombre de classes et leur nature n'ont pas été préalablement déterminés. En d'autres termes, ils n'ont pas été inclus dans l'ensemble de données initiales (Campbell, 2013; Rutkowski, 2010). La distance est souvent la mesure la plus fréquemment utilisée pour mesurer la similarité entre les groupes. Les problèmes de l'apprentissage non supervisé sont le problème du groupement automatique (partitionnement) et le problème des méthodes permettant de réduire la dimension de l'analyse en composantes principales ; ou des méthodes servant à estimer les densités de probabilité. Nous nous intéresserons au problème de la classification non supervisée, dont le but est d'identifier automatiquement les groupes partageant des profils communs à partir de différentes variables. Dans la prochaine section, nous utiliserons le terme « regroupement » ou son équivalent Anglais « clustering », pour faire référence au processus de classification sans supervision. De façon similaire, le résultat de ce processus sera parfois appelé "système de groupement" ou plus simplement "classification".

I.6.2.C *L'apprentissage semi-supervisé (semi-supervised learning)*

Les méthodes d'apprentissage semi-supervisées sont des techniques d'apprentissage automatique qui utilisent un ensemble de données étiquetées et quelques données non étiquetées pour le même ensemble de données (Maâmatou, 2017). L'apprentissage semi-supervisé se situe entre l'apprentissage supervisé et l'apprentissage non supervisé. Lorsque les deux types de données sont disponibles, les méthodes d'apprentissage semi-supervisées peuvent être utilisées pour tirer parti, ce qui produit une amélioration considérable de la précision de l'apprentissage (Wang, 2009).

I.6.3 ***Les algorithmes d'apprentissage automatisé***

I.6.3.A *Algorithmes de classification supervisés*

La classification supervisée, souvent appelée simplement classification, est l'une des techniques les plus utilisées dans l'analyse des bases de données. Elle consiste à trouver un modèle ou une fonction pour classer les objets ou les données de manière plus large, de sorte que les objets du même groupe (appelés classes) sont plus proches (dans le sens d'un critère de (dis) similarité choisie) les uns des autres que ceux des autres groupes. La construction du modèle relève de l'apprentissage automatique, l'ensemble des exemples constituant le corpus d'apprentissage étant annoté, c'est-à-dire qu'ils portent l'étiquette de leur classe donnée a priori. Pour ce faire, un algorithme reçoit une formation sur des données semblables ou très proches des données que nous voulons classer (Azencott, 2019). Cela se fait en deux étapes. La première étape d'apprentissage consiste à établir une règle de classification pour décrire les classes prédéfinies pour un ensemble de données. La construction du modèle d'apprentissage est fondée sur l'analyse d'éléments de données ou de concepts dont les affiliations de classe sont prédéterminées ou connues. La seconde étape consiste à appliquer le modèle appris à de nouveaux éléments de données pour prédire leur appartenance à une classe. La composante supervision de cette procédure en est à la phase de formation, qui permet au classificateur d'évaluer une mesure de la dépendance entre les attributs et les classes.

En classification supervisée :

- on a déjà fixé le nombre de groupes.
- on sait à quel groupe appartient à chaque observation.
- Nous souhaitons classer les observations dans les groupes appropriés en fonction des différentes variables.

Parmi les algorithmes les plus connus fonctionnant avec cette méthode d'apprentissage sont: réseaux de neurones, arbre de décision, support de vecteur de machine, régression logistique.

I.6.3.B *Les réseaux de neurones*

Définition: les réseaux neuronaux artificiels sont des réseaux de processeurs élémentaires qui fonctionnent en parallèle. Chaque processeur élémentaire calcule un seul résultat fondé sur l'information qu'il reçoit. Chaque structure hiérarchisée de réseau est bien entendue un réseau. Les réseaux

neuronaux sont des algorithmes d'apprentissage automatisés qui visent à modéliser le fonctionnement du cerveau humain (G Gupta, 2017; Parizeau, 2004). Parmi les propriétés d'un réseau de neurones, il y a la capacité d'apprendre de son environnement, d'améliorer ses performances au moyen d'un processus d'apprentissage (Parizeau, 2004). Un réseau de neurones est un ensemble de neurones fortement connectés entre eux. Les neurones sont comme des processeurs élémentaires opérants en parallèle ou dans une structure hiérarchisée. Un neurone est donc avant tout une unité de calcul élémentaire dont la valeur numérique se calcule par quelques lignes de logiciels (Dreyfus, 2002). Ils sont principalement utilisés pour résoudre les problèmes de classification, de reconnaissance de formes, d'identification, ... Etc.

I.6.3.C *Arbre de décision*

Les arbres de décision sont des algorithmes de prédictions qui fonctionnent en régression et en classification.En matière de classification, il permet d'exprimer un processus séquentiel dans lequel une correspondance est établie entre un objet décrit par un ensemble de caractéristiques (attributs), et un ensemble de classes disjointes. Chaque feuille de l'arbre indique une classe et chaque nœud intérieur un essai sur un ou plusieurs attributs, produisant un sous-arbre de décision pour chaque résultat possible du test. En théorie des graphes, un arbre est un graphe non orienté, acyclique et connexe (Cornuéjols, 2011).

Cette série de nœuds est divisée en trois catégories:

- Nœud racine: l'accès à l'arbre est via ce nœud.
- Nœuds internes : nœuds ayant des descendants.
- Nœuds terminaux (ou feuilles) : nœuds n'ayant aucune descendance.

Autrement dit un arbre de décision, c'est un arbre binaire dont les feuilles correspondent à des classes en relation avec la problématique. En partant de la racine, chaque embranchement de l'arbre représente une règle binaire. En faisant parcourir l'arbre depuis la racine à un candidat dans le respect des règles de chaque embranchement, nous obtenons une prédiction de l'appartenance du candidat à une classe du problème.cision, c'est un arbre binaire dont les feuilles correspondent à des classes en relation avec la problématique.

I.6.3.D *Les machines à vecteurs de supports*

Les machines à vecteurs supports (En anglais : Support vecteur Machines SVM) sont des classes d'algorithmes de reconnaissance de formes et de l'intelligence artificielle. Ils ont été développés par Vapnik sur la base de la théorie de l'apprentissage statistique. Le SVM a été utilisé dans plusieurs applications ; la reconnaissance manuscrite, la prédiction des séries chronologiques, la reconnaissance vocale et beaucoup d'autres. Ces modèles reposent sur une théorie mathématique solide conçue à l'origine pour la classification binaire et la régression. C'est-à-dire qu'on doit décider de la classe à laquelle appartient l'échantillon ou de la régression à laquelle on doit prévoir la valeur numérique d'une variable.

I.6.4 ***Algorithmes de régression supervisée***

L'algorithme en classification ou en régression utilise les mêmes principes avec seulement quelques différences mineures.

CHAPITRE II :

Sectorisation automatique des réseaux d'alimentation en eau potable en fonction de la théorie du graphe

II Sectorisation automatique des réseaux d'alimentation en eau potable en-fonction de la théorie du graphe

II.1 Introduction

La distribution rationnelle de l'eau dans les systèmes d'alimentation est une question complexe. Cette complexité augmente si le réseau est grand et que le but est d'assurer un approvisionnement régulier en eau potable à la pression requise par les consommateurs (Campbell, 2016). L'analyse de réseau (représentés par des graphes) est une composante importante pour comprendre le système très complexe issu de nombreuses disciplines telles que la biologie, la géographie, les réseaux neurones, les réseaux de génies ou la sociologie (Queyroi, 2013). Pour résoudre ces problèmes, nous avons recours à des méthodes et à des algorithmes issus de la théorie des graphes. Étudier des graphes composés d'une poignée de nœuds et des graphes avec des millions d'entre eux exige souvent des approches différentes. Un réseau complexe est un graphe composé de nœuds qui peuvent être (individu, objet) liés par des liens qui sont des interactions ou des relations. Un graphe est représenté visuellement avec ses nœuds affichés en points et ses liens qui sont des flèches ou des traits entre les points, ce schéma, qui est très utile pour développer une première idée du système, n'est pas très approprié pour faire des calculs, si nous étudions un système avec des millions de nœuds, nous sommes également forcés d'afficher un nombre très limité de nœuds à chaque fois. Un graphe est défini par une série de sommets et une série d'arêtes reliant des sommets en paires. Le graphe est donc construit avec un sommet qui représente chaque élément, les arêtes de ce graphe permettent de représenter les dépendances entre les éléments. Une approche complémentaire, et beaucoup plus utilisée en pratique, est de représenter le graphe avec sa matrice d'adjacence. Pour être capable d'appliquer un partitionnement de graphe à un problème donné, il faut qu'il soit modélisé. Il faut choisir une modélisation adaptée au problème pour que le partitionneur puisse créer une partition qui optimise les critères appropriés pour cette simulation. La méthode la plus courante est de modéliser les éléments à décomposer en graphes. Le graphe construit a pour but de modéliser les contraintes du calcul distribué équilibrage de charge et réduction de la communication (Vuchener, 2014). Le partitionnement est défini comme la division d'un graphe en sous-graphes autonomes et non vides dont l'union est égale au graphe initial. Nous essayons d'identifier des sous-graphes homogènes (faible hétérogénéité au sein de la classe) et des sous-graphes fortement diffé-

renciés (forte hétérogénéité entre les classes). Le principe est de confondre la définition d'une communauté avec un groupe de nœuds ayant des particularités topologiques communes (Gosalia, 2019). La simplification d'un graphe correspond à la suppression de certains sommets et/ou liens, sans conduire nécessairement à la partition du graphe. Ces deux opérations ont deux buts principaux, le premier est de révéler la structure profonde du réseau, surtout pour ceux de grande taille que la visualisation n'apporte rien, le second détermine les groupes uniformes à l'intérieur du réseau. Le partitionnement se distingue par le fait qu'il ne définit pas au préalable la nature des sous-ensembles sollicités (Ducruet, 2011).

II.2 **Théorie des graphes**

La théorie des graphes est née avec Léonard Euler (1736), la résolution du problème du pont de Königsberg est considérée comme le premier théorème de la théorie des graphes. Ensuite, en 1847, Kirchhoff a utilisé la théorie graphe pour analyser les circuits électriques. La théorie des graphes est l'un des domaines d'étude des mathématiques des ensembles (finit et infinie) souvent utilisée pour analyser les propriétés géométriques, spatiales et les relations entre les éléments qui constituent un graphe ou un réseau. Elle prendra de plus en plus d'espace avec le développement de réseaux dont l'utilisation a besoin d'être optimisée. À titre d'exemple. À titre d'exemple :

- réseaux informatiques.
- réseaux sociaux.
- graphe du web.
- Transports routiers.
- Les réseaux de distribution d'eau ont aussi leur propre structure graphique. Où les nœuds de consommation et d'alimentation (réservoirs) sont les sommets et les canalisations et les vannes sont les arêtes (arcs)(Zevnik, 2018). Le terme « graphe » désigne la représentation visuelle d'un tel ensemble, il permet de représenter la structure, les raccordements d'un ensemble complexe en exprimant les relations entre ses éléments.

II.3 **Les réseaux / les graphes**

Les graphes traités dans ce document sont des graphes non orientés et non pondérés. Le graphe est un ensemble G = (V, E) où V (de l'anglais vertex) désigne les sommets et E (de l'anglais edge)

les arêtes(Eusebio et al., 2015). C'est-à-dire, est un ensemble de points appelés sommets ou nœuds liés par des liens appelés arêtes ou arcs qui permettent de représenter les relations binaires entre les éléments d'un ensemble. Ces arêtes sont souvent représentées par des segments, mais il est possible qu'ils soient courbés, ils peuvent être orientés ou non, en outre une valeur peut être associée à chaque arête ou nœuds. Nombre de sommets du graphe appelé ordre des graphes. Le degré du sommet est le nombre d'arêtes qui y sont liées. Ces graphes sont présentés au moyen d'une matrice, d'une liste d'adjacences ou d'une matrice d'incidence (Maquin, 2003).

II.3.1 ***Types de graphes***

II.3.1.A *Graphe oriente*

Un graphe orienté (Figure II-2) est un graphe dont les bords ont un sens et une orientation Ils ont une origine et une fin Nous ne pouvons les parcourir que dans un sens, c'est-à-dire quand l'arête (i, j) est différente de l'arête (j, i).

II.3.1.B *Graphe non oriente*

Un graphe non orienté (Figure II-1) est un couple (V, E) où V est un ensemble fini non vide et E un ensemble de couples non ordonnés d'éléments de V. Un élément de V est appelé un sommet et un élément de E est appelé une arête. L'arête est représentée par deux sommets.

II.3.1.C *Graphe pondere*

Un graphe est pondéré (Figure II-3) si à chaque arête (conduite), ont été affectées d'un nombre réel positif (ou coût) appelé poids. (Par exemple, si les diamètres sont associés aux conduites). On appelle graphe pondéré, un graphe dont les arêtes ont été affectées d'un nombre appelé poids.

II.4 **Représentation matricielle des graphes**

Un réseau est un graphe dont les valeurs numériques ont été associées aux nœuds et/ou aux conduites. Il y a plusieurs manières de stocker un graphe (réseau) dans un ordinateur. Plusieurs pratiques de représentation peuvent être distinguées selon la nature des traitements que l'on souhaite appliquer au graphe considéré. Les matrices comptent parmi les structures de données utilisées en informatique pour représenter les graphes et les relations (Falcini et al., 2007).

II.4.1 ***Matrice d'adjacence***

Il s'agit d'une matrice d'incidence « sommet-sommet ». Prenons un réseau (Figure II-1)

G = (V, E) à n sommets. La matrice adjacente de G (Tableau II-1) correspond à la matrice U = (u_{ij}) de dimension n x n de sorte que le graphe G= (V,E) binaire sur un ensemble de nœuds (V= 1,.....n).

$$U_{ij} = \begin{cases} 1 \; si \; i \; et \; j \; sont \; adjacents \\ 0 \; si \; non \end{cases}$$

Cette matrice, qui ne contient que "0" et "1", est appelée matrice booléenne et sera donc néces sairement symétrique.

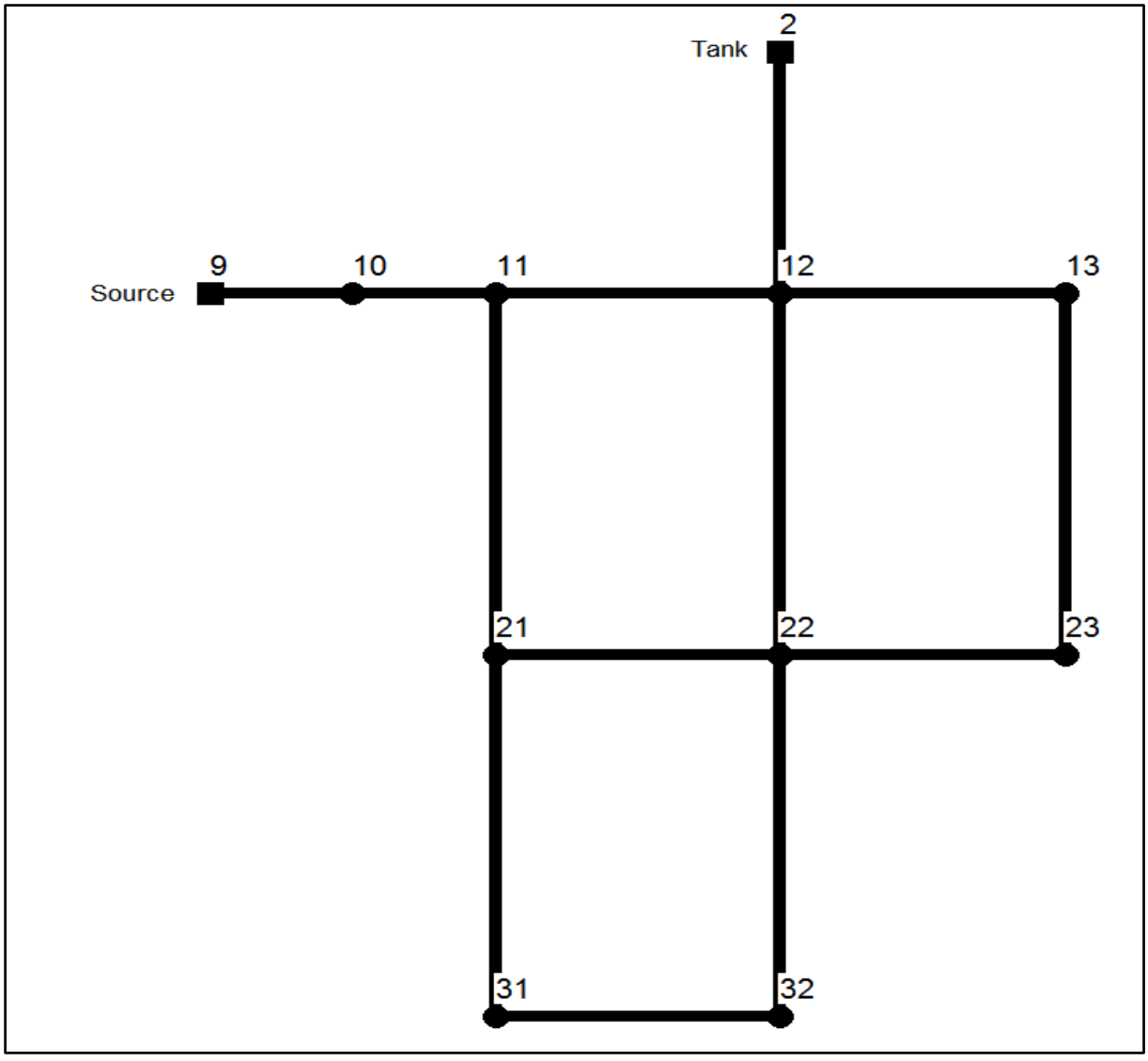


Figure II-1 Exemple du réseau Net1 non orienté

Table II-1 Matrice d'adjacence pour le graphe G de la figure II-1.

	2	9	10	11	12	13	21	22	23	31	32
2	0	0	0	0	1	0	0	0	0	0	0
9	0	0	1	0	0	0	0	0	0	0	0
10	0	1	0	1	0	0	0	0	0	0	0
11	0	0	1	0	1	0	1	0	0	0	0
12	1	0	0	1	0	1	0	1	0	0	0
13	0	0	0	0	1	0	0	0	1	0	0
21	0	0	0	1	0	0	0	1	0	1	0
22	0	0	0	0	1	0	1	0	1	0	1
23	0	0	0	0	0	1	0	1	0	0	0
31	0	0	0	0	0	0	1	0	0	0	1
32	0	0	0	0	0	0	0	1	0	1	0

II.4.2 *Degré d'un graphe*

Dans un graphe non orienté, Le degré deg (v) d'un sommet v est le nombre d'arêtes qui lui sont incidentes (Tableau II-2). Un sommet de degré 0 est appelé isolé.

Table II-2 Matrice des degrés du graphe G de la figure II-1.

	2	9	10	11	12	13	21	22	23	31	32
2	1	0	0	0	0	0	0	0	0	0	0
9	0	1	0	0	0	0	0	0	0	0	0
10	0	0	2	0	0	0	0	0	0	0	0
11	0	0	0	3	0	0	0	0	0	0	0
12	0	0	0	0	4	0	0	0	0	0	0
13	0	0	0	0	0	2	0	0	0	0	0
21	0	0	0	0	0	0	3	0	0	0	0
22	0	0	0	0	0	0	0	4	0	0	0
23	0	0	0	0	0	0	0	0	2	0	0
31	0	0	0	0	0	0	0	0	0	2	0
32	0	0	0	0	0	0	0	0	0	0	2

II.4.3 *Matrice d'incidence*

La seconde idée permettant une représentation matricielle d'un graphe (Tableau II-3) exploite la relation d'incidence entre arêtes et sommets. Est une matrice telle que chaque colonne correspond à un arc et chaque ligne à un sommet de G. Considérons un graphe orienté sans boucle G = (V, E) avec n sommets $v_1, v_2, \ldots\ldots v_n$ et E arêtes $e_1, e_2, \ldots\ldots e_m$ On appelle matrice d'incidence (aux arcs) de G la matrice M = m_{ij} des dimensions n x m tels que: 1 si v_i est l'extrémité initiale de e_j m_{ij} = -1 si v_i est l'extrémité terminale de e_j 0 si $v_i$xi n'est pas une extrémité de e_j.

$$m_{ij} = \begin{cases} 1 \text{ si } v_i \text{ est l'extrémité initiale de } e_j \\ -1 \text{ si } v_i \text{ est l'extrémité terminale de } e_j \\ 0 \text{ si } v_i \text{ n'est pas une extrémité de } e_j \end{cases}$$

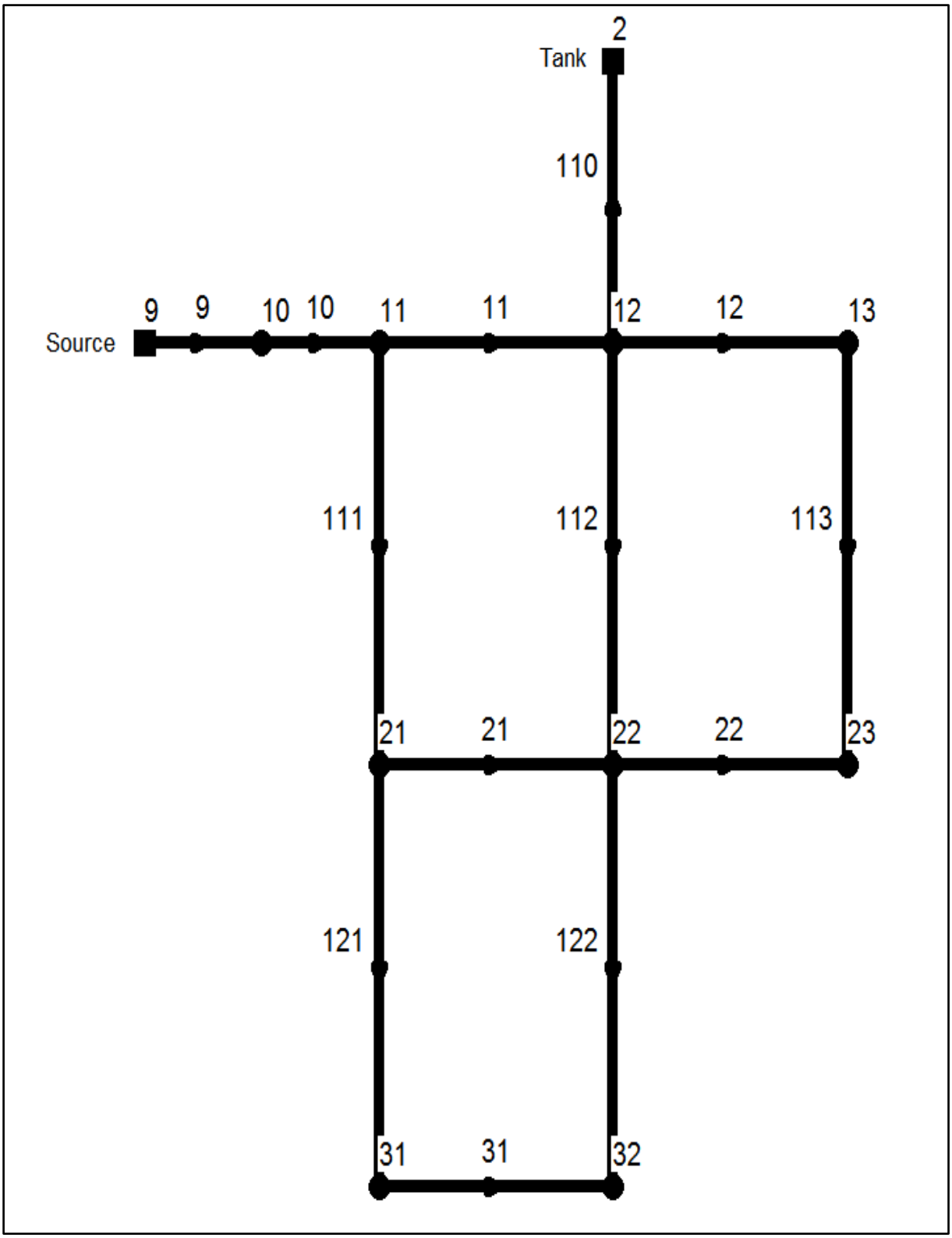


Figure II-2 Exemple du réseau Net1 orienté G

Table II-3 Matrice d'incidence du graphe orienté G de la Figure II-2

	9	10	11	12	21	22	31	110	111	112	113	121	122
2	0	0	0	0	0	0	0	-1	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0	0	0	0	0
10	-1	1	0	0	0	0	0	0	0	0	0	0	0
11	0	-1	1	0	0	0	0	0	1	0	0	0	0
12	0	0	-1	1	0	0	0	1	0	1	0	0	0
13	0	0	0	-1	0	0	0	0	0	0	1	0	0
21	0	0	0	0	1	0	0	0	-1	0	0	1	0
22	0	0	0	0	-1	1	0	0	0	-1	0	0	1
23	0	0	0	0	0	-1	0	0	0	0	-1	0	0
31	0	0	0	0	0	0	1	0	0	0	0	-1	0
32	0	0	0	0	0	0	-1	0	0	0	0	0	-1

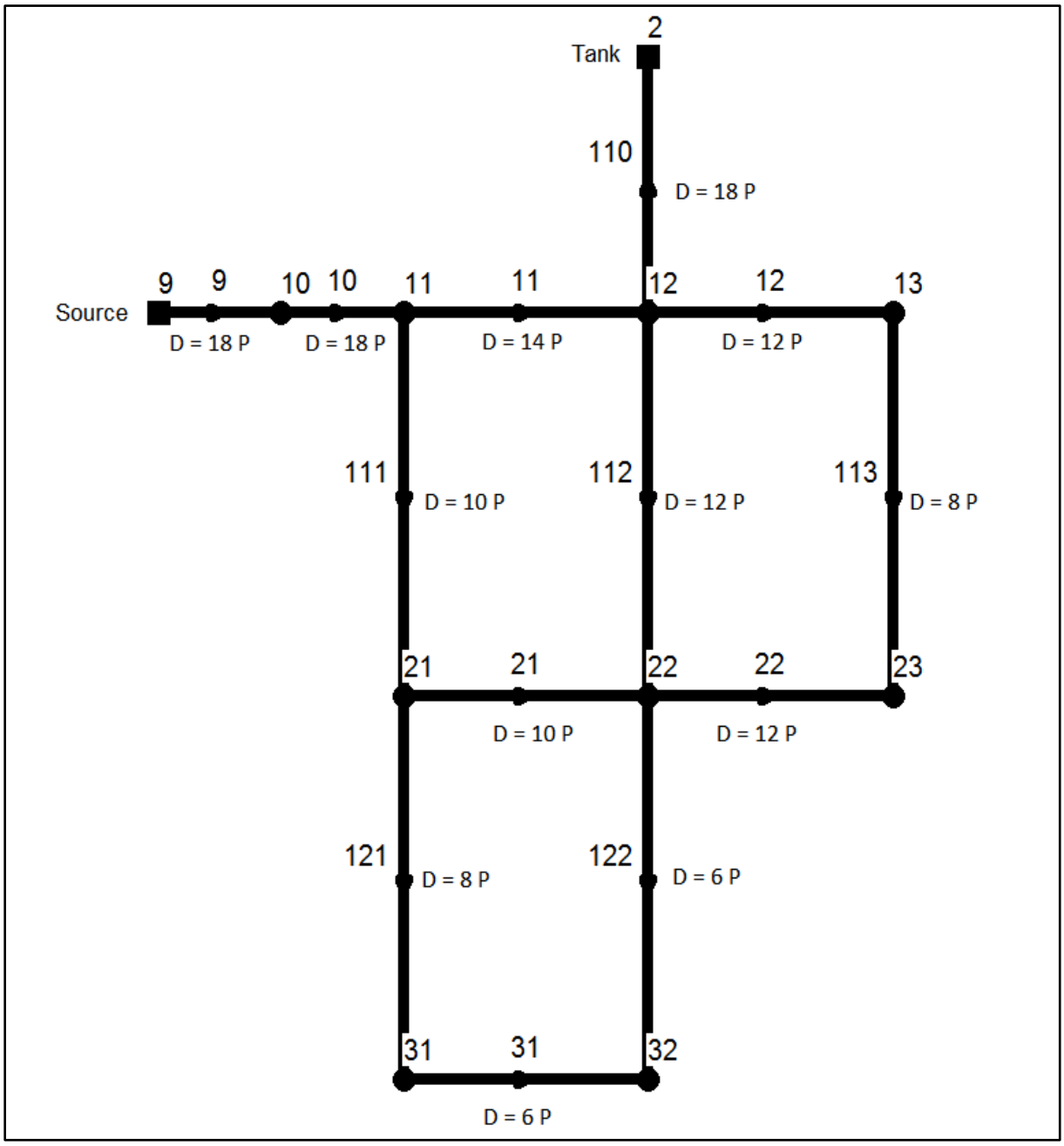


Figure II-3 Exemple du réseau Net1 du graphe G orienté est valué

(Le poids correspond au diamètre de la canalisation en pouces)

II.5 **Matrices laplaciennes de graphe**

Laplacian est une matrice associée à un graphe (Tableau II-4) dont les propriétés spectrales sont d'une grande utilité, c'est le principal outil de partitionnement spectral. Dans la littérature plusieurs formes de laplacien sont recensées. Nous supposons que le graphe de données G (V, E, W) est non orienté et sans pondération. La laplacienne est une matrice liée à un graphe dont les propriétés spectrales sont très utiles. Par exemple, nous verrons que les valeurs propres laplaciennes (Tableau II-6) nous donnent des informations sur le graphe (par exemple, s'il est lié ou pas). Avant cela donnons quelques définitions.

II.5.1 ***Matrice laplacienne non normalisée***

La matrice laplacienne non normalisée (Tableau II-4) est une combinaison linéaire de la matrice d'adjacence $\mathbb{W}$ et de la matrice de degré $\mathbb{D}$. Elle est définie de la façon suivante : $\mathbb{L} = \mathbb{D} - \mathbb{W}$.

Il résulte de cette définition que pour n'importe quel graphe, la somme des éléments dans chaque ligne de la matrice laplacienne est zéro(Matias, 2015). Cette matrice est symétrique et Semi-Définie positive, ce qui signifie que l'ensemble de ses valeurs propres est positif ou nul ($\lambda_i \geq 0$);elle présente des valeurs propres réelles non négatives, telles que.

$0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$.

1- L est symétrique (provient de la symétrie de D et W)

2- la plus petite valeur propre de L est. λ_1 =0 et son vecteur propre associé est un vecteur constant.

3- L est constitué de n valeurs propres réelles et non négatives.

Table II-4 Matrice laplacienne non normalisée.

	2	9	10	11	12	13	21	22	23	31	32
2	1	0	0	0	-1	0	0	0	0	0	0
9	0	1	-1	0	0	0	0	0	0	0	0
10	0	-1	2	-1	0	0	0	0	0	0	0
11	0	0	-1	3	-1	0	-1	0	0	0	0
12	-1	0	0	-1	4	-1	0	-1	0	0	0
13	0	0	0	0	-1	2	0	0	-1	0	0
21	0	0	0	-1	0	0	3	-1	0	-1	0
22	0	0	0	0	-1	0	-1	4	-1	0	-1
23	0	0	0	0	0	-1	0	-1	2	0	0
31	0	0	0	0	0	0	-1	0	0	2	-1
32	0	0	0	0	0	0	0	-1	0	-1	2

Table II-5 Valeurs propres du laplacien.

2	9	10	11	12	13	21	22	23	31	32
6.119	4.354	4.049	2.885	2.584	2.461	1.653	0.947	0.627	0.316	0.000

II.5.2 *Matrice laplacienne normalisée*

Dans la littérature, la matrice Laplacienne normalisée est généralement

$$\mathbf{L_1 = D^{-1} - L}$$

Il convient de souligner que les matrices laplaciennes (Tableau II-6) décrites ci-dessus sont semi-définies et ont N valeurs propres réelles non négatives (Tableau II-7), telles que:

0=0 = $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Ces propriétés sont les principales importances dans la théorie spectrale des graphes(Di Nardo, 2018).

Table II-6 Matrice Laplacienne normalisée.

	10	11	12	13	21	22	23	31	32	9	2
10	1.00	-0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.70	0.00
11	-0.40	1.00	-0.28	0.00	-0.33	0.00	0.00	0.00	0.00	0.00	0.00
12	0.00	-0.28	1.00	-0.35	0.00	-0.25	0.00	0.00	0.00	0.00	-0.5
13	0.00	0.00	-0.35	1.00	0.00	0.00	-0.50	0.00	0.00	0.00	0.00
21	0.00	-0.33	0.00	0.00	1.00	-0.28	0.00	-0.40	0.00	0.00	0.00
22	0.00	0.00	-0.25	0.00	-0.28	1.00	-0.35	0.00	-0.35	0.00	0.00
23	0.00	0.00	0.00	-0.50	0.00	-0.35	1.00	0.00	0.00	0.00	0.00
31	0.00	0.00	0.00	0.00	-0.40	0.00	0.00	1.00	-0.50	0.00	0.00
32	0.00	0.00	0.00	0.00	0.00	-0.35	0.00	-0.50	1.00	0.00	0.00
9	-0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
2	0.00	0.00	-.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table II-7 Valeurs propres du Laplacien.

2	9	10	11	12	13	21	22	23	31	32
2.000	1.797	1.677	1.410	1.187	1.000	0.812	0.589	0.322	0.202	0.000

II.5.3 *Chemin plus court*

Un chemin dans un graphe est une suite d'arcs successifs. La longueur d'un parcours correspond au nombre d'arcs qui composent le parcours (Tableau II-8).

Table II-8 Matrice chemin plus courts.

	2	9	10	11	12	13	21	22	23	31	32
2	0	4	3	2	1	2	3	2	3	4	3
9	4	0	1	2	3	4	3	4	5	4	5
10	3	1	0	1	2	3	2	3	4	3	4
11	2	2	1	0	1	2	1	2	3	2	3
12	1	3	2	1	0	1	2	1	2	3	2
13	2	4	3	2	1	0	3	2	1	4	3
21	3	3	2	1	2	3	0	1	2	1	2
22	2	4	3	2	1	2	1	0	1	2	1
23	3	5	4	3	2	1	2	1	0	3	2
31	4	4	3	2	3	4	1	2	3	0	1
32	3	5	4	3	2	3	2	1	2	1	0

II.6 **La classification automatique « Clustering »**

Sectoriser un réseau, c'est le subdiviser en zones interdépendantes (secteurs) et le doter de débitmètres pour tenir compte des volumes d'eau qui passe d'un secteur à l'autre. Deux problèmes surgissent donc : le premier est d'identifier les secteurs, le second de les isoler en déterminant les tuyaux à fermer et ceux à équiper de débitmètres. Si l'objectif principal de la sectorisation est le bilan hydrique et la détection de fuite de manière plus fine. Il y a beaucoup de questions qui doivent être prises en considération, à commencer par la sécurité de l'alimentation de l'abonné. De nombreuses études ont été réalisées sur la problématique de la sectorisation des systèmes de distribution d'eau. La plupart des auteurs abordent le problème avec des méthodes basées sur la théorie des

graphes : parcours de graphes, détection de communautés (Brentan, 2018) , algorithmes fondés sur la modularité (Liu, 2018) , partitionnement spectral (Di Nardo, 2018). Le clustering permet de grouper automatiquement les éléments. Cette approche vise à faire apparaître les groupes dans un ensemble d'éléments sans aucune information préalable dans ce cas, cette tâche s'appelle, selon le domaine, la classification non supervisée, la classification automatique ou clustering. Les groupes criés sont appelés clusters. Il en a été question dans de nombreux contextes et par des chercheurs dans de nombreuses disciplines ; c'est l'une des étapes les plus importantes en ce qui concerne l'exploration des données. L'objet de la détection communautés dans les graphes, consiste à créer une partition des sommets en tenant compte des relations qui existent entre les sommets dans le graphe, afin que les communautés soient constituées de sommets qui sont fortement liés et qui ont peu de liens entre eux (M. E. Newman, 2004). Parmi les plus importantes méthodes de détection communauté proposée dans la littérature ; on peut évoquer celles qui optimisent une fonction de qualité afin d'évaluer la qualité d'une partition donnée, telles que la modularité, les techniques hiérarchiques, les méthodes spectrales. Ces techniques de partitionnement des graphes sont très utiles pour détecter des composants hautement connectés dans un graphe (Combe, 2013).

II.7 **Théorie de la formation de clusters**

Les réseaux peuvent être modélisés à l'aide de graphe, lorsqu'un nœud représente un élément du système, et une arête représente un lien entre les nœuds en fonction d'une relation bien déterminée du système. La détection de communautés est souvent traitée comme un problème de regroupement (clustering) en fouille de données ou apprentissage machine et comme une situation de partitionnement de graphes en théorie des graphes.Le clustering (détection de communauté) est une tâche consistant à regrouper les nœuds en groupes (communautés ou clusters) selon une certaine notion de ressemblance. Dans une communauté disjointe, un nœud ne fait partie que d'une seule communauté. La communauté disjointe est également appelée affectation précise dans laquelle une relation binaire est établie entre un nœud et une communauté (Messaoudi, 2020).La détection de ces zones dites communautaires est un outil important qui permet de comprendre les structures et le fonctionnement de grands réseaux. L'idée est de structurer un ensemble de nœuds (physiques ou abstraits) dans différents groupes basés sur une certaine notion de ressemblance. Les nœuds qui sont considérés comme similaires, sont ainsi associés au même cluster alors que ceux qui sont considérés comme différents sont associés à des clusters distincts (M. E. Newman, Michelle, 2004).

C'est-à-dire découvrir des groupes au sein des données, de façon automatique. Ce domaine de recherche sur le clustring est étudié depuis de nombreuses années dans différentes communautés : l'apprentissage machine, extraction de données, statistiques, etc (Najma, 2014). Traditionnellement, il existe deux familles principales de méthodes de clustering: les méthodes hiérarchiques et de partition. Les méthodes hiérarchiques reposent sur des mesures de similitude entre sommets. Ils construisent des clusters en répartissant récursivement les instances de façon descendante ou ascendante. Le résultat de ce type d'algorithmes est un arbre de clusters nommé dendrogrammes, qui montre comment les clusters sont organisés. Pour les méthodes de partition, le résultat obtenu est une partition de l'espace objet, c'est-à-dire que chaque objet est associé à un seul cluster.

II.8 **Etat de l'art des techniques et technologies actuelles dans le domaine de la sectorisation**

Dans ce chapitre, nous décrivons les dernières avancées en matière de sectorisation automatique. Cette étape nous permettra de mieux comprendre les méthodes utilisées pour détecter les sous-réseaux dans les réseaux de l'AEP. Comme de nombreuses approches sont proposées, nous mettrons l'accent sur celles qui ont suscité le plus d'intérêt au sein de la communauté scientifique. Nous commençons par présenter les méthodes de partitionnement Fast Greedy, Random Walk (Liu, 2018) puis passons aux techniques spectrales (Di Nardo, 2018)

II.8.1 ***Présentation des réseaux à l'étude***

II.8.1.A *Exnet*

Tel qu'illustré dans la figure II-4(a), EXNET est un réseau de 400000 habitants. Le réseau compte 2416 conduites et 1891 nœuds. Les réservoirs surélevés alimentent l'ensemble du réseau (Liu, 2018).

II.8.1.B *C-town*

Le réseau de distribution d'eau de C-Town est l'un des modèles de référence classiques utilisés pour l'analyse (Pournaras et al., 2020). Il est basé sur un vrai réseau de taille moyenne avec 429 pipes et 388 noeuds figure II-4(b).

II.8.1.C *Oued El Ma*

Oued El Ma figure II-4(C) est une ville algérienne de taille moyenne ayant une population de 14000 habitants. Le réseau d'eau potable est constitué de 621 nœuds et de 630 tronçons.

Figure II-4 les réseaux de distribution d'eau Exnet(a), C-Town(b), Oued El Ma (c)

II.9 **Les methodes de partitionnement**

II.9.1 ***Travaux D'HAIXING LIU 2018***

En théorie, la sectorisation d'un réseau d'AEP se fait par regroupement de noeuds géographiquement proches les uns des autres. Les canalisations entre deux noeuds d'un même secteur font nécessairement partie de ce secteur alors que les canalisations aux frontières relient deux noeuds de différents secteurs. Nous introduisons les méthodes de partitionnement Fast Greedy, Random Walk, puis nous les mettrons à l'essai sur nos réseaux d'étude.

II.9.1.A *Fast Greedy*

La méthode de partitionnement Fast Greedy, développée par (Clauset, 2004). Est une méthode d'analyse de topologie basée sur la modularité. L'algorithme est basé sur la fonction de modularité Q. Chaque division possible du graphe possède sa propre valeur Q. Cette valeur est élevée pour une bonne division du graphe et faible si elle ne l'est pas. L'algorithme Fast Greedy optimise la valeur de modularité à travers toutes les partitions possibles pour trouver la meilleure partition du réseau de distribution d'eau (Liu, 2018). L'optimisation de la modularité, permet de rechercher directement le découpage en communautés correspondant à la valeur maximale de la modularité pour un graphe donné.

II.9.1.A.1 La modularite

La modularité est un indice de la qualité d'un partitionnement des nœuds d'un graphe, ou réseau, en communautés. Elle a été introduite par M. E. J. Newman. La meilleure structure de communautés est celle qui maximise la modularité (Talbi, 2013).

Modularité Q d'un graphe comme suit :

$$Q = \frac{1}{2m}\sum_{vw}\left[A_{vw} - \frac{K_v K_w}{2m}\right]\delta(C_v C_w)$$

Q est la valeur de modularité

m est le nombre d'arêtes dans le graphe

A_{vw}est un élément de la matrice d'adjacence du réseau

A_{vw}= 1 lorsque les sommets v et w sont connectés, sinon A_{vw} = 0

K_v est le degré d'un sommet v et est défini comme le nombre d'arêtes qui lui sont connectées

C_v et $C_w$sont les identifiants d'un cluster de réseau

δ est la fonction de la sommation des mêmes groupes (si C_v =C_w, alors δ = 1, et sinon δ = 0)

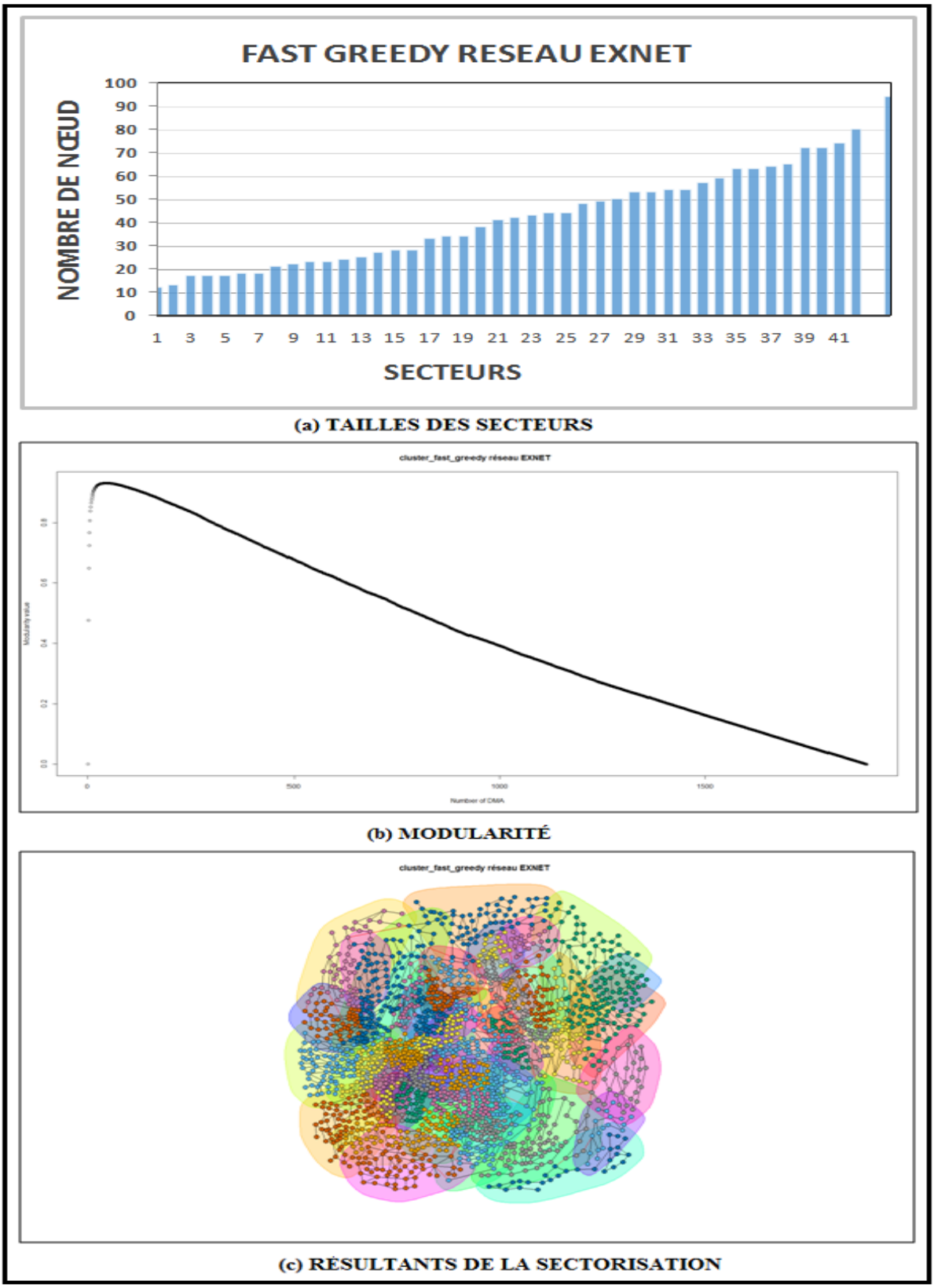


Figure II-5 Sectorisation du réseau d'AEP EXNET par Fast Greedy.

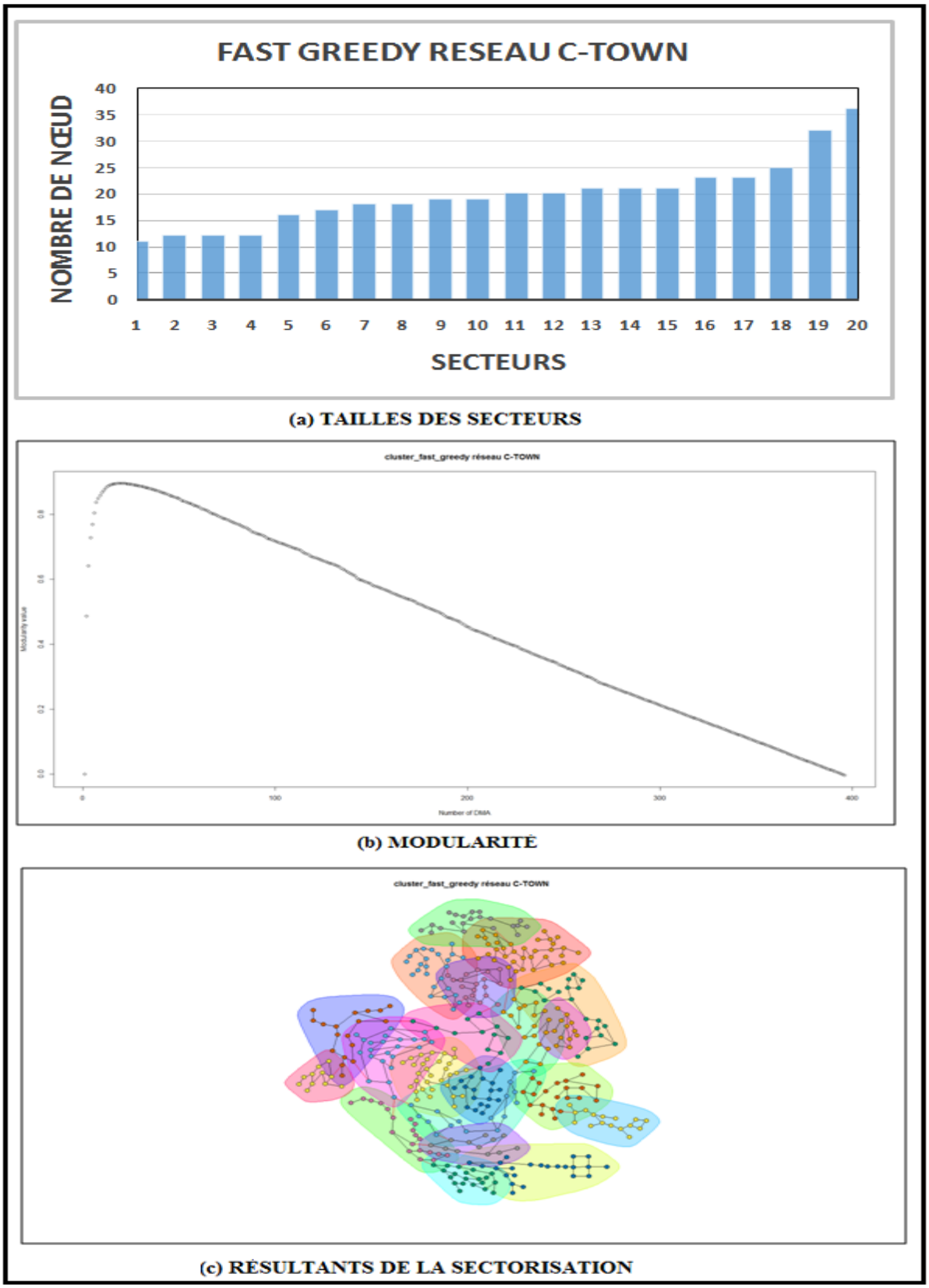


Figure II-6 Sectorisation du réseau d'AEP C-Town par Fast Greedy.

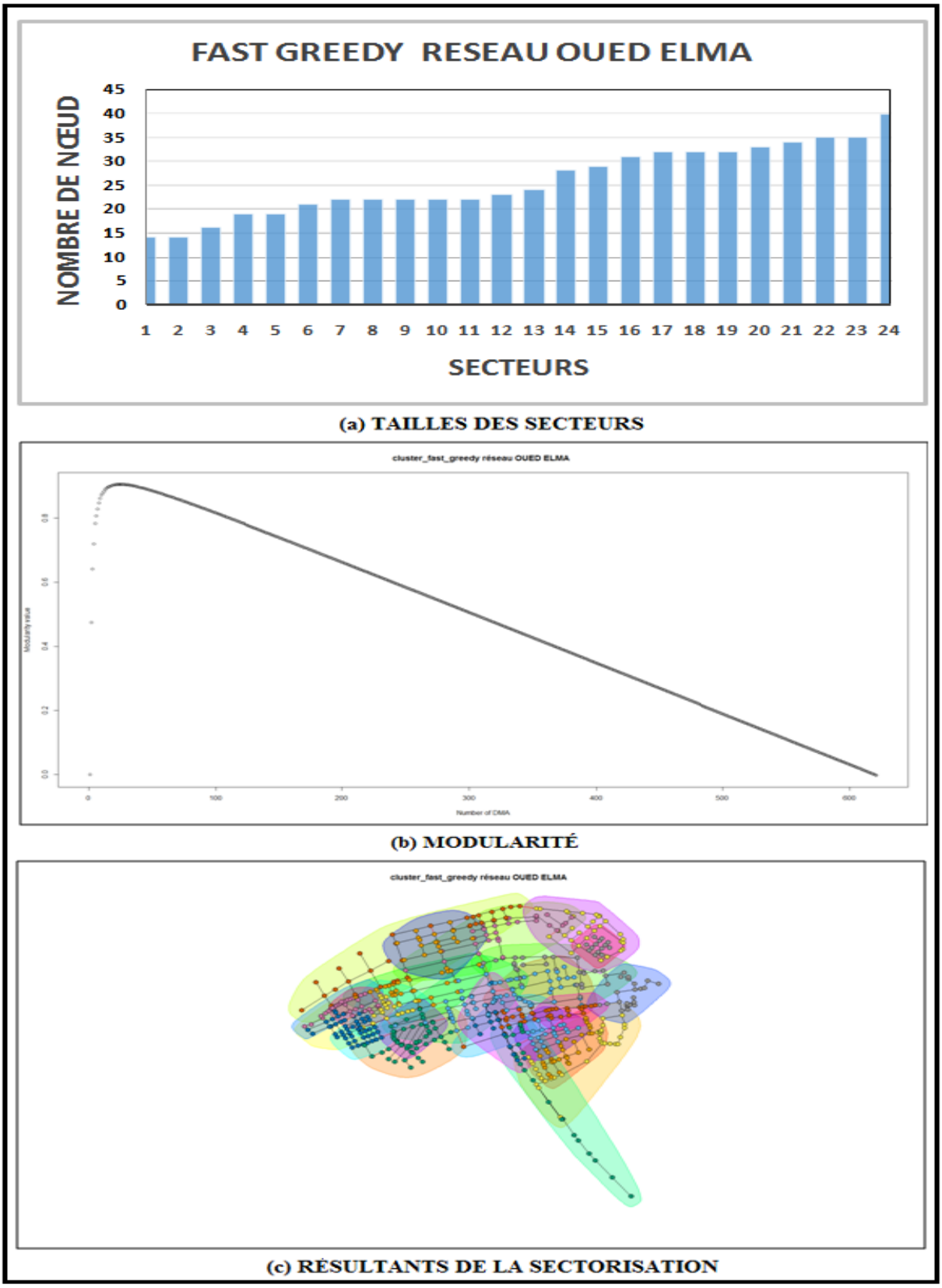


Figure II-7 Sectorisation du réseau d'AEP Oued El Ma par Fast Greedy.

II.9.1.B *Walktrap (Random Walk):*

C'est un algorithme qui utilise une distance entre sommets basés sur des marches aléatoires. Les marches aléatoires dans les graphes sont des processus aléatoires dans lesquels un marcheur est positionné sur un sommet du graphe et peut à chaque étape se déplacer vers un des sommets voisins (Liu, 2018). Le comportement des marches aléatoires est étroitement lié à la structure du graphe. Supposons que nous faisons une courte marche aléatoire sur le graphe à partir d'un nœud i la probabilité d'accéder à chacun de ses voisins dans une étape est 1/|Pi| (Talbi, 2013). Il est donc possible de calculer de la même manière la probabilité d'être au sommet j à partir d'i après avoir effectuait des k pas aléatoires. Cette probabilité permet de définir une distance entre les paires de nœuds du graphe dans laquelle deux nœuds i et j sont proches si leurs vecteurs de probabilité d'atteindre les autres nœuds sont similaires. Ainsi, la distance entre les résultats de deux marches aléatoires partant de deux sommets distincts révèle efficacement l'appartenance commune ou non de ces sommets à un même cluster (Kanawati, 2013). Une fois ces probabilités calculées pour toutes les paires de sommets, l'algorithme les utilise pour diviser le graphique via une méthode de classification hiérarchique. Commençant par n clusters ne contenant chacune qu'un seul sommet, l'algorithme cherche les deux clusters les plus proches, les fusionne, recalcule les distances, puis effectue une nouvelle fusion et ainsi de suite, jusqu'à n'obtenir qu'un seul cluster recouvrant tout le graphe

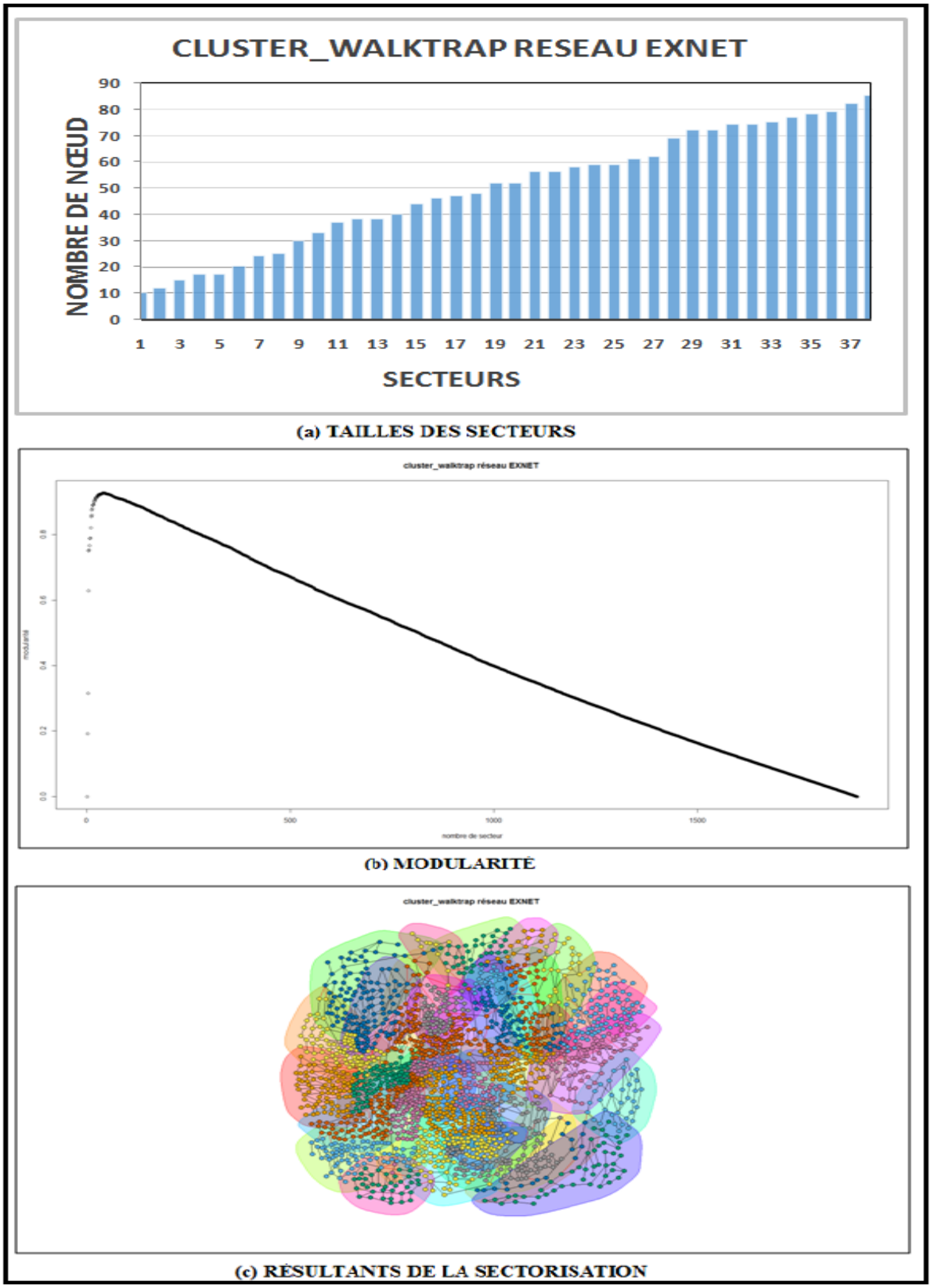


Figure II-8 Sectorisation du réseau d'AEP EXNET par Random Walk.

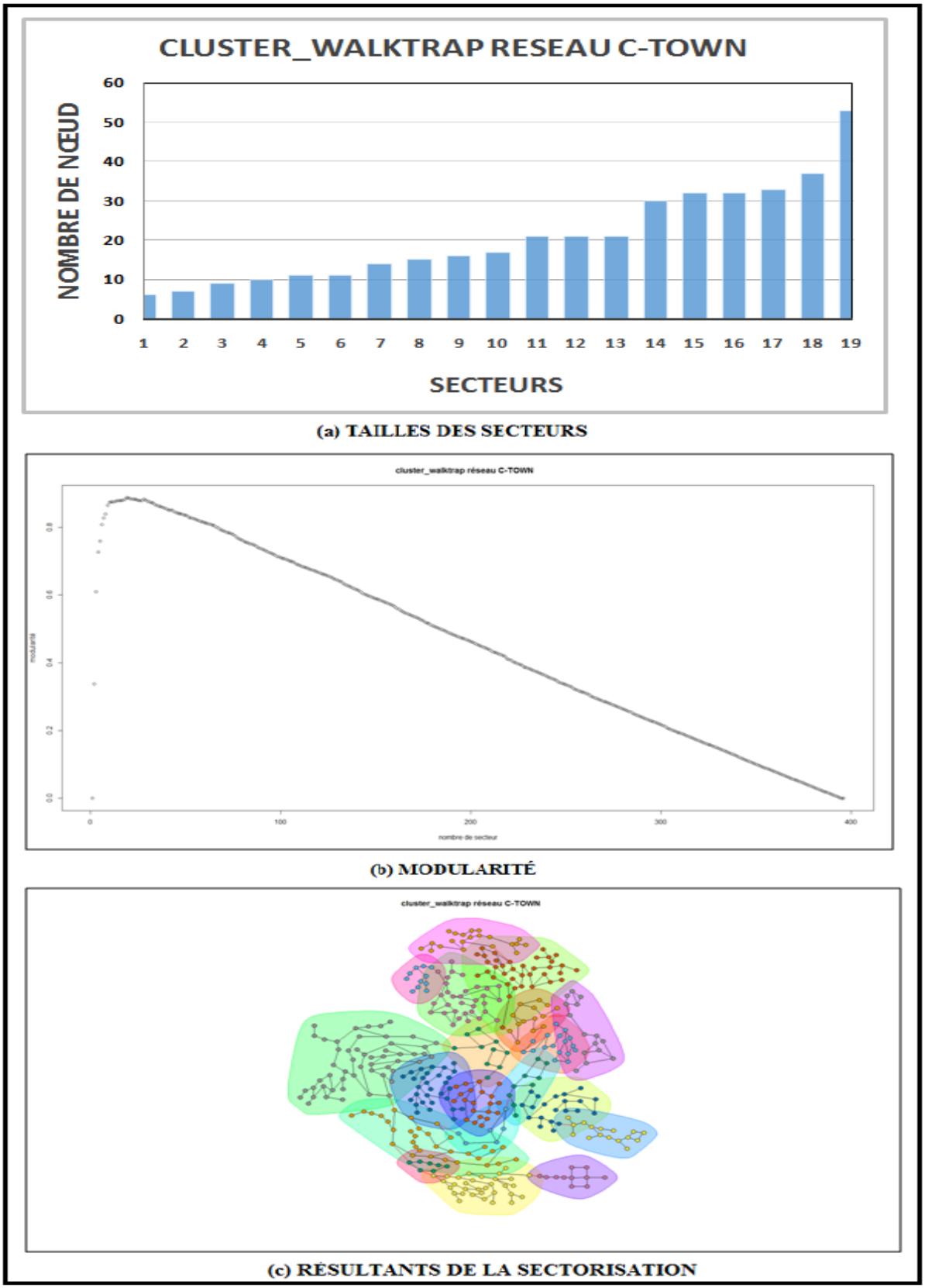


Figure II-9 Sectorisation du réseau d'AEP C-Town par Random Walk.

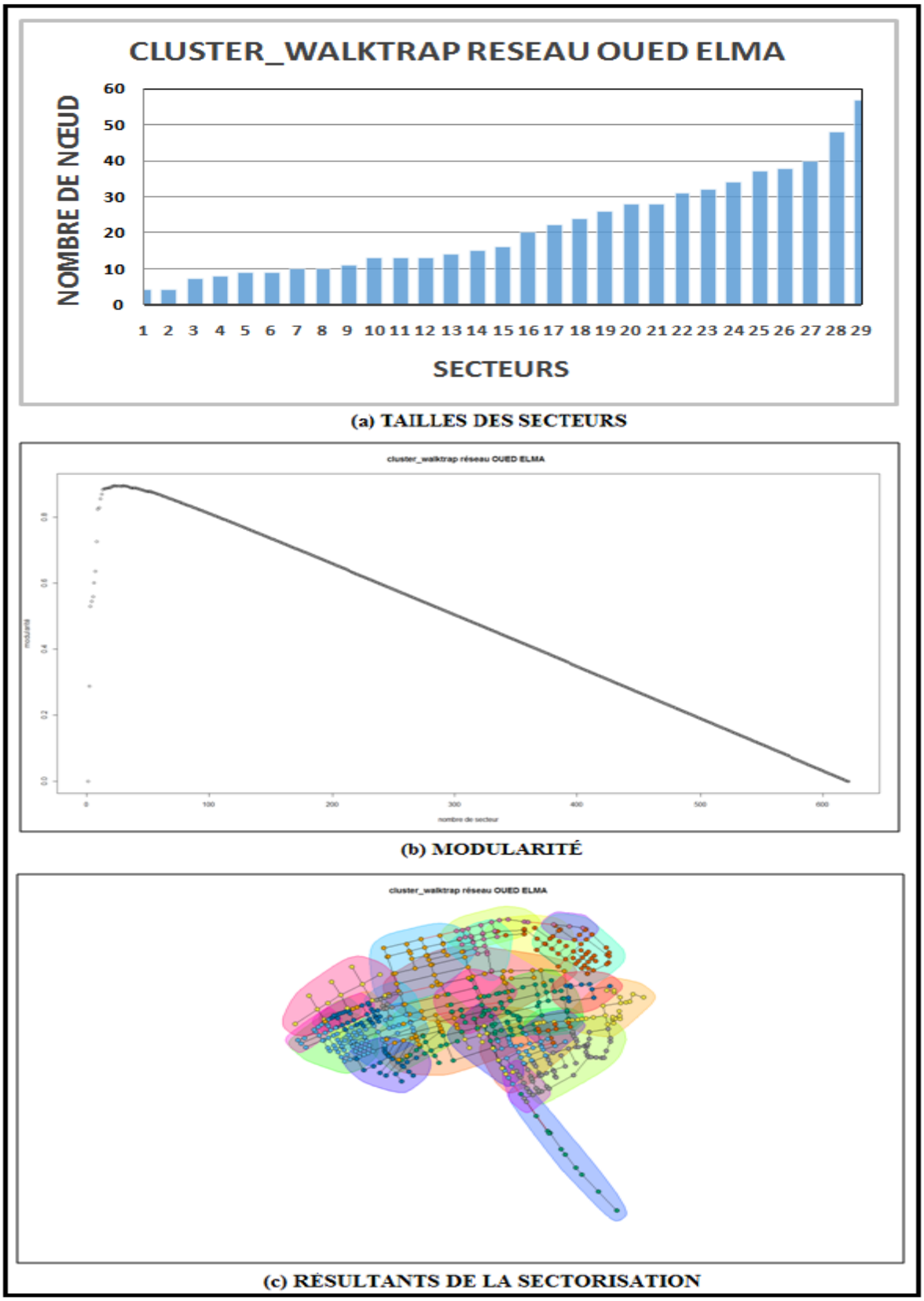


Figure II-10 Sectorisation du réseau d'AEP Oued El Ma par Random Walk.

II.9.2 *Resultats et analyse*

Le présent travail part de l'indice de modularité et l'utilise comme métrique pour la conception optimale de la segmentation pour les réseaux d'AEP. Cependant, l'utilisation directe de la formulation originale de l'indice de modularité dans les réseaux d'AEP, qui sont un type de réseau d'infrastructure, n'est pas recommandée (Giustolisi & Ridolfi, 2014). Le problème de la sectorisation repose toujours sur plusieurs éléments parmi ces éléments les contraintes techniques liées aux coûts d'investissement et de fonctionnement. Ce point est important, car il faut parvenir à une sectorisation optimale des réseaux d'AEP par rapport à la réduction du coût des équipements à installer (par exemple, des vannes d'isolement ou des dispositifs de mesure de débit). Nous constatons que l'application de ces méthodes entraîne la création d'un grand nombre de secteurs, donc un grand nombre de pièces et de débitmètres pour les séparer, ce qui suppose un investissement très important. Les figures (Figure II-5(a),10(a)) désigne le nombre et la taille de secteurs obtenus avec les méthodes Fast Greedy et Random Walk selon l'indice de modularité. Nous constatons que ce nombre est relativement élevé étant donné la taille des réseaux d'étude. Par exemple, le réseau Exnet se compose de 44 secteurs, le plus petit secteur ayant 12 nœuds et le secteur le plus important ayant 80 nœuds; d'autre part, le réseau C-town est constitué de 20 secteurs, le plus petit avec 11 nœuds et le plus grand avec 36 nœuds; et le réseau Oued El Ma comprend 24 secteurs, le plus petit secteur comptant 14 nœuds et le plus grand secteur comptant 35 nœuds. De même, les résultats de la méthode Random Walk sont presque identiques. Pour pouvoir recourir à ces méthodes, il convient de déterminer le nombre de secteurs, ce qui implique une réduction de l'indice de modularité figures (Figure II-5(b) ,10(b)).

II.9.3 *Travaux D'ARMANDO DI NARDO 2018*

Ce document propose une étude des possibilités qu'offrent les techniques spectrales des graphes. La classification spectrale est une méthode de classification en K groupes, basée sur le spectre de la matrice de similarités. L'analyse spectrale de graphes se base principalement sur l'étude des valeurs propres liées aux matrices laplaciennes des graphes. L'intérêt de cette application est de travailler sur un graphe qui représente précisément un réseau d'alimentation en eau. L'ensemble d'outils repose sur des informations topologiques et géométriques de la structure du réseau, aucun paramètre hydraulique (diamètre, rugosité, pression, etc.) n'est nécessaire. Il en résulte un nouvel

ensemble de techniques spectrales de graphe adaptées pour améliorer les tâches principales de la gestion de l'eau et de faciliter l'identification des pertes d'eau à travers la définition d'une partition optimale du réseau. L'importance de classer (partitionner) les nœuds dans un réseau de distribution d'eau permet d'identifier l'emplacement des vannes ou des capteurs. Les nœuds les plus influents ou importants ont également été obtenus. La proposition était donc particulièrement intéressante, parce que c'est une situation qui préoccupe souvent les services d'eau. Un autre avantage de la proposition qu'il fournit des paramètres utiles pour le contrôle de continuité en vérifiant s'il y a des parties non connectées du réseau d'eau.

II.9.3.A *Graphes et algorithmes de clustering spectral*

Ces dernières années, le clustering spectral est devenu l'un des algorithmes de clustering les plus populaires. Issu de la théorie des graphes et de l'analyse numérique. Il est de plus en plus utilisé aussi bien du fait de son efficacité; mais aussi pour sa simplicité d'exécution qui se résume par l'extraction du spectre (eigenvalues et eigenvectors) de sa matrice graphe associée (Di Nardo, 2018). l'approche est basée sur la décomposition spectrale laplacienne du graphe G = (V; E) que nous voulons trouver une partition de V dans des clusters ou des communautés (Rouvière, 2021). Le regroupement est le partitionnement d'un ensemble d'objets en groupes (clusters) de façon à ce que les objets d'un groupe soient plus semblables les uns aux autres qu'aux objets de groupes différents. La découverte de groupes par les spectres de graphes est utilisée dans réseaux sociaux, réseaux biologiques, réseaux d'information et réseaux technologiques (les réseaux technologiques sont des réseaux synthétiques conçus typiquement pour la distribution d'un certain produit ou ressource)

II.9.3.B *Algorithmes de clustering spectral*

Les algorithmes de clustering spectral minimisent le critère de coupe en résolvant un système de valeurs propres (ou un système de valeurs propres généralisé) grâce à l'extraction du spectre (un ensemble de valeurs propres) de la matrice laplacienne. Leurs processus de classification non supervisés se résument à quatre étapes Figure II-11 :

1. Prétraitement :

- la construction du graphique des données G(V, E).

2. Représentation spectrale.

- Construction de la matrice laplacienne normalisée associée au graphe G(V, E).
- extraction des valeurs et vecteurs propres de la matrice Laplacienne normalisée.
- La projection des objets dans l'espace spectral, basé sur le (s) vecteur (s) propre (s) retint (s).

3. Répéter le même travail au moyen du laplacien normalisé. Nous n'oublierons pas d'ajouter l'étape de normalisation qui se justifie par la théorie de la perturbation du spectre d'une matrice (Haghiri et al., 2017).

4. Partitionnement.

- Recherche de groupes dans l'espace spectral.
- Affectation des objets aux groupes.

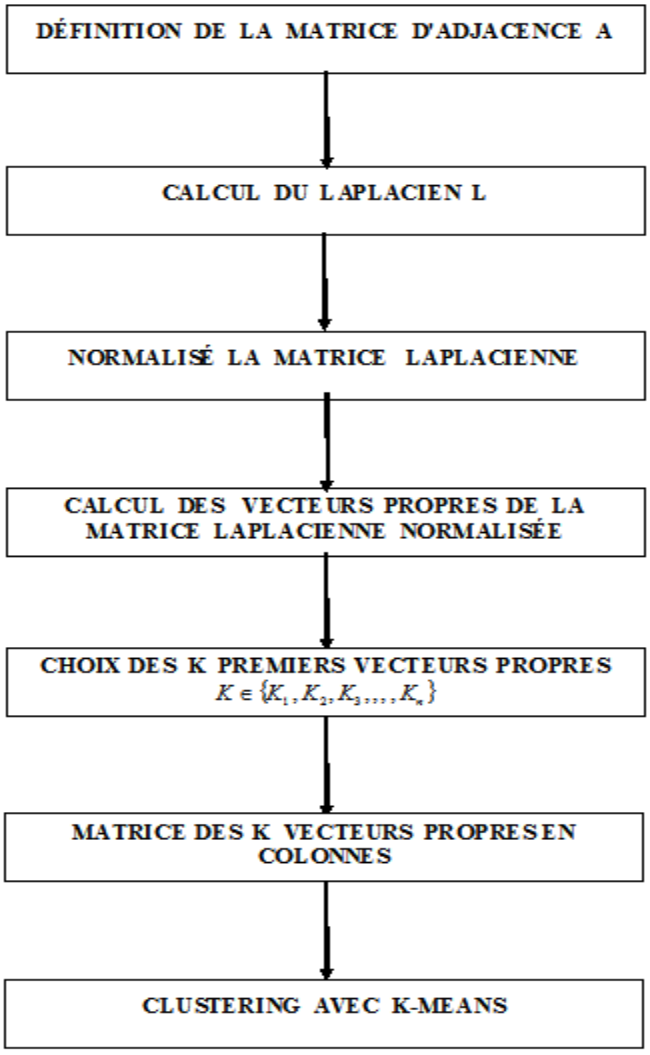


Figure II-11 Algorithme de clustering spectral.

II.9.3.C *Clustering K-means*

Le clustering K-means (Mac Queen 1967) est un algorithme d'apprentissage automatique non supervisé le plus couramment utilisé pour la partition d'un ensemble de données. Cet algorithme est employé sur de grands ensembles de données en raison de sa vitesse. Cette méthode vise à diviser les nœuds en partitions k dont chaque nœud appartient à la partition avec la moyenne la plus proche.

L'algorithme général de cette méthode :

Donnée : k le nombre maximum de classe désirée.

Début

(1) Choisir k objet (nœud) au hasard (comme centre des classes initiales)

(2) Affecter chaque objet au centre le plus proche

(3) Recalculer le centre de chacune de ces classes

(4) Répéter l'étape (2) et (3) jusqu'à stabilité des centres

(5) Éditer la partition obtenue

Fin

II.9.3.D *Valeurs propres, vecteurs propres*

La théorie spectrale des graphes consiste à formaliser les propriétés des valeurs propres et des vecteurs propres d'une matrice associée à un graphe. Cette matrice peut être une matrice représentative du graphe mais généralement le spectre étudié est celui de la matrice laplacienne. Le spectre du graphe est connu sous le nom de spectre de la matrice laplacienne.

II.9.3.E *Vecteur Fielder*

La deuxième valeur propre de L, qui est λ_2, est connue sous le nom de connectivité algébrique d'un graphe et le vecteur propre correspondant est communément appelé le vecteur Fiedler. Afin de tenir compte de la pertinence du vecteur Fielder, nous sommes intéressés par le problème de la division d'un graphique, dont l'objectif est de répartir un graphique connexe en deux sous-graphes (Di Nardo, 2018; Shi & Malik, 2000).

II.9.3.F *Le nombre optimal de secteurs k*

Afin d'obtenir un bon partitionnement d'un réseau (le nombre optimal de secteurs) où le nombre de secteurs est supérieur à 2, Di Nardo (Di Nardo, 2018) a utilisé des algorithmes de regroupement spectral, qui minimise le critère de coupe par la résolution d'un système de valeurs propres (ou un système de valeurs propres généralisé) par l'extraction du spectre (ensemble de valeurs propres) de la matrice laplacienne $\mathbf{L_1}$. Cette solution se caractérise par un nombre minimal de coupes et une taille de cluster (secteur) bien équilibrée. Le regroupement spectral est utilisé pour mettre en évidence le nombre de clusters. Nous recherchons un écart important entre les valeurs propres de la matrice de regroupement spectral, « eigengap » parce que la détermination de cette caractéristique

donne le nombre de groupes présents dans l'ensemble de données. La méthode heuristique Eigen-gap suggère que le nombre de clusters k est habituellement donné par la valeur de k qui optimise l'eigengap (différence entre les valeurs eigengap consécutives). (Si on laisse λ_1,…, λ_n valeur propre du laplacien, le but est de choisir k tel que λ_1,…, λ_k soient relativement petits, mais λ_{k+1}est relativement grand. Plus la différence entre les valeurs propres est grande, meilleure est la forme du regroupement.

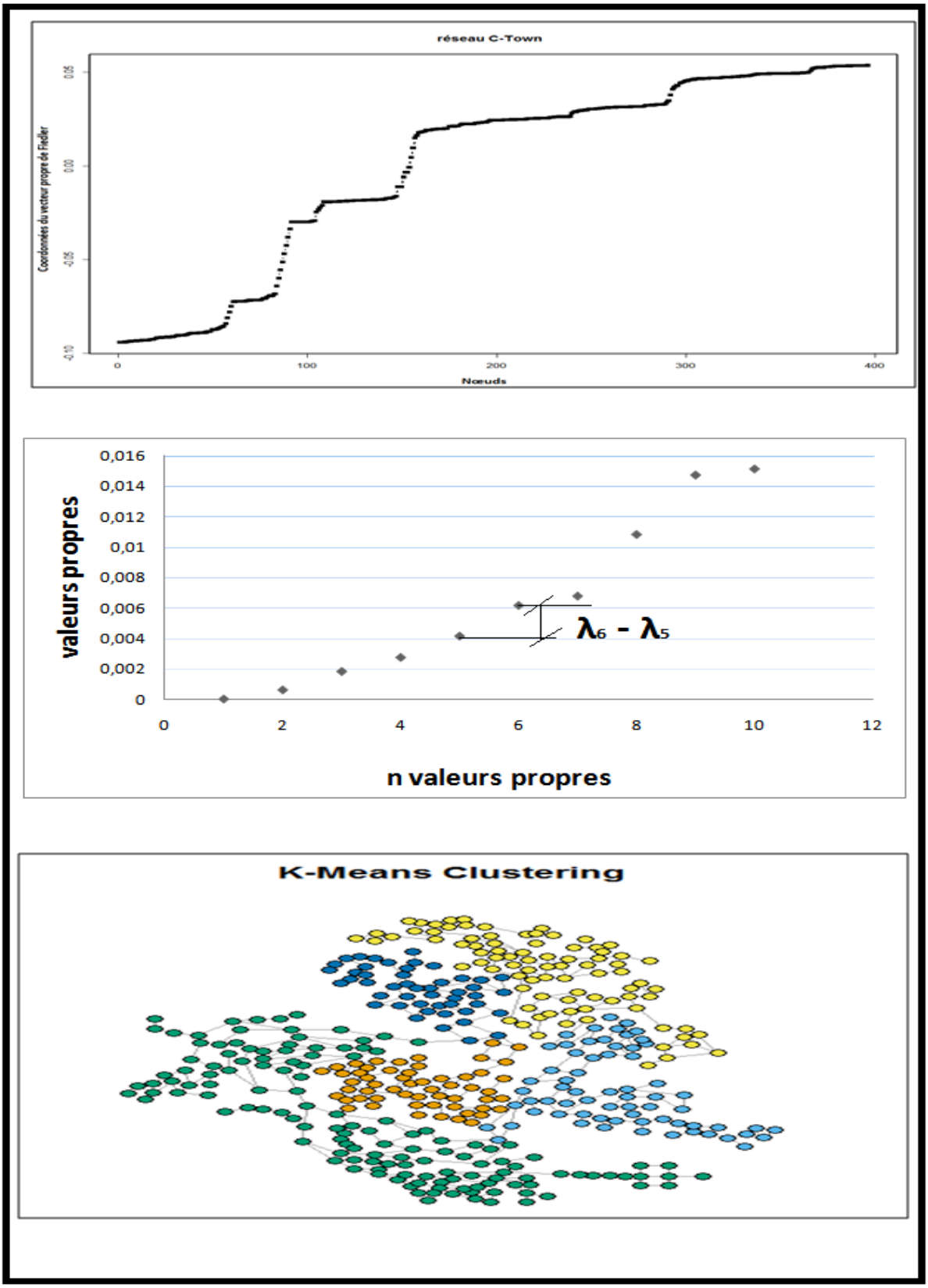


Figure II-12 Sectorisation du réseau d'AEP C-Town par k-means

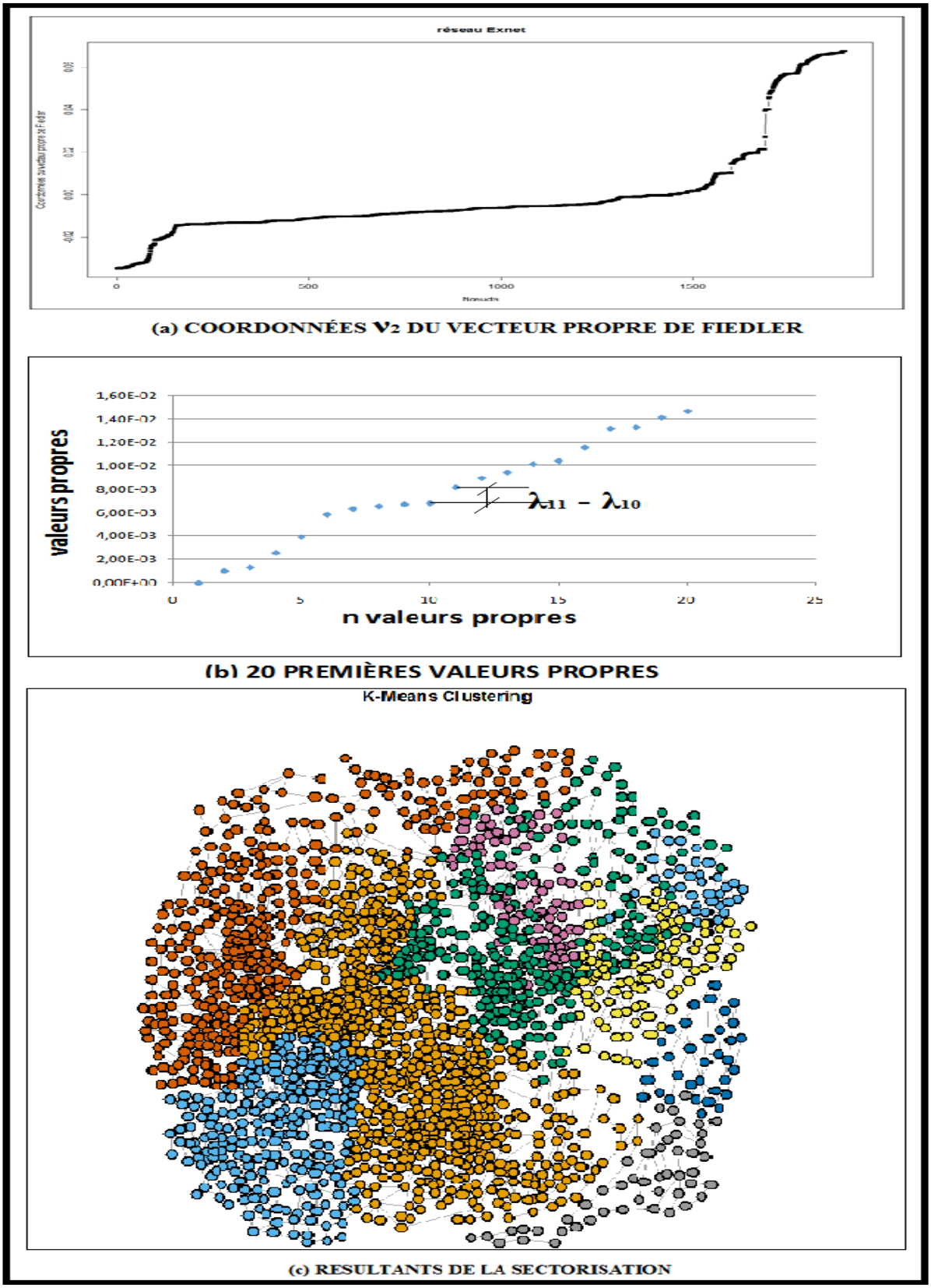


Figure II-13 Sectorisation du réseau d'AEP Exnet par k-means

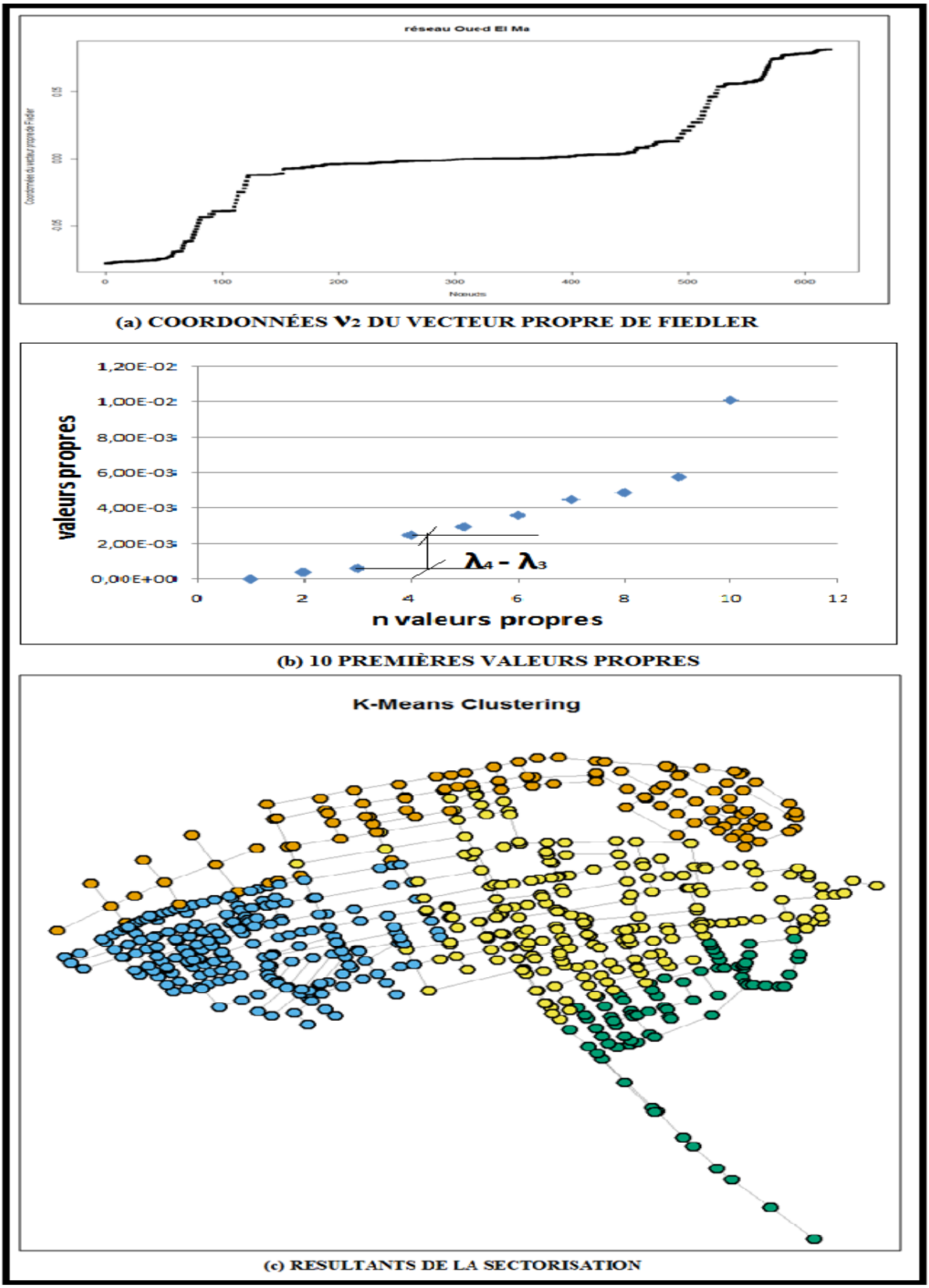


Figure II-14 Sectorisation du réseau d'AEP Oued El Ma par k-means.

II.9.4 ***Resultats et analyse :***

Le partitionnement spectral est une approche mathématique combinant à la fois l'algèbre linéaire et la théorie des graphes qui se résume à l'extraction du spectre. (Valeurs et vecteurs propres) des matrices associées aux graphes. Les méthodes spectrales sont des outils puissants pour réduire la dimension non linéaire et pour l'apprentissage. Leur fonctionnement repose sur la diagonalisation de matrices de "similitude" spécialement conçues, elle est définie à partir des graphes dont les sommets représentent les données à analyser et dont les arêtes indiquent des relations de voisinage. Il contient un ensemble complet de métriques et d'algorithmes, qui sont applicables au fonctionnement et à la gestion des systèmes d'approvisionnement en eau, comme le partage du réseau en secteurs par le processus de regroupement spectral. Parmi les propriétés de ces valeurs (vecteurs propres), nous pouvons déduire plusieurs paramètres comme le rayon spectral, l'écart spectral et Eigengap qui sera utilisé pour déterminer le nombre des groupes. Parmi les avantages de cette méthode est que le réseau d'AEP est modalisé sur la base d'informations topologiques et géométriques aucune donnée hydraulique (diamètre, rugosité, pression, etc.) n'est nécessaire; nous avons la possibilité de détecter le nombre de groupes. Les algorithmes de partitionnement spectral réduisent le critère de coupe optimal du graphe prédéfini, ce qui implique une réduction du nombre d'appareils requis pour sectoriser le réseau. Après application du partitionnement spectral est la démonstration de ses concepts fondamentaux, il a été conclu qu'il est facile à mettre en œuvre et donne de bons résultats.

CHAPITRE III :

Contribution: analyse comparative des méthodes de Clustering appliquées aux réseaux d'alimentation d'eau potable

III Contribution: analyse comparative des méthodes de Clustering appliquées aux réseaux d'alimentation d'eau potable

III.1 Introduction

Les villes sont dotées de nombreuses infrastructures hétérogènes et interconnectées pour fournir de l'eau potable aux consommateurs. En raison de cette complexité, des techniques numériques efficaces sont requises pour garantir le meilleur contrôle et la meilleure gestion des systèmes de distribution d'eau. D'autre part, l'évolution exponentielle de la puissance de calcul utilisé par les programmes de simulation nous a permis de réorienter l'analyse de conception et la gestion traditionnelle des réseaux d'alimentation. On peut considérer que la sectorisation des réseaux d'alimentation d'eau potable en sous-groupes homogènes constitue une stratégie de gestion ; qui consiste en la création de certains secteurs en insérant des vannes et des débitmètres sur les conduites du réseau. Cette partition a pour objectif de mieux gérer chaque sous-groupe (secteur) pour détecter des anomalies telles que : la détection des fuites d'eau, contrôle de la pression et protection des usagers contre les contaminations accidentelles et intentionnelles par une surveillance permanente des débits entrant dans chaque secteur. Récemment quelques techniques basées sur la théorie spectrale des graphes qui a simplifié le partitionnement et la sectorisation des réseaux et trouvé des solutions optimales ont été proposées dans la littérature (Di Nardo, 2018; A. N. Di Nardo, MicheleSantonastaso, Giovanni Francesco, 2014; Gutiérrez-Pérez et al., 2013; Herrera Fernández, 2011). Cette méthode est basée sur les propriétés spectrales des matrices associées à un graphe tel que la matrice adjacente et la matrice laplacienne. Elle consiste à projeter le graphe dans un espace métrique généré par les vecteurs d'une matrice liée à ce graphe en générant des données numériques. Puis nous appliquons les diverses méthodes de partition à ces données numériques pour partager le graphe. Le regroupement des nœuds du graphe est alors effectué par l'algorithme populaire K-MEANS qui est constamment utilisé. Il peut être utile de tenter de comparer les différents algorithmes existants (PAM, CLARA, HIERARCHICAL et DIANA) et de procéder à une analyse comparative. Nous avons sélectionné certains indicateurs de qualité généralement utilisés comme la modularité, l'indice interne (CONNECTIVITE, SILHOUETTE ET DUNN) et l'indice de stabilité (APN, AD, ADM AND FOM) afin de déterminer les algorithmes dominants. Les réseaux EXNET, C-TOWN et OUED EL MA font l'objet de comparaison. La sélection de ces réseaux s'est faite en fonction de leur nature et de leur taille.

Pour sectoriser un réseau d'AEP on va procéder de la manier suivante :

Algorithme Détection de communauté par la matrice de Laplacienne normalisée L_n

Données : Matrice d'adjacence A

Résultat : Matrice de partitionnement E

Calculer la matrice des degrés D

Calculer la matrice Laplacienne normalisée $L_n = D^{-1/2} LD^{-1/2}$

Calculer et ordonner les valeurs propres de L_n: $\lambda_1 \leq \cdots \leq \lambda_n$

Trouver K = arg max $(\lambda_{k+1} - \lambda_k)$

Former la matrice U à partir des K premiers vecteurs propres de L

Réaliser un clustering de type K-means sur les lignes de U

Affecter arbitrairement le résultat de ce clustering à E

III.2 **Principe general du clustering**

Le terme clustering fait référence à des méthodes de partitionnement des données. Il a pour but de diviser un ensemble de données en différents sous-ensembles en fonction de critères de proximité pour réaliser un bon partitionnement, deux critères doivent être considérés : d'une part, maximiser la proximité entre les éléments d'un même sous-ensemble et, d'autre part, minimiser la proximité entre les divers sous-ensembles.

III.3 **Methodes de clustering (ou regroupement)**

Plusieurs méthodes de Clustering ont été décrites et développées dans le cadre de la présente tâche. La littérature distingue différentes familles de méthodes telles que la hiérarchique et le partitionnement.

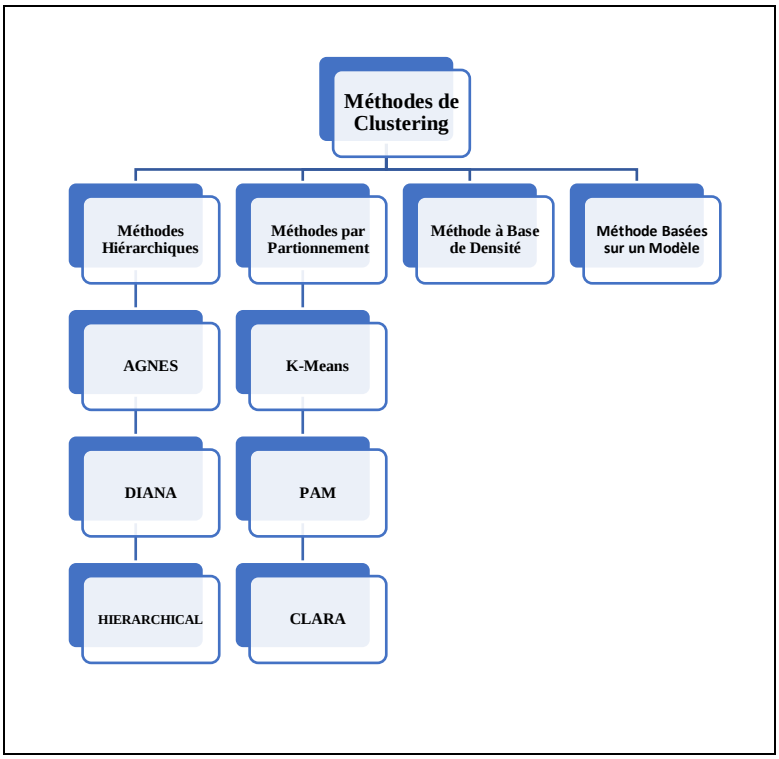


Figure III-1 Classification des méthodes de Clustering

III.3.1 *Methodes hierarchiques*

Les méthodes de clustering hiérarchique sont des méthodes visant à construire une hiérarchie de clusters ou, en d'autres termes, un arbre de clusters (plus communément connu comme Dendrogramme). Il s'agit d'un algorithme de clustering basé sur la connectivité. Pour parvenir à cette hiérarchie de cluster, il existe deux types d'approches : l'approche ascendante, connue sous le nom d'approche agglomérative (CHA Bottom up en anglais) et l'approche descendante (CHD Top down en anglais), appelée controversée (divisives). La méthode ascendante construit l'arbre depuis le bas vers le haut en commençant par autant de clusters que d'objets initiaux dans la base, ensuite fusionner successivement les clusters considérés comme les plus proches jusqu'à obtenir un seul cluster racine contenant tous les objets. En revanche, la méthode descendante construit l'arbre de haut en bas en commençant par un cluster unique contenant tous les objets de la base ; puis en divisant séquentiellement les clusters de telle manière que les clusters résultants soient aussi différents que possibles jusqu'à l'obtention des singletons (autant de clusters que d'objets en base). Voici certaines des méthodes hiérarchiques les mieux connues: DIANA, HIERARCHICAL.

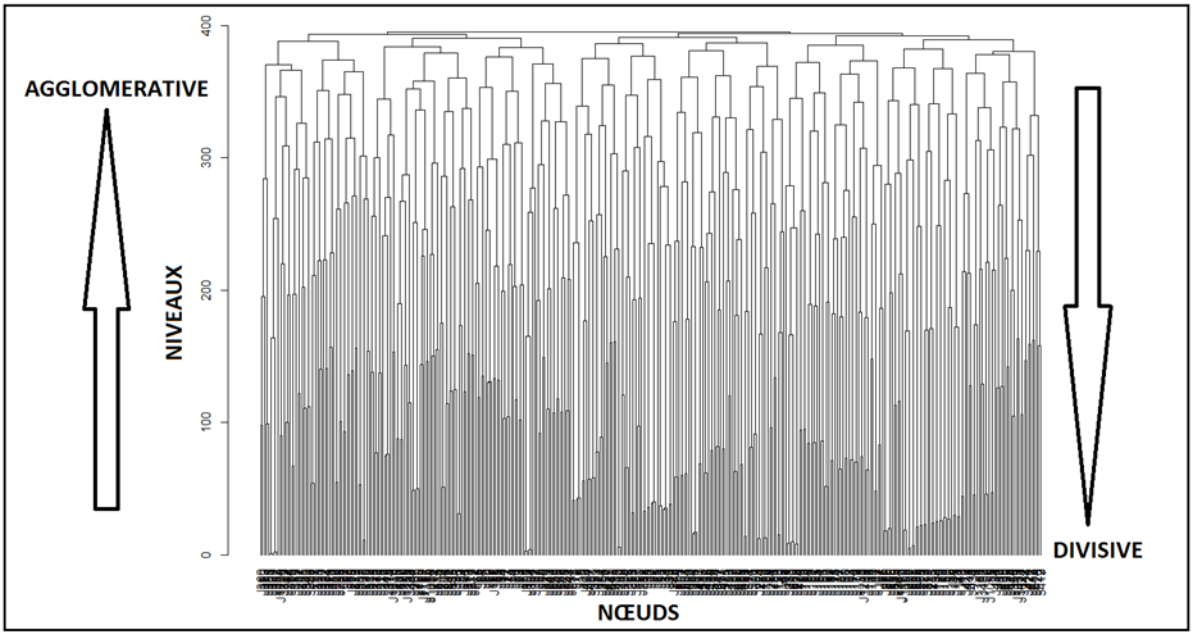


Figure III-2 Réseau C-Town déndrogramme

III.3.1.A *Algorithme Diana (Divisive Analysis)*

L'algorithme DIANA propose une structure descendante du dendrogram. Partant d'un cluster A non singleton et de plus grands diamètres (contenant initialement l'ensemble des objets $x_i \in X$), l'algorithme procède par divisions successives et itératives en deux parties A'et $\overline{A'}$ équilibrées. Le diamètre d'un cluster A est défini par:

$$Diam(A) = \max_{x_i \in A, x_j \in A} d(x_i, x_j)$$

À partir de $A = A'$ et $A = \emptyset$, la démarche est alors de transférer un ensemble d'objets.de A' à $\overline{A'}$ de façon à préserver l'équilibre entre ces deux ensembles. Nous choisissons de transférer à chaque étape l'objet $x_i \in X$ qui maximize

$$D(x_i, A\backslash\{x_i\}) = \frac{1}{\lceil A' \rceil - 1} \sum_{x_i \in A', x_j \neq x_i} d(x_i, x_j)$$

équivalent à une distance moyenne de l'objet x_i aux objets de $\backslash\{x_i\}$. Quand la quantité $\Delta(A', \overline{A'}, x_i) = D(x_i, A\backslash\{x_i\}) - D(x_i, \overline{A'})$ devient négative, le cluster x_i n'est alors pas déplacé et

le mécanisme de séparation de A s'arrête. Une nouvelle subdivision peut alors recommencer en choisissant à nouveau le cluster de plus grand diamètre entre A' et $\overline{A'}$.

III.3.1.B *Algorithme Hierarchical*

Le clustering hiérarchique est une méthode d'analyse de cluster qui vise à construire une hiérarchie de clusters, en d'autres termes, une arborescence basée sur la hiérarchie. Le nombre souhaité de clusters n'a pas besoin d'être défini à l'avance, car un dendrogramme peut être coupé à une hauteur particulière afin de générer un nombre spécifié de clusters. L'algorithme "hierarchical", commence par un cluster individuel distinct et fusionne les deux clusters les plus proches en un seul plus grand. Ce processus de fusion des deux clusters les plus proches se poursuit jusqu'à ce que tous les clusters soient combinés en un seul cluster. Il existe plusieurs méthodes pour calculer la distance entre les clusters, on peut mentionner la méthode de Ward où les clusters sont formés de manière à minimiser la somme des carrés dans les clusters. La distance entre deux clusters est l'accroissement de cette somme des carrés si les deux clusters sont fusionnés. Cette méthode utilise les critères ci-dessous pour recalculer la matrice des distances.

$$a(C_1, C_2) = \frac{|C_1| * |C_2|}{C_1 + C_2}\, d^2(gc_1, gc_2)$$

Avec : gc_1 est le centre de gravité du cluster C_1 et gc_2 est le centre de gravité du cluster C_2.

III.3.2 ***Methodes par partitionnement***

Dans les méthodes de partitionnement, les données sont divisées en un certain nombre de partitions dont chacune représente un cluster. Ce partitionnement en clusters est effectué de manière progressive. Cela signifie que le regroupement des données dans k cluster est effectué itérativement en améliorant un schéma initial, et en réaffectant ces données autour des centres mobiles (centroïdes). Ces clusters doivent donc remplir les conditions suivantes : dans un premier temps chaque groupe doit comporter au moins un objet, et en second chaque objet doit appartenir exactement à un groupe (cas de partitionnement strict). Quelques-unes des méthodes de partitionnement les plus mentionnées sont : K-Means, PAM, CLARA. Algorithme K-means

L'algorithme de k-means, ou l'algorithme de k-moyennes présenté à l'origine par (Mac Queen 1967). Il s'agit sans aucun doute de la méthode de partitionnement la plus connue et la plus répandue dans différents domaines d'application scientifique et industrielle. L'algorithme consiste à choisir de manière aléatoire des objets k qui représentent les centroïdes initiaux. Un objet est attribué au cluster pour lequel la distance de l'objet au centroïde est minimale. Le centroïde est ensuite recalculé et l'itération suivante est effectuée. La fonction objective généralement utilisée est:

$$F = \sum_{r}^{k} \sum_{x_i \in C_r} (x_i - g_r)^2$$

C_r : est le cluster numéro r.

g_r : est le centre du cluster C_r.

x_i : est un objet dans un cluste C_r.

III.3.2.A *Algorithme PAM (Partition Around Medoids)*

Il est introduit par Kaufman et Rousseew. L'algorithme PAM est une méthode de partitionnement basée sur les medoïdes pour la création de partitions (clusters). La médoïde est l'objet représentatif dans le cluster (au lieu de la moyenne). Le principe de cet algorithme consiste à commencer avec un ensemble de k médoïdes puis échanger le rôle entre un objet médoïde et un non-médoïde si cela permet de réduire la distance globale. Ce qui revient à minimiser la fonction couts. Le coût total de la permutation d'une médoïde $\overrightarrow{x_i}$ par un non-médoïde $\overrightarrow{x_h}$ est donné par :

$$TC_{ih} = \sum_{j=1}^{n} C_{jih} = \sum_{j=1}^{n} [dist(j,h) - dist(j,i)]$$

TC_{ih} représente le gain en distance globale que l'on va avoir en remplaçant h par j, Si TC_{ih} est négatif alors on va perdre en distance. Ça veut dire que les clusters seront plus compacts.

III.3.2.B *Algorithme Clara (Clustering Large Applications)*

L'algorithme CLARA a été mise en œuvre par Kaufman et Rousseeuw dans le but de réduire le coût de calcul de PAM (Candelieri et al., 2014; Díaz et al., 2008; Sublemontier, 2012). La méthode CLARA se fait en échantillonnant l'ensemble de données, en itérant à plusieurs reprises les étapes suivantes:

- tirage aléatoire de 40 + 2k $objets$[10]
- Appliquer l'algorithme PAM à l'échantillon obtenu.

Enfin, la partition fianale est obtenue en attribuant tous les objets aux médoïdes depuis le schéma le mieux généré.

III.4 **Techniques d'evaluation de la qualite du clustering**

Pour évaluer la qualité et la cohérence des clusters résultant de différents algorithmes de Clustering, deux mesures doivent être prises en considération: la première est la cohésion, qui mesure à quel point les objets sont étroitement liés dans un même cluster, la deuxième est la séparation, qui permet de mesurer la différence entre un cluster et d'autres. Pour évaluer et comparer les différents algorithmes de clustering, nous utiliserons des mesures de validité interne et de stabilité ainsi que de modularité.

III.4.1 ***Validation interne***

L'indice de validité interne est l'évaluation des résultats de Clustering à l'aide des caractéristiques du jeu de données ; cela signifie qu'il détermine l'adéquation entre les clusters et les données en utilisant uniquement les données elles-mêmes, plutôt qu'avec des informations à partir d'une bonne partition connue a priori. Certains des indices les mieux connus sont l'indice de connectivité, l'indice de Dunn, le coefficient de silhouette.

III.4.1.A *Indice de connectivité*

Cet indice utilise les distances entre les sommets pour mesurer la cohésion et la séparation des clusters. Elle est définie de la manière suivante (Brock et al., 2008; Handl & Knowles, 2005) :

$$Conn = \sum_{i=1}^{N} \sum_{j=1}^{L} \sum x_{i,nn_{i(j)}}$$

$$x = \begin{cases} \frac{1}{j}, IF \exists c_k : i \in c_k \wedge nn_{i_{(j)}} \in c_k \\ 0, autrement \end{cases}$$

Ou

K : est le nombre de clusters.

N : est le nombre total d'observations (lignes).

$nn_{i,(j)}$: est le $j^{ème}$ plus proche voisin du point de données i.

L : est le paramètre déterminant le nombre de voisins qui contribuent à la mesurede la connectivité. L'indice de connectivité à une valeur allant de zéro à ∞ ; et doit être minimisé.

III.4.1.B *Indice de Dunn*

Grâce à cet indice, il est possible d'identifier des clusters compacts et bien séparés. C'est le rapport entre la distance minimale entre les inter-clusters et la distance maximale entre les intra-clusters (Brock et al., 2008; Dunn, 1974). Pour calculer l'indice de Dunn, nous adoptons la formule suivante :

L'indice de Dunn généralisé

$$D = \frac{\min\limits_{C_k, C_1 \in C, C_k \neq C_l} \left(\min\limits_{i \in C_k, j \in C_l} dist(i,j) \right)}{\max\limits_{C_m \in C} ddiam(C_m)}$$

Où diam C_m est la distance maximale entre les observations dans le cluster C_m. L'indice Dunn a une valeur comprise entre zéro et 1 ; et doit être maximisé

III.4.1.C *Coefficient de silhouette*

Le coefficient de la silhouette mesure le degré de confiance dans l'attribution du clustering à un point particulier. Pour chaque point (i), son coefficient de silhouette représente la différence entre la distance moyenne avec les points du même cluster (cohésion) et la distance moyenne avec les points des autres clusters voisins (séparation) (Brock et al., 2008; Rousseeuw, 1987). Les points bien regroupés ayants des valeurs proches de 1 et les points mal regroupés ayants des valeurs proches de -1. La silhouette se définit comme :

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)}$$

b_i Est la distance moyenne de l'objet i aux autres objets du même cluster.

a_i Est le minimum des distances moyenne de l'objet i avec tous les autres objets du cluster le plus proche.

III.4.2 ***Validation externe***

Les mesures de stabilité comparent les résultats du Clustering basé sur les données complètes au Clustering basé sur la suppression de chaque colonne, une à la fois. Ils comprenaient la proportion moyenne de non-chevauchement (APN), la distance moyenne (AD), la distance moyenne entre les moyennes (ADM), et la figure de mérite (FOM).

III.4.2.A *La proportion moyenne de non-chevauchement (APN)*

L'APN mesure la proportion moyenne d'observations qui ne sont pas placées dans le même cluster en les regroupant en fonction de l'ensemble de données complet et en supprimant une seule colonne(Brock et al., 2008; Punitha, 2019). Cette mesure se présente sous la forme suivante :

$$APN(K) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\left(1 - \frac{n\left(C^{i,l} \cap C^{i,0}\right)}{n(C^{i,0})}\right)$$

K : est le nombre total de clusters.

N : est le nombre total de lignes (observation

$C^{i,0}$: représente le cluster qui contient l'observation i en utilisant le Clustering d'origine (basé sur toutes les données disponibles).

$C^{i,l}$: représente le cluster qui contient l'observation i où le Clustering est basé sur l'ensemble de données avec la colonne supprimée.

III.4.2.B *La distance moyenne (AD)*

AD calcule la distance moyenne entre les observations placées dans le même cluster dans les deux cas (l'ensemble complet des données et avec suppression d'une colonne) [Punitha, 2019]. La mesure AD se présente comme suit :

$$AD(K) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{l=1}^{M}\frac{1}{n(C^{i,1} \cap C^{i,0})}\left[\sum_{i\in C^{i,0}, j\in C^{i,l}}\left(dist(g_i, g_l)\right)\right]$$

$dist(g_i, g_j)$: est une distance (par exemple euclidienne, Manhattan, etc.) entre deux profils degènes d'expression i et j.

K : est le nombre total de clusters.

N : est le nombre total de lignes (observations).

M : est le nombre total de colonnes (attributs).

$C^{i,l}$: représente le cluster qui contient l'observation i en utilisant le Clustering d'origine (basé sur toutes les données disponibles).

III.4.2.C *La distance moyenne entre les moyennes (ADM)*

ADM calcule la distance moyenne entre les centres de cluster pour les observations situées dans le même cluster dans deux cas : celui où le clustering est effectué sur l'ensemble de données et celui effectué sur les données avec seulement une colonne supprimée (Brock et al., 2008; Punitha, 2019). Elle est définie comme suit :

$$ADM(K) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{l=1}^{M} dist\left(\bar{x}_{C^{i,l}}, \bar{x}_{C^{i,0}}\right)$$

N : est le nombre total de lignes (observations).

M : est le nombre total de colonnes (une collection d'échantillons, des points temporels).

$\bar{x}_{C^{i,l}}$: est la moyenne des observations dans le cluster qui contient l'observation i, lorsque le Clustering est basé sur l'ensemble de données avec la colonne supprimée. Actuellement, ADM utilise uniquement la distance euclidienne.

$\bar{x}_{C^{i,0}}$: est la moyenne des observations dans le cluster qui contient l'observation i, lorsque la classification est basée sur les données complètes.

III.4.2.D *La valeur du merite (FOM)*

FOM mesure la moyenne de la variance intra-cluster des observations de la colonne supprimée, dans lesquelles le Clustering est effectué en fonction des colonnes restantes (non supprimées)(Brock et al., 2008; Punitha, 2019). Pour une colonne spécifique, celle-ci est définie comme suit :

$$FOM(l,K) = \sqrt{\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in C_k(l)} dist\left(x_{i,l}, \bar{x}_{C_{k(l)}}\right)}$$

N : est le nombre total de lignes (observations).

K : est le nombre total de clusters.

$x_{i,l}$: est la valeur de l'observation $i^{ème}$ dans la colonne $l^{ème}$ du cluster.

$\bar{x}_{C_{k(l)}}$: est la moyenne du cluster $C_{k(l)}$. Actuellement, la seule distance disponible pour FOM est la distance Euclidienne.

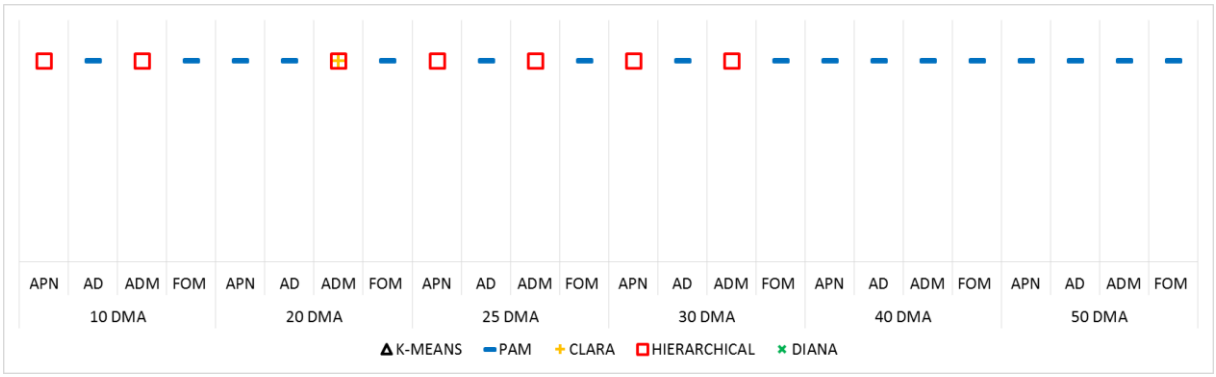
III.5 **Resultats & discussion**

Cette section présente les résultats commentés et discutés. Pour déterminer la performance des cinq algorithmes de clustering, une variation du nombre de clusters (secteurs) a été effectuée sur les réseaux en fonction de leur taille (Exnet de 10 à 50, C-town de trois à 10 et pour Oued el ma de trois à 12). Huit critères de performance sont calculés pour chacun des cinq algorithmes de clustering. L'ensemble de ces résultats est présenté dans les figures. III-3، III-4 et III-5 ci-dessous

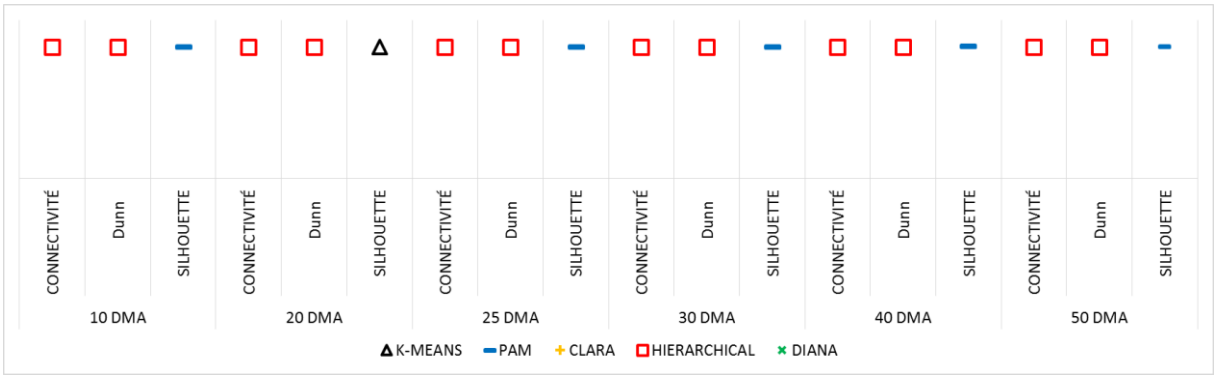
III.5.1 ***Évaluation des techniques de clustering pour le reseau Exnet***

D'après les résultats présentés dans la figure III-3, nous pouvons noter que l'algorithme Pam présente les meilleures performances avec les valeurs de mesure de la validation externe, il est dominant sur toutes les mesures pour les partitions (clusters) (50,40). Tout comme les mesures « AD et FOM » pour les partitions 10, 20, 25 et 30. Cette performance est suivie par l'algorithme hiérarchical qui affiche quelques meilleures performances telles que « APN et ADM » pour les partitions 10, 20 et 30.D'un autre côté, les indices de validité interne décrivent que l'algorithme hiérarchical a les meilleures performances sur tout avec des indices de connectivité et Dunn et pour toutes les

partitions. Pour la modularité Pam et dominante pour les partitions 10, 20 et 50 k_mens et performante pour les partitions 25 et 30 clusters.

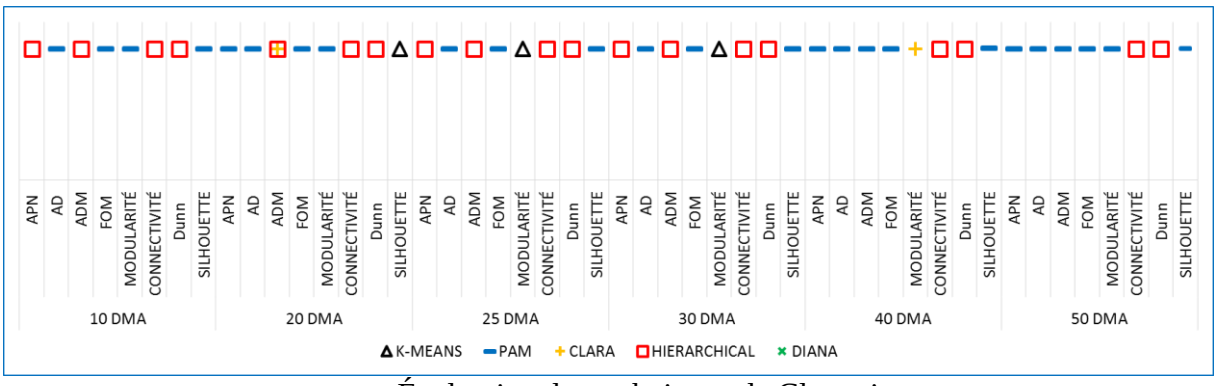


VALIDATION EXTERNE



VALIDATION INTERNE

MODULARITÉ



Évaluation des techniques de Clustering

Figure III-3 Évaluation des techniques de Clustering pour le réseau Exnet.

III.5.2 *Évaluation des techniques de clustering pour le reseau C-Town*

Compte tenu des résultats présentés à la figure III-4, on peut constater que l'algorithme Pam est le plus performant pour un nombre de clusters égal à (quatre et 10) elle domine sur l'ensemble des mesures de la validation externe. Cette performance est suivie par les algorithmes hiérarchical et Clara qui présentent les meilleures performances respectivement, pour un nombre de clusters égal à cinq (resp six et huit). Du point de vue des indices de stabilité, nous pouvons remarquer que les algorithmes hiérarchical et K-means fournissent des performances pour tous les indices de connectivité, Dunn et Silhouette pour la plupart des partitions. En ce qui concerne la modularité, nous constatons que l'algorithme K-means est efficace pour les partitions égales à 3 et 5; ainsi que Diana pour un nombre de clusters égal à quatre et huit, Tandique Pam s'impose pour un nombre de clusters égal à 5 et 10.

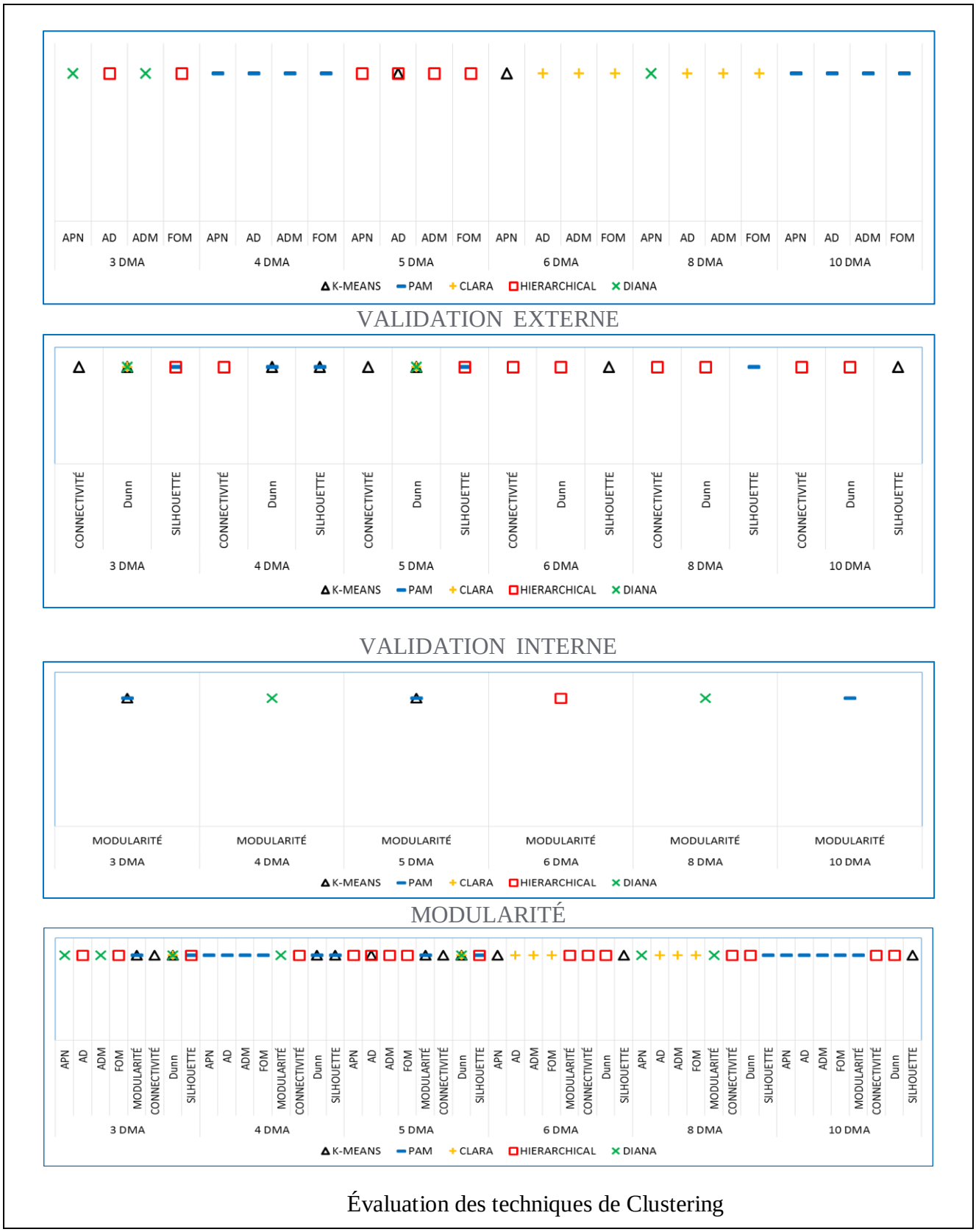


Figure III-4 Évaluation des techniques de Clustering pour le réseau C-town.

III.5.3 *Évaluation des techniques de clustering pour le reseau Ouel El Ma*

D'après les résultats présentés dans la figure III-5 ; Diana et K-means sont performants pour toutes les mesures de validation externe pour un nombre de clusters égal à 3 (resp huit). Les mesures de stabilité montrent aussi que les algorithmes de partitionnement, Clara produit également des meilleures performances avec AD, ADM et FOM pour un nombre de clusters égal quatre, six et dix, PAM affiche des performances avec les mesures AD, ADM et FOM pour un nombre optimal de 12. Hiérarchical obtient également un bon résultat avec les mesures APN et ADM pour un nombre de clusters égal à quatre et six. De l'autre côté, les indices de stabilité décrivent que Hierarchical fournit des performances pour tous les indices de connectivité, Dunn pour la plupart des partitions. Cette performance est suivie par l'algorithme K-means pour l'indice Silhouette pour un nombre de clusters égal à trois et huit. Sur le plan de la modularité, nous constatons que l'algorithme de Pam obtient de bons résultats pour les partitions supérieures à six.

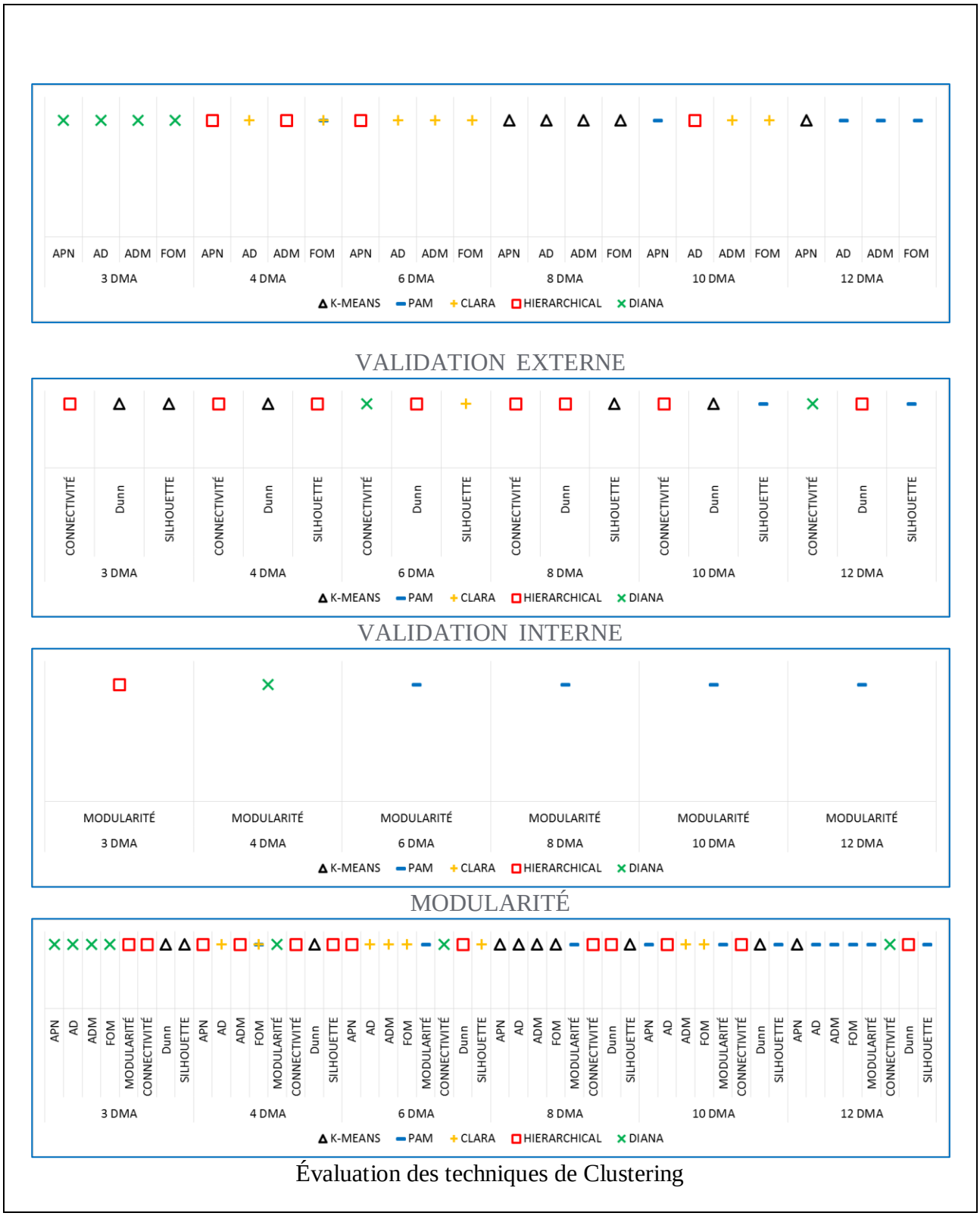


Figure III-5 Évaluation des techniques de Clustering pour le réseau Ouel El Ma.

III.6 **Conclusion**

On peut améliorer la gestion de réseau en les divisant en secteurs (DMA). L'une des méthodes les plus utilisées pour partitionner le réseau est le clustering spectral. Qui consiste à générer un espace spectral à partir des vecteurs propres des matrices laplaciennes associées au réseau. L'étape finale de ce clustering est souvent effectuée par l'algorithme K-means, qui n'est pas toujours performant. Il est donc utile d'examiner les performances d'autres algorithmes existants pour identifier les plus efficaces, en termes d'indices de qualité tels que la modularité, les indices internes (Connectivité, Silhouette et DUNN) ainsi que les mesures de validation externe tels que APN, AD, ADM et FOM. Une série de simulations a été réalisée pour évaluer la performance de 5 algorithmes de clustering sur 3 types de réseaux. Les réseaux étudiés ont été sélectionnés selon leur type et de leur taille. Nos résultats montrent que k-means n'est pas obligatoirement efficace, ni sur le plan du type de réseau, ni sur celui du nombre de secteurs. Notre étude a révélé que k-means ne fonctionne bien que dans certaines conditions, par exemple pour les petits réseaux dont le nombre de secteurs est très limité. Il n'est donc pas possible de confirmer la validité de l'utilisation des k-means en tant qu'outil de sectorisation sur les réseaux grands et moyens.

CHAPITRE IV:

DIAGNOSTIC DES DEFAILLANCES PAR SUPPORT VECTOR MACHINES

IV Diagnostic des defaillances par support vector machines

IV.1 **Introduction**

Le rôle d'un gestionnaire du réseau d'eau potable est de fournir aux usagers de l'eau en quantité suffisante et de meilleure qualité. À cet effet, il dispose d'installations apparentes telles que les usines de traitement, les réservoirs et les réseaux de conduites enfouis dans le sol, une fois construits, ils subissent une dégradation en raison du temps et doivent être surveillés, contrôlés, entretenus et rénovés. Le réseau de conduites ne fait pas exception au temps et à l'action de différents phénomènes (le sol corrosif, contraintes mécaniques, surpression, etc.) contribuent à la dégradation des canalisations et des accessoires du système. Une combinaison de ces phénomènes accélérera sans aucun doute la dégradation des canalisations. Dans la plupart des réseaux de distribution, une grande partie de l'eau est perdue durant le transport entre les installations de traitement et les points de consommation, le volume perdu et en général 20-30% de la quantité produite, dans certains réseaux particulièrement les plus âgés ces pertes peuvent atteindre 50 % (Campbell et al., 2016). Les pertes d'eau peuvent avoir plusieurs causes : fuites, erreurs de mesure, utilisation publique, vol, etc. La cause principale est en général les fuites. La tâche de la détection des fuites consiste à déterminer l'existence d'une fuite et à estimer le taux de fuite, ainsi que l'emplacement de la fuite, à savoir la possibilité de déterminer l'emplacement de la fuite au niveau du réseau. L'approche traditionnelle pour identifier les fuites dans la plupart des cas est passive, à savoir, une fuite est réparée seulement quand il devient visible, ce qui se traduit généralement par une perte énorme d'eau. Il est donc important de mettre au point des méthodes permettant de détecter rapidement les fuites. C'est pourquoi le développement de méthodes efficaces pour détecter et localiser les fuites est devenu un enjeu de recherche essentiel. Les méthodes courantes de détection et de repérage des fuites sont toutes fondées sur le bruit (acoustiques) émis par les fuites. Le bruit de la fuite peut être entendu soit par contact direct avec la conduite et tout ce qui est connecté à celui-ci (entrée de service, vanne, borne d'incendie) ou en écoutant au sol. Le son de la fuite provient du choc des molécules d'eau entre elles, de leur frottement contre les parois de l'orifice de fuite ou finalement du choc de l'eau sur le terrain. En écoutant et en analysant ce bruit, on peut établir une zone de détection des fuites plus ou moins importante. Les inconvénients de cette méthode l'interférence des sources extérieures, comme le trafic; la propagation du bruit d'une canalisation à l'autre, ainsi que la fatigue de l'opérateur. Afin de mieux détecter et localiser les fuites sur les canalisations,

il existe des techniques non acoustiques. Par exemple, l'analyse des signaux détectés par les capteurs, comme les capteurs de pression, de débit et de température, intégrés au réseau de canalisations. Cette méthode consiste à interpréter les lectures tirées des capteurs. Dans le but de vérifier est d'évaluer l'état des conduites dans le réseau selon les paramètres de pression et de débit mesurés à différents points et moments du réseau. Il s'agit d'un problème d'ingénierie inverse auquel les techniques d'extraction de données et de reconnaissance des modèles sont applicables. Parmi les méthodes fondées sur les données, il y a les méthodes statistiques. L'étude des caractéristiques d'un ensemble d'objets ou d'observations constitue un moyen essentiel de connaissance. En se basant sur ces données, ils fournissent un outil graphique qui peut être facilement manipulé et interprété par l'opérateur. Malheureusement, cela exige une connaissance complète du jeu de données, ce qui rend ces méthodes coûteuses, sinon impossibles dans certains cas. Il s'agit donc d'utiliser des statistiques déductives qui consistent à utiliser la théorie des probabilités. Afin d'induire les caractéristiques inconnues d'une donnée provenant de certains échantillons dérivés de ces données et présentant une certaine marge d'erreur, en d'autres mots, la détermination d'une loi de probabilité fondée sur une série d'expériences. Pour ce qui est des outils de traitement, les réseaux de neurones artificiels et la méthode SVM (Les machines à vecteurs de support) sont des outils particulièrement adaptés pour aider les spécialistes de la maintenance dans la détection et la classification des dysfonctionnements du réseau (Abdul Gaffoor, 2017; Kemba et al., 2017; Mashford et al., 2012). Le présent chapitre présente une méthode d'analyse SVM pour interpréter les données obtenues à partir d'un ensemble de capteurs de pression qui surveillent un réseau de canalisations, afin d'obtenir des informations sur l'emplacement et l'ampleur des fuites du réseau.

IV.2 **Le diagnostic de defaut de fonctionnement**

Le diagnostic de défaut de fonctionnement correspond à la détection et à l'estimation du défaut (Marref, 2013). La détection du défaut détermine le moment où un dysfonctionnement se produit au sein du système, tandis que l'isolement du défaut consiste à identifier les causes ou les sources du défaut pour déterminer s'il s'agit d'un défaut de type capteur, actionneur ou procédé, et l'identification du défaut donnent une estimation de certaines caractéristiques du défaut telles que son amplitude. Le diagnostic des défaillances a pour rôle de garantir la sécurité et le fonctionnement optimal du système. En indiquant toute condition indésirable (ou état hors de contrôle) ce qui évite

tout dommage ou accident. Les tâches de détection et de localisation sont accomplies en parallèle ou une tâche déclenche l'autre (Laouti, 2012).

IV.3 **Terminologie en matiere de diagnostic**

- Dysfonctionnement : irrégularité intermittente dans l'exécution du fonctionnement désiré du système.
- Détection des défauts : détermination des défauts présents dans un système et heure de détection.
- Isolation des défauts : détermination du type, de l'emplacement et du temps de détection d'un défaut. Cette étape fait suite à celle de détection des défauts.
- Identification du défaut : détermination de la taille et de la variation dans le temps due à la conduite d'un défaut. Cette étape suit l'étape de l'isolation de défaut.

IV.4 **SVM (support vecteur machines)**

Les machines à vecteurs de support (aussi appelées machines à vecteurs supports où SVM en Anglais support vecteur machine) sont l'une des méthodes de classification supervisée binaire basées sur les données (Dangeti, 2017). Elle provient de la théorie de l'apprentissage statistique. SVM est utilisé pour résoudre des problèmes de discrimination, qui est de décider de quelle classe un échantillon appartient, ou régression, qui est de prédire la valeur numérique d'une variable. Le précepte est donc de retrouver une fonction de classification ou de discrimination dont la généralisation est optimale (Azencott, 2019). La doctrine d'un SVM est de transformer l'espace dans lequel nous travaillons en un autre espace nommé grandes dimensions dans lequel nous allons établir une séparation en différentes classes. La transition d'un espace à un autre est effectuée selon certaines transformations mathématiques à l'aide de fonctions non linéaires, également appelées noyaux. Pour deux classes d'exemples données, la SVM a pour objectif de trouver un classificateur qui séparera les données et maximisera la distance entre ces deux classes. Avec la SVM, il s'agit d'un classificateur linéaire connu sous le nom d'hyperplan (Zaiz, 2010). Il faut préciser que dans ces méthodes les données sont traitées comme des produits scalaires. L'hyperplan optimal est celui qui en plus de Séparer les données dans les classes respectives maximise la distance entre elles et les points au sein de l'ensemble de données le plus proche. Autrement dit, c'est celui qui présente la plus forte marge par rapport aux deux catégories de données (GONZALEZ, 2013).

IV.4.1 *Principes de fonctionnement de SVM*

L'objectif de la méthode SVM est de trouver une bonne reconnaissance et une bonne séparation entre les différentes classes. Le classificateur linéaire qui réalisé une séparation de données et maximiser la distance entre ces deux classes est appelé hyperplan (Barra et al., 2021).

IV.4.2 *Les machines à vecteurs de support pour la classification (SVM)*

les modèles SVM pour la classification, il existe deux cas de séparations le premier est linéairement séparable et le deuxième non linéairement séparable.

IV.4.2.A *Séparateur linéaire*

Pour le cas de séparation linéaire, il existe une infinité d'hyperplans séparateurs permettant de séparer l'ensemble en deux classes sans erreur, mais il reste à trouver le meilleur séparateur. Il convient de rappeler que notre but est de trouver un hyperplan de séparation optimale parmi les nombreux autres qui séparent le mieux les deux classes d'exemples, en d'autres termes, cela maximise la distance euclidienne entre ces derniers. Cette distance est appelée "Marge". C'est pour cela qu'on parle de "séparateurs à vaste marge" puisqu'on essaie de maximiser cette marge(Laouti, 2012). Avoir une marge plus large donne plus de sécurité quand vous voulez classifier un nouvel exemple. D est considéré comme séparable linéairement s'il y a au moins un hyperplan dans $\mathcal{R}^n$ tel que, Tous les points positifs (étiquetés +1) sont d'un côté de cet hyperplan et tous les points négatifs (étiquetés -1) de l'autre.

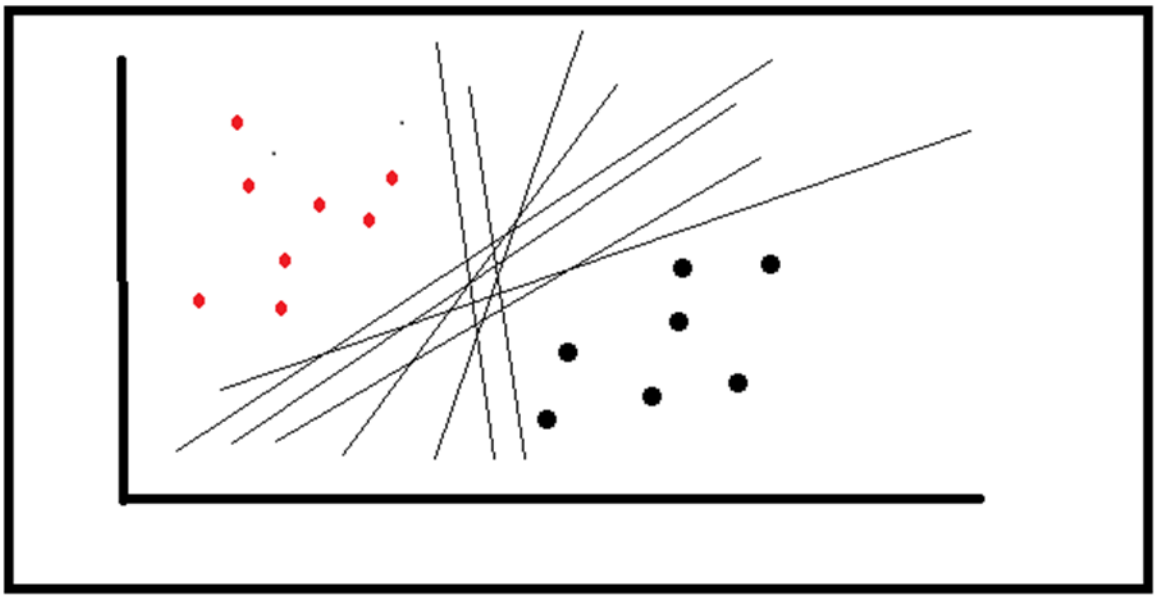


Figure IV-1 Cas Linéairement séparable

Notre espace d'entrée X correspond par conséquent à $\mathcal{R}^n$ ou n est le nombre de composants des vecteurs contenant les données.

$$\mathcal{F}(x) = \langle w.x \rangle + b = \sum_{j=1}^{n} w_j x_j + b$$

$w \in \mathcal{R}^n$ et $b \in R$ sont des paramètres, $x \in R$ est une variable.

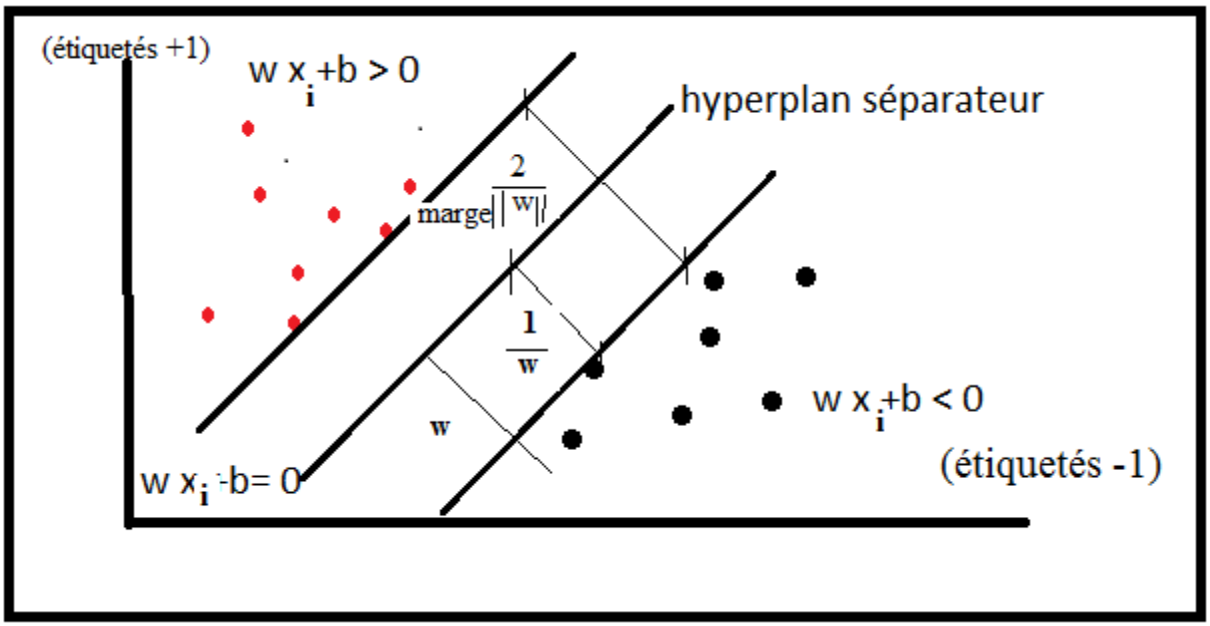


Figure IV-2 Séparation linéaire

IV.4.2.A.1 Maximisation de la marge

Dans le cas linéaire séparable, nous allons considérer les points les plus proches des hyperplans séparateurs appelés vecteurs porteurs. Le problème revient alors à trouver ***w*** et ***b*** tels que

$d = \frac{2}{\|w\|}$ est maximale $\forall(x_i, y_i)$.

La règle de classification d'une nouvelle observation x en fonction de l'hyperplan de marge maximale est donnée par :

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)(x_i, x) + b$$

Les coefficients $\alpha_i > 0$ sont appelés multiplicateurs de Lagrange ou encore variables duales

IV.4.2.B *Séparateur non-linéaire*

En général, les données ne sont pas séparables linéairement. Dans ce cas, quel que soit l'hyperplan de séparation, choisi quelques-uns des points seront mal classés, d'autres seront correctement classés, mais dans la zone d'indécision. Afin de surmonter ces inconvénients, l'idée des SVM est de

changer l'espace de données. La transformation de données non-linéaires peut permettre la séparation linéaire d'exemples dans un nouvel espace. Par conséquent, nous allons changer de dimension. On appelle cette nouvelle dimension l' « espace de redescription ». Étant donné que la dimension de l'espace de redescription est grande, il est possible de retrouver un hyperplan séparant les exemples. Alors, on a transformé un problème qui n'est pas séparable de façon linéaire dans l'espace de représentation bidimensionnelle, à un cas linéairement séparable dans un espace de plus grandes dimensions, c'est-à-dire l'espace de redescription. Cette transformation s'effectue par la fonction noyau (Kernel function)(Barra et al., 2021). Parmi les fonctions du noyau, on peut citer Linéaire, Polynomiale, sigmoïdaux et à fonction de base radiale (Radial Basis Function, RBF). La fonction de décision est :

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) k(x_i, x) + b$$

La fonction k est appelée fonction noyau (kernel function).

Exemple de kernels (noyaux) :

- Linéaire : $k(x, x') = \langle x, x' \rangle$;
- Polynomiale : $k(x, x') = (\gamma \langle x, x' \rangle + c)^d$
- Sigmoïdaux : $k(x, x') = tnah(\gamma \langle x, x' \rangle + c)$
- RBF : $k(x, x') = exp\left(\frac{-\|x - x'\|^2}{2\sigma^2}\right)$;

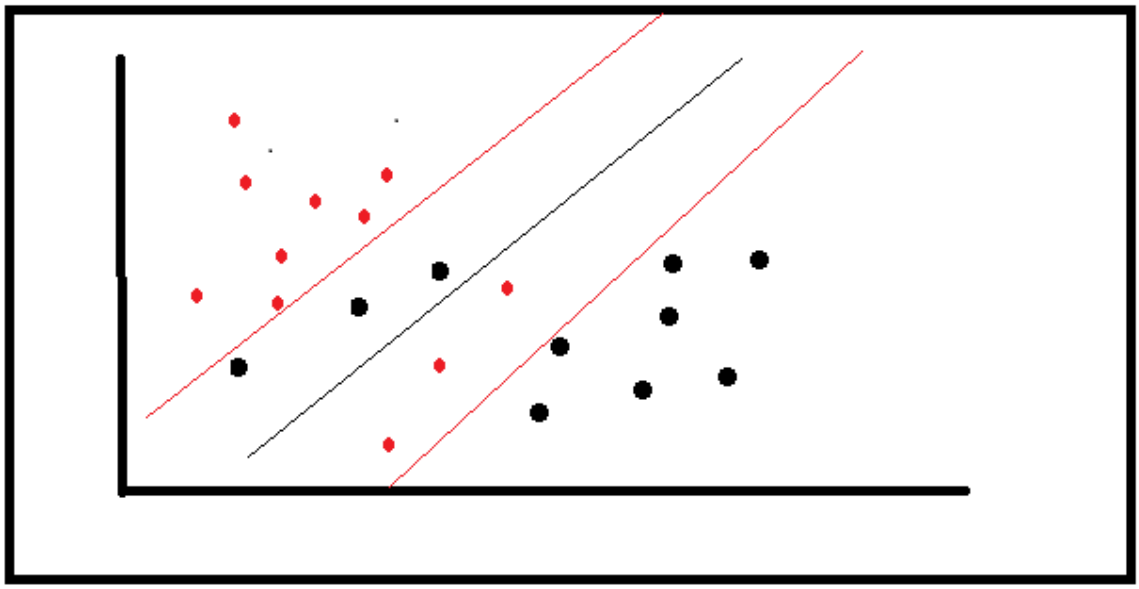


Figure IV-3 Cas Non Linéairement séparable.

IV.4.3 *SVM de regression*

L'algorithme de classification ou de régression emploie les mêmes principes et ne présente que des différences, mineures. Plutôt que de valoir une valeur de -1 ou 1, les étiquettes peuvent désormais prendre n'importe quelle valeur réelle(Marref, 2013). Une séparation linéaire est considérée comme correcte à ε près si :

$$\forall i | \langle w, x_i \rangle + b - y_i | \leq \varepsilon$$

IV.5 Detection de fuites dans reseau de distribution utilisant epanet et les machines a vecteurs de support

Une méthode de détection des fuites est appliquée au réseau C-town en utilisant les machines à vecteurs de support (SVM) est présentée. Les données de pression et de débit sont utilisées pour former et tester proviennent d'Epanet; un logiciel qui simule le comportement hydraulique et la qualité de l'eau dans les conduites sous pression. Le réseau de distribution d'eau de C-townt a été modélisé par EPANET. Le réseau de simulation est constitué de 429 tuyaux et de 388 nœuds. Dix nœuds ont été choisis comme nœuds de fuite. Les données relatives à la pression et au débit sont divisées en 1000 ensemble d'apprentissage et 500 ensembles de tests.

IV.5.1 *Établissement d'une base de données*

Il est possible d'utiliser SVM pour la régression ou la classification. Lorsqu'il s'agit de classifier, le résultat est une classe prédite associée à un modèle d'entrée, et lorsque vos faites une régression, le résultat est un nombre réel associé à un modèle d'entrée. Un SVM agissant en tant que régression se comporte en tant que fonction d'approche. Les SVM sont formés dans un ensemble d'apprentissage composé d'un certain nombre de modèles d'entrée et de valeurs ou de catégories de sortie associées. Ils peuvent ensuite être soumis à une série d'essais visant à déterminer la métrique de performance; autrement dit, la précision de la classification en tant que classificateurs ou l'erreur quadratique moyenne (MSE) et le coefficient de corrélation (R^2) en tant que régression(Mashford et al., 2012). L'approche proposée pour prédire l'ampleur et l'emplacement des fuites consiste à surveiller la pression et le débit à un certain nombre de nœuds de réseau de conduites, et transférer ces valeurs aux SVM formés afin de prévoir l'ampleur et l'emplacement de la fuite. Les SVM peuvent être formées sur la base d'un certain nombre de cas prévoyant des fuites de différentes ampleurs et endroits dans le réseau(Laouti, 2012). Puisque SVM exige des centaines peut-être des milliers de cas dans leurs jeux de données d'apprentissage ; il est impossible de produire des ensembles de données d'apprentissage en saisissant les fuites réelles dans le réseau de conduite. Il est possible de résoudre ce problème à l'aide d'un outil de simulation du comportement hydraulique des réseaux comme EPANET(Rossman, 2000). EPANET est un logiciel développé par l'agence pour la protection de l'environnement des États-Unis pour la simulation du comportement des réseaux de distribution d'eau d'un point de vue hydraulique ainsi que du point de vue de la qualité de l'eau. Il s'agit d'un logiciel gratuit accessible sur Internet. On peut simuler des fuites de différentes ampleurs dans EPANET et calculer les pressions et les débits résultants dans le réseau. Afin de produire le grand nombre de cas nécessaires pour l'ensemble d'apprentissage SVM (Kemba et al., 2017; Mashford et al., 2012; Mashford et al., 2009).

IV.5.1.A *Modelisation des fuites avec Epanet*

La modélisation du fonctionnement du réseau a pour objet de décrire le comportement hydraulique des différents dispositifs du réseau(Mashford et al., 2012). L'intérêt est de reproduire ce qui se passe réellement au sein du réseau en utilisant un modèle hydraulique. La représentation et la précision du modèle dépendent des objectifs du service de l'eau et des analyses attendues, de sorte que le degré de détail détermine les résultats de la modélisation. EPANET permet la réalisation d'un réseau d'eau potable virtuel et l'étude de l'influence de différents paramètres sur les pressions et

débits en différents points du réseau. Le logiciel américain, traduit en français, est distribué gratuitement par l'Environmental Protection Agency depuis le mois de septembre 1993.

IV.5.1.A.1 Emetteur

Le débit de fuite varie, selon la pression disponible. Ceci permet au service de l'eau de détecter les fuites au moyen de mesures de pression et de recourir à des vannes de contrôle de la pression ou des stabilisateurs pour réduire les débits des fuites et prolonger la durée de vie des canalisations. La formule de Torricelli est la base des rapports de pression - Débit de fuite sera probablement employés pour caractériser le débit à travers les orifices :

$$Q_f = C_d A\sqrt{2gp}$$

Q_f: Est le débit de fuite à travers l'orifice.

$C_{d:}$ Un coefficient de débit.

A: Est la surface de l'orifice.

g: L'accélération de la pesanteur.

p: La hauteur de pression à l'orifice.

Dans la pratique, le débit est décrit sous une forme plus générale de fonction puissance (c.-à-d. comme les émetteurs dans Epanet)(Mashford et al., 2012). Les émetteurs sont des dispositifs liés aux nœuds de demande. Ils sont utilisés pour modéliser l'écoulement à travers les systèmes d'irrigation, pour simuler une fuite dans un tuyau relié à un nœud (si on peut estimer un coefficient de débit et un exposant de pression pour la fuite) ou pour calculer le débit d'incendie au nœud (l'écoulement disponible à une certaine pression résiduelle minimale (Rossman, 2000). Le débit de l'émetteur s'exprime en fonction de la pression au nœud selon la formule :

$$Q = CP^{\gamma}$$

Dans laquelle Q est le débit, P la pression et γ *l*'exposant de pression. Pour la simulation des fuites, il est généralement pris égal à 0,5. C le coefficient de débit ou (coefficient de l'émetteur).

C'est la caractéristique principale qu'on va utiliser afin de simuler d'éventuelles fuites dans le réseau de la présente étude.

IV.6 **Choix du reseau d'etudes**

Notre choix est de continuer avec le réseau C-town. Afin de permettre à la comète scientifique de développer cette étude et de faire en sorte que les résultats soient plus visibles. Nous espérons avoir des analyses bien fondées pour le développement de notre recherche et pour la réalisation d'études futures. Les études déjà menées dans ce domaine sont expérimentées sur une petite partie du réseau global. Notre étude prendra en considération le réseau global. Premièrement, on a divisé le réseau en cinq secteurs, puis on a commencé à doter les secteurs d'instruments de mesure. L'emplacement du débitmètre et du manomètre sont installés à la sortie de chaque réservoir et aux limites des secteurs, afin de calculer le bilan hydrique et détecter la variation de pression dans chaque secteur. Deux fuites par secteur sont considérées, ils ont été sélectionnés de façon arbitraire. Les mesures de pression et de débit sont utilisées pour détecter les anomalies visant à prédire l'emplacement et le débit des fuites dans le réseau.

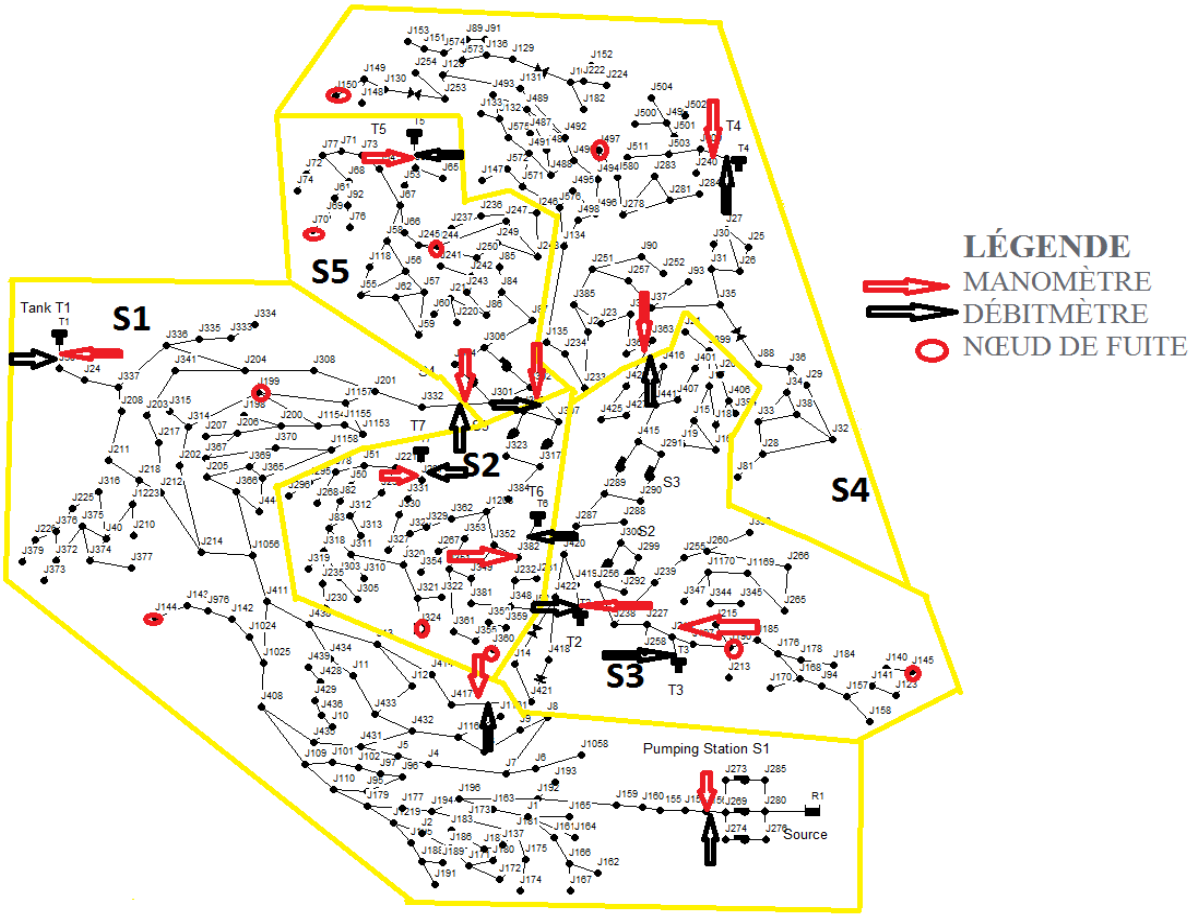


Figure IV-4 Emplacement des appareils de mesure réseau C-town

IV.7 **Simulation des fuites (coefficients d'emission) pour une regression lineaire**

Faire face aux contraintes de la méthode d'établissements de modèles de machines à vecteurs de support (SVM), chaque base de données utilisées doit être traitée au préalable en divisant tous les échantillons en deux sous-ensembles (apprentissage, tests).Le sous-ensemble apprentissage est utilisé pour créer un modèle. Ceci implique l'établissement des paramètres du modèle et la détermination des poids d'un réseau. Quand le modèle est formé avec des données d'apprentissage, nous avons besoin de savoir la réponse du modèle avec des données inconnues, le modèle devrait être utilisé avec un sous-ensemble (tests) différent non utilisé dans le processus d'apprentissage. Le premier test consistait à déterminer dans quelle mesure une régression SVM permet de prédire efficacement les coefficients d'émission en cas de fuite d'un nœud donné (j324). On a modélisé des fuites de zéro à un taux élevé d'environ (2 378 l/s). Le programme EPANET a servi à produire un ensemble de données de 300 cas avec l'exposant de pression égale à 0,5 et un coefficient d'émetteur variait entre 0.000 et 0.300 par incrément de 0.001. Parmi ces cas, 200 et 100 ont été sélectionnés aléatoirement pour former les données d'apprentissage et un ensemble de tests, respectivement. La SVM a été formée au moyen du noyau de la fonction de base radiale. Lorsque nous modélisons un comportement réel, nous nous demandons, après l'établissement du modèle, s'il est fiable et pertinent. Afin d'évaluer la performance d'un modèle, il est possible de recourir à des indicateurs de performance.

IV.7.1 ***Critere RMSE***

Racine de l'erreur quadratique moyenne ((RMSE) Root Mean Squared Error) est calculée à l'aide de la racine carrée des résidus. Il décrit l'ajustement absolu du modèle aux données indiquant à quel point les points de données observés sont actuellement proches des valeurs prédites du modèle. La RMSE peut également être interprétée comme l'écart-type de la variance inexpliquée et possède la propriété d'être dans les mêmes unités que la variable de réponse. Les valeurs les plus faibles de RMSE représentent un meilleur ajustement. La RMSE est une bonne mesure de la précision avec laquelle le modèle prédit la réponse(Khan & Noor, 2019; Kothari & Balamurugan, 2019).

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

Avec :

- $\hat{y}_i$et $y_i$sont respectivement la prévision et la valeur réelle.
- n est le nombre de points dans la série.

RMSE=0.005795586

IV.7.2 ***Critere MAE***

Erreur absolue moyenne (MAE) Mesure l'écart absolu entre la valeur réelle et la prédiction. Autrement dit, l'amplitude moyenne des erreurs sans prendre en considération leur direction(Laouti, 2012; Najwa Mohd Rizal et al., 2022). Tous les écarts individuels sont pondérés uniformément dans la moyenne par cette méthode. Quand les valeurs de l'indice MAE sont nulles, le modèle considéré sera le plus performant possible.

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

MAE= 0.004891305

IV.7.3 ***Critere R²***

Le R², où R-carré est appelé coefficient de détermination. C'est un indicateur utilisé en statistiques pour juger la qualité d'une régression linéaire, Qui sert à détecter de quelle façon des valeurs distinctes d'une variable peuvent être utilisées pour expliquer la différence d'une seconde variable. Le carré R a une caractéristique très importante que son échelle est intuitive, ce qui signifie qu'il passe de zéro à un. Si R² vaut 1, alors la régression détermine 100 % de la répartition des points(Mashford et al., 2012; Najwa Mohd Rizal et al., 2022). Un R carré est considéré comme élevé lorsque les valeurs se situent entre 0,85 et 1. Zéro illustrant le fait que le modèle proposé n'améliore pas la prédiction par rapport au modèle moyen et un signifié qu'il a une prédiction parfaite.

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

R²= 0.9958424

IV.7.4 ***Critere MSE***

L'erreur quadratique moyenne (MSE, Mean Squared Error) détermine la proximité d'une série de points par rapport à une ligne de régression. Cela se fait en prenant les distances entre l'ensemble de points et la droite de régression et en les mettant au carré. MSE sont toujours positives donc comprises dans l'intervalle [0; +∞ [. Un modèle considéré comme idéal lorsque MSE est proche de 0(Laouti, 2012; Najwa Mohd Rizal et al., 2022).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

MSE=3.358881e-05

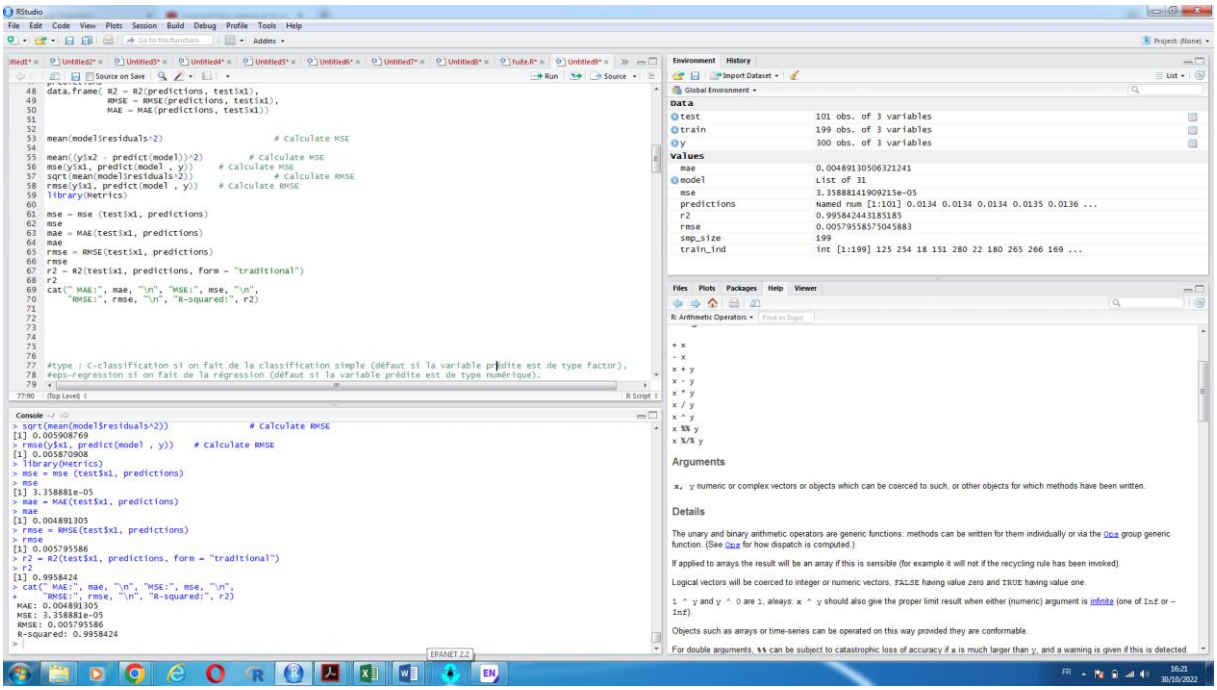


Figure IV-5 métriques de régression linéaire par Rstudio.

IV.8 **Resultats & discussion**

Afin d'évaluer la capacité de prédiction du coefficient d'émission à travers le modèle présenté dans cette recherche, on a appliqué des critères capables de présenter une prédiction quantitative du modèle. Même s'il existe différents indices statistiques, l'utilisation d'un seul indice statistique ne peut pas être considérée comme un critère suffisant pour étudier la précision de prédiction d'un

modèle(Mashford et al., 2012; Najwa Mohd Rizal et al., 2022). La précision de prédiction du modèle présenté ici a été étudiée à travers le coefficient de détermination (R^2), l'erreur quadratique moyenne (MSE), l'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE). Les résultats des essais sont considérés comme excellents. Nous avons obtenu un R^2= 0.9958424 qui est très proche de 1, une MSE qui vaut 3.358881e-05, RMSE=0.005795586 et MAE= 0.004891305 qui sont très faibles. C'est une prédiction SVM très satisfaisante.

IV.9 **Classification svm pour de detecter les fuites**

Les méthodes de classification jouent un rôle de premier plan dans l'analyse des données et sont utilisées dans de nombreuses applications scientifiques. La classification et la prédiction par machines à vecteurs de support (SVM) constituent une technique de classification supervisée largement utilisée et l'une des plus efficaces, en particulier pour les données à grande dimension (Kemba et al., 2017; Mashford et al., 2012). Dans cette section, nous allons faire appel à un classificateur SVM pour localiser les fuites. Dix nœuds ont été choisis dans le réseau C-town (deux nœuds par secteur) considéré comme des nœuds de fuite possible. Le programme de simulation EPANET a été utilisé pour générer des données représentant les pressions et les débits aux 24 points de surveillance. La simulation des fuites avec l'exposant de pression égale à 0,5 et un coefficient d'émetteur variait entre 0.000 et 0.300 par incrément de 0.002, a créé 10 ensembles de données de 150 cas chacun. Les ensembles de données ont ensuite été fusionnées et 1000 cas ont été sélectionnés pour former les données d'apprentissage. Les 500 cas restants ont été utilisés comme un ensemble de tests. Les cas d'apprentissage et d'essais comprennent les emplacements des fuites.

IV.9.1 ***Variables d'entrees-sorties du SVM :***

Les variables d'entrée Svm sont les valeurs de pression et de débit pendant les heures de pointe. Ces valeurs sont relevées à partir des débitmètres et des manomètres installés à chaque entrée dans un secteur et à la limite entre les secteurs. Et la variable de sortie correspond au nœud de fuite.

q/nJ 324	p/nJ3 24	Q/S1/ P310	Q/T1/ P15	Q/T2/ P215	Q/T3/ P787	Q/T4/ P347	Q/T5/ P1044	Q/T6/ P144	Q/T7/ P752	Q/J332 /P397	Q/J302 /P399	Q/J417 /P445	Q/J426 /P308	P/S1/ J269	P/T1/ J39	P/T2/ J419	P/T3/ J216	P/T4/ J509	P/T5/ J64	P/T6/ J382	P/T7/ J297	P/J33 2	P/J30 2	P/J41 7	P/J42 6	N F
0.87 3697	63.69 9010	193.1 99800	- 38.86 3680	21.60 5550	21.12 8030	7.594 641	- 17.333 190	3.984 458	5.568 594	66.127 940	30.688 900	105.14 0300	49.019 040	34.80 4270	28.42 2430	25.28 3440	71.61 1500	35.09 3700	24.97 8680	39.07 7590	5.532 788	21.31 4280	20.88 5030	29.78 8760	68.37 4380	nJ3 24

Les variables d'entrée

La variable de sortie

IV.9.2 *Analyse préliminaire des données*

Le traitement préalable consiste à corriger les problèmes de données avant de créer un modèle d'apprentissage automatique utilisant ces données. Les problèmes peuvent être de plusieurs types, notamment les valeurs manquantes, les attributs ayant une plage différente, etc figure IV-6.

```
   Q.S1.P310        Q.T1.P15        Q.T2.P215        Q.T3.P787        Q.T4.P347        Q.T5.P1044       Q.T6.P144
Min.   :193.2   Min.   :38.84   Min.   :21.02   Min.   :19.23   Min.   :5.961   Min.   :15.57   Min.   :2.890
1st Qu.:193.2   1st Qu.:38.87   1st Qu.:21.46   1st Qu.:21.13   1st Qu.:7.587   1st Qu.:17.32   1st Qu.:3.980
Median :193.2   Median :38.91   Median :21.59   Median :21.13   Median :7.593   Median :17.33   Median :3.984
Mean   :193.2   Mean   :39.09   Mean   :21.51   Mean   :20.94   Mean   :7.439   Mean   :17.16   Mean   :3.917
3rd Qu.:193.2   3rd Qu.:39.01   3rd Qu.:21.60   3rd Qu.:21.13   3rd Qu.:7.595   3rd Qu.:17.33   3rd Qu.:3.984
Max.   :193.4   Max.   :40.76   Max.   :21.61   Max.   :21.13   Max.   :7.595   Max.   :17.33   Max.   :3.984

   Q.T7.P752       Q.J332.P397      Q.J302.P399      Q.J417.P445      Q.J426.P308       P.S1.J269        P.T1.J39
Min.   :3.473   Min.   :66.08   Min.   :30.67   Min.   :104.8   Min.   :49.01   Min.   :34.75   Min.   :28.40
1st Qu.:5.559   1st Qu.:66.13   1st Qu.:30.68   1st Qu.:105.1   1st Qu.:49.02   1st Qu.:34.80   1st Qu.:28.42
Median :5.567   Median :66.13   Median :30.69   Median :105.1   Median :49.02   Median :34.80   Median :28.42
Mean   :5.397   Mean   :66.15   Mean   :30.70   Mean   :105.1   Mean   :49.02   Mean   :34.80   Mean   :28.42
3rd Qu.:5.568   3rd Qu.:66.16   3rd Qu.:30.69   3rd Qu.:105.1   3rd Qu.:49.02   3rd Qu.:34.80   3rd Qu.:28.42
Max.   :5.569   Max.   :66.37   Max.   :30.95   Max.   :105.2   Max.   :49.14   Max.   :34.80   Max.   :28.42

   P.T2.J419        P.T3.J216        P.T4.J509        P.T5.J64         P.T6.J382        P.T7.J297         P.J332
Min.   :25.27   Min.   :70.21   Min.   :35.08   Min.   :24.95   Min.   :33.65   Min.   :5.483   Min.   :21.24
1st Qu.:25.28   1st Qu.:71.61   1st Qu.:35.09   1st Qu.:24.98   1st Qu.:39.07   1st Qu.:5.533   1st Qu.:21.29
Median :25.28   Median :71.61   Median :35.09   Median :24.98   Median :39.08   Median :5.533   Median :21.31
Mean   :25.28   Mean   :71.47   Mean   :35.09   Mean   :24.98   Mean   :38.47   Mean   :5.529   Mean   :21.30
3rd Qu.:25.28   3rd Qu.:71.61   3rd Qu.:35.09   3rd Qu.:24.98   3rd Qu.:39.08   3rd Qu.:5.533   3rd Qu.:21.31
Max.   :25.28   Max.   :71.61   Max.   :35.09   Max.   :24.98   Max.   :39.08   Max.   :5.533   Max.   :21.32

    P.J302          P.J417          P.J426            gg
Min.   :20.81   Min.   :29.77   Min.   :68.11   j144   :150
1st Qu.:20.86   1st Qu.:29.78   1st Qu.:68.37   j145   :150
Median :20.88   Median :29.79   Median :68.37   J150   :150
Mean   :20.87   Mean   :29.79   Mean   :68.36   j190   :150
3rd Qu.:20.88   3rd Qu.:29.79   3rd Qu.:68.37   j199   :150
Max.   :20.89   Max.   :29.79   Max.   :68.37   J244   :150
                                                (Other):600
```

Figure IV-6 Détails sur les caractéristiques des données.

IV.9.3 *Formulation des groupes des donnees d'apprentissage et de test*

Epanet a servi à la création de la base de données (lecture de pression et débit aux points de mesure). Ce jeu de données est constitué de plusieurs variables et d'une variable cible (le résultat correspond à la localisation de la fuite.). Il faut à présent diviser l'ensemble de données qui contient 1500 observations de 25 variables et dix classes en un ensemble d'apprentissage et un ensemble de tests. 1000 cas ont été sélectionnés pour constituer les données d'apprentissage, les 500 autres observations ont été utilisées en tant que jeu de tests. Nous devons procéder au sous-échantillonnage suivant : nous mélangeons toutes nos données d'apprentissage et choisissons le même nombre d'échantillons par classe (Kemba et al., 2017; Mashford et al., 2012; Mashford et al., 2009).

IV.9.3.A *Ajustement des parametres c et γ*

Il existe deux paramètres lors de l'utilisation des noyaux radiaux C et γ. On ne connaît pas au préalable les "C" et les "γ" qui conviennent le mieux à l'apprentissage ; il est donc nécessaire de sélectionner un modèle (recherche de paramètres)(Kemba et al., 2017; Najwa Mohd Rizal et al., 2022). Le but est d'identifier le bon (C, γ) de sorte que le classificateur puisse prédire avec exactitude des données inconnues (c-à-d. des données de test) figure IV-7.

IV.9.3.B *Paramètre de pénalité C*

Puisque l'hyperplan peut être de dimensionnalité arbitraire, il peut être parfaitement ajusté pour correspondre au jeu de données d'apprentissage. Cependant, cela entraînerait un sur-ajustement extrême. Le paramètre de pénalité C permet à la SVM de mal classer les échantillons individuels, tout en les pénalisant. De grandes valeurs "C" se traduisent par un rétrécissement de la marge de l'hyperplan, ce qui peut aider à accroître le nombre d'échantillons correctement classé. En revanche, une petite valeur "C" se traduira par une marge hyperplan plus élevée, ce qui se traduira par des pénalités plus faibles et peut-être plus d'échantillons mal classés (Najwa Mohd Rizal et al., 2022).

IV.9.3.C *Parametre du noyau gamma* $\boldsymbol{\gamma}$

Le paramètre gamma définit l'étendue de l'influence d'un seul exemple d'un seul entraînement d'apprentissage, à savoir son rayon d'influence. Les valeurs faibles de gamma signifiant « loin » et les valeurs élevées signifiant « proches ». Par conséquent, les valeurs élevées produisent généralement des limites de décision très basses, tandis que les valeurs gammas faibles aboutissent souvent à une limite de décision plus linéaire (Laouti, 2012; Najwa Mohd Rizal et al., 2022). Examinons maintenant les différents paramètres du noyau radial et voyons quelle est la meilleure combinaison en termes de précision de validation croisée. Dans le tableau VI-1, nous avons une trame de résultats de 99 lignes (une pour chaque combinaison d'hyperparamètres). Nous nous intéressons à la troisième colonne (error) qui nous donne l'erreur de validation croisée pour une valeur spécifique de "cost" et "gamma". Nous recherchons la valeur minimale.

IV.9.4 ***Mesures d'evaluation***

La plupart des problèmes liés à l'apprentissage machine (ML) appartiennent à deux groupes : la classification et la régression. Pour évaluer la précision d'un modèle, on utilise souvent un ensemble de données de test qui n'a pas été utilisé dans le processus d'apprentissage du modèle. Bon nombre d'indicateurs de performance unidimensionnels peuvent être dérivés d'une matrice confusionnelle. Les mesures les plus importantes utilisées pour évaluer la performance des modèles de classification sont Précisions de classification (Accuracy), la précision, F1-Score et la sensibilité (Recall).

IV.9.4.A *Précision de classification (Accuracy)*

La précision de classification est donnée par la relation(El-Zahab et al., 2022) :

$$\text{Accuracy} = \text{SOMME (diag)} / N$$

IV.9.4.B *Précision*

Il s'agit du fait que tous ceux qui ont été classés dans une classe, appartiennent réellement à la classe (El-Zahab et al., 2022).

$$\text{Precision} = \text{Diag} / \text{NPC}$$

IV.9.4.C *La sensibilité (Recall)*

Il s'agit de la proportion entre le nombre d'éléments de la classe correctement classée dans la classe et le nombre total d'éléments de la classe. C'est-à-dire combien nous avons prédit correctement. Le Recall doit être élevé (El-Zahab et al., 2022).

$$\text{Recall} = \text{diag} / \text{NEC}$$

IV.9.4.D *F1-Score:*

Représente la moyenne harmonique entre Précision et Recall (Najwa Mohd Rizal et al., 2022; Suzuki, 2022).

F1-Score = 2 * (precision * recall) / (precision + recall)

Diag : nombre d'éléments correctement classées par classe

NEC : nombre d'éléments par classe

NPC: nombre de prédictions par classe

N : nombre d'éléments

IV.10 **L'implémentation du modèle svm (apprentissage)**

Afin de former le modèle SVM, nous devons choisir le noyau à utiliser. Nous pouvons effectuer des essais et des erreurs pour déterminer lequel est le meilleur noyau. Pour des raisons de simplicité, nous allons utiliser le noyau radial. Examinons maintenant les différents paramètres du noyau radial et voyons quelle est la meilleure combinaison en termes de précision de validation croisée. Dans le tableau VI-1, nous avons une trame de résultats de 99 lignes (une pour chaque combinaison d'hyperparamètres). Nous nous intéressons à la troisième colonne (error) qui nous donne l'erreur de validation croisée pour une valeur spécifique de "cost" et "gamma". Nous recherchons la valeur minimale.

Table IV-1 Performance des SVMs en fonction de C et gamma

Parameter tuning of 'svm':			
- sampling method: 10-fold cross validation			
- best parameters:			
	gamma	cost	
	0.004115226	3125	
- best performance: 0.092			
- Detailed performance results:			
	gamma	cost	error
1	0.0004572474	0.04	0.944
2	0.0013717421	0.04	0.944
3	0.0041152263	0.04	0.731
4	0.0123456790	0.04	0.448
5	0.0370370370	0.04	0.277
6	0.1111111111	0.04	0.219
7	0.3333333333	0.04	0.169
8	1.0000000000	0.04	0.211
9	3.0000000000	0.04	0.848

10	0.0004572474	0.20	0.874
11	0.0013717421	0.20	0.512
12	0.0041152263	0.20	0.320
13	0.0123456790	0.20	0.257
14	0.0370370370	0.20	0.210
15	0.1111111111	0.20	0.169
16	0.3333333333	0.20	0.144
17	1.0000000000	0.20	0.128
18	3.0000000000	0.20	0.130
19	0.0004572474	1.00	0.394
20	0.0013717421	1.00	0.291
21	0.0041152263	1.00	0.240
22	0.0123456790	1.00	0.184
23	0.0370370370	1.00	0.156
24	0.1111111111	1.00	0.129
25	0.3333333333	1.00	0.119
26	1.0000000000	1.00	0.118
27	3.0000000000	1.00	0.124
28	0.0004572474	5.00	0.264
29	0.0013717421	5.00	0.215
30	0.0041152263	5.00	0.169
31	0.0123456790	5.00	0.147
32	0.0370370370	5.00	0.118
33	0.1111111111	5.00	0.113
34	0.3333333333	5.00	0.112
35	1.0000000000	5.00	0.120
36	3.0000000000	5.00	0.124
37	0.0004572474	25.00	0.193
38	0.0013717421	25.00	0.159
39	0.0041152263	25.00	0.137
40	0.0123456790	25.00	0.112
41	0.0370370370	25.00	0.106
42	0.1111111111	25.00	0.106
43	0.3333333333	25.00	0.107
44	1.0000000000	25.00	0.118
45	3.0000000000	25.00	0.130
46	0.0004572474	125.00	0.153
47	0.0013717421	125.00	0.123
48	0.0041152263	125.00	0.113
49	0.0123456790	125.00	0.099
50	0.0370370370	125.00	0.096

51	0.1111111111	125.00	0.108
52	0.3333333333	125.00	0.109
53	1.0000000000	125.00	0.122
54	3.0000000000	125.00	0.135
55	0.0004572474	625.00	0.118
56	0.0013717421	625.00	0.113
57	0.0041152263	625.00	0.097
58	0.0123456790	625.00	0.095
59	0.0370370370	625.00	0.098
60	0.1111111111	625.00	0.108
61	0.3333333333	625.00	0.114
62	1.0000000000	625.00	0.123
63	3.0000000000	625.00	0.139
64	0.0004572474	3125.00	0.109
65	0.0013717421	3125.00	0.095
66	0.0041152263	3125.00	0.092
67	0.0123456790	3125.00	0.098
68	0.0370370370	3125.00	0.106
69	0.1111111111	3125.00	0.106
70	0.3333333333	3125.00	0.116
71	1.0000000000	3125.00	0.126
72	3.0000000000	3125.00	0.142
73	0.0004572474	15625.00	0.093
74	0.0013717421	15625.00	0.092
75	0.0041152263	15625.00	0.097
76	0.0123456790	15625.00	0.097
77	0.0370370370	15625.00	0.105
78	0.1111111111	15625.00	0.110
79	0.3333333333	15625.00	0.119
80	1.0000000000	15625.00	0.131
81	3.0000000000	15625.00	0.141
82	0.0004572474	78125.00	0.093
83	0.0013717421	78125.00	0.092
84	0.0041152263	78125.00	0.099
85	0.0123456790	78125.00	0.099
86	0.0370370370	78125.00	0.106
87	0.1111111111	78125.00	0.117
88	0.3333333333	78125.00	0.122
89	1.0000000000	78125.00	0.133
90	3.0000000000	78125.00	0.128
91	0.0004572474	390625.00	0.093

92 0.0013717421 390625.00 0.096
93 0.0041152263 390625.00 0.100
94 0.0123456790 390625.00 0.106
95 0.0370370370 390625.00 0.116
96 0.1111111111 390625.00 0.111
97 0.3333333333 390625.00 0.120
98 1.0000000000 390625.00 0.131
99 3.0000000000 390625.00 0.117

On a alors déterminé les paramètres de formation gamma (γ) = 0,0041152263 et C = 3 125,00 afin d'optimiser la précision. Le modèle est ajusté en fonction des données d'apprentissage. Le taux d'erreur du modèle est égal à 9,2 % (ce qui est assez bien) tableau IV-1.

Table IV-2 Paramètre du modèle SVM des données d'apprentissage.

Parameters:
SVM-Type: C-classification
SVM-Kernel: radial
cost: 3125
Number of Support Vectors: 250
(2 16 7 3 6 6 8 8 97 97)
Number of Classes: 10
Levels:
j144 j145 J150 j190 j199 J244 J360 J497 J70 nJ324

IV.11 **Validation du modele SVM**

Notre modèle a été créé par les valeurs gamma (γ) = 0,0041152263. C = 3 125,00 et le noyau radial. Nous sommes prêts à prédire les classes pour notre ensemble de tests. Il doit effectuer une classification afin de déterminer une règle de décision qui peut, sur la base des données de test, attribuer une fuite à l'un des nombreux nœuds présumés de fuite. Lorsque la SVM a été mise à l'essai, elle a obtenu une précision de classification (Accuracy) de 89,6 %. Le tableau qui suit compare les valeurs des prévisions aux valeurs réelles. Cela nous permet également de comprendre le taux d'erreur.

Table IV-3 Synthèse des résultats obtenus.

test_pred_grid	j144	j145	J150	j190	j199	J244	J360	J497	J70	nJ324
j144	48	0	0	0	0	0	0	0	0	0
j145	0	10	0	7	0	0	0	0	0	0
J150	0	0	50	0	0	0	0	0	0	0
j190	0	40	0	43	0	0	0	0	0	0
j199	2	0	0	0	50	0	0	0	0	0
J244	0	0	0	0	0	47	0	0	0	0
J360	0	0	0	0	0	0	50	0	0	0
J497	0	0	0	0	0	0	0	50	0	0
J70	0	0	0	0	0	3	0	0	50	0
nJ324	0	0	0	0	0	0	0	0	0	50

Table IV-4 Performance du modèle de classification.

	precision	recall	f1
j144	0.96	1.0000000	0.9795918
j145	0.20	0.5882353	0.2985075
J150	1.00	1.0000000	1.0000000
j190	0.86	0.5180723	0.6466165
j199	1.00	0.9615385	0.9803922
J244	0.94	1.0000000	0.9690722
J360	1.00	1.0000000	1.0000000
J497	1.00	1.0000000	1.0000000
J70	1.00	0.9433962	0.9708738
nJ324	1.00	1.0000000	1.0000000

IV.12 **Resultats & discussion**

On note que les 50 cas de fuites dans les nœuds « J150, J360, J497, J70 et nj324 » sont bien prédits. Les prédictions de fuites dans les nœuds "j144, J244 et j190 " sont bonnes ou au moins acceptables, bien que deux et trois cas de fuite au niveau des noeuds "j144" et "j244" soient estimés au niveau des noeuds "j199" et "j70", respectivement, le noeud "j190" dix cas de fuite sont estimés dans noeud "j145". Pour le noeud "j145", 40 cas de fuite sont des prédictions incorrectes ont été estimées dans le noeud "j190", seulement dix sont correctement prédites. Le taux d'erreur global est 0.896, cela veut dire que les prédictions faites sont bonnes dans 89.6 % des cas. Non seulement ça, cela veut dire aussi que nous avons bien choisi les paramètres : fonction noyau, gamma et cost.

CONCLUSION GENERALE

Le travail effectué dans cette thèse intitulée "Apprentissage artificiel pour la détection des fuites et gestion des réseaux d'alimentation en eau potable" porte sur l'utilisation de l'intelligence artificielle pour la gestion et la détection des fuites dans les réseaux d'alimentation en eau potable. Primitivement, afin de mieux gérer le réseau et de réduire l'aire d'inspection des fuites, nous avons recours à la sectorisation. Le processus de sectorisation est généralement réalisé sans suivi et sans rigueur scientifique et technique ; au contraire, il repose généralement sur une approche par essai et erreurs. Il s'agit de déterminer correctement les secteurs des réseaux en tenant compte de la grande quantité d'informations qui leur sont associées, il ne serait pas possible de mener un tel processus sans l'aide d'outils informatiques. Les chercheurs ont commencé à développer des procédures pour informatiser le processus de sectorisation dans le but de créer automatiquement les secteurs. Cette étape démontre l'applicabilité de l'apprentissage automatique à la tâche proposée. Il y a plusieurs méthodes de partitionnement des réseaux, les plus importantes, à savoir les méthodes spectrales. Le partitionnement spectral est une approche mathématique combinant à la fois l'algèbre linéaire et le graphe qui se résume en l'extraction du spectre (valeurs et vecteurs propres) des matrices associées aux graphes. Le partitionnement (clustering) des nœuds de graphe est ensuite effectué par l'algorithme populaire de K-MEANS qui est constamment employé. Notre contribution consiste à Effectuer une analyse comparative entre les différents algorithmes existants (PAM, CLARA, HIERARCHICAL et DIANA). Nous avons sélectionné un certain nombre d'indicateurs de performance comme la modularité, l'indice interne et l'indice de stabilité afin d'évaluer et de permettre la comparaison. Les réseaux EXNET, C-TOWN et OUED EL MA ont été sélectionnés sur la base de leur type et de leur dimension. Les résultats obtenus sur les divers exemples montrent que K-MEANS n'est pas toujours efficace ni en matière de type de réseau, ni en matière de nombre de secteurs. L'algorithme PAM montre de bonnes performances du point de vue de la modularité, alors que pour l'index interne (K-MEANS et HIERARCHICAL) sont très efficaces ; pour l'index de stabilité (PAM, HIERARCHICAL et CLARA) sont plus efficaces. Pour le réseau EXNET PAM est satisfaisante du point de vue de la modularité et de la stabilité, pour l'index interne HIERARCHICAL est plus appropriée. Le réseau C-TOWN, la modularité est idéale pour PAM, Algorithmes HIERARCHICAL et modérément pour K-MEAN; pour les indices de stabilité PAM, HIERARCHICAL et CLARA sont mieux adaptées, alors que pour les indices internes K-MEANS et HIE-

RARCHICAL sont très efficaces. Le réseau OUED EL MA K-MEANS, PAM et HIERARCHICAL fournissent les meilleurs résultats pour tous les indices de qualité, notamment l'indice de modularité pour PAM et CLARA. La deuxième étape consiste à appliquer une méthode permettant d'obtenir de l'information sur l'emplacement des fuites dans un réseau de conduites, en traitant des valeurs de pression et de débit obtenues à un certain nombre de points du réseau à l'aide de SVM. Les données relatives à la formation SVM ont été obtenues au moyen du système de modélisation hydraulique EPANET, qui permet de simuler la pression ou le débit d'une fuite donnée.Le taux d'erreur global est de 0,896, ce qui signifie que les prédictions faites sont exactes dans 89,6 % des cas.

Refrences bibliographiques

Abdul Gaffoor, T. (2017). *Real-time control and optimization of water supply and distribution infrastructure* University of Waterloo].

Azencott, C.-A. (2019). *Introduction au machine learning*. Dunod.

Barra, V., Cornuéjols, A., & Miclet, L. (2021). *Apprentissage artificiel-4e édition*. Eyrolles.

Brentan, B. M.-G., EnriqueGoulart, ThaisaManzi, DanielMeirelles, GustavoHerrera Fernández, Antonio ManuelIzquierdo Sebastián, JoaquínLuvizotto, Edevar. (2018). Social network community detection and hybrid optimization for dividing water supply into district metered areas. *Journal of Water Resources Planning and Management, 144*(5), 04018020-04018021-04018020-04018010.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R package for cluster validation. *Journal of Statistical Software, 25*, 1-22.

Burkov, A. (2020). *Machine learning engineering* (Vol. 1). True Positive Incorporated.

Campbell, E. (2013). *Propuesta para una metodología de sectorización de redes de abastecimiento de agua potable*

Campbell, E., Izquierdo, J., Montalvo, I., & Pérez-García, R. (2016). A novel water supply network sectorization methodology based on a complete economic analysis, including uncertainties. *Water, 8*(5), 179.

Campbell, E., JoaquínMontalvo, IdelPérez-García, Rafael. (2016). A novel water supply network sectorization methodology based on a complete economic analysis, including uncertainties. *Water, 8*(5), 179.

Candelieri, A., Conti, D., & Archetti, F. (2014). Improving analytics in urban water management: a spectral clustering-based approach for leakage localization. *Procedia-Social and Behavioral Sciences, 108*, 235-248.

Clauset, A., Mark EJMoore, Cristopher. (2004). Finding community structure in very large networks. *Physical review E, 70*(6), 066111.

Combe, D. (2013). *Détection de communautés dans les réseaux d'information utilisant liens et attributs*. David Combe.

Côme, E. (2009). *Apprentissage de modèles génératifs pour le diagnostic de systèmes complexes avec labellisation douce et contraintes spatiales* Compiègne].

Cornuéjols, A., Laurent. (2011). *Apprentissage artificiel: concepts et algorithmes*. Editions Eyrolles.

Dangeti, P. (2017). *Statistics for machine learning*. Packt Publishing Ltd.

Di Nardo, A., Giudicianni, C., Greco, R., Herrera, M., & Santonastaso, G. F. (2018). Applications of graph spectral techniques to water distribution network management. *Water, 10*(1), 45.

Di Nardo, A. N., Michele. (2010). A design support methodology for district metering of water supply networks. In *Water Distribution Systems Analysis 2010* (pp. 870-887).

Di Nardo, A. N., MicheleGargano, RGiudicianni, CarloGreco, RobertoSantonastaso, Giovanni Francesco. (2018). Performance of partitioned water distribution networks under spatial-temporal variability of water demand. *Environmental Modelling & Software, 101*, 128-136.

Di Nardo, A. N., MicheleGiudicianni, CarloGreco, RobertoSantonastaso, Giovanni Francesco. (2017). Weighted spectral clustering for water distribution network partitioning. *Applied network science, 2*(1), 1-16.

Di Nardo, A. N., MicheleSantonastaso, Giovanni Francesco. (2014). A comparison between different techniques for water network sectorization. *Water Science and Technology: Water Supply, 14*(6), 961-970.

Di Nardo, A. N., MicheleSantonastaso, Giovanni FrancescoVenticinque, Salvatore. (2011). Graph partitioning for automatic sectorization of a water distribution system. *Proceedings of computer and control in water industry CCWI*, 841-846.

Di Nardo, A. N., MicheleSantonastaso, Giovanni FTzatchkov, Velitchko GAlcocer-Yamanaka, Victor H. (2014). Water network sectorization based on graph theory and energy performance indices. *Journal of Water Resources Planning and Management, 140*(5), 620-629.

Diao, K., YuwenRauch, Wolfgang. (2013). Automated creation of district metered area boundaries in water distribution systems. *Journal of Water Resources Planning and Management, 139*(2), 184-190.

Díaz, J. L., Herrera, M., Izquierdo, J., Montalvo, I., & Pérez, R. (2008). A Particle Swarm Optimization derivative applied to cluster analysis.

Doiron, F. (2016). *Analyse comparative de méthodes de détection de communautés dans les graphes avec algorithmes exacts* HEC Montréal].

Dreyfus, G., JMSamuelides, MGordon, MBBadran, FThiria, SHérault, L. (2002). *Réseaux de neurones* (Vol. 39). Eyrolles Paris.

Ducruet, C. (2011). Simplification et partitionnement d'un graphe.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics, 4*(1), 95-104.

El-Zahab, S., Al-Sakkaf, A., Mohammed Abdelkader, E., & Zayed, T. (2022). A machine learning-based model for real-time leak pinpointing in buildings using accelerometers. *Journal of Vibration and Control*, 10775463211066247.

Eusebio, P., Levy, D., & Floch, J. M. (2015). Partitionnement et analyse de graphes. *Les douzièmes Journées de Méthodologie Statistique de l'Insee*.

Falcini, F., Cimaz, R., Ricci, L., Fanner, S., De Martino, M., & Ceruso, M. (2007). A boy with bizarre hands mimicking an inflammatory chronic disease. *Clinical and experimental rheumatology, 25*(5), 790.

Falcini, F., RRicci, LFanner, SDe Martino, MCeruso, M. (2007). A boy with bizarre hands mimicking an inflammatory chronic disease. *Clinical and experimental rheumatology, 25*(5), 790.

Giustolisi, O., & Ridolfi, L. (2014). New modularity-based approach to segmentation of water distribution networks. *Journal of Hydraulic Engineering, 140*(10), 04014049.

GONZALEZ, E. C. (2013). *Propuesta Para una Metodología de Sectorización de Redes de Abasteci-miento de Agua Potable* UNIVERSIDAD POLITECNICA DE VALENCIA].

Gosalia, K. (2019). *Introduction à*

l'apprentissage automatique.

Grayman, W. M., Deininger, R. A., & Males, R. M. (2001). *Design of early warning and predictive source-water monitoring systems*. American Water Works Association.

Grayman, W. M., Rolf AMales, Richard M. (2001). *Design of early warning and predictive source-water monitoring systems*. American Water Works Association.

[Record #57 is using a reference type undefined in this output style.]

Gupta, G. (2017). Monitoring Water Distribution Network using Machine Learning, EP242X. *Degree Project in Communication Networks, 66*.

Gutiérrez-Pérez, J. A., Herrera, M., Pérez-García, R., & Ramos-Martínez, E. (2013). Application of graph-spectral methods in the vulnerability assessment of water supply networks. *Mathematical and Computer Modelling, 57*(7-8), 1853-1859.

Haghiri, S., Ghoshdastidar, D., & von Luxburg, U. (2017). Comparison-based nearest neighbor search. Artificial Intelligence and Statistics,

Hajebi, S., SBarrett, SClarke, AClarke, S. (2014). Water distribution network sectorisation using structural graph partitioning and multi-objective optimization. *Procedia Engineering, 89*, 1144-1151.

Han, R., & Liu, J. (2017). Spectral clustering and genetic algorithm for design of district metered areas in water distribution systems. *Procedia Engineering, 186*, 152-159.

Handl, J., & Knowles, J. (2005). Exploiting the trade-off—the benefits of multiple objectives in data clustering. International conference on evolutionary multi-criterion optimization,

Herrera Fernández, A. M. (2011). *Improving water network management by efficient division into supply clusters* Universitat Politècnica de València].

Kanawati, R. (2013). Détection de communautés dans les grands graphes d'interactions (multiplexes): état de l'art.

Kemba, J., Gideon, K., & Nyirenda, C. N. (2017). Leakage detection in Tsumeb east water distribution network using EPANET and support vector regression. 2017 IST-Africa Week Conference (IST-Africa),

Khan, M., & Noor, S. (2019). Performance Analysis of Regression-Machine Learning Algorithms for Predication of Runoff Time. *Agrotechnology, 8*, 1-12.

Khoa Bui, X. M., MalvinKang, Doosun. (2020). Water network partitioning into district metered areas: a state-of-the-art review. *Water, 12*(4), 1002.

Kothari, A., & Balamurugan, M. (2019). An efficient scheme for water leakage detection using support vector machines (SVM)-Zig. *International Journal of Recent Technology and Engineering (IJRTE), 7*(64), 39-46.

Laouti, N. (2012). *Diagnostic de défauts par les Machines à Vecteurs Supports: application à différents systèmes mutivariables nonlinéaires* Université Claude Bernard-Lyon I].

Liu, H., MengkeZhang, ChiFu, Guangtao. (2018). Comparing topological partitioning methods for district metered areas in the water distribution network. *Water, 10*(4), 368.

Liu, H., Zhao, M., Zhang, C., & Fu, G. (2018). Comparing topological partitioning methods for district metered areas in the water distribution network. *Water, 10*(4), 368.

Maâmatou, H. (2017). *Apprentissage semi-supervisé pour la détection multi-objets dans des séquences vidéos: Application à l'analyse de flux urbains* Université de Sfax (Tunisie)].

Maquin, D. (2003). Eléments de théorie des graphes et programmation linéaire. *Cours, Institut National Polytechnique de Lorraine*.

Marref, N. (2013). *Apprentissage Incrémental & Machines à Vecteurs Supports* Université de Batna 2].

Mashford, J., De Silva, D., Burn, S., & Marney, D. (2012). Leak detection in simulated water pipe networks using SVM. *Applied Artificial Intelligence, 26*(5), 429-444.

Mashford, J., De Silva, D., Marney, D., & Burn, S. (2009). An approach to leak detection in pipe networks using analysis of monitored pressure values by support vector machine. 2009 Third International Conference on Network and System Security,

Matias, C. (2015). Notes de cours: Analyse statistique de graphes. *Université Pierre et Marie*.

Messaoudi, A. (2020). *Détection de communautés dans des réseaux complexes* Université du Québec en Outaouais].

Najma, H. (2014). Nouvelles techniques de recommandation et de détection des communautés.

Najwa Mohd Rizal, N., Hayder, G., Mnzool, M., Elnaim, B. M., Mohammed, A. O. Y., & Khayyat, M. M. (2022). Comparison between Regression Models, Support Vector Machine (SVM), and Artificial Neural Network (ANN) in River Water Quality Prediction. *Processes, 10*(8), 1652.

Nardo, A. D., Natale, M. D., Giudicianni, C., Greco, R., & Santonastaso, G. F. (2016). Water supply network partitioning based on weighted spectral clustering. International Workshop on Complex Networks and their Applications,

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical review E, 69*(6), 066133.

Newman, M. E., Michelle. (2004). Finding and evaluating community structure in networks. *Physical review E, 69*(2), 026113.

Parizeau, M. (2004). Réseaux de neurones. *GIF-21140 et GIF-64326, 124*.

Pournaras, E., Taormina, R., Thapa, M., Galelli, S., Palleti, V., & Kooij, R. (2020). Cascading failures in interconnected power-to-water networks. *ACM SIGMETRICS Performance Evaluation Review, 47*(4), 16-20.

Punitha, K. (2019). Extraction of Co-Expressed Degs From Parkinson Disease Microarray Dataset Using Partition Based Clustering Techniques. 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT),

Queyroi, F. (2013). *Partitionnement de grands graphes: mesures, algorithmes et visualisation* Université Sciences et Technologies-Bordeaux I].

Rossman, L. A. (2000). EPANET 2: users manual.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics, 20*, 53-65.

Rouvière, L. (2021). Graph Mining.

Rutkowski, L., RafałTadeusiewicz, RyszardZadeh, Lotfi AZurada, Jacek M. (2010). *Artificial Intelligence and Soft Computing, Part I: 10th International Conference, ICAISC 2010, Zakopane, Poland, June13-17, 2010, Part I* (Vol. 6113). Springer.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence, 22*(8), 888-905.

Slimani, Y., & Drif, A. (2016). Découverte de communautés dans les réseaux complexes.

Sturm, R., J. (2005). Proactive leakage management using District Metered Areas (DMA) and pressure management—Is it applicable in North America. IWA Leakage 2005 conference proceedings,

Sublemontier, J.-H. (2012). *Classification non supervisée: de la multiplicité des données à la multiplicité des analyses* Université d'Orléans].

Suzuki, J. (2022). *Kernel Methods for Machine Learning with Math and R*

Talbi, M. (2013). *Une nouvelle approche de détection de communautés dans les réseaux sociaux* Université du Québec en Outaouais].

Tzatchkov, V. G.-Y., Victor HBourguett Ortíz, Víctor. (2008). Graph theory based algorithms for water distribution network sectorization projects. Water Distribution Systems Analysis Symposium 2006,

Vuchener, C. (2014). *Equilibrage de charges dynamique avec un nombre variable de processeurs basé sur des méthodes de partitionnement de graphe* Bordeaux].

Wang, J., XiaotongPan, Wei. (2009). On Efficient Large Margin Semisupervised Learning: Method and Theory. *Journal of Machine Learning Research, 10*(3).

Wright, R., EdoParpas, PanosStoianov, Ivan. (2015). Control of water distribution networks with dynamic DMA topology using strictly feasible sequential convex programming. *Water Resources Research, 51*(12), 9925-9941.

Zaiz, F. (2010). *Les Supports Vecteurs Machines (SVM) pour la reconnaissance des caractères manuscrits arabes* Université Mohamed Khider Biskra].

Zevnik, J. (2018). *Učinkovita metoda za avtomatsko vzpostavitev merilnih območij v vodovodnih omrežjih: magistrsko delo* Univerza v Ljubljani, Fakulteta za gradbeništvo in geodezijo].