POLYTECH TOURS
Département
Aménagement et Environnement

citeres
UMR 7324

université de TOURS

Projet de Fin d'Etudes (PFE) 2021-2022

Intelligence Artificielle et Énergétique Urbaine

Mindjid Maïzia (sous la direction de) **Côme Geoffray**

2022

AVERTISSEMENT

Cette recherche a fait appel à des lectures, enquêtes et interviews. Tout emprunt à des contenus d'interviews, des écrits autres que strictement personnel, toute reproduction et citation, font systématiquement l'objet d'un référencement.

L'auteur (les auteurs) de cette recherche a (ont) signé une attestation sur l'honneur de non plagiat.

Formation par la recherche, Projet de Fin d'Etudes en génie de l'aménagement et de l'environnement

La formation au génie de l'aménagement et de l'environnement, assurée par le département aménagement et environnement de l'Ecole Polytechnique de l'Université de Tours, associe dans le champ de l'urbanisme, de l'aménagement des espaces fortement à faiblement anthropisés, l'acquisition de connaissances fondamentales, l'acquisition de techniques et de savoir faire, la formation à la pratique professionnelle et la formation par la recherche. Cette dernière ne vise pas à former les seuls futurs élèves désireux de prolonger leur formation par les études doctorales, mais tout en ouvrant à cette voie, elle vise tout d'abord à favoriser la capacité des futurs ingénieurs à :

- Accroître leurs compétences en matière de pratique professionnelle par la mobilisation de connaissances et de techniques, dont les fondements et contenus ont été explorés le plus finement possible afin d'en assurer une bonne maîtrise intellectuelle et pratique,
- Accroître la capacité des ingénieurs en génie de l'aménagement et de l'environnement à innover tant en matière de méthodes que d'outils, mobilisables pour affronter et résoudre les problèmes complexes posés par l'organisation et la gestion des espaces.

La formation par la recherche inclut un exercice individuel de recherche, le projet de fin d'études (P.F.E.), situé en dernière année de formation des élèves ingénieurs. Cet exercice correspond à un stage d'une durée minimum de trois mois, en laboratoire de recherche, principalement au sein de l'équipe Dynamiques et Actions Territoriales et Environnementales de l'UMR 7324 CITERES à laquelle appartiennent les enseignants-chercheurs du département aménagement.

Le travail de recherche, dont l'objectif de base est d'acquérir une compétence méthodologique en matière de recherche, doit répondre à l'un des deux grands objectifs :

- Développer toute ou partie d'une méthode ou d'un outil nouveau permettant le traitement innovant d'un problème d'aménagement
- Approfondir les connaissances de base pour mieux affronter une question complexe en matière d'aménagement.

Afin de valoriser ce travail de recherche nous avons décidé de mettre en ligne sur la base du Système Universitaire de Documentation (SUDOC), les mémoires à partir de la mention bien.

REMERCIEMENTS

Je tiens à remercier toutes les personnes qui ont contribué au succès de mes études et qui m'ont aidé lors de la rédaction de ce mémoire.

Je voudrais dans un premier temps remercier, mon directeur de projet de fin d'études M.MAIZIA, enseignant chercheur à l'université de Tours, pour sa patience, sa disponibilité et surtout son suivi régulier tout au long de ce projet. Ses conseils et son exigence m'ont permis d'organiser mon temps de travail régulièrement sans perdre de vue mes objectifs finaux.

Je tiens à témoigner toute ma reconnaissance aux personnes suivantes, pour leur aide dans la réalisation de ce projet de fin d'études :

Monsieur Thomas Masse, pour le sérieux et la rigueur qu'il m'a inculqué des années durant, tout particulièrement en cette période difficile. Ses questionnements et remise en question m'ont poussé à fournir un travail de qualité et à essayer de me dépasser – le dépasser ?- tout au long de mes études, défaisant systématiquement l'attrait que j'ai pu avoir pour l'abandon.

Mademoiselle Albane Arthuis, qui brille par son absence mais sans qui je n'aurais probablement jamais eu l'audace de choisir ce sujet technique et distant de ma formation.

Monsieur Benjamin Prunier, pour son soutien sans faille et sa bonne humeur malgré les affres rencontrées durant ses études.

Monsieur Diego Delcastillo, pour toute l'aide personnelle comme professionnelle apportée au cours de mes études, en classe préparatoire tout particulièrement.

Mes parents, pour leur soutien constant et leur enthousiasme.

SOMMAIRE

Introduction

Les méthodes actuelles d'estimation et de gestion des consommations énergétiques en milieu urbain restent trop souvent empiriques, définies à même le terrain[1]. Or, face au défi que s'est lancé l'Europe de réduire les émissions de gaz à effet de serre, la maîtrise des consommations en énergie est un levier potentiel puissant. L'enjeu est donc de disposer de bons outils de prédiction et de gestion des consommations pour faciliter la prise de décision en la matière. En effet, de bons modèles de prédictions peuvent par exemple faciliter le dimensionnement de systèmes de récupération des énergies fatales ou, tout simplement, révéler les variables dont l'impact sur le système global est important. Aujourd'hui, des outils de prédiction basés sur l'intelligence artificielle voient le jour dans de nombreux domaines. C'est pourquoi nous étudierons la problématique suivante : « Pouvons-nous prévoir la consommation en énergie électrique d'une région urbaine grâce à l'intelligence artificielle ? ». Nous essayerons dans ce projet de fin d'étude de montrer l'efficacité de l'intelligence artificielle à déterminer la consommation en énergie électrique d'une région urbaine.

Définition des termes du sujet

Dans cette première partie, nous allons nous intéresser à la définition des termes de notre problématique.

Explorons tout d'abord le concept de « l'intelligence artificielle » qui est au cœur de notre sujet. Nous nous intéresserons à la signification de chacun des deux mots qui composent « intelligence artificielle » avant de proposer une définition globale. L'intelligence, du latin « *intellĕgō* » est composée du préfixe « *inter* » qui signifie « entre » ou « parmi » et du radical « *ego* » qui désigne l'acte de « choisir », « recueillir ». En associant les deux sens nous obtenons « choisir entre » ou « recueillir parmi ». On peut donc au sens premier considérer l'intelligence comme la faculté à s'approprier des éléments choisis parmi d'autres. On a bien la notion de sélection qui s'ajoute à l'action de s'approprier un objet. Le Larousse, lui, nous apprend que l'intelligence est la faculté de comprendre par la pensée. Bien que l'action de sélection ne se retrouve pas dans le mot « comprendre » donné ici, cette définition permet d'affiner l'approche étymologique en précisant la nature de l'objet. Le Larousse nous propose aussi une seconde définition : « Aptitude à s'adapter à une situation, à choisir en fonction des circonstances : capacité de comprendre, de donner un sens à telle ou telle chose. ». Cette deuxième définition enrichi la première en ajoutant la création de sens.

[1] Marijana Zekić-Sušac, Rudolf Scitovski, et Adela Has, « Cluster analysis and artificial neural networks in predicting energy efficiency of public building as a cost saving approach », *Croation Review of Economic, Business and Social Statistics (CREBSS)* 4, nº 2 (2018), https://doi.org/10.1515/crebss.

Le mot « artificiel » vient du latin « *artificialis* » qui signifie « fait avec art ou fait selon l'art ». *Le mot art ici se rapporte à l'ensemble des connaissances et des règles d'action dans un domaine particulier*. L'adjectif artificiel se rapporterait donc à un quelque chose qui a été fabriqué par l'homme selon des règles établies. Le Larousse lui, nous donne une première définition : « Produit par une technique humaine, et non par la nature ». Cette première définition est bien plus large que ce que nous avons déduit de la simple étymologie puisque l'on se rapporte ici à tout objet produit par l'homme.

En articulant les définitions de ces deux termes, nous pouvons donc essayer de construire une première définition de l'intelligence artificielle : Il s'agit donc d'un objet créé par des méthodes humaines capable de comprendre et de donner du sens aux choses. Etant donné le cadre de notre sujet, nous pouvons dès lors recentrer cette première définition de la façon suivante : un objet créé par des méthodes humaines capable de comprendre et de donner du sens aux données qu'on lui propose. Il apparaît donc dès lors que l'intelligence artificielle doit être capable de sélectionner les prédicateurs importants dans un grand jeu de variables à sa disposition afin d'obtenir les meilleures prévisions. On retrouvera d'ailleurs cet aspect dans les différents articles que nous étudierons par la suite.

Le Larousse nous propose cependant une définition de l'intelligence artificielle : c'est un « ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine ». Nous pouvons alors grâce à cette dernière définition construire celle que nous utiliserons dans l'intégralité de ce projet de fin d'études : « L'intelligence artificielle désigne l'ensemble des théories et techniques mises en œuvre en vue de réaliser une machine capable de sélectionner et comprendre les données que l'on lui propose». On note ici que notre définition n'a pas été enrichie de sources venant de la philosophie ou de la littérature en général, afin de ne pas perdre de vue l'aspect concret et matériel de notre sujet.

Nous pouvons ensuite brièvement définir ce qu'est la consommation en énergie électrique d'une région urbaine. Il s'agit en fait de la quantité d'électricité consommée en un temps donné à l'intérieur d'une aire urbaine spatialement finie. On peut obtenir cette valeur par la différence entre l'électricité sortant du territoire et la somme de l'électricité entrante avec l'électricité produite sur ce territoire.

Nous conserverons ces deux définitions des termes de notre sujet pour la suite.

État de l'art

1. Machine Learning et arbres de décision

Commençons par un point sur les techniques d'intelligence artificielles. Une intelligence artificielle peut être créée à partir de techniques variées, en fonction des données qu'elle sera amenée à traiter et des objectifs qui lui seront attribués. Nous allons ici nous intéresser tout particulièrement au *machine learning* ou « apprentissage automatique » en français.

Il s'agit de technologies d'intelligence artificielle basées sur un « apprentissage » grâce à de grands jeux de données dans le but de pouvoir par la suite anticiper ou classer des données nouvelles.

L'apprentissage peut se faire sur un jeu de données *étiquetées*, c'est-à-dire qu'il s'effectue sur des données pour lesquelles on sait déjà ce que l'intelligence artificielle devrait renvoyer après avoir effectué ce pour quoi elle a été conçue. Dans notre cas, nous travaillerons avec des données sous forme de vecteurs. L'intelligence artificielle s'entraine donc sur des données entièrement connues, des exemples qu'elle apprendra à classer pour pouvoir ensuite y ranger des données nouvelles. On parlera alors ici *d'apprentissage supervisé*.

Cependant, l'apprentissage peut aussi être *non supervisé* si les données ne sont pas *étiquetées*. L'intelligence artificielle doit alors découvrir elle-même la logique intrinsèque aux données qui lui sont présentées. Mais le *machine learning* englobe encore lui-même de nombreux algorithmes basés sur des techniques variées, différentes architectures de programme.

L'une de ces architectures types est appelée « *Decision Tree* » et nous la désignerons par la suite « DT ». Un DT est une forme d'arbre de décision logique formé grâce à un entrainement supervisé, aléatoire : « by a stochastic process »[2]. Chaque nœud de l'arbre correspond à un test sur une variable d'entrée des données[3] (le test renvoi donc un booléen). Seul un sous-ensemble aléatoire et de proportion prédéterminée de l'ensemble des données est sélectionné pour la construction de l'arbre en lui-même[4]. Chaque nœud mènera donc à deux chemin différents selon le résultat du test. Enfin, il convient d'utiliser des « pruning techniques »[5] afin de réduire l'arbre (largement éployé) obtenu de la sorte. Une fois la construction terminée, le programme est donc en mesure de classer de nouvelles données et de leur fournir une estimation de la valeur d'une des variables (dans notre cas, la variable à estimer sera toujours celle de la consommation électrique).

---

[2] Anh-Duc Pham et al., « Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability », *Journal of Cleaner Production*, n° 260 (2020), https://www.sciencedirect.com/science/article/abs/pii/S095965262031129X.

[3] Anh-Duc Pham et al.

[4] Anh-Duc Pham et al.

[5] Anh-Duc Pham et al.

La « *Random Forest* » est une seconde architecture que l'on notera donc « RF » dans le reste de ce document. La RF est en fait un modèle qui intègre de multiples arbres de décision, de multiples modèles DT en d'autres termes. Chaque arbre de la forêt est entrainé sur un échantillon aléatoire différent créé à partir du jeu de données d'entrainement[6]. Une fois la RF entrainée arbre par arbre, on peut utiliser le modèle de prédiction ainsi développé. Chaque arbre renvoi alors sa prédiction pour la donnée proposée. La prédiction globale du modèle peut alors être obtenue de deux façons différentes. Par vote majoritaire si l'on souhaite une classification, ou par moyenne de tous les résultats si l'on souhaite une régression : « Finally, when making predictions, the final decisions are made based on the prediction of each tree by averaging for regression problems or majority vote for classification problems. »[7] Enfin, l'architecture RF a pour particularité de permettre d'estimer l'importance des variables d'entrées pour la construction du résultat. « One important feature of RF is that it can calculate feature importance.»[8] Il existe encore bien d'autres modèles dont nous parlerons pour certains plus loin mais ce sont ces deux derniers qui nous intéressent réellement.

2. Un modèle RF efficace

Des outils de prédiction basés sur des intelligences artificielles ont déjà fait l'objet de recherches dans le domaine de l'énergétique. Marijana Zekić-Sušac, Saša Mitrović et Adela Has, dans un article de 2018 traitant de la prédiction des consommations énergétique de bâtiments publics grâce à des réseaux de neurones artificiels ont réussi à mettre en avant le potentiel de l'intelligence artificielle dans le domaine de l'énergétique urbaine. Ils ont, dans cette recherche, testé l'efficacité de trois intelligences artificielles différentes à prédire la consommation d'énergie (en kWh/m².a avec « a » qui est ici l'aire au sol chauffée) de bâtiments publics Croates grâce à un vaste jeu de données de cinq cent soixante-quinze bâtiments recensant quatre-vingt-deux variables différentes. (Des variables diversifiées qui recensent des attributs traitant des conditions géospatiales, de la construction des bâtiments ou encore des systèmes de chauffage…). Dans un autre article daté de 2021, Marijana Zekić-Sušac , Adela Has et Marinela Knežević confrontent à nouveaux les mêmes architectures d'intelligence artificielles à ce jeu de données publiques pour trouver une méthode de sélection des variables importantes. Dans ces deux articles, l'équipe de Marijana Zekić-Sušac a utilisé trois types d'intelligences artificielles relevant de l'apprentissage automatique : un modèle de type RF, plusieurs arbres de décisions ainsi qu'un modèle de type « *Deep Artificial Neural Network* » que l'on abrègera en « DNN ».

Le modèle DNN correspond à des méthodes de « *deep learning* » majoritairement basées sur la construction d'un réseau de neurones artificiels permettant un

[6] Anh-Duc Pham et al.
[7] Zhongnan Ye et al., « Identifying critical building-oriented features in city-block-level building energy consumption: A data-driven machine learning approach », *Applied Energy* 301 (1 novembre 2021), https://www.sciencedirect.com/science/article/pii/S0306261921008436.
[8] Zhongnan Ye et al.

apprentissage non supervisé. Ces différentes intelligences artificielles, une fois entrainées sont donc testées sur des données étiquetées provenant du même jeu Croate. Ainsi, connaissant déjà les valeurs réelles de la consommation d'énergie des différents bâtiments, les auteurs peuvent calculer le « *Symmetric Mean Absolute Percentage Error* » ou « SMAPE » afin de quantifier et comparer l'efficacité des différents modèles entre eux. Ils sont sans appel sur les résultats : « The results show that the most accurate model on validation data was the Random forest, which has produced the SMAPE of 13.5875% showing a potential of machine learning methods in energy management in the public sector »[9]. Selon eux, ce SMAPE de 13.5875% obtenu par leur modèle RF construit avec 500 arbres différents confirme donc bien le potentiel de l'intelligence artificielle. La seconde étude plus récente de Marijana Zekić-Sušac et al. confirme une nouvelle fois le potentiel du machine learning « Such findings confirm the potential of hybrid machine learning methods »[10] et affirme l'efficacité d'une sélection des variables en entrée au moyen d'un algorithme de boruta couplé à l'utilisation d'un modèle RF.

3. Efficace malgré des protocoles différents

D'autres travaux confirment et plaident en faveur du potentiel de l'intelligence artificielle dans le domaine de l'énergétique urbaine. Anh-Duc Pham et ses collaborateurs, dans leur article de 2020 intitulé « Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability » concluent eux aussi à l'efficacité de l'intelligence artificielle pour estimer les consommations énergétiques : « The evaluation results revealed that the RF model was effective in predicting hourly building energy consumption »[11].

Pour parvenir à ces résultats, ils ont contrairement à Marijana Zekić-Sušac et al., comparé des modèles de type RF et DT cette fois à un modèle dit « *M5 Model tree* » ou « M5P ». Le modèle M5P est une forme d'arbre de décision binaire possédant des fonctions linéaires sur les derniers nœuds (appelés « *feuilles* »)[12]. C'est-à-dire qu'en fonction des choix faits pour la construction du modèle, l'arbre propose en sortie une valeur obtenue après transformation par une application qui transformera le vecteur que forment les différentes variables de notre donnée en un nombre réel.

Anh-Duc Pham et al. dans cet article s'intéressent à des prédictions à très court terme : entre une et vingt-quatre heures. Ceci les distingue grandement de Zekić-Sušac et al. Qui, eux, ont travaillé sur une consommation énergétique correspondant à une période de onze années (entre 2006 et 2017).

---

[9] Marijana Zekić-Sušac, Saša Mitrović, et Adela Has, « Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities », *International Journal of Information Management* 58 (2021), https://www.sciencedirect.com/science/article/pii/S0268401219302968.

[10] Marijana Zekić-Sušac, Adela Has, et Marinela Knežević, « Predicting energy cost of public buildings by artificial neural networks, CART, and random forest », *Neurocomputing* 439 (10 février 2021): 223- 33.

[11] Anh-Duc Pham et al., « Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability ».

[12] Anh-Duc Pham et al.

Les deux études se distinguent aussi par les indicateurs d'erreur utilisés : Le SMAPE et le MAPE. Cette différence empêche d'en comparer les résultats. Ainsi, bien qu'elles n'utilisent pas la même méthodologie, ces deux équipes affirment et défendent l'efficacité de l'apprentissage artificiel - tout particulièrement celle du modèle RF- pour la prédiction de consommations énergétiques de bâtiments. Si l'on arrive à prévoir la consommation énergétique de plusieurs bâtiments, il parait probable que l'on puisse calculer la consommation d'une région urbaine.

4. Appliqué à un ensemble de bâtiments

Au mois de novembre dernier, l'équipe de chercheurs composée de Zhongnan Ye, Kuangly Cheng, Shu-Chien, Hsi-Hsien Wei et Clara Man Cheung décida de s'intéresser à la consommation énergétique de « city-blocks »[13] assimilables à des pâtés de maison dans l'optique d'appliquer les méthodes étudiées précédemment à une échelle régionale. « Little attention has been paid to geospatial patterns of building energy consumption from a regional perspective »[14]. Ils proposent ainsi dans leur article d'identifier les caractéristiques clés des bâtiments pour prévoir la consommation d'énergie à l'échelle d'un pâté de maisons. Toutefois, Zhongnan Ye et al. dans leur article se placent dans un problème de classification RF plutôt que dans un problème de régression. Les consommations énergétiques de pâtés de maisons de Taipei qu'ils utilisent sont en fait classées en deux catégories de consommation. La première correspond à une consommation inférieure ou égale à 186.2 kWh/m² et la seconde à des consommations supérieures. Leurs résultats sont donc binaires et l'évaluation de leurs performances ne peut pas être comparée à celle des articles précédents. Ceci dit, l'équipe de Zhongnan Ye et al. conclue aussi sur une performance supérieure du modèle RF avec une précision atteignant 80.31% dans le meilleur des cas. « The RF model is found to outperform other machine learning models »[15]
C'est donc grâce aux conclusions de ces articles que nous pouvons formuler l'hypothèse de ce projet de fin d'étude : « l'intelligence artificielle peut permettre de déterminer la consommation en énergie électrique d'une région urbaine ».

[13] Zhongnan Ye et al., « Identifying critical building-oriented features in city-block-level building energy consumption: A data-driven machine learning approach ».
[14] Zhongnan Ye et al.
[15] Zhongnan Ye et al.

Données et échantillonnage

1. Constituer un jeu de données

Afin d'entrainer nos modèles d'intelligence artificielles, il est nécessaire d'avoir un vaste jeu de données cohérentes. Un jeu de données réelles a été utilisé pour ce projet de fin d'études. Le site internet français « data.gouv.fr » est une plateforme ouverte des données publiques françaises émanant du Premier ministre. Le jeu de données « Consommation annuelle d'électricité et gaz par secteur d'activité et par commune » est sélectionné dans un premier temps pour caractériser notre finalité à savoir la consommation des communes. La note de méthodologie de la collecte de ces données est disponible en ligne sur la même page que les données. Quatre variables en sont extraites pour construire notre jeu de données.

Code variable	Libellé variable	Description
annee	Année	Année
code_commune	Code Commune	Code INSEE de la commune
id_filiere	Identifiant de la filière	Si électricité: "100"; si gaz : "200"
consototale	Consommation totale	Consommation totale (MWh)

Tableau 1: Code variable, Libellé variable et Description des variables extraites du premier jeu de données.

Ce jeu de données contient des enregistrements pour des années comprises entre 2014 et 2018. Le choix est fait de ne conserver que les enregistrements de 2018 de la filière électrique.

Afin d'expliquer les consommations électriques observées dans ce premier jeu de données il convient d'avoir d'autres données contenues dans des variables décrivant concrètement les régions urbaines. Ces variables seront appelées par la suite « *descripteurs* ». L'INSEE émane du ministère chargé des finances et propose de nombreux jeux de données décrivant le territoire français. Le dossier complet de l'INSEE est un vaste jeu de données contenant 1890 descripteurs pour chacune des 35 993 communes de France. Ces descripteurs décrivent l'état des communes en 2013, 2016 et 2018. Il est décidé de manière arbitraire d'extraire 22 descripteurs de cette table concernant tous l'année 2018 afin de correspondre avec le premier jeu de données. Le champ « CODEGEO » contenant le code INSEE de la commune est aussi extrait pour servir d'identifiant. Il convient de noter ici que les termes « commune » et « région urbaine » seront confondus dans ce projet de fin d'étude en raison de la forme des données employées.

Code variable	Libellé variable	THEME	SOURCE
P18_POP	Population en 2018	Évolution et structure de la population	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
P18_POP0014	Nombre de personnes de 0 à 14 ans en 2018	Évolution et structure de la population	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
C18_POP15P	Nombre de personnes de 15 ans ou plus en 2018	Évolution et structure de la population	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
C18_POP15P_CS7	Nombre de personnes de 15 ans ou plus Retraités en 2018	Évolution et structure de la population	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
P18_NBPI_RPMAISON	Nombre de pèces des résidences principales de type maison en 2018	Logement	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
P18_RPAPPART	Nombre de résidences principales de type appartement	Logement	Insee, RP2008, RP2013 et RP2018, géographie
P18_RP_ACH45	Nombre de résidences principales construites de 1919 à 1945 en 2018	Logement	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
P18_RP_ACH70	Nombre de résidences principales construites de 1946 à 1970 en 2018	Logement	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
P18_RP_ACH90	Nombre de résidences principales construites de 1971 à 1990 en 2018	Logement	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
P18_RP_ACH05	Nombre de résidences principales construites de 1991 à 2005 en 2018	Logement	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
P18_RP_ACH15	Nombre de résidences principales construites de 2006 à 2015 en 2018	Logement	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
P18_RP_CCCOLL	Nombre de résidences principales avec chauffage central collectif en 2018	Logement	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
P18_RP_CCIND	Nombre de résidences principales avec chauffage central individuel en 2018	Logement	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
P18_RP_CINDELEC	Nombre de résidences principales avec chauffage individuel électrique en 2018	Logement	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
C18_EMPLT_AGRI	Nombre d'emplois au lieu de travail dans l'agriculture en 2018	Caractéristiques de l'emploi au sens du recensement	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
C18_EMPLT_INDUS	Nombre d'emplois au lieu de travail dans l'industrie en 2018	Caractéristiques de l'emploi au sens du recensement	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
C18_EMPLT_CONST	Nombre d'emplois au lieu de travail dans la construction en 2018	Caractéristiques de l'emploi au sens du recensement	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
C18_EMPLT_CTS	Nombre d'emplois au lieu de travail dans le commerce, les transports et les services divers en 2018	Caractéristiques de l'emploi au sens du recensement	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
C18_EMPLT_APESAS	Nombre d'emplois au lieu de travail dans l'administration publique, l'enseignement, la santé humaine et l'action sociale en 2018	Caractéristiques de l'emploi au sens du recensement	Insee, RP2008, RP2013 et RP2018, géographie au 01/01/2021
HTCH21	Nombre de chambres dans les hÃ´tels en 2021	Tourisme	Insee, partenaires territoriaux en géographie au 01/01/2021
CPGE21	Nombre d'emplacements de camping en 2021	Tourisme	Insee, partenaires territoriaux en géographie au 01/01/2021
VVLIT21	Nombre total de places lit dans les Villages vacances - Maisons familiales en 2021	Tourisme	Insee, partenaires territoriaux en géographie au 01/01/2021

Tableau 2: Descripteurs provenant du Dossier complet de l'INSEE retenus arbitrairement

Afin d'obtenir un seul et même jeu de données reliant descripteurs et consommations totales des régions urbaines, ces deux tables sont jointes grâce au champ du code commune INSEE présent dans les deux jeux.

2. Nettoyage du jeu de données

Pour obtenir un jeu cohérent duquel les intelligences artificielles pourront apprendre facilement, les données sont brièvement « nettoyées ». Les enregistrements ne renseignant pas de valeur pour l'un des champs sont supprimés. Ce choix est pris en raison de la faible proportion d'enregistrements concernés par ce défaut. Nous aurions autrement pu utiliser la méthode des moindres carrés pour conserver et compléter les enregistrements incomplets[16]. Les communes ayant héritées lors de la jointure de plusieurs consommations électriques différentes voient tous leurs enregistrements supprimés en raison de mon incapacité à départager les différents résultats. Finalement, seuls les 22 descripteurs de l'INSEE ainsi que le champ de la

[16] Marijana Zekić-Sušac, Rudolf Scitovski, et Adela Has, « Cluster analysis and artificial neural networks in predicting energy efficiency of public building as a cost saving approach ».

consommation totale sont conservés. Après ce nettoyage, le jeu de données ne contient plus que 33 881 enregistrements.

Il est important de noter à ce stade que les communes ne sont pas réparties également par leur consommation électrique. Cette particularité est tout à fait logique étant donné que les grandes communes sont bien moins nombreuses que les petites communes.

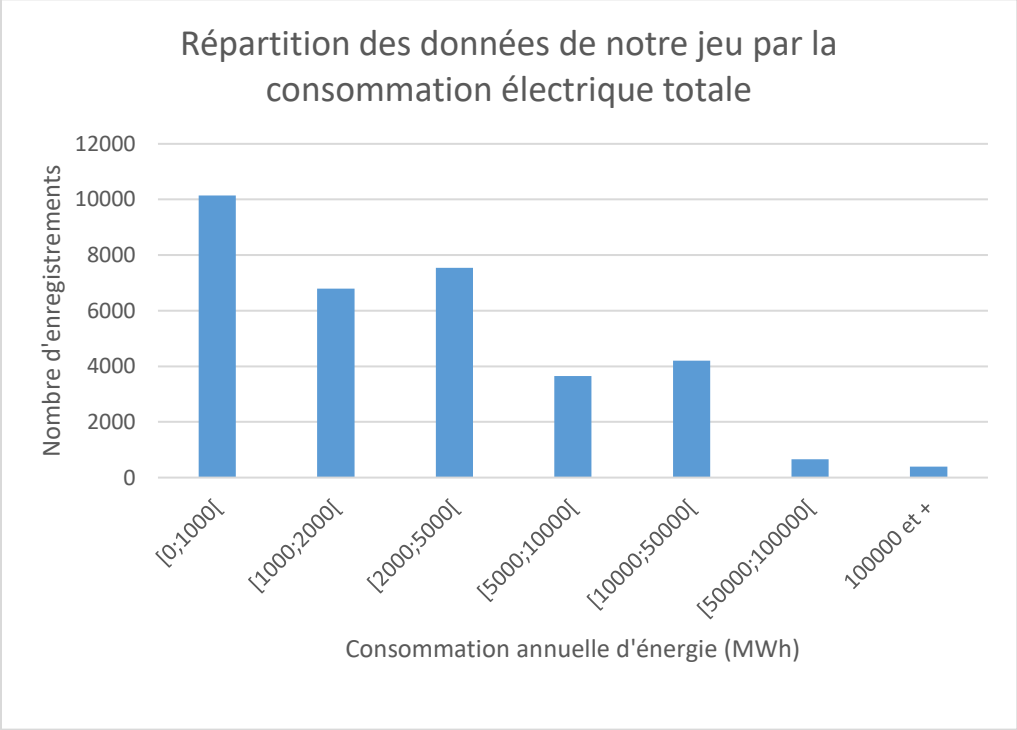

Figure 1: Répartition des communes par consommation électrique

3. Échantillonnage

Afin de pouvoir obtenir par la suite une estimation de l'efficacité des prédictions du modèle, il faut séparer le jeu de données en deux échantillons. L'un sera consacré à l'entrainement des modèles de prédiction et l'autre sera consacré à l'évaluation de ces derniers. Par ailleurs, l'efficacité de l'intelligence artificielle ne semble pas être influencée positivement par une procédure de regroupement des données avant l'entrainement[17]. Nous n'utiliserons donc pas de telles méthodes dans le cadre de notre étude. En se basant sur les proportions utilisées dans la littérature, il est choisi d'utiliser 70% des données pour l'entrainement et de ne conserver que 30% pour la validation des résultats[18]. Ces échantillons sont tirés aléatoirement avant d'être fixés pour pouvoir être réutilisé par la suite.

[17] Marijana Zekić-Sušac, Rudolf Scitovski, et Adela Has.
[18] Marijana Zekić-Sušac, Saša Mitrović, et Adela Has, « Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities »; Marijana Zekić-Sušac, Rudolf Scitovski, et Adela Has, « Cluster analysis and artificial neural networks in predicting energy efficiency of public building as a cost saving approach »; Marijana Zekić-Sušac, Adela Has, et Marinela Knežević, « Predicting energy cost of public buildings by artificial neural networks, CART, and random forest ».

Méthode

1. Paramétrage pour l'entrainement

Les programmes sont réalisés sous python grâce à la bibliothèque Scikit-Learn permettant de construire des modèles d'intelligences artificielles. Bien que plusieurs articles testent et utilisent des réseaux de neurones artificiels (Artificial Neural Networks – ANN - ou Deep Artificial Neural network), les résultats semblent peu probants face au modèle RF[19]. Les modèles ANN n'obtiennent pas un SMAPE inférieur à 35.48% sur le jeu de données Croate sur lesquelles Zekić-Sušac et al. avaient obtenu des résultats plus de deux fois meilleurs grâce au modèle RF[20]. En posant l'hypothèse que les performances relatives de ces modèles seront les mêmes en étudiant des régions urbaines, nous choisissons donc d'utiliser seulement deux architectures différentes : DT et RF.

Le modèle DT est entrainé selon les paramètres par défaut de la classe « *DecisionTreeRegressor* » de la bibliothèque Scikit-Learn étant donné le peu d'informations trouvées sur le sujet dans la littérature. On retiendra que l'arbre de régression une fois construit selon les paramètres par défaut aura une « *profondeur* » maximum égale à 48 pour un total de 23363 « *feuilles* ». Pour explication, la profondeur d'un arbre de décision correspond au nombre maximum de nœuds rencontrés verticalement. Je rappelle ici qu'une feuille est un nœud en sortie de l'arbre. Le nombre de feuilles correspond donc au nombre de sorties différentes que peut proposer l'arbre de décision.

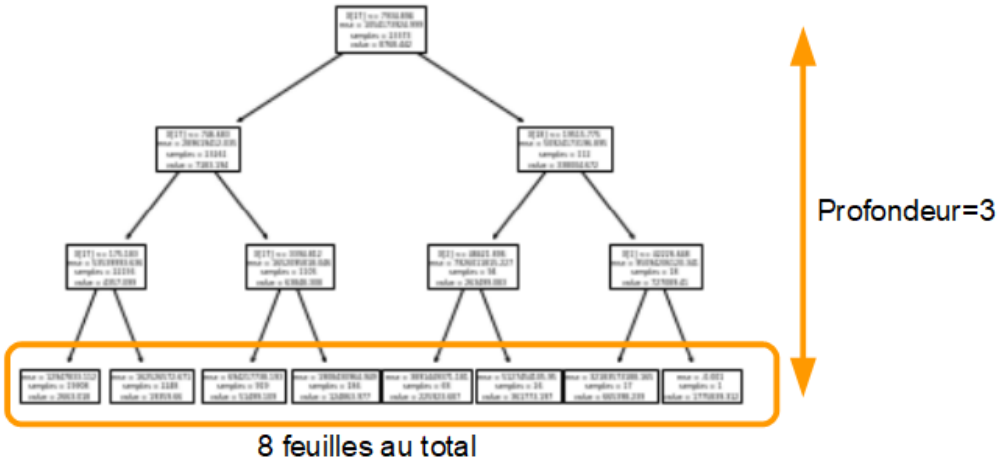

Figure 2: Schéma illustrant l'architecture d'un arbre de décision

[19] Marijana Zekić-Sušac, Adela Has, et Marinela Knežević, « Predicting energy cost of public buildings by artificial neural networks, CART, and random forest ».
[20] Marijana Zekić-Sušac, Rudolf Scitovski, et Adela Has, « Cluster analysis and artificial neural networks in predicting energy efficiency of public building as a cost saving approach ».

Le modèle RF, supposé fournir les résultats les plus probants d'après nos hypothèses se doit d'être paramétré plus précisément. Tout d'abord, il est constaté que les articles utilisent un nombre d'arbres « $n_{estimators}$ » pour les modèles RF compris entre 50[21] et 500[22]. Il est choisi ici d'entrainer trois forêts différentes quant au nombre total d'arbre par modèle soit de50, 100 et 450 arbres chacune. A l'instar de deux articles de notre bibliographie, la profondeur maximale des arbres est placée à l'infini[23] ; il n'y a donc pas de profondeur maximale. Enfin, le nombre de descripteurs considérés pour chercher la meilleure division à chaque nœud « $max_{features}$ » est paramétré en log2 comme l'ont fait Anh-Duc Pham et al. ou encore Marijana Zekić-Sušac et al. dans leurs études. Les autres paramètres sont choisis par défaut. Tous les paramètres sont disponibles en annexe n°1.

2. Validation des modèles

Pour évaluer les performances de nos modèles, il est nécessaire d'utiliser un indicateur qui puisse d'une part donner en un chiffre une idée de la performance générale du modèle de prédiction et d'autre part servir d'outil de comparaison entre nos résultats et ceux de la littérature. A cette fin, nous choisissons d'utiliser le SMAPE rencontré très régulièrement dans la littérature. La moyenne symétrique des écarts en valeur absolue par rapport aux valeurs observées a pour avantage d'être influencée de manière équivalente par une prédiction inférieure à la valeur réelle que par une prédiction supérieure à cette dernière. Ce point le distingue du « *Mean Absolute Percentage Error* » qui pénalise plus durement les erreurs supérieures à la valeur réelle que celles inférieures. Le SMAPE est calculé de la manière suivante :

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

Où n est le nombre total d'enregistrements de l'échantillon de test, A_t la valeur réelle observée et F_t la valeur produite par le modèle.

[21] Zhongnan Ye et al., « Identifying critical building-oriented features in city-block-level building energy consumption: A data-driven machine learning approach ».
[22] Marijana Zekić-Sušac, Saša Mitrović, et Adela Has, « Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities »; Marijana Zekić-Sušac, Adela Has, et Marinela Knežević, « Predicting energy cost of public buildings by artificial neural networks, CART, and random forest ».
[23] Zhongnan Ye et al., « Identifying critical building-oriented features in city-block-level building energy consumption: A data-driven machine learning approach »; Anh-Duc Pham et al., « Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability ».

Résultats et discussion

1. Résultats

Les résultats obtenus après validation des modèles sur l'échantillon prévu à cet effet sont encourageants et démontrent le potentiel du *machine learning* à l'échelle de la région urbaine. Le meilleur modèle est caractérisé par un SMAPE égal à 22,82% et l'ensemble des résultats sont présentés dans le tableau n°3.

Machine learning modèle	n_estimators	SMAPE
RT	1	29,30%
RF	50	23,13%
RF	100	22,94%
RF	450	22,83%

Tableau 3: Résultats des modèles de machine learning

Ces prédictions présentent toutes des écarts moyens à la réalité plus ou moins grands. Ces imprécisions sont en partie dues au modèle RF qui ne fait qu'extrapoler des lois « absurdes » grâce aux données fournies. En effet, chaque feuille de l'arbre propose une valeur de la consommation électrique définie pour un volume de l'espace à 22 dimensions que forment nos différents descripteurs. L'arbre ne fait que diviser cet espace en régions dont il considère le comportement comme homogène. Chaque division/ nœud sépare une des dimensions de l'espace de manière binaire (une borne fixe pour un des descripteurs). Le modèle ne cherche pas une équation globale rattachée à des lois de physique permettant de décrire la consommation électrique en fonction des descripteurs. C'est pourquoi ces prédictions ne peuvent qu'approcher la réalité et conserveront toujours un certain bruit.

La figure n°3 présente la dispersion des différents modèles autours des valeurs réelles. Bien que les résultats ne sont pas exactement les mêmes, on peut visuellement apparenter les trois profils de dispersion des modèles RF entre eux.

Globalement, ces modèles présentent des prédictions plus précises pour des valeurs faibles de la consommation électrique. En effet, le fonctionnement des arbres de décision favorise nécessairement les prédictions pour des communes aux caractéristiques « classiques ». L'espace évoqué précédemment et formé par les descripteurs n'est en fait pas « régulièrement découpé ». Une région moins riche en données d'entrainement se retrouve moins précisément découpée qu'une région largement fournie en données. Les communes aux caractéristiques peu banales seront donc associées à des communes aux caractéristiques plus éloignées et partageront la même valeur en sortie. Comme vus avec la figure n°1, nos données sont plus nombreuses pour de faibles consommations électriques. On peut donc

expliquer de la sorte la dispersion des prédictions pour les fortes consommations électriques.

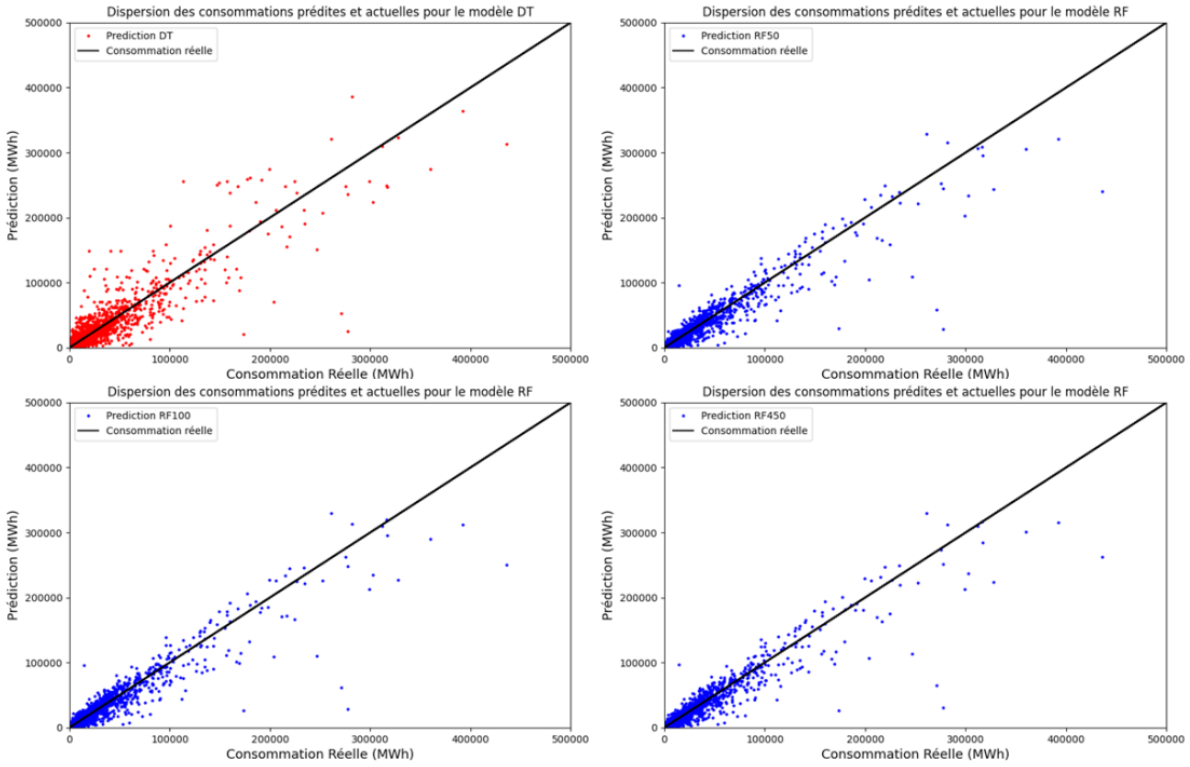

Figure 3: Dispersion des prédictions et des valeurs réelles pour les 4 modèles utilisés

On remarque ensuite que les écarts aux valeurs réelles constatés sont majoritairement des erreurs positives (valeur de la prédiction inférieure à la valeur réelle). Cette tendance s'avère d'autant plus remarquable pour de grandes consommations. Cette tendance est explicable par les descripteurs utilisés. Pour expliquer ce constat, il faut noter que les descripteurs retenus consultables dans le tableau n°2 décrivent des paramètres caractérisant toutes les communes, centrés sur le logement et la population. L'activité économique n'est que peu décrite par ces données. Par ailleurs, on sait aussi que les petites communes ne présentent que peu ou pas de structures particulières consommant de l'énergie (usines, infrastructures de transports électriques etc…). Or, le modèle fourni une répartition plus centrée autour des prédictions de communes enregistrant de faibles consommations électriques. Donc, les descripteurs retenus sont plus pertinents pour décrire de petites villes peu ou pas fournies structures particulières consommant de l'énergie. De cette première conclusion, on déduit donc que l'existence de sites particuliers de consommations énergétique non représentés par les descripteurs utilisés est possible. L'existence de tels sites entrainerait nécessairement une augmentation de la consommation en énergie. Cela explique donc enfin la tendance observée : les consommations réelles sont majoritairement supérieures aux consommations prédites.

Enfin, on note une différence flagrante de performance entre le modèle DT et les 3 modèles RF. Cette différence est due à la structure de l'algorithme. Les prédictions des modèles RF sont obtenues au moyen de nombreux modèles DT construits sur des

échantillons différents provenant des données dédiées à l'entrainement (ces échantillons sont appelés « *bootstrap samples* »). La moyenne des résultats obtenus par chacun des arbres de la forêt permet d'améliorer les résultats du modèle[24].

2. Mise en perspective

Bien qu'encourageants et confirmant la thèse de ce projet de fin d'études, le meilleur SMAPE de 22,83% obtenu par le modèle RF pour $n_{estimators}$=450 reste bien inférieur au modèle RF développé par Marijana Zekić-Sušac et al. qui était de 13,5875%[25]. Une raison probable à leur meilleure performance est qu'ils ont utilisé $n_{estimators}$=500 comme paramètre pour leur modèle. Comme vu dans notre table de résultats, augmenter $n_{estimators}$ tend à diminuer le SMAPE et donc à améliorer les performances du modèle. Toutefois, il n'est pas utile d'augmenter indéfiniment $n_{estimators}$ étant donné qu'il s'agit d'un nombre de tirages aléatoires, augmenter le nombre de tirage ne peut que tendre à stabiliser les performances aux alentours d'un SMAPE minimum inhérent au modèle. Ce n'est donc pas la seule explication à l'écart de performance entre leurs modèle et le nôtre.

Finalement, il faut tout de même noter qu'étudier la consommation d'une région urbaine n'est pas tout à fait la même chose qu'étudier la consommation d'un bâtiment public. Les descripteurs utilisés par l'équipe de chercheurs sont des caractéristiques techniques de bâtiments ayant un lien « physique » si je puis dire avec la consommation électrique. Mon projet de fin d'études, lui, utilise des variables caractérisant la population d'une région urbaine. Même s'il existe nécessairement un lien entre cette population et le bâti de la région, de nombreux facteurs sont totalement ignorés (on pensera particulièrement à la date de construction des bâtiments par exemple) par mon jeu de données et n'entrent donc pas dans le modèle que nous avons construit. « This type of model can reflect the macroeconomic and socioeconomic impacts on regional energy consumption. […] As a result, the accuracy of the predicted results from the top-down model is limited partially due to the simplification of the methodology »[26]. En parlant de ces « *top-down* » modèles l'équipe de Zhongnan Ye et al. décrit en fait les modèles d'intelligence artificielle qui comme le nôtre cherchent à obtenir des prédictions grâce à des descripteurs géo-spatiaux très généraux. Notre méthode s'oppose à leur modèle « *bottom-up* » qui utilise plutôt des descripteurs techniques des bâtiments qui décrivent des détails de la zone urbaine plutôt que sa globalité.

---

[24] Anh-Duc Pham et al., « Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability », *Journal of Cleaner Production*, nº 260 (2020), https://www.sciencedirect.com/science/article/abs/pii/S095965262031129X.

[25] Marijana Zekić-Sušac, Saša Mitrović, et Adela Has, « Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities », *International Journal of Information Management* 58 (2021), https://www.sciencedirect.com/science/article/pii/S0268401219302968.

[26] Zhongnan Ye et al., « Identifying critical building-oriented features in city-block-level building energy consumption: A data-driven machine learning approach », *Applied Energy* 301 (1 novembre 2021), https://www.sciencedirect.com/science/article/pii/S0306261921008436.

Conclusion :

Pour conclure, les modèles développés dans ce projet de fin d'étude confirment l'efficacité et le potentiel du machine learning pour prédire la consommation d'énergie de régions urbaines. On pourrait ainsi imaginer la création d'outils d'aide à la décision permettant aux collectivités de suivre l'impact des décisions de planification urbaine sur les futures consommations d'énergie. Qui plus est, ces outils pourraient être facilement automatisés et mis à jours grâces aux bases de données émanant d'organismes publics. Toutefois, plusieurs pistes d'améliorations sont à envisager. Tout d'abord, la sélection non arbitraire de descripteurs appropriés au moyen d'un modèle Random Forest couplé avec un algorithme de Boruta semble être une étape préalable indispensable à l'amélioration de notre modèle[27]. Enfin, préciser le choix des paramètres de la forêt aléatoire afin de maximiser la précision du modèle serait aussi une étape essentielle.

[27] Marijana Zekić-Sušac, Adela Has, et Marinela Knežević, « Predicting energy cost of public buildings by artificial neural networks, CART, and random forest », *Neurocomputing* 439 (10 février 2021): 223- 33.

Bibliographie

Anh-Duc Pham, Ngoc-Tri Ngo, Thi Thu Ha Truong, Nhat-To Huynh, et Ngoc-Son Truong. « Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability ». *Journal of Cleaner Production*, nº 260 (2020). https://www.sciencedirect.com/science/article/abs/pii/S095965262031129X.

Marijana Zekić-Sušac, Adela Has, et Marinela Knežević. « Predicting energy cost of public buildings by artificial neural networks, CART, and random forest ». *Neurocomputing* 439 (10 février 2021): 223-33.

Marijana Zekić-Sušac, Rudolf Scitovski, et Adela Has. « Cluster analysis and artificial neural networks in predicting energy efficiency of public building as a cost saving approach ». *Croation Review of Economic, Business and Social Statistics (CREBSS)* 4, nº 2 (2018). https://doi.org/10.1515/crebss.

Marijana Zekić-Sušac, Saša Mitrović, et Adela Has. « Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities ». *International Journal of Information Management* 58 (2021). https://www.sciencedirect.com/science/article/pii/S0268401219302968.

Zhongnan Ye, Kuangly Cheng, Shu-Chien Hsu, Hsi-Hsien Wei, et Clara Man Cheung. « Identifying critical building-oriented features in city-block-level building energy consumption: A data-driven machine learning approach ». *Applied Energy* 301 (1 novembre 2021). https://www.sciencedirect.com/science/article/pii/S0306261921008436.

CITERES *UMR 7324*
Cités, Territoires,

Environnement et Sociétés

POLYTECH TOURS
Département
Aménagement et Environnement

Directeur de recherche :

Mindjid Maïzia

Côme Geoffray

PFE/DAE5

UIT/RESEAU

2021-2022

Intelligence Artificielle et Énergétique Urbaine : sous-titre (minuscule)

Résumé : Les méthodes actuelles d'estimation et de gestion des consommations énergétiques en milieu urbain restent trop souvent empiriques, définies à même le terrain. Or, face au défi que s'est lancé l'Europe de réduire les émissions de gaz à effet de serre, la maîtrise des consommations en énergie est un levier potentiel puissant. L'enjeu est donc de disposer de bons outils de prédiction et de gestion des consommations pour faciliter la prise de décision en la matière. En effet, de bons modèles de prédictions peuvent par exemple faciliter le dimensionnement de systèmes de récupération des énergies fatales ou, tout simplement, révéler les variables dont l'impact sur le système global est important. Aujourd'hui, des outils de prédiction basés sur l'intelligence artificielle voient le jour dans de nombreux domaines. C'est pourquoi nous étudierons la problématique suivante : « Pouvons-nous prévoir la consommation en énergie électrique d'une région urbaine grâce à l'intelligence artificielle ? ». Nous verrons dans ce projet de fin d'étude que l'intelligence artificielle permet effectivement de déterminer la consommation en énergie électrique d'une région urbaine. Pour ce faire, quatre modèles d'intelligences urbaines sont entrainées et testées sur un jeu de données gouvernementales décrivant les communes françaises.

Mots Clés : Intelligence artificielle, consommation électrique, ville, communes, énergie, prévision, random forest, decision tree, machine learning.

Annexes 1: Paramétrage des modèles d'IA

Modèle DT :

Paramètre	Valeur
ccp_alpha	0
criterion	mse
max_depth	None
max_features	None
max_leaf_nodes	None
min_impurity_decrease	0
min_impurity_split	None
min_samples_leaf	1
min_samples_split	2
min_weight_fraction_leaf	0
random_state	25678
splitter	best

Modèle RF :

Paramètre	Valeur
bootstrap	True
ccp_alpha	0
criterion	mse
max_depth	None
max_feature	log2
max_leaf_nc	None
max_sample	None
min_impurit	0
min_impurit	None
min_sample	1
min_sample	2
min_weight_	0
n_estimator	50-100-450
n_jobs	None
oob_score	False
random_stat	25678
verbose	0
warm_start	False

Annexes 2: Code Python

```
# -*- coding: utf-8 -*-
"""
Spyder Editor

This is a temporary script file.
"""
# import libraries
from sklearn.model_selection import train_test_split
from sklearn import tree
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_percentage_error
import sklearn
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import random as rd
import time


#Définition de la fonction SMAPE
def SMAPE(A, F):
    return 100/len(A) * np.sum(2 * np.abs(F - A) / (np.abs(A) + np.abs(F)))



#données random inventées pour mettre à l'épreuve mon programme
#données =pd.DataFrame({"A": pd.Series([rd.randrange(1,1000) for i in range (30000)]),
"B":pd.Series([rd.randrange(1,1000) for i in range (30000)]), "C":pd.Series([rd.randrange(1,1000) for i in range
(30000)]), "D":pd.Series([rd.randrange(1,1000) for i in range (30000)]), "Consommation":
pd.Series([rd.randrange(1,1000) for i in range (30000)])})
#X=données[['A','B','C','D']]
#Y=données[['Consommation']]


# On insère directement grâce au module Pandas nos données comme on veut en forme. IL FAUT PENSER A
TRANSFORMER LES "," en "." DANS LE CSV
#Y=pd.read_csv('E:/cours/PFE/5A/Données/csv/Premier_jeu.csv',sep=';', usecols=[22], dtype=float)
```

```
#X=pd.read_csv('E:/cours/PFE/5A/Données/csv/Premier_jeu.csv',sep=';', usecols=[i for i in range(22)], dtype=float)


X_train=pd.read_csv('E:/cours/PFE/5A/Données/csv/Echantillons_Premier_jeu/X_train.csv',sep=';', dtype=float)
X_test=pd.read_csv('E:/cours/PFE/5A/Données/csv/Echantillons_Premier_jeu/X_test.csv',sep=';',  dtype=float)
Y_train=pd.read_csv('E:/cours/PFE/5A/Données/csv/Echantillons_Premier_jeu/Y_train.csv',sep=';',  dtype=float)
Y_test=pd.read_csv('E:/cours/PFE/5A/Données/csv/Echantillons_Premier_jeu/Y_test.csv',sep=';',  dtype=float)




def Echantillonnage(X,Y):
    X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3) # Paramètre= % pour validation
    X_train.to_csv('E:/cours/PFE/5A/Données/csv/Echantillons_Premier_jeu/X_train.csv',sep=';', index=False)
    X_test.to_csv('E:/cours/PFE/5A/Données/csv/Echantillons_Premier_jeu/X_test.csv',sep=';', index=False)
    Y_train.to_csv('E:/cours/PFE/5A/Données/csv/Echantillons_Premier_jeu/Y_train.csv',sep=';', index=False)
    Y_test.to_csv('E:/cours/PFE/5A/Données/csv/Echantillons_Premier_jeu/Y_test.csv',sep=';', index=False)
    return(X_train,X_test,Y_train,Y_test)


#entrainement des modèles d'IA


def Entrainement(X_train,Y_train):


    #Arbre de décision classique
    Decision_Tree= tree.DecisionTreeRegressor(random_state=25678)
    Decision_Tree = Decision_Tree.fit(X_train, Y_train)


    #Random Forest à dpth=50
    RF_model50= RandomForestRegressor(n_estimators=50,max_depth=None,max_features="log2",random_state=25678)#PARAMETRES A DETERMINER
    RF_model50.fit(X_train, np.ravel(Y_train))


    #Random Forest à dpth=100
    RF_model100= RandomForestRegressor(n_estimators=100,max_depth=None,max_features="log2",random_state=25678)#PARAMETRES A DETERMINER
    RF_model100.fit(X_train, np.ravel(Y_train))
```

```
    #Random Forest à dpth=450
    RF_model450=
RandomForestRegressor(n_estimators=450,max_depth=None,max_features="log2",random_state=25678)#PAR
AMETRES A DETERMINER
    RF_model450.fit(X_train, np.ravel(Y_train))


    return(Decision_Tree,RF_model50,RF_model100,RF_model450)



#Application des modèles d'IA

def Prediction(Decision_Tree,RF_model50,RF_model100,RF_model450,X_test,Y_test):

    #Arbre de décision classique
    Y_pred_DT=Decision_Tree.predict(X_test)
    #RF50
    Y_pred_RF50=RF_model50.predict(X_test)
    #RF100
    Y_pred_RF100=RF_model100.predict(X_test)
    #RF450
    Y_pred_RF450=RF_model450.predict(X_test)

#Production des indicateurs

    #Arbre de décision classique
    SMAPE_DT=SMAPE(Y_test["consototale"],Y_pred_DT)
    #Modèle Random Forest50
    SMAPE_RF50=SMAPE(Y_test["consototale"],Y_pred_RF50)
    #Modèle Random Forest100
    SMAPE_RF100=SMAPE(Y_test["consototale"],Y_pred_RF100)
    #Modèle Random Forest450
    SMAPE_RF450=SMAPE(Y_test["consototale"],Y_pred_RF450)


return(Y_pred_DT,Y_pred_RF50,Y_pred_RF100,Y_pred_RF450,SMAPE_DT,SMAPE_RF50,SMAPE_RF100,SM
APE_RF450)
```

```
def Stabilisation(X_train, X_test, Y_train, Y_test):
    start = time.time()
    #X_train, X_test, Y_train, Y_test=Echantillonnage(X,Y)
    Decision_Tree,RF_model50,RF_model100,RF_model450=Entrainement(X_train,Y_train)

Y_pred_DT,Y_pred_RF50,Y_pred_RF100,Y_pred_RF450,SMAPE_DT,SMAPE_RF50,SMAPE_RF100,SMAPE_
RF450=Prediction(Decision_Tree,RF_model50,RF_model100,RF_model450,X_test,Y_test)




    plt.figure(figsize=[9, 6], dpi=100)
    plt.plot(Y_test["consototale"],Y_pred_DT, 'ro', markersize=2, label='Prediction DT')
    plt.plot(Y_test["consototale"],Y_test["consototale"], 'black', markersize=1, label='Consommation réelle')
    plt.xlabel("Consommation Réelle (MWh)", size = 13)
    plt.xlim(0, 500000)
    plt.ylim(0, 500000)
    plt.ylabel("Prédiction (MWh)", size = 13)
    plt.title("Dispersion des consommations prédites et actuelles pour le modèle DT")
    plt.legend()
    plt.show()


    plt.figure(figsize=[9, 6], dpi=100)
    plt.plot(Y_test["consototale"],Y_pred_RF50, 'bo', markersize=2, label='Prediction RF50 ')
    plt.plot(Y_test["consototale"],Y_test["consototale"], 'black', markersize=1, label='Consommation réelle')
    plt.xlabel("Consommation Réelle (MWh)", size = 13)
    plt.xlim(0, 500000)
    plt.ylim(0, 500000)
    plt.ylabel("Prédiction (MWh)", size = 13)
    plt.title("Dispersion des consommations prédites et actuelles pour le modèle RF")
    plt.legend()
    plt.show()


    plt.figure(figsize=[9, 6], dpi=100)
    plt.plot(Y_test["consototale"],Y_pred_RF100, 'bo', markersize=2, label='Prediction RF100 ')
    plt.plot(Y_test["consototale"],Y_test["consototale"], 'black', markersize=1, label='Consommation réelle')
    plt.xlabel("Consommation Réelle (MWh)", size = 13)
```

```
    plt.xlim(0, 500000)
    plt.ylim(0, 500000)
    plt.ylabel("Prédiction (MWh)", size = 13)
    plt.title("Dispersion des consommations prédites et actuelles pour le modèle RF")
    plt.legend()
    plt.show()


    plt.figure(figsize=[9, 6], dpi=100)
    plt.plot(Y_test["consototale"],Y_pred_RF450, 'bo', markersize=2, label='Prediction RF450 ')
    plt.plot(Y_test["consototale"],Y_test["consototale"], 'black', markersize=1, label='Consommation réelle')
    plt.xlabel("Consommation Réelle (MWh)", size = 13)
    plt.xlim(0, 500000)
    plt.ylim(0, 500000)
    plt.ylabel("Prédiction (MWh)", size = 13)
    plt.title("Dispersion des consommations prédites et actuelles pour le modèle RF")
    plt.legend()
    plt.show()




    end = time.time()
    elapsed = end - start
    print('Temps d\'exécution :'+str(elapsed)+'s')

print("SMAPE_DT="+str(SMAPE_DT)+"\nSMAPE_RF50="+str(SMAPE_RF50)+"\nSMAPE_RF100="+str(SMAPE
_RF100)+"\nSMAPE_RF1450="+str(SMAPE_RF450))


    return(Y_pred_DT,Y_pred_RF50,Y_pred_RF100,Y_pred_RF450)


def Importance(Arbre):
    importances= Arbre.feature_importances_
    std = np.std([tree.feature_importances_ for tree in Arbre.estimators_], axis=0)
    forest_importances = pd.Series(importances, index=X_train.columns)
    fig, ax = plt.subplots()
    forest_importances.plot.bar(yerr=std, ax=ax)
    ax.set_title("Feature importances using MDI")
```

```
    ax.set_ylabel("Mean decrease in impurity")
    fig.tight_layout()
    plt.show()
    return()
```