# Optimisation de la Gestion de l'Énergie par l'Apprentissage par Renforcement et le Clustering de Séries Temporelles pour la Génération de Politiques Individualisées

Théo Zangato*, Aomar Osmani*
Pegah Alizadeh*

*LIPN - UMR - CNRS, 7030
Université Paris Sorbonne Nord,
Paris, France
{zangato, ao, alizadeh}@lipn.univ-paris13.fr

**Résumé.** Cet article propose une méthode innovante pour optimiser les cycles de charge des unités de stockage d'énergie des bâtiments, en réponse à la demande croissante d'énergie et aux préoccupations environnementales. La technique, se base sur l'apprentissage par renforcement pour générer des politiques individualisées. Elle utilise le clustering de courbes de charge des bâtiments pour identifier les modèles communs, intègre les connaissances du domaine dans l'algorithme d'apprentissage, et prédit les observations futures pour des décisions en temps réel. Les résultats sur des données réelles démontrent une efficacité significative, réduisant les coûts énergétiques jusqu'à 15%, en limitant la consommation pendant les périodes de pointe et en s'adaptant aux différents profils de consommation des bâtiments par rapport à la référence.

## 1 Introduction

La consommation mondiale d'énergie a considérablement augmenté, passant de 5 268 TWh en 1974 à 22 315 TWh récemment. Les bâtiments représentent 30% de cette consommation, contribuant ainsi aux émissions de gaz à effet de serre. Ce phénomène est particulièrement marqué dans les zones urbaines, où résidait 54% de la population mondiale en 2018. Cette tendance a des répercussions financières, avec une part de l'énergie atteignant 25% du budget des ménages européens en 2021 [1]. Les bâtiments, y compris les maisons individuelles, jouent un rôle crucial dans la transition énergétique visant à contrer le changement climatique. L'intégration des énergies renouvelables dans les réseaux électriques requiert une adaptabilité pour gérer les variations des flux énergétiques. Les avancées technologiques, telles que les panneaux solaires et les batteries, rendent les énergies renouvelables plus accessibles, soulignant ainsi le besoin d'outils de gestion énergétique. En exploitant le stockage domestique et l'énergie renouvelable auto-produite, les bâtiments peuvent réduire leur empreinte carbone, diminuer les

1. source : `http://data.europa.eu/88u/dataset/e3td1ejcprfbhotlntxwa`

coûts énergétiques et atténuer la pression sur le réseau pendant les heures de pointe. L'utilisation d'électricité auto-produite, notamment l'énergie solaire, permet aux bâtiments de diminuer leur dépendance vis-à-vis des sources émettrices de carbone, réduisant ainsi leurs empreintes carbone. Au cours de la dernière décennie, les systèmes de gestion de l'énergie (EMS) ont gagné en importance (Gelazanskas et Gamage (2014)). Des chercheurs tels que Huang et al. (2015) ou Lee et Cheng (2016) ont proposé des approches pour intégrer les énergies renouvelables dans les architectures de réseau existantes et améliorer l'efficacité énergétique. Les EMS permettent la surveillance, le contrôle et l'optimisation de la consommation d'énergie des bâtiments, réduisant la consommation d'énergie. La gestion efficace des cycles de charge et de décharge des unités de stockage d'énergie (ESU) constitue un défi clé pour les EMS, étant donné leur rôle pour compenser la nature intermittente des énergies renouvelables (de Sisternes et al. (2016)).

L'objectif principal de cet article est de proposer une méthode efficace pour la gestion de ces cycles. Nous améliorons le processus d'apprentissage des agents d'EMS en utilisant l'apprentissage par renforcement (RL), en abordant les problèmes d'inefficacité d'échantillons dénotés par Buckman et al. (2018); Zhang et al. (2021); Yu (2018). Notre solution implique l'incorporation de connaissances préalables dans le modèle d'apprentissage, permettant aux agents de se concentrer sur l'optimisation des tâches plutôt que d'apprendre les contraintes environnementales à partir de zéro. Cette approche favorise la généralisation à des sous-groupes spécifiques, réduisant la nécessité de s'adapter à une large gamme de sous-groupes en fonction des besoins énergétiques variables. Dans les sections 2 et 3, nous introduisons le problème de gestion de l'énergie ainsi que sa modélisation en tant que problème RL. Les sections 4 et 5 détaillent notre méthode ainsi que les résultats de nos expériences. Pour assurer la reproductibilité, le code source est disponible sur un dépôt GitHub [2].

# 2 Description du problème

Les EMS jouent un rôle essentiel dans la régulation de la consommation d'énergie au sein des bâtiments et visent principalement à réaliser des économies de coûts dans l'utilisation de l'énergie (Kurte et al. (2023)). Ils opèrent soit dans un ensemble de bâtiments constituant un micro-réseau, soit dans un bâtiment individuel. Ces structures hébergent divers systèmes de consommation d'énergie, et l'approvisionnement énergétique provient d'un fournisseur d'énergie via le réseau électrique. Les coûts associés peuvent être de nature financière et/ou environnementale, et sont sujets à des tarifs horaires. Parmi les demandes énergétiques d'un bâtiment, on distingue deux catégories distinctes : les charges déplaçables, qui peuvent être retardées sans perturber les opérations (ex : four, machine à laver), et les charges non déplaçables (NSL), qui ne peuvent pas être reprogrammées. Pour atteindre ses objectifs, un EMS gère efficacement ces dernières qui peuvent être optimisées en fonction des tarifs, des conditions environnementales (ex : température, humidité, exposition au soleil) et de l'heure de la journée. Cela implique qu'un contrôleur prenne des décisions sur l'activation de dispositifs spécifiques en fonction des informations disponibles. De plus, les bâtiments peuvent intégrer des dispositifs de génération d'énergie renouvelable tels que des panneaux solaires et des ESU qui fournissent des capacités de stockage pour l'énergie électrique ou thermique précédemment générée.

2. https ://github.com/TheoZan/CL

Diverses approches d'EMS ont été proposées, notamment via le Model Predictive Control (MPC) (Mariano-Hernández et al. (2021)), le Rule-Based Control (RBC), et le Fuzzy Logic Control (FLC) (Motevasel et Seifi (2014)). Récemment, l'apprentissage par renforcement (RL) a attiré l'attention, en particulier grâce son succès dans le domaine des jeux (Silver et al. (2018)), entraînant l'émergence d'EMS basé sur le RL (Yu et al. (2021)). Ces approches s'étendent aux systèmes de HVAC dans divers contextes, tels que les logements personnels et les bureaux. Par exemple, dans Xu et al. (2022), les auteurs ont combiné des fonctions d'experts avec l'apprentissage par renforcement, tandis que dans Ren et al. (2022), un réseau d'apprentissage profond Dueling-double Q-learning avec un module de prédiction LSTM assisté par corr-entropie généralisée a été utilisé pour les systèmes de HVAC des logements personnels.

# 3 Formulation du problème

Dans notre étude, nous examinons un ensemble de $N$ bâtiments $B = \{b_1, b_2, \ldots, b_N\}$. Chaque bâtiment est équipé d'un contrôleur responsable de réguler le flux d'énergie de chaque ESU. Ce mécanisme de contrôle influence dynamiquement l'énergie prélevée sur le réseau, l'augmentant lors du stockage d'énergie dans les unités et la diminuant lors de la libération d'énergie précédemment stockée. Pour un bâtiment spécifique on note : $L_t$ les charges non déplaçables (NSL) en $kWh$ au temps $t$, $E_t^{th}$ la consommation électrique en $kWh$ associée aux besoins thermiques (ex : eau chaude, chauffage), $E_t^{ESU}$ les transferts d'énergie en $kWh$ se produisant dans toutes les ESU. Si une ESU stocke de l'énergie, alors $E_t^{ESU} > 0$, et si elle en libère, alors $E_t^{ESU} < 0$. De même, $E_t^{pv}$ est l'énergie produite en $kWh$ par des sources renouvelables, dans notre cas grâce à des panneaux solaires, $H_t$ est l'état de charge (SoC) normalisée de toute ESU par rapport à sa capacité $v$. La consommation totale d'énergie du bâtiment est définie par : $E_t = L_t + E_t^{th} + E_t^{ESU} + E_t^{pv}$.

À tout moment, la consommation du bâtiment doit être satisfaite, ce qui est réalisé par une combinaison de l'utilisation de l'énergie renouvelable auto-produite, la libération d'énergie stockée depuis les unités de stockage et l'acquisition d'énergie depuis le réseau. De plus, l'énergie renouvelable est toujours priorisée, ainsi : $E_t \geq L_t + E_t^{th} - E_t^{pv}$. Selon l'action choisie par le contrôleur, l'énergie restante $E^r > 0$ est acquise depuis le réseau.

## 3.1 Fonctions objectif

Notre étude de cas propose un EMS pour bâtiment donné avec plusieurs objectifs, se concentrant spécifiquement sur l'optimisation des cycles de charge/décharge de la batterie du bâtiment. Les principaux objectifs sont de minimiser les dépenses énergétiques tout en limitant simultanément les émissions de gaz à effet de serre. L'optimisation est guidée par deux objectifs clés, notés $C^1$ (1) et $C^2$ (2), mesurés en dollars. Ces objectifs varient sur une séquence temporelle spécifiée $T$ et dépendent des coûts financiers et environnementaux en temps réel d'une unité d'énergie du réseau $E_t^r$, représentés par $c_t^1$ et $c_t^2$ au temps $t$.

$$C^1(t) = \sum_{t=0}^{T} E_t^r \times c_t^1 \qquad (1) \qquad\qquad C^2(t) = \sum_{t=0}^{T} E_t^r \times c_t^2 \qquad (2)$$

## 3.2 Apprentissage par renforcement : modélisation

En définissant le contrôle des flux d'énergie de manière horaire, on peut considérer l'EMS comme un problème de décision séquentielle. Afin de résoudre ce problème, chaque bâtiment $b_i$ et son environnement peuvent être modélisés comme un Processus de Décision Markovien (MDP) : $M =< S, A, P, R, \gamma >$ dans lequel $S, A$ sont les ensembles des états et des actions, $P$ la probabilité de choisir une action dans un état donné et de passer à un autre état, $R$ la fonction de récompense et $\gamma \in [0, 1]$ le facteur d'escompte. Dans un MDP, l'objectif principal est de trouver une politique optimale $\pi : S \longrightarrow A$ qui maximise (minimise) la somme escomptée des récompenses (pénalités) : $J = \mathbb{E}\tau \left[\sum^{T-1} t = 0\gamma^t r_t\right]$, où $\tau$ est une trajectoire générée par la politique $\pi$ et $T$ la longueur maximale de l'épisode.

### 3.2.1 Etats et actions

Chaque $s_t \in S$ est défini pour un bâtiment donné comme : $s_t = \left\{C_t^{\text{grid}}, \{H_t\}, E_t^{\text{pv}}, L_t, \{K_t\}\right\}$. Il comprend le coût d'une unité d'énergie du réseau $C_t^{\text{grid}}$, qui est l'agrégation de $c_t^1$ et $c_t^2$, un ensemble de valeurs de SoC de l'ESU, la production d'énergie renouvelable, la charge non déplaçable du bâtiment et un ensemble de caractéristiques temporelles $\{K_t\}$ représentant le mois, le jour et l'heure.

L'agent gère l'ESU du bâtiment à travers une seule action continue, notée $a_t \in [-1, 1]$. Cela représente la quantité d'énergie, en $kWh$, à stocker ou à libérer en proportion de la capacité totale du dispositif. Les limites de cet espace ne prennent en compte que la capacité total de l'ESU. L'espace d'actions réellement valide à chaque instant est déduit de la modélisation des différents composants et des contraintes physiques générales. Par exemple, si $H_t = 0$, alors $a_t \geq 0$. L'espace d'actions valide à chaque instant pour une ESU donnée $d$ est défini par $[0, \frac{v_d - H_t}{v_d}]$ si $E^{pv} \geq E_t$ et $[-\frac{\max(E_t, H_t)}{v_d}, \frac{v_d - H_t}{v_d}]$ sinon.

### 3.2.2 Fonction de récompense

La fonction de récompense tient compte de deux coûts : 1) $R^1$, lié à l'utilisation d'énergie à partir des ESU, et 2) $R^2$, résultant de l'utilisation de l'énergie du réseau pour la consommation résiduelle. Le calcul de $R^1$ implique le coût de l'unité d'énergie de l'ESU $C^d$, calculé en suivant la quantité et le coût associé de l'énergie transférée vers le dispositif, $E^d$ :

$$C_t^d = \frac{C_{t-1}^d \times H_{t-1} + E^d \times C_t^{\text{grid}}}{H_t}. \tag{3}$$

Pour les périodes de charge, nous introduisons l'hyperparamètre $\zeta$ dans la fonction de coût. Ce paramètre ajuste la prise en compte du coût de l'utilisation de l'énergie stockée entre les périodes de charge et de décharge. Une valeur de $\zeta$ de 0 ou 1 signifie la prise en compte du coût uniquement sur la période de décharge ou de charge, respectivement. La deuxième partie de l'équation quantifie le coût de la réponse aux demandes restantes en utilisant le réseau.

$$\begin{aligned} R(t) &= R_t^1 + R_t^2 \\ &= (1 - \zeta) C_t^d \left(\max(0, H_{t-1} - H_t)(1 - \alpha)\right) + \zeta C_t^{\text{grid}} \sum_{e \in E^{th}} E_t^r + E^{ESU} \end{aligned} \tag{4}$$

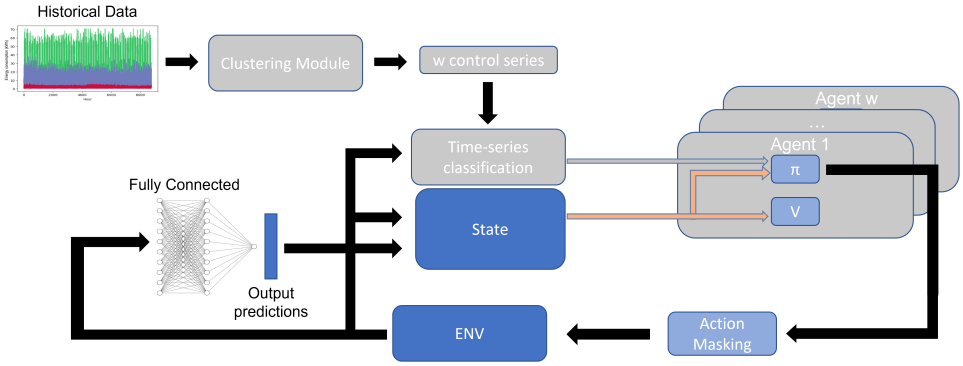



FIG. 1 – Le cadre général de l'approche proposée. La partie grise représente le mapping de politiques basé sur des modèles de consommation pré-identifiés.

# 4 Solution algorithmique

L'analyse du jeu de données d'entraînement montre différents profils de consommation, suggérant la présence de différents types de bâtiments. Le défi consiste à trouver une solution qui peut être appliquée à ces groupes de bâtiments différents, chacun ayant ses propres exigences.

## 4.1 Identification des comportements par analyse des séries temporelles

### 4.1.1 Clustering de séries temporelles

Pour regrouper la consommation en clusters, nous utilisons une méthode de clustering similaire à celle présentée par Łuczak (2016). Soit $\mathbf{X} \in \mathbb{R}^{n \times m}$ la matrice des séries temporelles, où $n$ est le nombre de séries et $m$ est la longueur de chaque série. Pour chaque série temporelle $\mathbf{x}_i$ dans $\mathbf{X}$, nous calculons la dérivée $\mathbf{x'}_i$ comme suit, qui reflète la tendance de la série sur tous les points : $\mathbf{x'}_i(t) = \frac{d}{dt}\mathbf{x}_i(t)$. Ensuite, nous effectuons la transformation de Fourier (Sneddon (1995)) sur chaque série dérivée $\mathbf{x'}_i$ pour obtenir la représentation dans le domaine fréquentiel, que nous notons $\hat{\mathbf{x}}_i$, donnée par :

$$\hat{\mathbf{x}}_i(\omega) = \mathcal{F}[\mathbf{x'}_i(t)] = \int_{-\infty}^{\infty} \mathbf{x'}_i(t)\, e^{-j\omega t}\, dt \tag{5}$$

Nous utilisons la transformée comme extracteur de caractéristiques pour accentuer les motifs significatifs liés à la fréquence dans les données. Les séries résultantes deviennent moins complexes, améliorant l'efficacité de l'algorithme Dynamic Time Warping (DTW), ce qui est particulièrement avantageux lorsqu'il s'agit de séries temporelles longues. Cette approche permet de se concentrer de manière ciblée sur les motifs qui présentent des dépendances fréquentielles prononcées. Cela est pertinent dans le contexte de la consommation électrique dans les bâtiments, connue pour présenter des motifs distincts liés au temps.

Nous transformons notre matrice une dernière fois via le DTW (Müller (2007)). Cette technique permet la comparaison de similarité entre des séries temporelles tout en limitant les

effets de décalage et de distorsion et permet de détecter des motifs et des tendances communs même s'ils sont déphasés. Il existe trois étapes principales. La première consiste à construire la matrice des distances représentant l'ensemble des distances entre deux séries $X$ et $Y$ dans un certain espace de caractéristiques $\Phi$.

$$C_l \in R^{N \times M} : c_{i,j} = \|x_i - y_i\|, i \in [1:N], j \in [1:M] \tag{6}$$

La deuxième étape consiste à parcourir la matrice résultante afin de sélectionner l'alignement. L'algorithme l'exploite pour minimiser le coût entre les points des séquences tout en satisfaisant des conditions de limites, de monotonie et de taille de pas. La fonction de coût est notée :

$$c_p(X,Y) = \sum_{l}^{L} = c(x_{n_l}, y_{m_l}) \tag{7}$$

La solution finale est ensuite générée en trouvant le chemin avec le coût le plus bas à l'aide de l'algorithme de programmation dynamique. La fonction de distance DTW résultante est notée :

$$DTW(X,Y) = c_{p*}(X,Y) = \min\left\{c_p(X,Y), p \in P^{N \times M}\right\} \tag{8}$$

Les clusters sont construits à partir de la matrice résultante en utilisant le regroupement hiérarchique. Le clustering hiérarchique agglomératif est utilisé pour construire des clusters imbriqués en fusionnant chacun d'eux de manière récursive. La mesure de distance choisie est la distance euclidienne calculée sur la base de notre matrice définie entre deux points $a$ et $b$ par : $d(a,b) = \sqrt{(a-b)^2}$. La variance entre les clusters fusionnés est minimisée en tant que critère de liaison appelé critère de liaison de Ward, défini par :

$$\frac{|A|.|B|}{|A| \cup |B|}\|\mu_A - \mu_B\|^2 = \sum_{x \in A \cup B} ||x - \mu_{A \cup B}\|^2 - \sum_{x \in A} ||x - \mu_A\|^2 - \sum_{x \in B} ||x - \mu_B\|^2 \tag{9}$$

### 4.1.2 Classification de séries temporelles

Pour appliquer une politique à un nouveau bâtiment, nous devons déterminer à quel cluster il s'apparente. Pour classer chaque bâtiment dans un type de cluster connu, et ainsi lui associer la bonne politique, nous procédons à une classification du bâtiment en utilisant la méthode décrite précédemment pour calculer la distance entre les séries temporelles.

La première étape consiste à récupérer une série temporelle de contrôle pour chacun de nos clusters connus. Nous ajoutons ensuite la nouvelle série que nous voulons classifier à notre matrice de séries temporelles $\mathbf{M}$ de taille $(w+1, m)$, où $w$ est le nombre de clusters. Nous calculons le vecteur de dissimilarité $V = [V_1, V_2, \ldots, V_k]$ représentant la dissimilarité entre la nouvelle série temporelle et les $w$ séquences de contrôle. Nous trouvons ensuite l'index $j$ tel que $V_j$ est la valeur minimale parmi les éléments de $V$ tel que :

$$j = \underset{j}{\operatorname{argmin}}(V). \tag{10}$$

La séquence de contrôle la plus analogue à la nouvelle série temporelle correspond à l'index $j$. Par conséquent, nous pouvons mapper la politique de $V_j$ sur le nouveau bâtiment, en tirant parti des idées et des stratégies dérivées de la séquence de contrôle similaire identifiée.

### 4.1.3 Prévisions

La dépendance exclusive de l'agent aux données de l'heure précédente pour orienter ses actions empêche de prévoir des stratégies plus efficaces à long terme. Une telle contrainte entrave sa prise de décision, surtout lorsque la situation actuelle diverge significativement de la précédente. Des facteurs tels que la volatilité des prix de l'énergie, les changements dans les habitudes de consommation des utilisateurs ou les variations de la production d'énergie solaire imminente peuvent compliquer davantage le scénario. Pour relever ce défi, nous avons conçu un module prédictif.

Nous avons choisi un perceptrons multicouches (MLP) comme modèle de prédiction, avec 3 couches cachées de 50 neurones suivies de fonctions d'activation ReLu. Le modèle prend en entrée les caractéristiques temporelles données par l'environnement et les 5 dernières observations de la valeur à prédire, et il est entraîné à l'aide de l'optimiseur Adam avec un taux d'apprentissage décroissant.

## 4.2 Algorithme d'apprentissage

La complexité du problème provient des contraintes dépendantes de l'état sur l'espace d'actions. Pour élaborer une politique de prise de décision optimale, on intègre ces contraintes en tant que connaissances a priori dans le MDP en utilisant la version masquable (Huang et Ontañón (2020)) de l'algorithme PPO (Schulman et al. (2017)).

### 4.2.1 PPO

L'algorithme Proximal Policy Optimization agit comme une méthode de gradient de politique, où un estimateur du gradient de la politique $\nabla_\theta J(\theta)$ est calculé à l'aide d'un algorithme d'ascension stochastique du gradient. Il s'appuie sur le travail de Schulman et al. (2015) qui ont proposé l'algorithme TRPO qui mettait déjà en œuvre l'idée d'une contrainte de région de confiance de la divergence entre l'ancienne et la nouvelle politique.

Le gradient est de la forme suivante : $\hat{g} = \mathbb{E}_t \left[ \nabla_\theta \log \pi_\theta(a_t|s_t) \hat{A}_t \right]$, avec $\pi_\theta$ une politique stochastique et $\hat{A}_t$ un estimateur de la fonction avantage à $t$ défini par : $A(s_t, a_t) = Q(s_t, a_t) - V(s_t)$. Le gradient de la fonction objective associée correspond à l'estimateur du gradient de la politique tel que :

$$L^{PG}(\theta) = \mathbb{E}_t \left[ \log \pi_\theta(a_t|s_t) \hat{A}_t \right]. \tag{11}$$

L'algorithme introduit la *clipped surrogate objective function* qui contraint le changement de politique à l'aide d'une fonction de troncature pour éviter des mises à jour de poids destructrices qui sont trop éloignées de la politique actuelle, garantissant des mises à jour stables :

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[ \min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right], \tag{12}$$

avec $r_t(\theta)$ étant le rapport de probabilité entre les anciennes et nouvelles politiques défini comme : $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{old}}(a_t|s_t)$.

#### 4.2.2 Réduction de l'espace des actions

Nous déterminons les limites valides $\beta^1$ et $\beta^2$ de notre espace, qui sont ensuite utilisées dans l'échantillonnage et la prédiction des actions. Les limites peuvent être calculées en utilisant $\phi$ la puissance de sortie maximale de la source chargeant l'ESU, $\Gamma$ et $\eta$ respectivement la puissance maximale d'entrée/sortie et l'efficacité de l'ESU

$$\beta^1 = -\frac{1}{v_d} \max\left(\Gamma_t, \eta_t H_t, E_t\right) \tag{13}$$

$$\beta^2 = \frac{1}{v_d} \min\left(\Gamma_t, \frac{v_d - H_t}{\eta_t}, \phi_t - E_t\right) \tag{14}$$

Pour filtrer les actions invalides, nous utilisons un masque d'actions, garantissant : 1) que les trajectoires $T$ ne contiennent que des actions valides et 2) que seules les actions valides sont utilisées pour le calcul du gradient. Les actions invalides sont masquées en attribuant des valeurs de logit négatives, rendant leur probabilité d'échantillonnage nulle par la fonction *softmax* du réseau.

# 5 Expérimentations

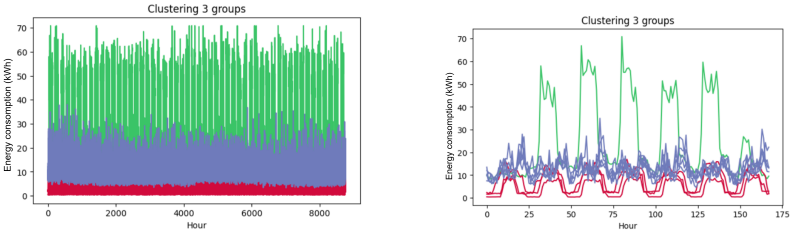
## 5.1 Citylearn

Notre approche a été développée dans l'environnement CityLearn, un environnement Gym open source conçu pour la création d'agents spécialisés dans la coordination de l'énergie des bâtiments et la réponse à la demande dans des contextes urbains (Vázquez-Canteli et al. (2019)). Connu pour sa scalabilité, CityLearn couvre plus de 60 bâtiments situés dans quatre zones climatiques distinctes aux États-Unis. Le modèle de bâtiment dans CityLearn adopte une approche hybride, combinant des modèles basés sur les principes fondamentaux et des modèles basés sur les données pour simuler à la fois les comportements thermiques et électriques, tenant compte des préférences des occupants et des caractéristiques physiques des bâtiments. Cet environnement polyvalent englobe divers systèmes, y compris des ESU et des dispositifs d'énergie renouvelable, chaque bâtiment étant équipé de manière unique.

## 5.2 Résultas

#### 5.2.1 Clustering et classification

Nous avons appliqué la méthode de clustering sur les séries temporelles de consommation annuelle provenant des bâtiments de l'ensemble d'entraînement. Nous avons utilisé $w = 3$ clusters, en choisissant $w$ en considérant les scores de silhouette et d'incohérence pour différents nombres de clusters. Bien que le score de silhouette favorise légèrement 2 clusters, les faibles scores d'incohérence pour 3 clusters et l'observation d'un plateau dans le graphique d'incohérence jusqu'à 6 clusters suggèrent que 3 clusters offrent un compromis raisonnable entre la qualité des clusters et la cohérence dans la structure hiérarchique. Cela concorde avec les connaissances du domaine du jeu de données qui suggèrent plus de deux types de motifs dans les données.




(a) 1 an (m=8760) (b) 1 semaine (m=168)

FIG. 2 – Clustering de consommation électrique en 3 groupes sur a) 1 an et b) 1 semaine.

TAB. 1 – Evaluation metrics comparison of MLP predictor vs. 1-hour lag observations (Base).

| **Observation predicted** | **Evaluation results** | |
|---|---|---|
| | ***MLP*** | ***Base*** |
| Financial cost | **0.007**[a], **0.995**[b] | 0.702, 0.5 |
| Solar Generation | **0.128**, **0.952** | 0.56, 0.798 |
| Non Shiftable Load | **0.117**, **0.964** | 0.220, 0.871 |

[a]RMSE.[b] $R^2$.

Pour associer rapidement un bâtiment inconnu à un groupe connu, nous évaluons la méthode de clustering en ne conservant que $m = 168$ (une semaine ou 1.9% d'une année). Les deux instances de clustering ont identifié des clusters identiques, validant ainsi une semaine de données par rapport aux données annuelles. Après une simulation d'une semaine, nous pouvons attribuer précisément le bâtiment à un groupe et optimiser la réponse de l'agent. La Figure 2 montre les résultats pour des bâtiments représentatifs sur une année et une semaine. Sur la base de ces résultats, nous classifions chaque bâtiment de l'ensemble de test en utilisant $m = 168$.

### 5.2.2 Prévisions

Les prévisions effectuées comportent : la production d'énergie solaire en $W/kW$, qui aide à dériver le coût environnemental ; le coût financier, ainsi que les charges non déplaçables du bâtiment en $W/kW$. Le Tableau 1 présente les résultats de notre approche pour chaque valeur prédite, comparée au cas de base, représenté par une observation avec un décalage d'une heure.

### 5.2.3 Politique

Nous avons formé 3 agents en fonction des groupes identifiés précédemment, chacun apprenant une politique spécifique basée sur le profil de consommation du groupe. Sur l'ensemble de test, notre approche réduit le coût opérationnel du bâtiment d'environ 5% par rapport aux scénarios sans stockage d'énergie. On compare notre méthode à un algorithme RL continu (Haarnoja et al. (2018)) et on note que notre approche converge plus rapidement vers une solution plus performante (Fig 3). On compare aussi notre approche à une heuristique basée sur des règles. L'impact d'une gestion désordonnée de la capacité de stockage sur les coûts d'exploitation est démontré à l'aide d'une politique aléatoire. Les résultats dans le Tableau 2

TAB. 2 – Évaluation des algorithmes sur les bâtiments de l'ensemble de test.

| **groupe identifié** | **Résultats d'évaluation avec comparaison des approches étudiées.** | | | |
|---|---|---|---|---|
| | ***Ours*** | ***SAC*** | ***Aléatoire*** | ***RBC*** |
| 1 | **1.0**[a], **0.937**[b] | 1.018, 0.951 | 1.159, 1.163 | 1.004,0.998 |
| 2 | **1.013**, **0.858** | 1.029, 0.876 | 1.186, 1.171 | 1.017, 0.998 |
| 3 | 1.02, **0.887** | 1.018, 0.935 | 1.071, 1.062 | **1.004**, 0.999 |

[a]yearly carbon emissions.[b]Yearly price cost.

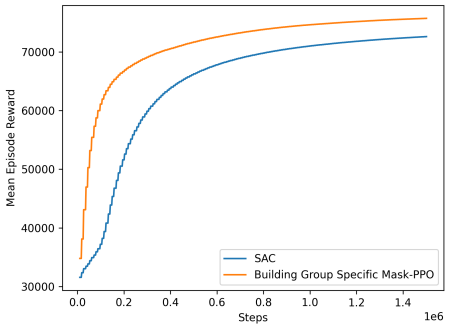



FIG. 3 – Convergence des deux algorithmes comparés sur un groupe donné de bâtiments.

montrent des métriques pour les émissions de carbone et les coûts d'exploitation du bâtiment sur la durée $T$, normalisés par rapport aux scénarios sans stockage. Une valeur de 1.1 indique une augmentation de 10%. Notre approche réduit les coûts d'exploitation financiers jusqu'à 15%, selon le type de bâtiment, tout en maintenant des coûts environnementaux ainsi qu'une consommation globale d'électricité stable.

# 6 Conclusion

Cette étude aborde la gestion séquentielle du stockage de l'énergie dans une grande variété de types de bâtiments, en tenant compte de leurs caractéristiques de consommation uniques. En étudiant ces caractéristiques en détail, nous avons construit un ensemble concis mais exhaustif de politiques qui optimisent le comportement énergétique dans tous les types de bâtiments, ce qui renforce l'intérêt des méthodes RL pour l'EMS. Les méthodes basées sur les MDP permettent de modéliser des contraintes importantes directement dans le cadre d'apprentissage, ce qui permet une exploration sûre, un entraînement sans risque et un déploiement pratique. Notre recherche s'appuie sur des données et des modèles de consommation spécifiques pour améliorer de manière significative l'efficacité de la gestion de l'énergie sans modifier les algorithmes existants, uniquement par le biais d'améliorations de modélisation. En outre, nous avons développé un module de classification robuste, conçu pour associer efficacement de nouvelles données inédites aux politiques existantes en utilisant seulement un petit nombre de points

de données. Cela nous permet d'utiliser efficacement les politiques formées sans avoir besoin d'apprendre de nouveaux comportements pour chaque nouveau bâtiment rencontré.

Les améliorations futures consisteraient à résoudre les problèmes d'accès aux données. L'objectif serait de concevoir une solution qui nécessite un minimum de données pour s'adapter efficacement à des profils de consommation inconnus jusqu'alors en tirant parti du méta-apprentissage.

# Références

Buckman, J., D. Hafner, G. Tucker, E. Brevdo, et H. Lee (2018). Sample-efficient reinforcement learning with stochastic ensemble value expansion. In *NeurIPS 2018*.

de Sisternes, F. J., J. D. Jenkins, et A. Botterud (2016). The value of energy storage in decarbonizing the electricity sector. *Applied Energy*.

Gelazanskas, L. et K. A. Gamage (2014). Demand side management in smart grid : A review and proposals for future direction. *Sustainable Cities and Society*.

Haarnoja, T., A. Zhou, P. Abbeel, et S. Levine (2018). Soft actor-critic : Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR.

Huang, S. et S. Ontañón (2020). A closer look at invalid action masking in policy gradient algorithms. *CoRR abs/2006.14171*.

Huang, Y., H. Tian, et L. Wang (2015). Demand response for home energy management system. *International Journal of Electrical Power & Energy Systems*.

Kurte, K., K. Amasyali, J. Munk, et H. Zandi (2023). Deep reinforcement learning based hvac control for reducing carbon footprint of buildings. In *2023 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*.

Lee, D. et C.-C. Cheng (2016). Energy savings by energy management systems : A review. *Renewable and Sustainable Energy Reviews*.

Mariano-Hernández, D., L. Hernández-Callejo, A. Zorita-Lamadrid, O. Duque-Pérez, et F. S. García (2021). A review of strategies for building energy management system : Model predictive control, demand side management, optimization, and fault detect & diagnosis. *Journal of Building Engineering*.

Motevasel, M. et A. R. Seifi (2014). Expert energy management of a micro-grid considering wind energy uncertainty. *Energy Conversion and Management 83*, 58–72.

Müller, M. (2007). Dynamic time warping. *Information retrieval for music and motion*, 69–84.

Ren, M., X. Liu, Z. Yang, J. Zhang, Y. Guo, et Y. Jia (2022). A novel forecasting based scheduling method for household energy management system based on deep reinforcement learning. *Sustainable Cities and Society 76*, 103207.

Schulman, J., S. Levine, P. Abbeel, M. Jordan, et P. Moritz (2015). Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR.

Schulman, J., F. Wolski, P. Dhariwal, A. Radford, et O. Klimov (2017). Proximal policy optimization algorithms. *CoRR abs/1707.06347*.

Silver, D., T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science 362*(6419), 1140–1144.

Sneddon, I. N. (1995). *Fourier transforms*. Courier Corporation.

Vázquez-Canteli, J. R., J. Kämpf, G. Henze, et Z. Nagy (2019). Citylearn v1.0 : An openai gym environment for demand response with deep reinforcement learning. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pp. 356–357.

Xu, S., Y. Fu, Y. Wang, Z. Yang, Z. O'Neill, Z. Wang, et Q. Zhu (2022). Accelerate online reinforcement learning for building HVAC control with heterogeneous expert guidances. In *International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2022*, pp. 89–98.

Yu, L., S. Qin, M. Zhang, C. Shen, T. Jiang, et X. Guan (2021). A review of deep reinforcement learning for smart building energy management. *IEEE Internet Things J. 8*(15), 12046–12063.

Yu, Y. (2018). Towards sample efficient reinforcement learning. In J. Lang (Ed.), *IJCAI*.

Zhang, J., J. Kim, B. O'Donoghue, et S. P. Boyd (2021). Sample efficient reinforcement learning with REINFORCE. AAAI Press.

Łuczak, M. (2016). Hierarchical clustering of time series data with parametric derivative dynamic time warping. *Expert Systems with Applications 62*, 116–130.

# Summary

In response to escalating energy demands and environmental concerns, the imperative promotion of sustainable practices is explored in this paper. The focus is on employing RL techniques to optimize energy consumption and associated costs, within energy management systems. A three-step approach is introduced to efficiently manage charging cycles in building energy storage units. The strategy involves clustering load curves, incorporating domain knowledge into the learning algorithm, and predicting future observations for real-time decision-making. The method enables controlled exploration and efficient training of EMS agents. In comparison to the benchmark, our model reduces energy costs by up to 15%, decreases consumption during peak periods, and showcases adaptability across diverse building consumption profiles.