

Jornadas de Automática

Aprendizaje por refuerzo para reparto de energía comunitaria: análisis coste-equidad

Andrés A. Pabón Pérez^a, Juan A. Méndez Pérez^{a,*}, José M. González Cava^a, Benjamín González Díaz^a, Santiago Torres Álvarez^a, A. Marcos Trujillo Trujillo^a

^a Universidad de La Laguna (ULL), Camino San Francisco de Paula, 19. 38200 La Laguna (Tenerife), España.

To cite this article: Andrés A. Pabón-Pérez, Juan A. Méndez-Pérez, José M. González-Cava, Benjamín González-Díaz, Santiago Torres-Álvarez, A. Marcos Trujillo-Trujillo. 2026. Aprendizaje por refuerzo para reparto fotovoltaico comunitario: análisis coste-equidad. Jornadas de Automática, 47.
<https://doi.org/10.17979/ja-cea.2026.47.13697>

Resumen

Este trabajo propone una estrategia de aprendizaje por refuerzo para repartir energía fotovoltaica (FV) en comunidades energéticas con autoconsumo colectivo. Como caso de estudio se considera una comunidad de tres miembros con consumos y derechos de participación desiguales. La propuesta emplea optimización proximal de políticas (Proximal Policy Optimization, PPO) y se compara con reparto estático, un controlador basado en reglas (RBC) y control predictivo basado en modelo (MPC). Todas las estrategias deciden los mismos coeficientes de reparto y mantienen fija la capa de control local. La evaluación considera el coste medio y la equidad del cumplimiento acumulado de derechos FV. PPO y MPC comparten una ventana predictiva de 8 h, mientras que RBC no emplea predicción. Las simulaciones muestran que RBC obtiene el menor coste admisible medio, MPC la mayor equidad dinámica y PPO un compromiso entre ambas. Los resultados evidencian el potencial del aprendizaje por refuerzo para el reparto de energía en comunidades energéticas.

Palabras clave: Comunidades energéticas, Reparto fotovoltaico, Aprendizaje por refuerzo, Control predictivo, Autoconsumo colectivo

Reinforcement learning for community energy sharing: a cost-fairness analysis

Abstract

This work proposes a reinforcement learning strategy for allocating photovoltaic (PV) energy in energy communities with collective self-consumption. A three-member community with unequal consumption profiles and participation rights is considered as a case study. The proposed approach employs Proximal Policy Optimization (PPO) and is compared with static allocation, a rule-based controller (RBC), and model predictive control (MPC). All strategies determine the same allocation coefficients while keeping the local control layer fixed. The evaluation considers the average cost and the fairness of cumulative compliance with PV participation rights. PPO and MPC share the same 8 h prediction horizon, whereas RBC does not use forecasts. Simulation results show that RBC achieves the lowest average admissible cost, MPC provides the highest dynamic fairness, and PPO offers a compromise between the two. These results demonstrate the potential of reinforcement learning for energy allocation in energy communities.

Keywords: Energy communities, Photovoltaic sharing, Reinforcement learning, Model predictive control, Collective self-consumption

*Autor para correspondencia: jamendez@ull.edu.es

1. Introducción

Las comunidades energéticas permiten coordinar generación renovable distribuida, almacenamiento y consumo flexible a escala local. Su despliegue se apoya en el marco europeo de participación ciudadana en energía renovable (European Parliament and Council of the European Union, 2018, 2023). En la operación diaria, sin embargo, la minimización de la factura agregada puede entrar en conflicto con el reparto de beneficios entre participantes. Cuando los miembros poseen derechos distintos sobre la instalación fotovoltaica, una política económicamente eficiente puede asignar sistemáticamente más energía a los miembros con mayor capacidad de autoconsumo, dificultando que otros participantes reciban, a lo largo del tiempo, la proporción de energía que les corresponde.

Este trabajo estudia dicho conflicto en una comunidad de tres miembros. Un coordinador comunitario decide los coeficientes de reparto $\beta_i(t)$ de la producción fotovoltaica común, mientras cada miembro opera internamente con el mismo controlador local basado en reglas. En este marco, el reparto fotovoltaico (FV) se aborda como una decisión secuencial. La política observa el estado operativo de la comunidad, selecciona coeficientes de reparto y recibe una recompensa asociada al coste y a la evolución del cumplimiento acumulado de los derechos FV. La estrategia propuesta se basa en aprendizaje por refuerzo y se implementa mediante optimización proximal de políticas (Proximal Policy Optimization, PPO). Su rendimiento se compara con un controlador basado en reglas (RBC) y un controlador predictivo basado en modelo (MPC) manteniendo común la variable de decisión, la capa de control local y el criterio de evaluación. En particular, PPO y MPC se comparan bajo la misma información predictiva, mientras RBC proporciona una referencia heurística sin predicción.

Si una política aprendida, una heurística RBC y un MPC no operan sobre las mismas variables, información y capa local de control, las diferencias de coste observadas no pueden atribuirse exclusivamente a la política de coordinación. Para garantizar una comparación homogénea, este trabajo restringe la evaluación a la asignación global de la generación FV: todas las estrategias comparten la misma capa de control local y se comparan con las mismas magnitudes externas de coste y equidad.

2. Estado del arte y contribución

La literatura reciente sobre comunidades energéticas destaca su potencial para activar flexibilidad local, aumentar autoconsumo renovable, compartir energía y articular nuevos modelos de negocio (Lopez et al., 2024; Ahmed et al., 2024). En comunidades con propiedad compartida o derechos de participación, el criterio de reparto constituye además una variable de gobernanza: la reducción del coste agregado debe evaluarse junto con la coherencia entre energía asignada y derechos pactados. Este aspecto se relaciona con trabajos recientes sobre equidad en sistemas energéticos locales y comunidades energéticas (Soares et al., 2024; Gasca et al., 2025), así como con enfoques que incorporan la equidad directamente en el diseño de problemas de asignación energética compartida (Fornier et al., 2025).

Los controladores basados en reglas siguen siendo atractivos por su bajo coste computacional y facilidad de despliegue, aunque su rendimiento depende de heurísticas fijadas manualmente. El MPC proporciona una referencia habitual en sistemas energéticos porque optimiza sobre un horizonte con restricciones explícitas (Joshua and Gupta, 2023), pero requiere modelo y predicciones. El aprendizaje por refuerzo formaliza la toma de decisiones secuencial mediante interacción agente-entorno (Sutton and Barto, 2018). En energía, trabajos recientes lo han aplicado a gestión con almacenamiento, comunidades energéticas, microrredes comunitarias y mercados locales (Pinthurat et al., 2024; Palma et al., 2024; Li et al., 2024; May et al., 2024). La optimización proximal de políticas es una alternativa robusta para entrenar políticas con acciones continuas (Schulman et al., 2017). Frente a los métodos de optimización en línea, el aprendizaje por refuerzo permite entrenar fuera de línea una política de decisión para un problema secuencial y aplicarla posteriormente mediante una evaluación directa de la política, sin resolver un problema de optimización en cada intervalo de control. A cambio, requiere un entorno de entrenamiento suficientemente representativo y una formulación adecuada de la recompensa y de las restricciones. PPO es un método actor-crítico de gradiente de política que limita cambios excesivos entre actualizaciones mediante un objetivo recortado, favoreciendo un entrenamiento estable con espacios de acción continuos. En este trabajo se seleccionan RBC y MPC como referencias complementarias: RBC representa una solución interpretable y desplegable sin predicción, mientras que MPC representa una referencia predictiva basada en optimización con restricciones explícitas. No se pretende comparar exhaustivamente todas las familias de optimización existentes, sino aislar el efecto del coordinador comunitario manteniendo comunes la variable de decisión, la capa de control local y la información predictiva disponible para los métodos dinámicos.

Cuando las comparaciones entre aprendizaje por refuerzo y MPC en sistemas energéticos difieren en la variable de acción, la información predictiva o la lógica local de control, resulta difícil atribuir las diferencias observadas exclusivamente al coordinador. Este trabajo aborda la necesidad de evaluar políticas de reparto FV comunitario bajo una comparación controlada, centrada exclusivamente en la variable de asignación. A diferencia de formulaciones más amplias que optimizan también restricciones supervisoras o decisiones internas de batería, aquí se aísla una única decisión: el reparto FV global. La contribución es, por tanto, metodológica-aplicada: la propuesta se implementa mediante PPO y se evalúa frente a RBC y MPC con una acción común, una capa local fija y una evaluación bicriterio común de coste y equidad.

En este contexto, la contribución del trabajo se concreta en tres puntos. Primero, se propone una estrategia de reparto fotovoltaico basada en aprendizaje por refuerzo mediante PPO. Segundo, se define un protocolo de evaluación controlado en una comunidad heterogénea de tres miembros, manteniendo comunes la variable de acción, la capa de control local y la información predictiva disponible para los métodos dinámicos. Tercero, se compara la política aprendida con MPC, RBC y dos referencias estáticas, Uniforme y Proporcional, mediante una evaluación bicriterio de coste y equidad basada en el cumplimiento acumulado de los derechos FV.

3. Métodos

3.1. Modelo comunitario

La Figura 1 resume la arquitectura de evaluación: el coordinador selecciona únicamente $\beta(t)$ y la operación interna de cada miembro permanece fija.

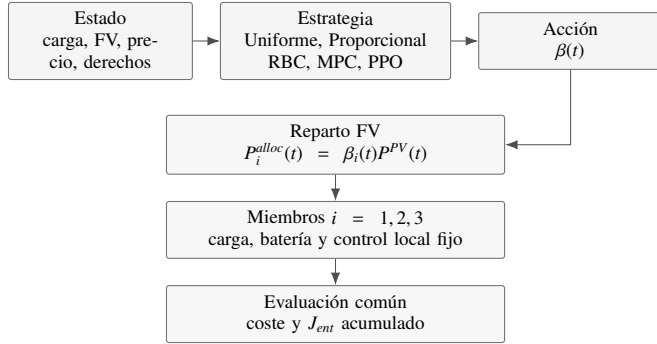


Figura 1: Esquema de coordinación comunitaria evaluado.

En la Figura 1, J_{ent} denota el índice de Jain aplicado al cumplimiento relativo acumulado de los derechos FV. Su definición formal y el procedimiento de cálculo se presentan en la Sección 3.2. La comunidad tiene $N = 3$ miembros. Los subíndices i y j recorren los miembros de la comunidad, mientras t denota el instante de simulación. Las variables con acento circunflejo representan magnitudes previstas dentro del horizonte de control. En particular, $\hat{C}_H$ y $\hat{J}_{ent,H}$ representan, respectivamente, el coste y la equidad estimados a partir de las predicciones de carga y generación FV disponibles durante las próximas 8 h. En cada instante t , el coordinador asigna la producción fotovoltaica agregada $P^{PV}(t)$ mediante

$$P_i^{alloc}(t) = \beta_i(t)P^{PV}(t), \quad \sum_i \beta_i(t) = 1, \quad \beta_i(t) \geq \beta_i^{min}, \quad \beta_i^{min} = \rho r_i, \quad \rho = 0,20. \quad (1)$$

En (1), $P_i^{alloc}(t)$ es la potencia FV asignada al miembro i , $\beta_i(t)$ el coeficiente de reparto, r_i el derecho FV nominal y β_i^{min} la cota inferior, definida con el factor $\rho = 0,20$. Los coeficientes $\beta_i(t)$ son adimensionales y representan fracciones de la producción común. La restricción de suma impone que toda la generación FV se reparte entre los miembros, y la cota inferior evita que un miembro quede sin asignación instantánea. En particular, se adopta $\rho = 0,20$ como salvaguarda operativa para garantizar que cada miembro conserve, en cada instante, al menos el 20 % de su derecho FV nominal y evitar su exclusión temporal total del reparto. Este valor se mantiene fijo en el presente estudio y no se realiza un análisis sistemático de sensibilidad, que se plantea como una extensión futura. El coste comunitario evaluado es

$$C = \sum_t \sum_i p(t)P_{g,i}^+(t)\Delta t, \quad (2)$$

donde C es el coste acumulado, $p(t)$ el precio de energía, $P_{g,i}^+$ la potencia importada de red por el miembro i y Δt el paso temporal de simulación. Sólo se contabiliza la importación positiva de red, por lo que la métrica mide el coste de compra de energía. El escenario principal representa una comunidad heterogénea de tres miembros: escalados

de carga [1,35; 0,75; 1,0], desfases [0; 8; 16] h, derechos FV $r = [0,20; 0,50; 0,30]$ y horizonte de predicción de 8 h.

Tabla 1: Configuración del escenario experimental.

Miembro	Escala carga	Desfase [h]	Derecho FV
1	1,35	0	0,20
2	0,75	8	0,50
3	1,00	16	0,30

La escala de carga multiplica el perfil base de demanda de cada miembro, mientras que el desfase desplaza temporalmente dicho perfil para representar hábitos de consumo distintos. El derecho FV indica la fracción nominal de la producción fotovoltaica comunitaria que corresponde a cada participante. El horizonte de 8 h corresponde a la información futura disponible para MPC y PPO. La configuración induce una tensión operativa entre coste y equidad: el primer miembro tiene mayor demanda relativa y menor derecho FV, mientras el segundo tiene menor demanda y mayor derecho. En ausencia de restricciones de equidad, una política orientada sólo a coste tiende a desplazar energía hacia el miembro con mayor capacidad de autoconsumo. El escenario permite evaluar el conflicto entre eficiencia económica y derechos de participación sin modificar la operación local de los miembros.

3.2. Equidad por derechos FV

La equidad se evalúa sobre energía FV acumulada, en lugar de hacerlo sobre la beta media. Para cada miembro:

$$A_i = \sum_t \beta_i(t)P^{PV}(t)\Delta t, \quad a_i = \frac{A_i}{\sum_j A_j}, \quad f_i = \frac{a_i}{r_i}. \quad (3)$$

En (3), A_i es la energía FV acumulada asignada al miembro i , a_i su cuota acumulada sobre el total asignado y f_i el cumplimiento relativo respecto a su derecho r_i . Un valor $f_i = 1$ indica cumplimiento exacto del derecho acumulado, mientras valores mayores o menores que uno reflejan sobreasignación o déficit relativo. La métrica principal es el índice de Jain (Jain et al., 1984) aplicado al cumplimiento relativo:

$$J_{ent} = \frac{(\sum_i f_i)^2}{N \sum_i f_i^2}. \quad (4)$$

En (4), J_{ent} resume la uniformidad de los cumplimientos relativos entre miembros. El índice toma valores cercanos a uno cuando todos los f_i son similares, pero no identifica por sí solo qué miembro queda por encima o por debajo de su derecho. Se adopta $\bar{J}_{ent} \geq 0,9$ como umbral operativo mínimo de admisibilidad media, no como criterio normativo universal. Umbrales más exigentes modificarían qué estrategias se consideran admisibles y, por tanto, la interpretación operativa de la comparación. Por ello se reportan los valores continuos de $\bar{J}_{ent}$ y del cumplimiento medio por miembro. La estrategia proporcional representa el reparto justo por construcción, mientras que la uniforme representa una referencia estática orientada a igualdad de coeficientes y no garantiza cumplimiento de derechos.

3.3. Estrategias comparadas

3.3.1. Referencias estáticas

Uniforme fija $\beta_i = 1/N$, mientras que Proporcional fija $\beta_i = r_i$. Ambas estrategias mantienen los coeficientes constantes durante toda la semana de simulación y se emplean como

referencias para cuantificar, respectivamente, el efecto de ignorar o respetar de forma fija los derechos FV nominales.

3.3.2. Control basado en reglas (RBC)

La estrategia RBC es una regla dinámica con un mecanismo de corrección de equidad. Calcula una distribución de oportunidad

$$o(t) = \text{norm}(w_n n(t) + w_s m^{SOC}(t)), \quad (5)$$

donde $o(t)$ es la distribución de oportunidad, n_i mide déficit neto normalizado, m_i^{SOC} el margen de carga de la batería y w_n , w_s ponderan ambos términos. En paralelo calcula un objetivo de corrección:

$$\begin{aligned} u(t) &= \text{norm}(\max(r - a(t), 0)), \\ q(t) &= \text{norm}((1 - \lambda)r + \lambda u(t)). \end{aligned} \quad (6)$$

En (6), $a(t)$ y r son los vectores de cuotas acumuladas y derechos nominales. La función $\max(\cdot, 0)$ se aplica componente a componente, $u(t)$ prioriza a los miembros con menor cuota acumulada que su derecho y $q(t)$ combina derechos nominales con dicha corrección mediante λ . El peso de la corrección es

$$\begin{aligned} g(t) &= \text{clip}\left(\frac{J_{high} - J_{ent}(t)}{J_{high} - J_{low}}, 0, 1\right), \\ \beta(t) &= \Pi((1 - g)o + gq). \end{aligned} \quad (7)$$

En (7), J_{low} y J_{high} delimitan la banda de activación, $g(t)$ activa gradualmente la corrección según $J_{ent}(t)$, $\text{clip}(\cdot)$ acota el valor al intervalo $[0, 1]$ y Π proyecta la mezcla sobre las restricciones de (1). Los parámetros utilizados son $J_{low} = 0,88$, $J_{high} = 0,96$, $\lambda = 0,3$, $w_n = 0,75$ y $w_s = 0,25$ y se fijaron empíricamente durante el desarrollo del escenario, antes de la evaluación final, y no se reajustaron sobre las diez semanas de validación. La operación $\text{norm}(\cdot)$ normaliza vectores no negativos a suma uno y usa r como respaldo si la suma es nula. Todas las operaciones sobre $o(t)$, $u(t)$, $q(t)$ y $\beta(t)$ son vectoriales y se aplican sobre los N miembros de la comunidad.

Algoritmo 1. RBC con corrección de equidad

Entrada: déficits $n(t)$, márgenes $m^{SOC}(t)$, cuota acumulada $a(t)$, derechos r y $J_{ent}(t)$.

1. Calcular oportunidad instantánea $o(t)$ con déficit neto y margen de carga.
2. Calcular corrección $u(t)$ para miembros con menor cumplimiento que su derecho.
3. Mezclar derechos nominales y corrección en $q(t)$.
4. Activar la corrección $g(t)$ cuando $J_{ent}(t)$ se aproxima al umbral.
5. Proyectar la mezcla $(1 - g)o + gq$ sobre el conjunto factible de coeficientes de reparto definido por (1).

Salida: vector de reparto $\beta(t)$.

3.3.3. Control predictivo basado en modelo (MPC)

MPC resuelve en cada decisión un problema de horizonte recedente:

$$\begin{aligned} \min_{\beta^{(0)}} \quad & \hat{C}_H(\beta^{(0)}, r_{tail}) + w_f(1 - \hat{J}_{ent,H}) \\ \text{sujeto a} \quad & (1). \end{aligned} \quad (8)$$

En (8), $\beta^{(0)}$ es el reparto optimizado para el primer bloque de control, H el horizonte de 8 h, $\hat{C}_H$ el coste previsto, $\hat{J}_{ent,H}$ la

equidad prevista, r_{tail} el reparto proporcional usado en la cola del horizonte y w_f el peso interno de equidad. En cada iteración se aplica únicamente el primer bloque. La optimización se resuelve mediante SLSQP con 100 iteraciones máximas y tolerancia 10^{-7} . La configuración evaluada utiliza $w_f = 1000$, seleccionado mediante un barrido discreto de sensibilidad como un punto operativo admisible orientado a alta equidad. El MPC emplea predicciones perturbadas de carga y FV, sin acceso al futuro realizado.

3.3.4. Aprendizaje por refuerzo mediante PPO

La estrategia propuesta se formula como una política de aprendizaje por refuerzo. En cada decisión, el agente recibe una observación con el estado operativo disponible: carga, generación FV, precio, derechos de participación y variables acumuladas de equidad. La acción es el vector continuo de coeficientes de reparto $\beta(t)$. La recompensa combina el ahorro frente al reparto proporcional con la variación incremental de equidad. La política se entrena mediante PPO sobre las mismas betas que emplean el resto de estrategias dinámicas. Su recompensa por paso es

$$r_t = (C_t^{prop} - C_t) - w_f[(1 - J_{ent}^{t+1}) - (1 - J_{ent}^t)], \quad (9)$$

donde r_t es la recompensa, C_t^{prop} el coste instantáneo del reparto proporcional, C_t el coste de la acción aplicada y J_{ent}^t , J_{ent}^{t+1} la equidad antes y después de la decisión. El primer término es positivo cuando la acción reduce coste frente a Proporcional. El segundo término penaliza aumentos de la desigualdad acumulada y premia mejoras relativas de equidad. La política reportada utiliza $w_f = 500$ y se entrena con tres inicializaciones aleatorias independientes, $3 \cdot 10^6$ pasos por inicialización, 4 entornos paralelos, red [128, 64] para actor y crítico, tasa de aprendizaje lineal inicial $3 \cdot 10^{-4}$, $n_{steps} = 512$, lote 128, 5 épocas, $\gamma = 0,99$, $\lambda_{GAE} = 0,95$, recorte PPO 0,2, coeficientes de entropía y valor 0,01 y 0,5, recorte de gradiente 0,5 y normalización de observaciones y recompensa.

Los pesos w_f son parámetros internos de cada estrategia y no se comparan directamente. Las configuraciones reportadas corresponden a puntos operativos calibrados en el escenario estudiado. Por tanto, los resultados se interpretan como una comparación empírica de compromisos coste-equidad y no como una selección independiente óptima para escenarios no observados.

3.4. Validación

Todas las estrategias se evalúan sobre las mismas diez semanas de validación, excluidas del entrenamiento. Los perfiles base de carga, FV y precio proceden del conjunto de datos experimental del proyecto y se remuestran con paso de 15 min. La heterogeneidad entre miembros se induce mediante los escalados y desfases de la Tabla 1. Las decisiones comunitarias se actualizan cada hora y la simulación interna avanza con paso de 15 min. Las referencias estáticas fijan betas constantes durante toda la semana. RBC, MPC y PPO actualizan la beta en línea. RBC usa únicamente información disponible en el instante de decisión: déficit neto, margen de carga, cuota acumulada y derechos FV. PPO y MPC reciben la misma trayectoria de precios y las mismas predicciones perturbadas de carga y generación FV para las 8 h siguientes. Ninguno de los dos métodos accede al futuro realizado. MPC utiliza esta

información en una optimización recedente, mientras PPO la incorpora en su observación de estado.

La validación se interpreta como una comparación bicriterio:

$$\pi \mapsto (\bar{C}(\pi), \bar{J}_{ent}(\pi)), \quad \pi \text{ admisible si } \bar{J}_{ent}(\pi) \geq 0,9. \quad (10)$$

En (10), π denota una estrategia evaluada, mientras $\bar{C}$ y $\bar{J}_{ent}$ son las medias sobre las semanas de validación. Una estrategia se considera admisible en la evaluación agregada si su índice medio satisface $\bar{J}_{ent} \geq 0,9$. El umbral separa soluciones admisibles y no admisibles en media, pero no convierte la equidad adicional por encima de 0,9 en irrelevante: el análisis mantiene simultáneamente coste y $\bar{J}_{ent}$. Este criterio no exige que todas las semanas superen individualmente el umbral, sino que resume el comportamiento agregado del periodo de evaluación. Para PPO, las métricas semanales se calculan promediando primero las tres inicializaciones evaluadas y después agregando las diez semanas. La comparación final se realiza con las magnitudes externas $\bar{C}$ y $\bar{J}_{ent}$. La comparación directa entre PPO y MPC se apoya en que ambos métodos operan sobre la misma acción, disponen de la misma ventana de información predictiva y se someten a la misma condición de admisibilidad media. RBC se incluye como referencia heurística desplegable basada únicamente en información disponible en línea.

La evaluación distingue tres tipos de estrategias. Uniforme y Proporcional actúan como referencias estáticas: el primero cuantifica el efecto de ignorar derechos y el segundo el coste de respetarlos de forma fija. RBC representa una heurística desplegable sin entrenamiento ni predicción. MPC actúa como referencia predictiva basada en optimización y orientada a alta equidad. La política aprendida, implementada mediante PPO, se evalúa bajo el mismo criterio de admisibilidad media.

4. Resultados y discusión

La Figura 2 representa el compromiso entre coste y equidad obtenido por las estrategias evaluadas en el experimento principal. La línea discontinua marca $\bar{J}_{ent} = 0,9$. Las barras indican la desviación estándar entre las diez semanas de evaluación. Para PPO, los valores semanales se obtienen promediando previamente las tres inicializaciones evaluadas. Esta representación separa el coste económico de la admisibilidad media por derechos FV, de modo que las estrategias de menor coste sólo resultan aceptables si superan el umbral de admisibilidad media.

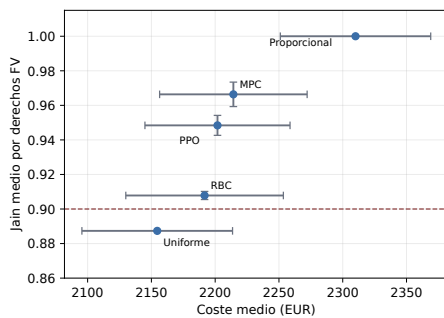


Figura 2: Coste medio e índice medio de equidad por derechos FV. Las barras indican la desviación estándar entre las diez semanas de evaluación. Para PPO, los valores semanales se obtienen promediando previamente las tres inicializaciones evaluadas.

Tabla 2: Resultados del experimento principal sobre 10 semanas.

Estrategia	Coste [EUR]	Δ prop. [%]	$\bar{J}_{ent}$	Admisible
Uniforme	2154,6	-6,73	0,8873	No
Proporcional	2310,0	0,00	1,0000	Sí
RBC	2191,7	-5,13	0,9078	Sí
MPC	2214,3	-4,14	0,9664	Sí
PPO	2201,8	-4,68	0,9484	Sí

La columna “Admisible” de la Tabla 2 indica si la estrategia satisface el umbral de equidad media $\bar{J}_{ent} \geq 0,9$. La Tabla 3 desagrega el cumplimiento relativo medio por miembro. Esta información complementa a $\bar{J}_{ent}$, que resume la dispersión del cumplimiento pero no indica qué participante recibe más o menos energía respecto a su derecho.

Tabla 3: Cumplimiento relativo medio por miembro, $\bar{f}_i$.

Estrategia	$\bar{f}_1$	$\bar{f}_2$	$\bar{f}_3$
Uniforme	1,667	0,667	1,111
Proporcional	1,000	1,000	1,000
RBC	1,603	0,825	0,890
MPC	1,333	0,859	1,013
PPO	1,420	0,825	1,012

Estos resultados muestran que superar el umbral agregado de admisibilidad no implica que cada miembro reciba exactamente su cuota nominal. Por ello, el índice debe interpretarse junto con el cumplimiento individual. Dado que el índice de Jain es no lineal, $\bar{J}_{ent}$ no coincide necesariamente con el valor obtenido al aplicar el índice a los cumplimientos medios $\bar{f}_i$.

El controlador uniforme obtiene 2154,6 EUR y queda por debajo del umbral de equidad medio ($\bar{J}_{ent} = 0,8873$). En este escenario, repartir a partes iguales asigna más FV del derecho nominal al miembro con mayor demanda, reduciendo importación de red a costa de incumplir la distribución pactada. El proporcional representa la referencia estática opuesta: equidad perfecta con un coste de 2310,0 EUR.

RBC incorpora una corrección explícita de equidad sobre la heurística de oportunidad. Bajo el criterio de admisibilidad media adoptado, presenta el menor coste entre las estrategias admisibles: 2191,7 EUR y $\bar{J}_{ent} = 0,9078$. MPC alcanza la mayor equidad entre las estrategias dinámicas, $\bar{J}_{ent} = 0,9664$, con 2214,3 EUR. La configuración de MPC evaluada se selecciona por su orientación hacia una mayor equidad y por cumplir el umbral de admisibilidad media, no por minimizar el coste sin restricciones. La política aprendida ocupa un punto intermedio en el compromiso observado: incrementa $\bar{J}_{ent}$ respecto a RBC al reducir la sobreasignación del miembro 1 y corregir prácticamente el déficit del miembro 3, aunque el cumplimiento relativo medio del miembro 2 permanece en $\bar{f}_2 = 0,825$. A cambio, mantiene un coste inferior al del MPC orientado a equidad.

Desde el punto de vista operativo, los resultados muestran tres comportamientos diferenciados. Uniforme obtiene el menor coste observado, pero queda fuera del conjunto admisible en media. Proporcional garantiza equidad perfecta mediante un reparto fijo, aunque con el mayor coste. Entre las estrategias dinámicas admisibles en media, RBC ofrece el menor coste, MPC alcanza la mayor equidad media y la propuesta de aprendizaje por refuerzo proporciona una solución intermedia

con menor coste que MPC y mayor equidad que RBC. Por tanto, el resultado no recomienda una estrategia única, sino que cuantifica alternativas de operación según la preferencia coste-equidad.

La comparación directa entre la política implementada mediante PPO y MPC no concede ventajas informativas al aprendizaje por refuerzo. El agente se evalúa sobre semanas no usadas en entrenamiento y actúa sobre la misma variable que MPC. MPC tampoco usa el futuro realizado, sino predicciones perturbadas dentro de la misma ventana operativa de 8 h. RBC, por su parte, se interpreta como referencia heurística sin predicción. Las tres inicializaciones evaluadas de PPO presentan una dispersión reducida en el escenario considerado: $\bar{J}_{ent} = 0,9495, 0,9516$ y $0,9441$, con costes medios de 2206,5, 2204,2 y 2194,7 EUR.

5. Conclusiones

Se ha diseñado y evaluado una estrategia de reparto fotovoltaico basada en aprendizaje por refuerzo para una comunidad energética de tres miembros. Al limitar todos los métodos a la decisión global $\beta(t)$ y mantener fija la operación local, la comparación se centra en la política de coordinación comunitaria. Además, la política implementada mediante PPO y el MPC emplean la misma información predictiva.

Los resultados evidencian un compromiso claro entre coste y equidad en las estrategias evaluadas. Bajo el criterio de admisibilidad media adoptado, RBC proporciona el menor coste entre las estrategias admisibles. La estrategia propuesta no sustituye como referencia única a RBC ni a MPC, pero alcanza una solución aprendida intermedia: mayor equidad media que RBC y menor coste que el MPC orientado a equidad. Esta interpretación evita convertir el umbral de admisibilidad en una ordenación total de estrategias y conserva la lectura bicriterio del experimento. Como líneas futuras, resulta de interés evaluar distintas ponderaciones de recompensa, umbrales de admisibilidad, comunidades de mayor tamaño, predicciones más realistas y políticas condicionadas por preferencias de operación.

La aportación metodológica se extiende más allá de este caso de tres miembros: la comparación entre estrategias de coordinación debe fijar la variable controlada, la información disponible y el criterio que define una solución admisible. Los valores cuantitativos, sin embargo, corresponden al escenario evaluado y requieren validación en comunidades mayores y configuraciones adicionales. Bajo este diseño experimental, la propuesta resulta competitiva como compromiso coste-equidad aprendido, mientras que RBC y MPC delimitan referencias útiles de bajo coste y alta equidad, respectivamente.

Agradecimientos

Este trabajo ha sido desarrollado en el marco del proyecto Sustainable Atlantic Communities (SATComm, EAPA

0019/2022), cofinanciado por la Unión Europea a través del programa Interreg Atlantic Area.

Referencias

- Ahmed, S., Ali, A., D'Angola, A., 2024. A review of renewable energy communities: Concepts, scope, progress, challenges, and recommendations. *Sustainability* 16 (5), 1749.
DOI: 10.3390/su16051749
- European Parliament and Council of the European Union, 2018. Directive (EU) 2018/2001 on the promotion of the use of energy from renewable sources. *Official Journal of the European Union*, L 328, 82–209.
- European Parliament and Council of the European Union, 2023. Directive (EU) 2023/2413 amending Directive (EU) 2018/2001 as regards the promotion of energy from renewable sources. *Official Journal of the European Union*, L 2023/2413.
URL: <https://eur-lex.europa.eu/eli/dir/2023/2413/oj>
- Fornier, Z., Leclère, V., Pinson, P., 2025. Fairness by design in shared-energy allocation problems. *Computational Management Science* 22 (2), 11.
DOI: 10.1007/s10287-025-00532-7
- Gasca, M.-V., Rigo-Mariani, R., Debusschere, V., Sidqi, Y., 2025. Fairness in energy communities: Centralized and decentralized frameworks. *Renewable and Sustainable Energy Reviews* 208, 115054.
DOI: 10.1016/j.rser.2024.115054
- Jain, R. K., Chiu, D.-M. W., Hawe, W. R., 1984. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. *Tech. Rep. DEC-TR-301*, Digital Equipment Corporation.
- Joshal, K. S., Gupta, N., 2023. Microgrids with model predictive control: A critical review. *Energies* 16 (13), 4851.
DOI: 10.3390/en16134851
- Li, J., Jiang, Z., Chen, Z., Liu, J., Cheng, L., 2024. CuEMS: Deep reinforcement learning for community control of energy management systems in microgrids. *Energy and Buildings* 304, 113865.
DOI: 10.1016/j.enbuild.2023.113865
- Lopez, I., Goitia-Zabaleta, N., Milo, A., Gomez-Cornejo, J., Aranzabal, I., Gaztanaga, H., Fernandez, E., 2024. European energy communities: Characteristics, trends, business models and legal framework. *Renewable and Sustainable Energy Reviews* 197, 114403.
DOI: 10.1016/j.rser.2024.114403
- May, D. C., Taylor, M., Musilek, P., 2024. Decentralized coordination of distributed energy resources through local energy markets and deep reinforcement learning. *Energy and AI* 18, 100446.
DOI: 10.1016/j.egyai.2024.100446
- Palma, G., Guiducci, L., Stentati, M., Rizzo, A., Paoletti, S., 2024. Reinforcement learning for energy community management: A european-scale study. *Energies* 17 (5), 1249.
DOI: 10.3390/en17051249
- Pinthurat, W., Surinkaew, T., Hredzak, B., 2024. An overview of reinforcement learning-based approaches for smart home energy management systems with energy storages. *Renewable and Sustainable Energy Reviews* 202, 114648.
DOI: 10.1016/j.rser.2024.114648
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms.
DOI: 10.48550/arXiv.1707.06347
- Soares, J., Lezama, F., Faia, R., Limmer, S., Dietrich, M., Rodemann, T., Ramos, S., Vale, Z., 2024. Review on fairness in local energy systems. *Applied Energy* 374, 123933.
DOI: 10.1016/j.apenergy.2024.123933
- Sutton, R. S., Barto, A. G., 2018. *Reinforcement Learning: An Introduction*, 2nd Edition. MIT Press, Cambridge, MA.
URL: <https://mitpress.mit.edu/9780262039246/reinforcement-learning/>