Estrategia de control local tensión-reactiva en sistemas de distribución a partir de bancos de capacitores y cambia tomas en transformadores utilizando aprendizaje por refuerzo

Aurelio Artur Álvarez Castillo y Julián David Campos Soto

Trabajo de grado para optar al titulo de ingeniero electricista

Director

César Antonio Duarte Gualdrón

Doctor en ingenieria electrica

Codirector

Alan Ferney Lizarazo Maldonado

Ingeniero electricista

Universidad Industrial de Santander

Facultad de Ingenierías Fisico mecanicas

Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones

Programa de ingenieria electrica

Bucaramanga

2026

Tabla de contenido

Lista de apéndices

Resumen

Titulo: Estrategia de control local tensión-reactiva en sistemas de distribución a partir de bancos de capacitores y cambia tomas en transformadores utilizando aprendizaje por refuerzo

Autores: Aurelio Artur Álvarez Castillo y Julián David Campos Soto

Palabras clave: control tensión-reactiva, aprendizaje profundo por refuerzo, Soft Actor-Critic, sistemas multiagente, redes de distribución, bancos de capacitores, OLTC.

Descripción

El control tensión-reactiva (Volt/Var Control, VVC) es una función esencial para mantener la calidad del servicio en redes de distribución, especialmente ante la creciente variabilidad introducida por cambios en la demanda y la incorporación de recursos energéticos distribuidos. En este trabajo se evalúa una estrategia de control local basada en aprendizaje profundo por refuerzo para coordinar dispositivos discretos convencionales en una red de distribución. La propuesta utiliza un esquema *Multi-Agent Soft Actor-Critic* (MSAC) con entrenamiento centralizado y ejecución descentralizada, de manera que cada dispositivo decide con información local durante la operación, mientras que la coordinación entre agentes se favorece durante el aprendizaje. La metodología se implementa sobre el alimentador IEEE de 33 nodos en *pandapower*, con un transformador OLTC y siete bancos de capacitores.El desempeño del método se compara con un esquema centralizado basado en *Discrete Coordinate Descent* (DCD) y se evalúa tanto en un escenario nominal como en escenarios estocásticos generados con distribución de Weibull. Frente al método DCD, el MSAC ofrece un mejor control del perfil de tensión, mientras que el DCD presenta mayor eficiencia en la reducción de la potencia aparente global.

Trabajo de grado facultad de ingenierías fisicomécanicas. Escuela de ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Director César Antonio Duarte Gualdrón. Codirector Alan Ferney Lizarazo Maldonado

Abstract

Title: Local Volt/Var control strategy in distribution systems based on capacitor banks and transformer tap changers using reinforcement learning

Authors: Aurelio Artur Álvarez Castillo and Julián David Campos Soto

Keywords: Volt/Var control, deep reinforcement learning, Soft Actor-Critic, multi-agent systems, distribution networks, capacitor banks, OLTC.

Description

Volt/Var Control (VVC) is an essential function for maintaining service quality in distribution networks, especially given the increasing variability introduced by changes in demand and the incorporation of distributed energy resources. In this work, a local control strategy based on deep reinforcement learning is evaluated to coordinate conventional discrete devices in a distribution network. The proposal uses a *Multi-Agent Soft Actor-Critic* (MSAC) scheme with centralized training and decentralized execution, so that each device decides with local information during operation, while coordination among agents is favored during learning. The methodology is implemented on the IEEE 33-bus feeder in *pandapower*, with one OLTC transformer and seven capacitor banks. The method's performance is compared with a centralized scheme based on *Discrete Coordinate Descent* (DCD) and is evaluated in both a nominal scenario and stochastic scenarios generated with a Weibull distribution. Compared to the DCD method, MSAC offers better voltage profile control, while DCD presents higher efficiency in reducing the global apparent power.

Degree project, Faculty of Physico-Mechanical Engineering. School of Electrical, Electronic and Telecommunications Engineering. Director César Antonio Duarte Gualdrón. Co-director Alan Ferney Lizarazo Maldonado

Introducción

El control tensión-reactiva (Volt/Var Control, VVC) y la optimización tensión-reactiva (Optimización Volt/Var, VVO) constituyen una función operativa esencial en redes de distribución, ya que contribuye a mantener los perfiles de tensión dentro de rangos admisibles y a coordinar el uso de los recursos de compensación reactiva disponibles en el sistema. Este problema adquiere mayor relevancia cuando la red opera bajo condiciones variables, asociadas tanto a cambios en la demanda como a la incorporación de recursos energéticos distribuidos, los cuales pueden modificar de forma significativa los flujos de potencia y el comportamiento de tensión a lo largo del alimentador (Allahmoradi et al., 2024; Wu et al., 2022). En este contexto, el VVC sigue siendo un problema de interés práctico y académico, especialmente en redes de distribución donde coexisten restricciones operativas, actuadores discretos y condiciones de operación cambiantes.

El problema de control tensión-reactiva según la literatura se ha abordado mediante esquemas centralizados, distribuidos, descentralizados y jerárquicos. Un enfoque centralizado puede ofrecer una visión mas global del problema, pero también requiere suficiente infraestructura de medición y comunicación para operación. Recurrir a un enfoque distribuido o jerárquico esta orientado a mejorar la escalabilidad o a reducir la dependencia de comunicación a costa de una visión global del sistema (Hai et al., 2022; Zhang et al., 2024). Dentro de este panorama, resulta de interés estudiar esquemas con ejecución local, en los que cada dispositivo actúa con información propia durante la operación, pero donde la coordinación entre decisiones sigue siendo un desafío técnico relevante.

En los últimos años, los enfoques basados en datos han ganado atención en aplicaciones de VVC y VVO, particularmente aquellos apoyados en aprendizaje profundo por refuerzo. Este interés se explica por la capacidad del Deep Reinforcement Learning (DRL) para tratar problemas de decisión secuencial en entornos no lineales y variables, sin requerir que el problema de control se resuelva completamente en línea mediante un optimizador

convencional (Allahmoradi et al., 2024; Zhang et al., 2024). En particular, se han reportado resultados promisorios tanto para formulaciones centralizadas con dispositivos discretos como para esquemas multiagente con entrenamiento centralizado y ejecución descentralizada (Cao et al., 2021; S. Wang et al., 2020; W. Wang et al., 2019).

La estrategia *centralized training and decentralized execution* (CTDE) es utilizada en el aprendizaje por refuerzo multi agente, donde busca coordinar las acciones de todos los agentes dividiendo el proceso en dos etapas. Durante el entrenamiento (Centralizado) los criticos del algoritmo como el Soft Actor-Critic (SAC) pueden observar el estado de toda la red, las acciones de todos los agentes y la recompensa global, esto permite que el modelo aprenda cómo las acciones de un agente afectan o ayudan a los demás. Durante la ejecución (Descentralizada), una vez que los agentes han aprendido, ya no necesitan ver toda la información de la red ni comunicarse entre ellos, el actor de cada agente solo recibe su observación local.

En este trabajo se evalúa un esquema multiagente de control tensión-reactiva basado en SAC bajo la filosofía de CTDE, aplicado a dispositivos discretos convencionales: un transformador con cambia tomas bajo carga (On-Load Tap Changer,OLTC) y bancos de capacitores. La implementación se realiza sobre el alimentador IEEE de 33 nodos en `pandapower`, y su desempeño se compara con un método centralizado de referencia basado en *Discrete Coordinate Descent* (DCD). La evaluación se desarrolla en un escenario nominal y en escenarios estocásticos construidos con distribución de Weibull, con el fin de analizar el alcance y las limitaciones de una estrategia con ejecución local y coordinación inducida durante el entrenamiento. De esta manera, el trabajo busca aportar una evaluación técnica del comportamiento de este tipo de esquema en un entorno de prueba de interés para estudios de pregrado en control de tensión en redes de distribución.

Objetivos

Objetivo General

Evaluar una estrategia de control local tensión-reactiva en un sistema de distribución, a partir de banco de capacitores y cambia tomas en transformadores, utilizando DRL.

Objetivos específicos

Identificar las estrategias de control tensión-reactiva en sistemas de distribución partiendo de la búsqueda de artículos científicos que involucren el uso de aprendizaje profundo por refuerzo. Analizar los criterios y restricciones que condicionan el sistema de recompensa en un modelo de control basado en máquina de aprendizaje profundo por refuerzo. Realizar la simulación del control local de tensión reactiva utilizando aprendizaje profundo por refuerzo en un modelo de prueba IEEE. Evaluar el desempeño del método implementado con respecto a una estrategia identificada en la literatura revisada y en trabajos de pregrado realizados en el grupo de investigación.

1. Marco teórico y revisión de literatura

Este capítulo establece los fundamentos conceptuales necesarios para comprender el problema de VVC y la VVO, así como las diversas estrategias, dispositivos y enfoques de modelado que se han empleado para su solución. Se realizará una revisión de la literatura reciente, prestando especial atención a los métodos basados en aprendizaje por refuerzo profundo (DRL) y a las arquitecturas multiagente, con el fin de identificar las tendencias actuales y delimitar la investigación que este trabajo busca abordar.

1.1 Fundamentos del control Volt/Var en redes de distribución

1.1.1 Distinción entre VVC y VVO

El VVC y la VVO tienen fines relacionados, pero no equivalentes. En términos generales, el VVC agrupa las acciones orientadas a mantener los niveles de tensión dentro de rangos operativos aceptables mediante la coordinación de dispositivos que inyectan o absorben potencia reactiva o modifican la relación de transformación. La VVO, en cambio, incorpora además un criterio explícito de optimización sobre una o varias funciones objetivo, por ejemplo pérdidas, perfil de tensión o costo operativo, lo que normalmente exige una mayor disponibilidad de información del sistema y de capacidades de medición y comunicación (Allahmoradi et al., 2024).

1.1.2 Objetivos de control en VVC y VVO

Las funciones objetivo más frecuentes en VVC y VVO pueden resumirse en cuatro grupos:

- **Conservación de energía y reducción de demanda**: la reducción conservativa

de tensión (CVR) busca operar lo más cerca posible del límite inferior permitido sin comprometer la calidad del servicio, con el fin de reducir consumo y demanda pico cuando la naturaleza de la carga lo permite.

- **Reducción de pérdidas**: en redes de distribución, donde la relación R/X es alta y las variaciones de carga son significativas, la reducción de pérdidas activas constituye un objetivo habitual del control Volt/Var.
- **Regulación del perfil de tensión**: uno de los propósitos centrales del VVC es mantener la magnitud de tensión de los usuarios dentro de los límites operativos permitidos a lo largo del alimentador y a través del tiempo.
- **Gestión de potencia reactiva**: la inyección o absorción de potencia reactiva se utiliza como medio para sostener el perfil de tensión y coordinar el funcionamiento de los dispositivos disponibles.

(IEEE, 2022)

1.1.3 Estrategias de control

Desde la perspectiva de la coordinación y del intercambio de información, en la literatura se distinguen principalmente estrategias **centralizadas**, **distribuidas o descentralizadas** y **jerárquicas** (Hai et al., 2022). En este trabajo, además, se utiliza el término **control local** para referirse a una ejecución en la que cada dispositivo decide a partir de mediciones propias, sin depender de comunicación en línea con otros agentes.

- **Control centralizado**: un controlador único recopila información del sistema completo y calcula acciones coordinadas para todos los dispositivos. Este enfoque puede aproximarse a un óptimo global, pero su desempeño depende de la disponibilidad de una infraestructura robusta de medición, comunicación y procesamiento de datos (Hai et al., 2022).

- **Control distribuido o descentralizado**: el problema se reparte entre varios controladores o agentes, usualmente asociados a zonas del sistema o a conjuntos de dispositivos. Este enfoque reduce la carga computacional por unidad y mejora la escalabilidad, aunque puede sacrificar parte de la optimalidad global si la coordinación entre agentes es limitada (Hai et al., 2022).

- **Control jerárquico**: organiza la toma de decisiones en varios niveles. Los niveles superiores fijan consignas o referencias y los niveles inferiores ejecutan acciones locales o de respuesta rápida. Este tipo de arquitectura resulta útil cuando coexisten dispositivos con distintas escalas temporales de operación (Hai et al., 2022).

Para el problema abordado en este trabajo, el interés se centra en una estrategia de control local, entrenada con apoyo de información centralizada. Esta filosofía busca reducir la dependencia de una comunicación permanente durante la operación, sin renunciar por completo a la coordinación entre dispositivos en la fase de aprendizaje.

1.1.4 Dispositivos de control

Los dispositivos utilizados en VVC y VVO pueden clasificarse, de forma general, según su escala de tiempo de actuación.

- **Dispositivos de respuesta lenta**: incluyen transformadores con cambia tomas bajo carga (OLTC), bancos de capacitores (CBs) y reguladores de voltaje (VR). Son apropiados para variaciones relativamente lentas de la demanda y del perfil de tensión. Su operación es discreta y depende de componentes electromecánicos, por lo que una conmutación excesiva puede incrementar el desgaste y el costo operativo (Hai et al., 2022).

- **Dispositivos de respuesta rápida**: incluyen inversores inteligentes (SIs), compensadores estáticos de potencia reactiva (STATCOM) y compensadores estáticos de

tensión (SVC). Debido a su naturaleza electrónica y continua, ofrecen una respuesta más rápida frente a fluctuaciones abruptas y son especialmente útiles en presencia de generación distribuida variable (Hai et al., 2022).

La elección de dispositivos condiciona tanto la formulación del problema como la naturaleza del espacio de acciones. En este trabajo se seleccionan OLTC y CBs porque siguen siendo dispositivos ampliamente desplegados en redes de distribución y, además, constituyen un caso de interés al tratarse de actuadores discretos con costos operativos asociados a la conmutación.

1.2 Enfoques de modelado y aprendizaje aplicados a VVC

1.2.1 Control basado en modelos y control basado en datos

Las soluciones para VVC y VVO pueden agruparse en dos grandes enfoques: **basado en modelos** y **basado en datos**. El primero se apoya en una representación explícita del sistema eléctrico, sus restricciones y sus variables de decisión. El segundo aprende relaciones útiles de control a partir de mediciones históricas, simulaciones o interacción con el entorno (Allahmoradi et al., 2024).

El control basado en modelos puede ofrecer interpretabilidad física y un manejo riguroso de restricciones, pero su aplicación en redes de distribución extensas o cambiantes exige un conocimiento suficientemente preciso de parámetros, topología, estados operativos y disponibilidad de mediciones. Por ello, la complejidad del problema aumenta a medida que crece el tamaño del sistema o la incertidumbre operativa.

El control basado en datos, por su parte, reduce la dependencia de un modelo detallado y puede adaptarse mejor a patrones complejos de operación. Sin embargo, su desempeño depende de la calidad de los datos, del diseño del proceso de entrenamiento y de la capacidad del método para generalizar a escenarios no vistos. Por esta razón, la elec-

ción entre ambos enfoques no es excluyente, sino que responde al nivel de información disponible, al horizonte temporal de control y a los objetivos perseguidos.

La Tabla 1 resume las diferencias fundamentales entre ambos enfoques, destacando las ventajas que el aprendizaje por refuerzo aporta en términos de velocidad de respuesta y adaptabilidad.

Tabla 1

Diferencias entre el enfoque basado en modelos y el basado en datos (DRL)

Criterio	**VVO basado en modelos**	**VVO basado en datos (DRL)**
Conocimiento del sistema	Requiere topología y parámetros físicos precisos.	Puede operar como *model-free* mediante interacción con el sistema.
Resolución del problema	Optimización matemática iterativa en línea.	Ejecución directa de la política.
Escalabilidad	Se dificulta ante alta complejidad del sistema.	Alta, especialmente con arquitecturas multiagente.
Adaptabilidad	Limitada ante cambios no modelados de la red.	Capacidad de generalizar patrones ante variabilidad.
Requerimientos de datos	Mediciones de estado completas en tiempo real.	Datos históricos o simulados para entrenamiento.

Nota. Adaptado de (Allahmoradi et al., 2024).

1.2.2 Aprendizaje por refuerzo y DRL aplicado a VVC

En VVC y VVO, el DRL resulta atractivo porque permite aprender políticas de control a partir de la interacción con un modelo simulado del sistema, sin requerir necesariamente una formulación cerrada y exacta del problema de optimización en línea. Su

utilidad aumenta cuando el sistema presenta variabilidad, incertidumbre o una elevada dimensión del espacio de estados.(Zhang et al., 2024) No obstante, su desempeño depende de decisiones de diseño críticas, como la definición del estado, el espacio de acciones, la función de recompensa, la estrategia de exploración y la representatividad de los escenarios de entrenamiento.(Lapan, 2020)

1.3 Revisión de literatura

1.3.1 Alcance y criterios de selección

Con el fin de identificar antecedentes directamente relacionados con el objetivo del trabajo, se realizó una búsqueda bibliográfica de carácter exploratorio en bases y portales de uso frecuente en ingeniería eléctrica, entre ellos IEEE Xplore, ScienceDirect, Scopus y Google Scholar. La búsqueda se enfocó principalmente en publicaciones recientes y utilizó combinaciones de términos como *"Volt/Var control"*, *"distribution system"*, *"distribution network"*, *"deep reinforcement learning"* y *"DRL"*.

En una primera etapa se preseleccionaron cincuenta documentos. Posteriormente se priorizaron aquellos trabajos que describían de manera suficientemente clara el problema de control, los dispositivos considerados, la estrategia de coordinación, el algoritmo empleado y el sistema de prueba utilizado. Aunque esta revisión no pretende ser una revisión sistemática exhaustiva, sí busca construir una base argumentativa suficiente para ubicar la propuesta del presente trabajo dentro del estado del arte más cercano.

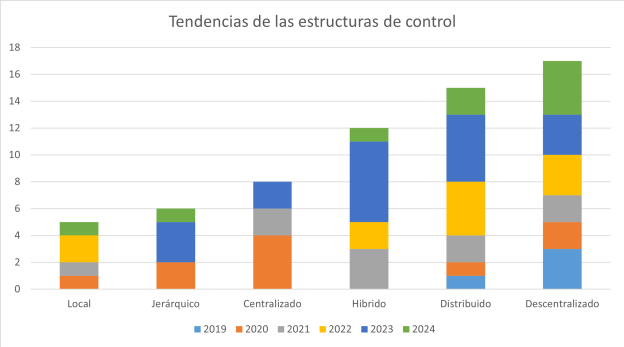
1.3.2 Tendencias observadas en la literatura

La revisión muestra que las estrategias más reportadas en VVC y VVO corresponden a esquemas centralizados, hibridos, distribuidos y descentralizados con un creciente

interés en enfoques multiagente que emplean entrenamiento centralizado y ejecución descentralizada (CTDE), particularmente cuando se busca coordinar un número elevado de dispositivos o integrar generación distribuida.(Hein et al., 2026)

Figura 1

Tendencias observadas en la literatura



Nota. Elaboración propia.

En cuanto a los dispositivos, los trabajos recientes combinan cada vez con mayor frecuencia actuadores de respuesta lenta, como OLTC y CBs, con recursos de respuesta rápida como inversores inteligentes, SVC o STATCOM. Esto permite desacoplar acciones en diferentes escalas temporales: los dispositivos discretos se reservan para ajustes más estructurales del perfil de tensión, mientras que los recursos electrónicos compensan fluctuaciones rápidas. Aun así, los dispositivos convencionales continúan siendo relevantes porque están ampliamente instalados en sistemas reales y representan un caso operativo de interés práctico.

1.3.3 Trabajos representativos

Entre los trabajos más cercanos al presente estudio se encuentran los siguientes:

- W. Wang et al. (2019) proponen un esquema de VVC libre de modelo formulado como un proceso de decisión de Márkov con restricciones. El problema se resuelve

mediante CSAC con espacio de acción discreto y dispositivos como OLTC, CBs y reguladores de voltaje. El objetivo combina restricciones de tensión, reducción de pérdidas y costos de operación.

- S. Wang et al. (2020) presentan un esquema multiagente con entrenamiento centralizado y ejecución descentralizada. La red se divide en zonas y cada agente actúa a partir de información local durante la ejecución, mientras que el entrenamiento aprovecha información global para actualizar la red crítica. Esta filosofía es particularmente relevante para el enfoque adoptado en este trabajo.

- Sun y Qiu (2021) desarrollan una estrategia de dos etapas para coordinar dispositivos lentos y rápidos. Los OLTC y CBs operan en una capa lenta, mientras que los recursos fotovoltaicos con capacidad de control reactivo atienden fluctuaciones rápidas de tensión. El trabajo destaca la conveniencia de separar acciones según la escala temporal del dispositivo.

- Cao et al. (2021) proponen un enfoque multiagente, libre de modelo y basado en datos, con entrenamiento centralizado y ejecución descentralizada para redes con alta penetración de generación distribuida. Su principal aporte es mostrar que la coordinación multiagente puede sostener perfiles de tensión aceptables sin depender de un modelo exacto de la red.

En conjunto, estos trabajos evidencian tres tendencias relevantes: la creciente adopción de DRL en problemas Volt/Var, el uso extendido de arquitecturas multiagente con información global en entrenamiento y local en ejecución, y la preferencia por combinar dispositivos de distinta escala temporal cuando el sistema incluye generación distribuida variable.

1.4 Brecha de investigación y justificación del trabajo

La revisión realizada permite identificar que el uso de DRL en VVC ha crecido de forma sostenida, especialmente en arquitecturas centralizadas o distribuidas y en problemas donde participan recursos de respuesta rápida. Sin embargo, siguen siendo menos frecuentes los estudios que examinan estrategias de **control local** apoyadas en aprendizaje por refuerzo para dispositivos discretos convencionales como OLTC y bancos de capacitores, particularmente cuando se busca mantener la operación sin depender de una coordinación en línea entre dispositivos.

Esta observación justifica el enfoque del presente trabajo. Dada la importancia de las estrategias de control local y la necesidad de entender su desempeño en entornos reales y su comparativa con métodos consolidados, esta investigación se centra en evaluar un esquema multiagente basado en la filosofía CTDE para OLTC y bancos de capacitores.

De esta manera, el capítulo metodológico se orienta a formular e implementar un esquema multiagente aplicable a OLTC y CBs. El objetivo es evaluar hasta qué punto una estrategia de este tipo puede mitigar violaciones de tensión y mejorar el comportamiento operativo del sistema, contrastando su rendimiento con una alternativa de referencia ya estudiada en un trabajo anterior sobre el método DCD (Cristancho Castro, 2023). Esta comparación permitirá cuantificar las ventajas y limitaciones del enfoque de aprendizaje por refuerzo en un contexto de control local.

2. Fundamentos del DRL

El aprendizaje por refuerzo profundo (*Deep Reinforcement Learning*, DRL) constituye un marco adecuado para problemas de control secuencial en los que un agente debe seleccionar acciones a partir del estado del entorno y mejorar su comportamiento con base en una señal de desempeño. En el contexto del control Volt/Var, este enfoque resulta

de interés porque permite aprender políticas de decisión a partir de la interacción con un simulador del sistema, sin requerir que el problema operativo se resuelva en línea mediante una optimización completa en cada instante. Sin embargo, su desempeño depende de manera crítica del algoritmo seleccionado y de la forma en que se construyen el estado, las acciones y la recompensa.

El propósito de este capítulo es presentar los elementos conceptuales para justificar las decisiones metodológicas adoptadas. En particular, se sigue el planteamiento de W. Wang et al. (2019) para el control de OLTC y CBs en espacios discretos, y se toma como base la estructura multiagente bajo la filosofía CTDE propuesta por Cao et al. (2021).

2.1 DRL aplicado a problemas de control

En un problema de DRL, el agente observa el entorno, ejecuta una acción y recibe una recompensa que orienta el aprendizaje de una política. Cuando esta política se aproxima con redes neuronales profundas, el método puede manejar relaciones no lineales y espacios de estado de alta dimensión.(Zhang et al., 2024)

Para el caso de VVC, el interés del DRL radica en que el controlador puede aprender a coordinar dispositivos a partir de trayectorias de operación simuladas, sin depender de una formulación cerrada que deba resolverse en línea en cada instante. No obstante, esta ventaja no elimina la necesidad de incorporar conocimiento del problema: la calidad de la política aprendida sigue dependiendo de la representatividad de los escenarios de entrenamiento y del diseño de los componentes de la formulación. En consecuencia, el DRL no reemplaza la comprensión física del sistema, sino que ofrece una forma distinta de construir la ley de control.

2.2 Soft Actor-Critic

Dentro de los algoritmos *actor-critic*, SAC resulta especialmente atractivo por tres razones. En primer lugar, es un método *off-policy*, lo que permite reutilizar experiencias almacenadas previamente y mejora la eficiencia muestral del entrenamiento. En segundo lugar, incorpora un criterio de máxima entropía que favorece una exploración más estable y reduce la tendencia a converger prematuramente a políticas pobres. En tercer lugar, utiliza dos críticos y redes objetivo, lo que ayuda a mitigar problemas de sobreestimación y mejora la estabilidad del aprendizaje (Haarnoja et al., 2018).

La idea central de SAC es maximizar simultáneamente la recompensa esperada y la entropía de la política. De forma compacta, su objetivo puede expresarse como

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}\left[r(s_t, a_t) + \alpha\, \mathcal{H}\left(\pi(\cdot \mid s_t)\right)\right], \tag{1}$$

donde $\mathbb{E}$ denota el promedio ponderado de los resultados posibles, T representa la duración del episodio, $r(s_t, a_t)$ representa la recompensa asociada a la acción ejecutada (a_t) en el estado actual (s_t), $\mathcal{H}(\pi(\cdot \mid s_t))$ es la entropía de la política y α controla el balance entre explotación y exploración. En términos prácticos, este criterio incentiva políticas que no solo obtengan buen desempeño, sino que también mantengan suficiente diversidad de acciones durante el aprendizaje.

Para el presente trabajo, los elementos de SAC que resultan realmente relevantes son los siguientes:

- **Política estocástica**: permite representar decisiones probabilísticas sobre el espacio de acciones y favorece una exploración más robusta durante el entrenamiento.
- **Entrenamiento *off-policy***: posibilita reutilizar trayectorias almacenadas en el *replay buffer*, lo que mejora la eficiencia del aprendizaje.

- **Críticos gemelos y redes objetivo**: contribuyen a estabilizar la estimación del valor y a reducir sesgos por sobreestimación.
- **Regularización por entropía**: introduce una exploración controlada, especialmente útil en problemas con múltiples decisiones localmente razonables.

Aunque SAC fue introducido originalmente para espacios de acción continuos (Haarnoja et al., 2018), su lógica puede adaptarse a problemas discretos mediante políticas categóricas y salidas tipo *softmax*, como se ha mostrado en variantes aplicadas a VVC con dispositivos discretos (W. Wang et al., 2019). Esta adaptación es particularmente pertinente en este trabajo, ya que tanto el OLTC como los bancos de capacitores operan mediante cambios discretos de posición o etapa.

2.3 Extensión a esquema multiagente con CTDE

La formulación adoptada en este manuscrito no corresponde a un controlador único, sino a un esquema en el que varios dispositivos deben aprender políticas coordinadas. En este contexto, la idea de *centralized training and decentralized execution* (CTDE) resulta especialmente útil. Bajo esta filosofía, el entrenamiento aprovecha información más amplia del sistema para facilitar la coordinación entre agentes, mientras que durante la operación cada agente actúa únicamente con la información disponible en su entorno local.

Esta separación responde de forma natural al problema estudiado. Durante la fase de aprendizaje, disponer de información global o ampliada permite que los críticos evalúen mejor el efecto conjunto de las acciones de múltiples dispositivos sobre el perfil de tensión del sistema. En cambio, durante la ejecución se busca preservar una lógica de control local, reduciendo la dependencia de comunicación continua y acercando el esquema a una implementación más realista para redes de distribución.

Desde el punto de vista del algoritmo, la transición desde SAC hacia un esquema multiagente no exige redefinir por completo sus principios, sino reorganizar cómo se dis-

tribuye la información entre actores y críticos. Los actores se especializan en decisiones locales asociadas a cada dispositivo, mientras que los críticos incorporan información global para orientar el aprendizaje conjunto. Esta idea ha sido utilizada en trabajos recientes de control Volt/Var con múltiples agentes (Cao et al., 2021; S. Wang et al., 2020) y constituye la base conceptual del enfoque seguido aquí.

2.4 Relación con la formulación adoptada en este trabajo

Con base en lo anterior, el presente trabajo utiliza SAC no como un fin en sí mismo, sino como fundamento de una formulación multiagente orientada al control de dispositivos discretos convencionales. En términos concretos, los aspectos del algoritmo que se emplean de manera directa en la metodología son:

- un criterio de aprendizaje basado en **recompensa más entropía**, que favorece estabilidad y exploración durante el entrenamiento.
- un ***experience replay buffer*** para almacenar transiciones y entrenar de forma *off-policy*;
- una **política estocástica discreta** para seleccionar acciones sobre OLTC y CBs;
- **críticos con información más amplia que la disponible para cada actor**, coherentes con la filosofía CTDE;

3. Metodología y configuración experimental

En este capítulo se presenta la formulación matemática del problema de control, la arquitectura del esquema multiagente propuesto y la configuración experimental empleada para su entrenamiento y evaluación. En coherencia con el capítulo anterior, aquí se aterrizan los elementos de la arquitectura multiagente basada en SAC bajo la filosofía

CTDE, especificando el sistema de prueba, la formulación del problema como proceso de decisión de Markov descentralizado, el estado y las acciones de cada agente, la función de recompensa, las ecuaciones de actualización del algoritmo, la arquitectura neuronal, la configuración del entrenamiento y los escenarios de evaluación considerados.

3.1 Especificaciones y criterios de diseño de la solución

Siguiendo los lineamientos de una experiencia de diseño en ingeniería, la solución propuesta se fundamenta en la satisfacción de requerimientos técnicos y regulatorios específicos, derivados de la problemática de la regulación de tensión en redes activas.

3.1.1 Requerimientos y restricciones de diseño

Para que el sistema de control se considere viable, el diseño debe satisfacer los siguientes requerimientos fundamentales:

- **Regulación de tensión (R1):** Mantener la magnitud de tensión en todos los nodos del sistema dentro del rango de $\pm 5\,\%$ del valor nominal (0.95 a 1.05 p.u.).
- **Operación con dispositivos discretos (R2):** La estrategia debe ser capaz de gestionar actuadores mecánicos (OLTC y CBs) que poseen estados de operación discretos y finitos.
- **Descentralización de la ejecución (R3):** El diseño final debe permitir que cada agente tome decisiones autónomas basadas únicamente en mediciones locales, eliminando la dependencia de una infraestructura de comunicación de baja latencia en tiempo real.
- **Comparabilidad y validación (R4):** La elección del sistema de prueba (IEEE 33 nodos), junto con la distribución y parámetros técnicos de los equipos de control,

responde a la necesidad de garantizar un escenario de comparación directa y válida con el método de referencia DCD reportado en (Cristancho Castro, 2023).

3.1.2 Marco regulatorio

El diseño atiende a los estándares internacionales de calidad de la potencia, específicamente la norma **ANSI C84.1**, que define los rangos de utilización de voltaje en sistemas eléctricos. Asimismo, se consideran los lineamientos de la normativa local colombiana (**Resolución CREG 015 de 2018**), que establece las penalizaciones y límites operativos para la inyección de energía reactiva y la calidad del perfil de tensión en sistemas de distribución.

3.1.3 Estudio de alternativas de solución

Antes de proceder con el diseño final basado en MSAC, se evaluaron tres alternativas tecnológicas principales:

- **Deep Q-Network (DQN):** Un algoritmo fundamental de DRL que utiliza redes neuronales para aproximar la función de valor Q. Aunque efectivo para problemas con espacios de acción discretos, su formulación es inherentemente para un solo agente y puede sufrir de sobreestimación de valores, lo que limita su aplicación directa a problemas multiagente coordinados.
- **Double Deep Q-Network (2DQN):** Una mejora sobre DQN que aborda el problema de la sobreestimación de valores mediante el uso de dos redes Q separadas. Si bien mejora la estabilidad del aprendizaje, mantiene una arquitectura de agente único y no está diseñado para la coordinación intrínseca de múltiples actuadores en un sistema complejo.

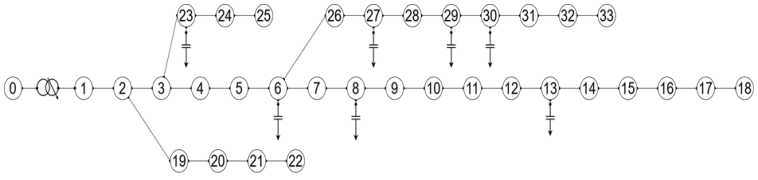
- **Aprendizaje por Refuerzo Multi-Agente (MSAC - Seleccionada):** Esta alternativa combina la capacidad de optimización de los métodos basados en datos con la robustez de la ejecución local. Su filosofía de Entrenamiento Centralizado y Ejecución Descentralizada (CTDE) permite aprender políticas coordinadas para múltiples agentes, lo que lo hace idóneo para cumplir con el requerimiento R3 sin sacrificar significativamente el desempeño técnico frente a alternativas centralizadas, superando las limitaciones de los enfoques de agente único como DQN y 2DQN.

3.2 Sistema de prueba, dispositivos y agentes de control

El sistema de prueba adoptado corresponde al alimentador IEEE de 33 nodos, implementado en la librería `pandapower` para `Python`. La elección de este sistema se justifica por la necesidad de garantizar un escenario de comparación directa y válida con el método de referencia DCD reportado en (Cristancho Castro, 2023); la ubicación y los parámetros de los dispositivos controlados se toman íntegramente de dicho trabajo para asegurar la consistencia de los resultados comparativos. La topología y los datos de carga corresponden a la formulación clásica de Baran y Wu (1989), y la topología general empleada se muestra en la Figura 2.

Figura 2

Topología de la red de pruebas



Nota. Elaboración propia.

El esquema de control considera dos tipos de actuadores discretos: un transformador con cambia tomas bajo carga (OLTC), ubicado entre los nodos 0 y 1, y siete bancos de

capacitores (CBs) distribuidos a lo largo del alimentador. En consecuencia, la formulación multiagente se construye con un total de ocho agentes, uno asociado a cada dispositivo de control. Esta decisión permite mantener una lógica de ejecución local por dispositivo, sin recurrir a un controlador único durante la operación.

Los parámetros del transformador se presentan en la Tabla 2. Se dispone de 33 posiciones de tap y la relación 1:1 corresponde al tap 16.

Tabla 2

Parámetros del transformador

Característica	Valor
Fases	3
Devanados	2
Reactancia de dispersión	8 %
Voltaje nominal	110/12.66 kV
Potencia	5500 kVA
Conexión	YDyn1

Nota. Elaboración propia.

La Tabla 3 muestra la ubicación y capacidad nominal de los bancos de capacitores. Los bancos de 150 kVAr operan con etapas discretas en el conjunto $\{0, 1, 2, 3\}$, mientras que los bancos de 300 kVAr lo hacen en $\{0, 1, 2, 3, 4, 5, 6\}$. En ambos casos, cada incremento de etapa equivale a 50 kVAr y el estado cero representa el banco desconectado.

La Figura 3 define el perfil temporal utilizado para escalar la demanda base del sistema a lo largo de un día de operación. El valor máximo se presenta a la hora 20, lo que introduce un escenario exigente para la regulación de tensión en los tramos más alejados del alimentador. Cada paso de simulación equivale a una hora; por tanto, un episodio completo representa $T = 24$ pasos horarios de operación.

La Tabla 4 presenta la demanda base asignada a cada barra a partir del caso de Baran y Wu (1989).

Tabla 3

Ubicación y potencia de los bancos de capacitores

Nodo	Potencia [kVAr]
6	150
8	150
13	150
23	300
27	150
29	300
30	150

Nota. Elaboración propia.

Figura 3

Perfil de demanda horaria del sistema de distribución (Cristancho Castro, 2023)

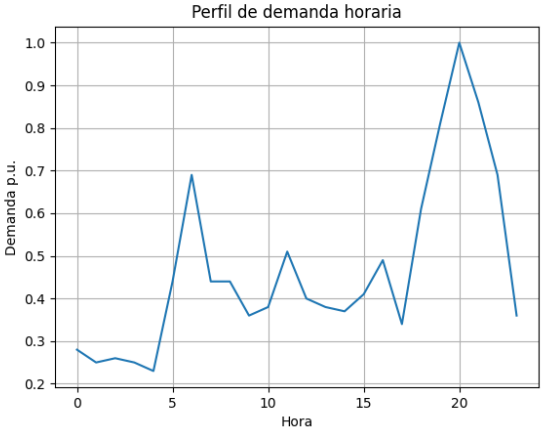


Tabla 4

Demanda por nodos

Nodo	P [kW]	Q [kVAr]	Nodo	P [kW]	Q [kVAr]
1	100	60	17	90	40
2	90	40	18	90	40
3	120	80	19	90	40
4	60	30	20	90	40
5	60	20	21	90	40
6	200	100	22	90	50
7	200	100	23	420	200
8	60	20	24	420	200
9	60	20	25	60	25
10	45	30	26	60	20
11	60	35	27	60	20
12	60	35	28	120	70
13	120	80	29	200	600
14	60	10	30	150	70
15	60	20	31	210	100
16	60	20	32	60	40

Nota. Elaboración propia.

3.3 Formulación del problema como proceso de decisión de Markov

3.3.1 Definición del MDP

El problema de control se formula como un proceso de decisión de Markov (*MDP*) (**oliehoek2016concise**), definido por la tupla

$$\langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \mathcal{T}, \mathcal{R}, \{\Omega_i\}_{i\in\mathcal{N}}, \{\mathcal{O}_i\}_{i\in\mathcal{N}}, \gamma\rangle, \tag{2}$$

donde:

- $\mathcal{N} = \{1, \ldots, 8\}$ es el conjunto de agentes, con $i = 1$ para el OLTC e $i = 2, \ldots, 8$ para los siete bancos de capacitores.
- $\mathcal{S}$ es el espacio de estados globales del sistema, cuyo elemento genérico es el vector de

tensiones, potencias activas, potencias reactivas y posiciones de todos los dispositivos en un instante t.

- $\mathcal{A}_i$ es el espacio de acciones discretas del agente i.
- $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ es la función de transición de estado, determinada por el flujo de potencia AC del sistema.
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ es la función de recompensa compartida.
- Ω_i es el espacio de observaciones locales del agente i, y $\mathcal{O}_i : \mathcal{S} \to \Omega_i$ es la función de observación que mapea el estado global a la observación local del agente i.
- $\gamma \in (0, 1)$ es el factor de descuento intertemporal.

El objetivo de cada agente es encontrar una política local $\pi_{\theta_i} : \Omega_i \to \Delta(\mathcal{A}_i)$ que maximice el retorno esperado acumulado

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t R_t \right], \tag{3}$$

donde $\Delta(\mathcal{A}_i)$ denota el simplex de probabilidad sobre $\mathcal{A}_i$ y R_t es la recompensa global obtenida en el paso t.

3.3.2 Función de transición de estado y modelo de la red

El entorno de simulación opera en régimen estacionario: dada la acción conjunta $a_t = (a_{1,t}, \ldots, a_{8,t})$ y el estado de la red en el paso t, el nuevo estado s_{t+1} se obtiene resolviendo el flujo de potencia AC. Este flujo se modela mediante las ecuaciones de balance de potencia nodal (Baran & Wu, 1989):

$$P_n = V_n \sum_{m \in \mathcal{N}_n} V_m \Big(G_{nm} \cos \theta_{nm} + B_{nm} \sin \theta_{nm} \Big), \tag{4}$$

$$Q_n = V_n \sum_{m \in \mathcal{N}_n} V_m \Big(G_{nm} \sin \theta_{nm} - B_{nm} \cos \theta_{nm} \Big), \tag{5}$$

donde V_n y V_m son las magnitudes de tensión en los nodos n y m, $\theta_{nm} = \theta_n - \theta_m$ es la diferencia de ángulos, y G_{nm} y B_{nm} son la conductancia y la susceptancia del elemento que conecta ambos nodos. La acción del OLTC modifica la relación de transformación τ del transformador, alterando los valores efectivos de G_{nm} y B_{nm} de la rama correspondiente. La acción de un CB modifica directamente la inyección de potencia reactiva en el nodo de conexión. Bajo la escala temporal horaria adoptada, estos cambios se aplican al inicio de cada paso y el flujo de potencia resultante determina el vector de tensiones $\mathbf{V}_{t+1}$, que forma la base del estado observado en el siguiente paso.

3.3.3 Estado local de los actores y estado de los críticos

Sea $i \in \mathcal{N}$ el índice del agente. El estado local observado por el actor del agente i en el instante t se define como

$$s_{i,t} = \Big[V_{i,t},\ P_{i,t},\ Q_{i,t},\ z_{i,t} \Big] \in \mathbb{R}^4, \tag{6}$$

donde $V_{i,t} \in \mathbb{R}^+$ es la magnitud de tensión en el nodo asociado al agente i, $P_{i,t}$ y $Q_{i,t}$ son las potencias activa y reactiva medidas en dicho nodo, y $z_{i,t} \in \mathbb{Z}$ es el estado del actuador local. Para el agente del OLTC, $z_{1,t} \in \{1, \ldots, 33\}$ corresponde a la posición del tap; para los agentes de CBs, $z_{i,t}$ corresponde a la etapa activa del banco. Con esta definición, el estado del actor es estrictamente local: el agente decide únicamente a partir de la información de su propio nodo y de su dispositivo, sin requerir comunicación con otros agentes durante la ejecución.

En contraste, el crítico de cada agente recibe información de alcance global durante el entrenamiento, coherente con la filosofía CTDE. Su entrada se construye concatenando los estados locales de todos los agentes y la acción conjunta:

$$x_t^{\text{critic}} = \Big[s_{1,t},\ s_{2,t},\ \ldots,\ s_{8,t},\ a_t\Big] \in \mathbb{R}^{40}, \tag{7}$$

donde la acción conjunta es

$$a_t = \Big[a_{1,t},\ a_{2,t},\ \ldots,\ a_{8,t}\Big] \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_8. \tag{8}$$

De esta manera, los críticos pueden valorar el efecto conjunto de las decisiones sobre el sistema completo, mientras los actores conservan una lógica de observación local.

3.3.4 Espacio de acciones discretas y remapeo de límites físicos

Las acciones son incrementales: representan el cambio aplicado al estado del actuador, no su posición absoluta. Para el OLTC:

$$\mathcal{A}^{\text{OLTC}} = \{-2, -1, 0, +1, +2\}, \tag{9}$$

y para cada banco de capacitores:

$$\mathcal{A}^{\text{CB}} = \{-1, 0, +1\}. \tag{10}$$

El OLTC admite desplazamientos de hasta dos posiciones por paso para acelerar la corrección ante cambios horarios pronunciados del perfil de carga. Los CBs adoptan una lógica más conservadora de una etapa por paso, lo que evita variaciones bruscas del soporte reactivo.

Cuando la acción propuesta conduce a un estado físicamente no admisible, el entorno aplica un remapeo al límite operativo más cercano:

$$z_{i,t+1} = \text{clip}\Big(z_{i,t} + \Delta a_{i,t},\ z_i^{\text{mín}},\ z_i^{\text{máx}}\Big), \tag{11}$$

donde $z_i^{\text{mín}}$ y $z_i^{\text{máx}}$ son los límites inferior y superior del dispositivo i, $\Delta a_{i,t} \in \mathcal{A}_i$ es la acción emitida por la política del agente, y la función $\text{clip}(x, l, u) = \text{máx}(l, \text{mín}(u, x))$ devuelve el valor más cercano dentro del rango $[l, u]$.

3.3.5 Función de recompensa

La recompensa se diseña para reflejar dos objetivos operativos: penalizar violaciones de tensión fuera de la banda admisible y desalentar un alto consumo de potencia aparente desde la barra *slack*. Para evitar ambigüedad con el factor de descuento del algoritmo, el peso asociado a la penalización por violaciones de tensión se denota por λ_v. El valor de este parámetro se adopta de la función objetivo propuesta en (Cristancho Castro, 2023), donde se utiliza una magnitud de 220000/16. Esta elección equilibra el orden de magnitud de la penalización con los términos de potencia del sistema, asegurando que la regulación de tensión sea el objetivo dominante de la política aprendida.

Se definen las magnitudes de subtensión y sobretensión en cada barra n:

$$U_{n,t} = \text{máx}\Big(V^{\text{mín}} - V_{n,t},\ 0\Big), \qquad V^{\text{mín}} = 0{,}95 \text{ p.u.}, \tag{12}$$

$$O_{n,t} = \text{máx}\Big(V_{n,t} - V^{\text{máx}},\ 0\Big), \qquad V^{\text{máx}} = 1{,}05 \text{ p.u.}, \tag{13}$$

donde los límites $V^{\text{mín}}$ y $V^{\text{máx}}$ corresponden a la banda operativa del sistema de prueba, coherente con los criterios de calidad de potencia aplicables a sistemas de distribución (W. Wang et al., 2019). La penalización por violaciones de tensión es

$$R_{v,t} = -\lambda_v \sum_{n=1}^{N} \Big(U_{n,t} + O_{n,t}\Big). \tag{14}$$

El segundo término penaliza la potencia aparente suministrada por la barra de alimentación:

$$S_t = -\sqrt{P_{\text{slack},t}^2 + Q_{\text{slack},t}^2}, \tag{15}$$

donde $P_{\text{slack},t}$ y $Q_{\text{slack},t}$ son la potencia activa y reactiva inyectadas desde la barra *slack*. La recompensa total en cada paso se define como

$$R_t = R_{v,t} + S_t. \tag{16}$$

No se introducen términos de penalización por conmutación en la función de recompensa; el costo operativo asociado al número de maniobras se analiza como métrica de desempeño en el capítulo de resultados.

3.4 Supuestos de entorno e implementación

El entorno opera en régimen estacionario, de modo que cada acción de control se evalúa mediante un flujo de potencia AC en el estado horario correspondiente. Bajo esta escala temporal, el OLTC y los CBs se modelan como actuadores de respuesta lenta, adecuados para ajustar el perfil de tensión ante cambios horarios de demanda.

Durante los primeros 100 episodios de entrenamiento se emplea una etapa de *warmup* en la que las acciones se generan aleatoriamente dentro del conjunto de acciones válidas de cada dispositivo, con el fin de poblar el *replay buffer* con trayectorias iniciales del entorno.

En la implementación reportada se realiza una sola ejecución del entrenamiento con semilla fija igual a 42. El proceso no incorpora criterio de parada temprana y el modelo evaluado en el capítulo de resultados corresponde al último estado almacenado tras los 1000 episodios. Esta configuración responde a la reproducibilidad experimental propia del alcance de un trabajo de pregrado y no a una búsqueda exhaustiva de robustez estadística entre múltiples ejecuciones independientes.

3.5 Algoritmo de entrenamiento MSAC

3.5.1 Funciones de valor y actualización de los críticos

Cada agente dispone de un par de críticos gemelos $Q_{\phi_i^1}$ y $Q_{\phi_i^2}$, parametrizados por ϕ_i^1 y ϕ_i^2, que reciben la entrada global x_t^{critic} y estiman la función de valor acción-estado. Para mitigar el sesgo de sobreestimación propio de los métodos *actor-critic*, el valor objetivo se construye tomando el mínimo entre ambos críticos (Haarnoja et al., 2018):

$$Q_{\phi_i}^{\text{mín}}(x_t, a_t) = \text{mín}\Big(Q_{\phi_i^1}(x_t, a_t),\ Q_{\phi_i^2}(x_t, a_t)\Big). \tag{17}$$

El valor objetivo para la actualización de los críticos en el paso t es

$$y_t = R_t + \gamma \sum_{a' \in \mathcal{A}_i} \pi_{\theta_i}(a' \mid s_{i,t+1}) \Big[Q_{\bar{\phi}_i}^{\text{mín}}(x_{t+1}, a') - \alpha \log \pi_{\theta_i}(a' \mid s_{i,t+1}) \Big], \tag{18}$$

donde $\bar{\phi}_i$ denota los parámetros de las redes objetivo (*target networks*), actualizadas mediante una media móvil exponencial con tasa τ:

$$\bar{\phi}_i \leftarrow \tau\, \phi_i + (1 - \tau)\, \bar{\phi}_i. \tag{19}$$

La función de pérdida de cada crítico se minimiza por descenso de gradiente estocástico:

$$\mathcal{L}(\phi_i^k) = \mathbb{E}_{(x_t, a_t, R_t, x_{t+1}) \sim \mathcal{D}} \left[\left(Q_{\phi_i^k}(x_t, a_t) - y_t \right)^2 \right], \quad k \in \{1, 2\}, \tag{20}$$

donde $\mathcal{D}$ denota el *experience replay buffer*.

3.5.2 Actualización del actor

Los parámetros del actor se actualizan maximizando el objetivo de máxima entropía de SAC (Haarnoja et al., 2018), adaptado al caso discreto (W. Wang et al., 2019):

$$\mathcal{L}(\theta_i) = -\mathbb{E}_{s_{i,t}\sim\mathcal{D}}\left[\sum_{a\in\mathcal{A}_i} \pi_{\theta_i}(a \mid s_{i,t})\Big(Q_{\phi_i}^{\text{mín}}(x_t, a) - \alpha \log \pi_{\theta_i}(a \mid s_{i,t})\Big)\right]. \tag{21}$$

3.5.3 Ajuste automático del coeficiente de entropía

El coeficiente α se ajusta automáticamente para que la entropía de la política no caiga por debajo de un umbral mínimo $\mathcal{H}_{\text{tgt}} = -\log(1/|\mathcal{A}_i|)$ (Haarnoja et al., 2018):

$$\mathcal{L}(\alpha) = \mathbb{E}_{s_{i,t}\sim\mathcal{D}}\Big[-\alpha\Big(\log \pi_{\theta_i}(a^*_{i,t} \mid s_{i,t}) + \mathcal{H}_{\text{tgt}}\Big)\Big], \tag{22}$$

donde $a^*_{i,t} = \arg\max_a \pi_{\theta_i}(a \mid s_{i,t})$ es la acción de mayor probabilidad según la política actual.

3.5.4 Experience replay buffer

El entrenamiento es *off-policy*: cada agente i almacena transiciones en su propio buffer $\mathcal{D}_i$. Cada transición tiene la forma

$$\Big(s_{i,t},\ x_t^{\text{critic}},\ a_{i,t},\ R_t,\ s_{i,t+1},\ x_{t+1}^{\text{critic}}\Big) \in \mathcal{D}_i, \tag{23}$$

donde se almacena tanto el estado local del agente (para la actualización del actor) como la entrada global del crítico (para la actualización de los críticos), lo que permite recuperar de forma consistente todos los componentes necesarios para cada función de pérdida.

3.5.5 Ciclo de interacción entre etapas del algoritmo

La Figura 14 muestra el diagrama de flujo del entrenamiento para un paso (una hora) y para un episodio (un día). En cada paso t, la interacción entre las etapas del esquema MSAC sigue la secuencia:

1. **Observación**: cada agente i lee su estado local $s_{i,t} = [V_{i,t}, P_{i,t}, Q_{i,t}, z_{i,t}]$ desde el simulador de red.

2. **Selección de acción**: la política $\pi_{\theta_i}(\cdot \mid s_{i,t})$, evaluada mediante la **??**, produce una distribución categórica sobre $\mathcal{A}_i$ y se muestrea la acción $a_{i,t}$.

3. **Remapeo**: la acción ejecutada se obtiene aplicando la Ecuación (11), garantizando que $z_{i,t+1}$ permanezca dentro del rango físico del dispositivo.

4. **Ejecución y transición**: las acciones remapeadas de todos los agentes se aplican al sistema y se resuelve el flujo de potencia AC (Ecuaciones (4) y (5)), obteniendo el nuevo vector de tensiones $\mathbf{V}_{t+1}$ y el estado s_{t+1}.

5. **Recompensa**: se calcula R_t según las Ecuaciones (14) a (16).

6. **Almacenamiento**: la transición completa se guarda en $\mathcal{D}_i$ según la Ecuación (23).

7. **Actualización**: si el buffer supera el tamaño mínimo de lote, se extraen muestras aleatorias y se actualizan los parámetros ϕ_i^1, ϕ_i^2, θ_i y α minimizando las pérdidas de las Ecuaciones (20) a (22), y se actualizan las redes objetivo $\bar{\phi}_i$ según la Ecuación (19).

3.6 Arquitectura neuronal del actor y del crítico

Cada agente cuenta con su propio actor y con su propio par de críticos; las redes no comparten parámetros, aunque sí comparten la misma estructura general y el mismo valor

de `hidden_dim`. Esta decisión mantiene una arquitectura homogénea entre dispositivos pero permite que cada agente especialice su política y sus estimadores de valor según el tipo de actuador y la zona de la red donde opera.

La red del actor recibe como entrada $s_{i,t} \in \mathbb{R}^4$, procesa la información a través de capas ocultas con función de activación $ReLU$ de ancho `hidden_dim`, y produce en la capa de salida una distribución categórica, con $|\mathcal{A}_i|$ salidas: cinco para el OLTC y tres para cada CB.

Las redes de los críticos reciben la entrada global $x_t^{\text{critic}} \in \mathbb{R}^{40}$, compuesta por la concatenación de los ocho estados locales ($8 \times 4 = 32$ componentes) y las ocho acciones conjuntas. Sus capas internas también emplean $ReLU$ con el mismo `hidden_dim`, y la capa de salida produce un escalar Q, estimación de la función de valor acción-estado.

3.7 Configuración del entrenamiento

En total se entrenan ocho agentes de manera conjunta: uno asociado al OLTC y siete a los bancos de capacitores. La Tabla 5 resume los hiperparámetros utilizados.

Tabla 5

Hiperparámetros de entrenamiento

Hiperparámetro	**Valor**
`hidden_dim`	512
`buffer_capacity`	50000
`batch_size`	1024
γ	0.99
τ	0.005
`learning_rate_actor`	5×10^{-5}
`learning_rate_critic`	1×10^{-3}
`learning_rate_alpha`	3×10^{-4}

Nota. Elaboración propia.

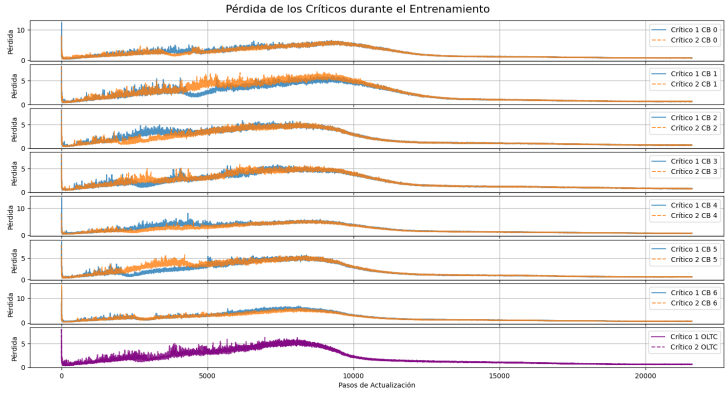
La selección de estos valores toma como punto de partida los hiperparámetros reportados por W. Wang et al. (2019) y se ajusta empíricamente para estabilizar el entrena-

miento, evitando oscilaciones extremas en las pérdidas e inestabilidades numéricas en la actualización de las redes. El entrenamiento se ejecuta durante 1000 episodios; los primeros 100 corresponden a la fase de *warmup*. El modelo evaluado corresponde al último estado almacenado tras completar los 1000 episodios.

Las Figuras 4 y 5 presentan la evolución de la pérdida de los críticos y de los actores durante el entrenamiento, respectivamente.

Figura 4

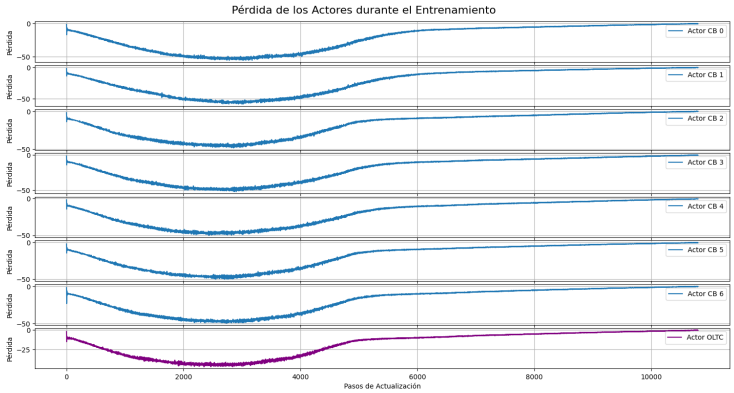
Pérdida de los críticos durante el entrenamiento



Nota. Datos recopilados durante la fase de pruebas.

Figura 5

Pérdida de los actores durante el entrenamiento



Nota. Datos recopilados durante la fase de pruebas.

3.8 Escenarios de evaluación

La evaluación del esquema propuesto se realiza en tres escenarios complementarios: un caso base sin control, un escenario nominal con control y un conjunto de escenarios estocásticos construidos a partir de una perturbación Weibull sobre la demanda activa.

3.8.1 Escenario sin control

En el escenario sin control, el sistema opera con el perfil de demanda de la Figura 3 y la demanda base de la Tabla 4, pero sin modificar el estado del OLTC ni el de los CBs a lo largo del día. Este caso sirve como línea base para cuantificar la magnitud de las violaciones de tensión en ausencia de acciones correctivas.

3.8.2 Escenario con control

En el escenario con control se activa la política aprendida por el esquema MSAC. Cada agente observa su estado local $s_{i,t}$, muestrea una acción desde $\pi_{\theta_i}(\cdot \mid s_{i,t})$ y la aplica dentro de los límites físicos del dispositivo mediante la Ecuación (11). Los resultados se comparan con el método centralizado DCD de referencia reportado en (Cristancho Castro, 2023).

3.8.3 Escenarios estocásticos con Weibull

Con el fin de analizar la robustez del controlador ante variabilidad de carga, se construyen escenarios estocásticos multiplicando la demanda activa de cada nodo por un factor w_t muestreado de una distribución de Weibull. La elección de esta distribución se justifica por su soporte estrictamente no negativo ($w \geq 0$), lo cual garantiza la consistencia

física del modelo de carga. Dado que la perturbación se aplica sobre la potencia activa nominal, el uso de una distribución Normal resultaría inadecuado en escenarios de alta dispersión, ya que podría generar factores negativos; esto implicaría erróneamente que la red actúa como fuente de generación en lugar de consumo, alterando la naturaleza del problema de distribución abordado.

$$P_{n,t}^{\text{est}} = w_t \cdot P_{n,t}, \qquad w_t \sim \text{Weibull}(k, \lambda), \tag{24}$$

cuya función de densidad de probabilidad es

$$f(w;\, k, \lambda) = \frac{k}{\lambda}\Big(\frac{w}{\lambda}\Big)^{k-1} \exp\Big[-\Big(\frac{w}{\lambda}\Big)^{k}\Big], \quad w \geq 0, \tag{25}$$

donde $k > 0$ es el parámetro de forma y $\lambda > 0$ es el parámetro de escala. El coeficiente de variación $CV = \sigma/\mu$ del factor w_t determina el nivel de dispersión alrededor del perfil nominal. Los valores de k y λ para cada nivel de variabilidad se obtienen a partir de la relación analítica entre CV, k y λ establecida por la distribución de Weibull, y se presentan en la Tabla 6.

Tabla 6

Parámetros de la distribución Weibull para cada nivel de variabilidad

CV	k	λ
0.10	29.296	1.019
0.30	5.795	1.080
0.60	2.085	1.129

Nota. Elaboración propia.

El factor w_t se aplica de manera uniforme a toda la red y se generan 10 escenarios para cada nivel de dispersión, manteniendo constante la estructura de red y los dispositivos de control. Esta formulación modela variaciones agregadas de carga activa y no recursos energéticos distribuidos ni generación fotovoltaica.

3.8.4 Métricas de evaluación

El desempeño del método se evalúa con base en métricas eléctricas y operativas coherentes con la formulación del problema. En primer lugar, se analiza el perfil de tensión a lo largo del alimentador y su permanencia dentro de la banda $[V^{\text{mín}}, V^{\text{máx}}]$. En segundo lugar, se examina la energía aparente acumulada suministrada por la barra *slack*, en coherencia con el término S_t de la función de recompensa. En tercer lugar, se considera el factor de potencia en la barra de alimentación como indicador complementario del comportamiento global del sistema. Además, se reporta el número total de conmutaciones realizadas por el OLTC y por los CBs como métrica operativa relevante.

4. Resultados

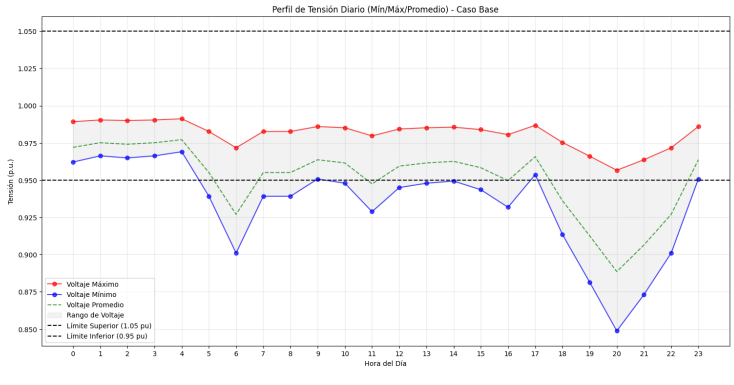
En este capítulo se presentan y analizan los resultados obtenidos tras la implementación del esquema de control propuesto. La evaluación se divide en tres etapas: primero, un análisis del desempeño nominal del algoritmo MSAC frente a un caso base sin control; segundo, una comparación sistemática con el método de descenso de coordenadas discretas (DCD); y finalmente, una evaluación de robustez bajo escenarios de carga estocásticos mediante distribuciones de Weibull.

4.1 Caso base

El escenario base representa la operación del alimentador IEEE de 33 nodos sin la intervención de los dispositivos de control (OLTC en posición neutra y bancos desconectados). En la Figura 6 se observa el perfil de tensión para todos los nodos durante un ciclo de 24 horas. Los resultados, sintetizados en la Tabla 7, evidencian una degradación severa del perfil de tensión, especialmente durante las horas de mayor demanda.

Figura 6

Perfil de tensión en el escenario base (sin control)



Nota. Datos recopilados durante la fase de pruebas.

Tabla 7

Síntesis de resultados del caso base sin control

Aspecto	Nodo(s)	Valor / Observación
Horas críticas	–	6 y 20
Tensión mínima registrada	18	0.8500 p.u. (t=20)
Nodos con violaciones sostenidas	16, 17, 18	16 horas/día
Total de nodos fuera de rango	–	28 nodos

Nota. Elaboración propia.

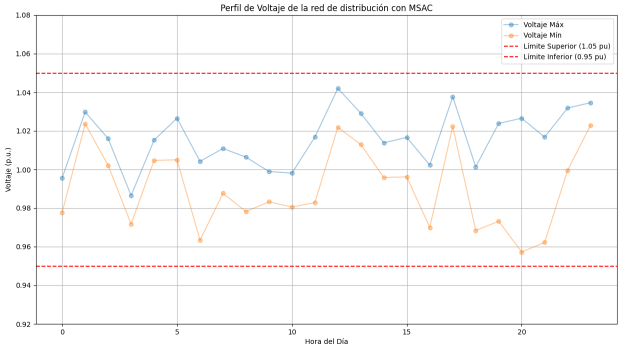
4.2 Caso MSAC

Al aplicar la política de control multiagente (MSAC), el sistema logra corregir las desviaciones de tensión dentro de la banda operativa permitida para el escenario nominal. La Figura 7 muestra el comportamiento resultante, donde la tensión mínima se eleva a 0.9573 p.u. en el nodo 18 (hora 20), mientras que la Figura 16 presenta las acciones del OLTC durante el día de operación. Esta mejora se fundamenta en la coordinación de

los ocho agentes, cuyas acciones incrementalmente ajustan la compensación reactiva y la relación de transformación. La Tabla 8 resume las conmutaciones realizadas por cada dispositivo y la Tabla 9 sintetiza los valores extremos de tensión bajo control MSAC.

Figura 7

Perfil de tensión bajo control MSAC (Escenario nominal)



Nota. Datos recopilados durante la fase de pruebas.

Tabla 8

Conmutaciones realizadas por los dispositivos de control

Dispositivo	Total acciones
OLTC	26
CB0	12
CB1	2
CB2	2
CB3	11
CB4	6
CB5	13
CB6	8

Nota. Datos recopilados durante la fase de pruebas.

Tabla 9

Síntesis de resultados del caso MSAC

Aspecto	Nodo(s)	Valor / Observación
Tensión máxima	2	1.0421 (hora 12)
Tensión mínima	18	0.9573 (hora 20)

Nota. Elaboración propia.

4.3 Comparación con DCD

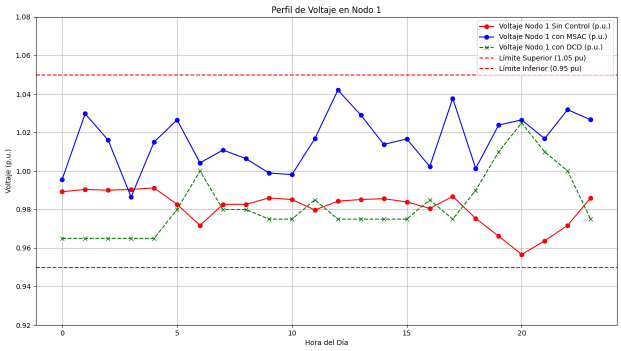
Para validar el desempeño del MSAC, se comparan sus métricas con el algoritmo de Descenso de Coordenadas Discretas (DCD) reportado en (Cristancho Castro, 2023). Esta comparación permite evaluar la efectividad de una estrategia de aprendizaje por refuerzo frente a una técnica de optimización tradicional.

4.3.1 Tensión nodo 1

La Figura 8 presenta el comportamiento de la tensión en la barra de cabecera (Nodo 1). Se observa que el MSAC logra una desviación promedio de 0.0173 p.u. respecto al valor ideal, superando ligeramente la precisión del DCD en este punto específico del sistema. La Tabla 10 resume las desviaciones promedio, mínima y máxima para los tres casos comparados.

Figura 8

Tensión en el Nodo 1: Comparativa Base, DCD y MSAC



Nota. Datos recopilados durante la fase de pruebas.

Tabla 10

Desviaciones en nodo 1 (respecto a 1 p.u.)

Caso	Promedio	Mínima	Máxima
Base	0.0188	0.0088	0.0433
DCD	0.0217	0.0000	0.0350
MSAC	0.0173	0.0010	0.0421

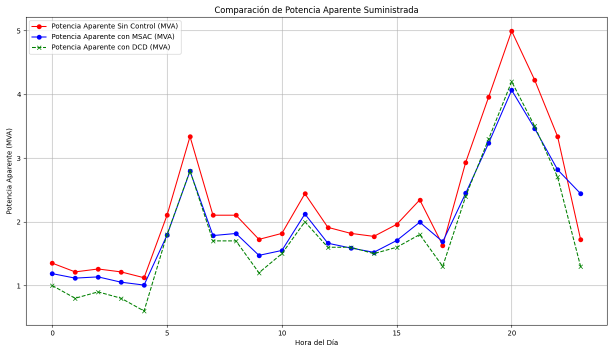
Nota. Elaboración propia.

4.3.2 Potencia aparente

El impacto del control sobre la eficiencia energética se evalúa mediante la potencia aparente suministrada desde la barra *slack*. Como se ilustra en la Figura 9, el MSAC reduce el consumo acumulado de 54.39 MVAh a 47.45 MVAh. La Tabla 11 resume la energía aparente acumulada para cada caso. No obstante, el DCD presenta una mayor eficacia en la reducción de potencia aparente, alcanzando los 43.6 MVAh, lo que sugiere que la política local del MSAC prioriza el cumplimiento de tensión sobre la optimización extrema de pérdidas.

Figura 9

Potencia aparente en la barra de alimentación (Slack)



Nota. Datos recopilados durante la fase de pruebas.

Tabla 11

Comparativa de energía aparente acumulada (24 h)

Caso	S [MVAh]
Base	54.3918
DCD	43.6
MSAC	47.4553

Nota. Elaboración propia.

4.3.3 Factor de potencia

El comportamiento del factor de potencia (FP) revela una particularidad operativa del esquema MSAC durante los periodos de baja demanda. Como se observa en la Figura 10, existe una degradación pronunciada del FP hacia el final del ciclo (hora 23), alcanzando un valor de 0.62. Este fenómeno no responde a una falla del control, sino a una decisión de la política aprendida ante la reducción de la carga activa.

El análisis detallado de la Tabla 12 permite identificar que, mientras la demanda activa (P) cae a su punto mínimo, los bancos de capacitores mantienen una configura-

ción de inyección elevada (posiciones [2 0 3 3 2 3 1]). Al ser el FP la relación entre P y la potencia aparente (S), la persistencia de una componente reactiva capacitiva dominante ($Q = -1{,}9074$ MVAr) frente a una carga activa pequeña provoca que el ángulo de fase se desvíe significativamente, reduciendo el FP.

Esta condición de sobrecompensación es un subproducto de la estructura de la recompensa y la naturaleza del aprendizaje multiagente:

1. **Priorización de la tensión:** Dado que la penalización por violaciones de tensión (λ_v) es crítica, los agentes prefieren mantener un soporte reactivo preventivo para asegurar que ningún nodo caiga por debajo de 0.95 p.u., incluso si esto penaliza la eficiencia en la barra slack.

2. **Inercia de la política local:** En la transición despues del pico de demanda (horas 22-24), la política local de los agentes de los bancos de capacitores prioriza el cumplimiento normativo de tensión sobre la optimización del factor de potencia, un parámetro que no está explícitamente penalizado de forma individual en la función de recompensa.

La inyección neta de potencia reactiva hacia la red se confirma en la Figura 11, donde el flujo reactivo del MSAC se vuelve negativo (capacitivo) al finalizar el día, evidenciando que el sistema se comporta como un generador de reactivos. Esta condición de

Tabla 12

Parámetros operativos detallados - Hora 23 (MSAC)

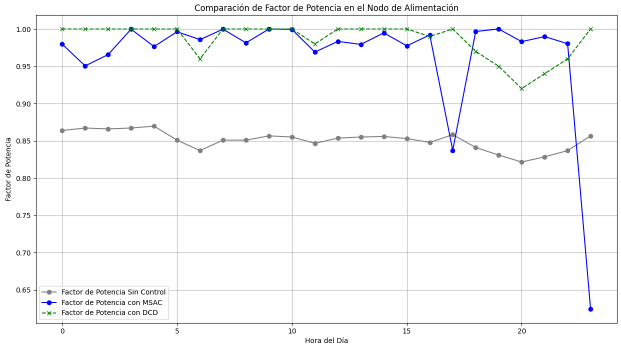
Parametro	Valor
Potencia activa	1.5238 MW
Potencia reactiva	-1,9074 MVAr
Potencia aparente	2.4414 MVA
Factor de potencia	0.6242
Tap OLTC	16
Posiciones CBs	[2 0 3 3 2 3 1]

Nota. Elaboración propia.

sobrecompensación es un subproducto de la política aprendida para asegurar el perfil de tensión en el periodo de transición post-pico (horas 18-22). La inyección neta de potencia reactiva hacia la red se confirma en la Figura 11, donde el flujo reactivo del MSAC se vuelve negativo (capacitivo) al finalizar el ciclo diario.

Figura 10

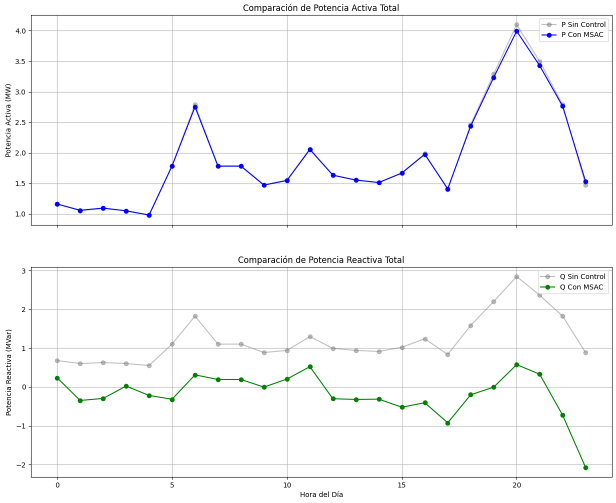
Evolución horaria del factor de potencia en la barra slack



Nota. Datos recopilados durante la fase de pruebas.

Figura 11

Comparativa de flujos de potencia activa (P) y reactiva (Q)



Nota. Datos recopilados durante la fase de pruebas.

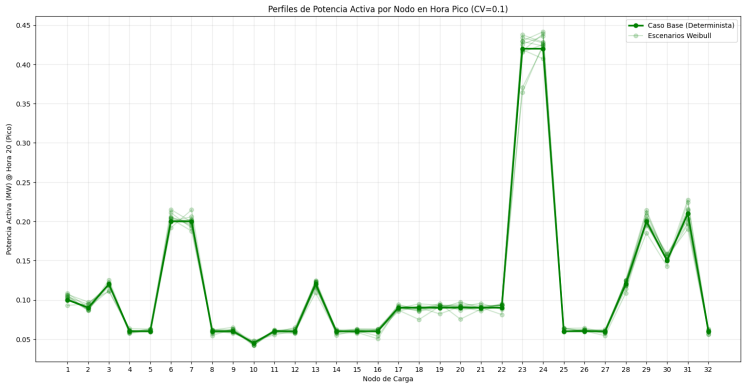
4.4 Escenarios Weibull

Para evaluar la capacidad de generalización del modelo, se sometió al controlador a variaciones estocásticas de carga utilizando distribuciones de Weibull con coeficientes de variación (C_v) de 10 %, 30 % y 60 %. A diferencia del caso nominal, estos escenarios introducen condiciones extremas no vistas durante el entrenamiento estándar.

Los resultados muestran que, bajo una variación del 10 %, el sistema experimenta picos de sobretensión que superan los 1.10 p.u. en nodos alejados de la cabecera. Este comportamiento indica que, aunque la estrategia local es robusta frente a cambios moderados, la falta de una coordinación centralizada en tiempo real limita la capacidad de respuesta ante fluctuaciones de carga simultáneas y de gran magnitud en múltiples nodos.

Figura 12

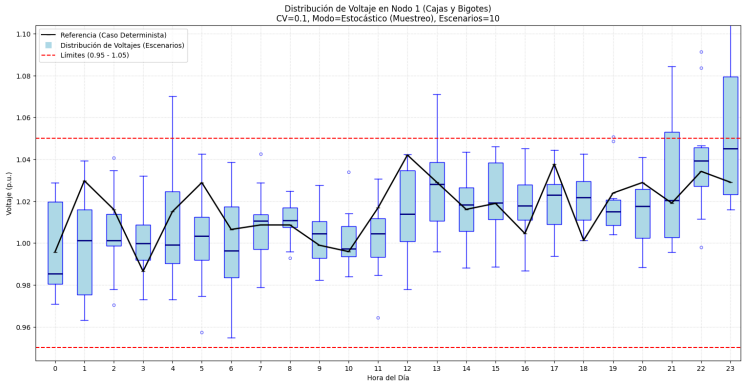
Escenario Weibull para 10 % de CV



Nota. Datos recopilados durante la fase de pruebas.

Figura 13

Evaluación con Weibull para 10 % de CV



Nota. Datos recopilados durante la fase de pruebas.

4.5 Discusión

Los hallazgos de este capítulo permiten extraer tres conclusiones fundamentales sobre la implementación del MSAC para el control Volt/Var:

1. **Eficacia en escenarios nominales:** El MSAC demuestra una capacidad superior

para eliminar violaciones de tensión en condiciones predecibles, logrando perfiles más planos que el escenario base e incluso mayor precisión en nodos específicos que el DCD.

2. **Compromiso entre tensión y pérdidas:** El algoritmo prioriza el cumplimiento de los límites normativos de tensión mediante una gestión agresiva de reactivos. Sin embargo, esto conlleva una menor eficiencia en la reducción de potencia aparente global comparado con métodos de optimización centralizada como el DCD. La sobre-compensación observada al final del día refuerza esta observación.

3. **Límites de la descentralización:** La evaluación estocástica que se hizo con los escenarios Weibull no es suficiente para determinar que el control puramente local, aunque robusto ante fallas de comunicación, presenta dificultades para mitigar fenómenos sistémicos de sobretensión bajo incertidumbre elevada, debido a que la evaluación se hizo solo teniendo en cuenta el nodo 1, sin tener en cuenta los demás nodos del sistema. Es necesario evaluar esta variación de manera global en el sistema para determinar el rendimiento del algoritmo ante variaciones en la demanda.

5. Conclusiones

- La revisión de literatura muestra una tendencia creciente hacia enfoques basados en datos y DRL. Sin embargo, existe poca exploración en estrategias de control local para dispositivos discretos tradicionales, pese a que dichas estrategias prometen menores costos de comunicación e infraestructura. En este contexto, arquitecturas CTDE permiten aprovechar las ventajas del control local con reducidos requerimientos de comunicación.

- El sistema de recompensa de un modelo de control basado en DRL se debe relacionar con el objetivo de control, donde la ponderación de las diferentes variables que se

relacionan con los objetivos debe estar debidamente balanceada y modelada conforme se espere el comportamiento del sistema controlado, recompensando las acciones que llevan en mayor medida a un estado optimo.

- En la evaluación comparativa de la estrategia de control local del MSAC Frente al método DCD centralizado se observa que, en el aspecto de la entrega de potencia aparente al sistema, el MSAC presenta un mejor rendimiento respecto al caso sin control, sin embargo, el DCD presenta un mejor rendimiento que el MSAC en un número mayor de eventos de control, esto mismo se puede observar respecto al factor de potencia en el nodo de alimentación.

- Fuera de la comparativa con el DCD se observa que empleando el método MSAC la tensión de todo el sistema se encuentra dentro de los valores permisibles de tensión, esto comprueba que el método MSAC para la estrategia de control local con entrenamiento centralizado realiza coordinación de los dispositivos de control de tal manera que logra este objetivo de control.

Referencias

Allahmoradi, S., Afrasiabi, S., Liang, X., Zhao, J., & Shahidehpour, M. (2024). Data-driven volt/var optimization for modern distribution networks: A review. *IEEE Access*, *12*, 71184-71204. https://doi.org/10.1109/ACCESS.2024.3403035

Baran, M., & Wu, F. (1989). Network reconfiguration in distribution systems for loss reduction and load balancing. *IEEE Transactions on Power Delivery*, *4*(2), 1401-1407. https://doi.org/10.1109/61.25627

Cao, D., Zhao, J., Hu, W., Ding, F., Huang, Q., Chen, Z., & Blaabjerg, F. (2021). Data-driven multi-agent deep reinforcement learning for distribution system decentralized voltage control with high penetration of PVs. *IEEE Transactions on Smart Grid*, *12*(5), 4137-4150.

Cristancho Castro, C. A. (2023, 23 de octubre). *Estrategia de control tensión-reactiva en sistemas de distribución a partir de bancos de condensadores y cambia tomas en transformadores* [Trabajo de grado]. Universidad Industrial de Santander.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., & Abbeel, P. (2018). Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*.

Hai, D., Zhu, T., Duan, S., Huang, W., & Li, W. (2022). Deep reinforcement learning for Volt/Var control in distribution systems: A review. *2022 5th International Conference on Energy, Electrical and Power Engineering (CEEPE)*, 596-601.

Hein, H., Zhu, J., Yu, L., Liu, Z., Liu, X., Xu, S., & Hao, Y. (2026). Voltage Control Strategies in Distribution Systems With High Distributed Generation Penetration: Models, Methods and Future Research. *IET Energy Systems Integration*, *8*(1), e70036.

IEEE. (2022). *IEEE Guide for Assessing, Measuring, and Verifying Volt-Var Control and Optimization on Distribution Systems*. IEEE. https://doi.org/10.1109/IEEESTD.2022.9828004

Lapan, M. (2020). *Deep Reinforcement Learning Hands-On: Apply modern RL methods to practical problems of chatbots, robotics, discrete optimization, web automation, and more*. Packt Publishing Ltd.

Sun, X., & Qiu, J. (2021). Two-stage volt/var control in active distribution networks with multi-agent deep reinforcement learning method. *IEEE Transactions on Smart Grid*, *12*(4), 2903-2912.

Wang, S., Duan, J., Shi, D., Xu, C., Li, H., Diao, R., & Wang, Z. (2020). A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning. *IEEE Transactions on Power Systems*, *35*(6), 4644-4654.

Wang, W., Yu, N., Gao, Y., & Shi, J. (2019). Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems. *IEEE Transactions on Smart Grid*, *11*(4), 3008-3018.

Wu, M., Hong, L., Wang, Y., Yan, Z., & Chen, Z. (2022). Volt-VAR control for distribution networks with high penetration of DGs: An overview. *The Electricity Journal*, *35*(5), 107130.

Zhang, X., Wu, Z., Sun, Q., Gu, W., Zheng, S., & Zhao, J. (2024). Application and progress of artificial intelligence technology in the field of distribution network voltage Control: A review. *Renewable and Sustainable Energy Reviews*, *192*, 114282.

6. Apéndices

Los recursos adicionales, código fuente y datos del proyecto se encuentran disponibles en el Repositorio Institucional

Apéndice A.Síntesis comparativa de trabajos representativos

La Tabla 13 presenta una síntesis comparativa de los trabajos representativos revisados.

Tabla 13

Síntesis de trabajos representativos revisados

Artículo	**Objetivo principal**	**Estrategia de control**	**Dispositivos**	**Método**
A Two-Layer Volt-Var Control Method in Rural Distribution Networks Considering Utilization of Photovoltaic Power	Mantener restricciones de tensión y reducir pérdidas	Control en dos capas, con coordinación por particiones y acciones locales	Inversores inteligentes	Optimización
Energy Savings Estimation of a Distribution System in Presence of Intelligent Volt-VAr Control Based on IEEE Std. 1547-2018	Mejorar perfil de tensión y flujo reactivo	Ajuste dinámico de tensión e inyección/absorción de reactivos	Reguladores, capacitores conmutables, inversores inteligentes	Optimización y coordinación
Método basado en máquinas de soporte vectorial para el control tensión-reactiva en sistemas de distribución de energía eléctrica a partir de cambia tomas en transformadores y bancos de condensadores	Reducir variaciones de tensión y potencia aparente suministrada	Selección de configuración óptima con información de varios nodos	OLTC y capacitores shunt	SVM
Multi-Stage Volt/VAR Support in Distribution Grids: Risk-Aware Scheduling with Real-Time Reinforcement Learning Control	Cumplir restricciones de tensión y reducir costo operativo	Dos etapas: dispositivos lentos por horario y recursos rápidos en tiempo real	OLTC, CBs, almacenamiento, inversores inteligentes	DDPG

Tabla 13

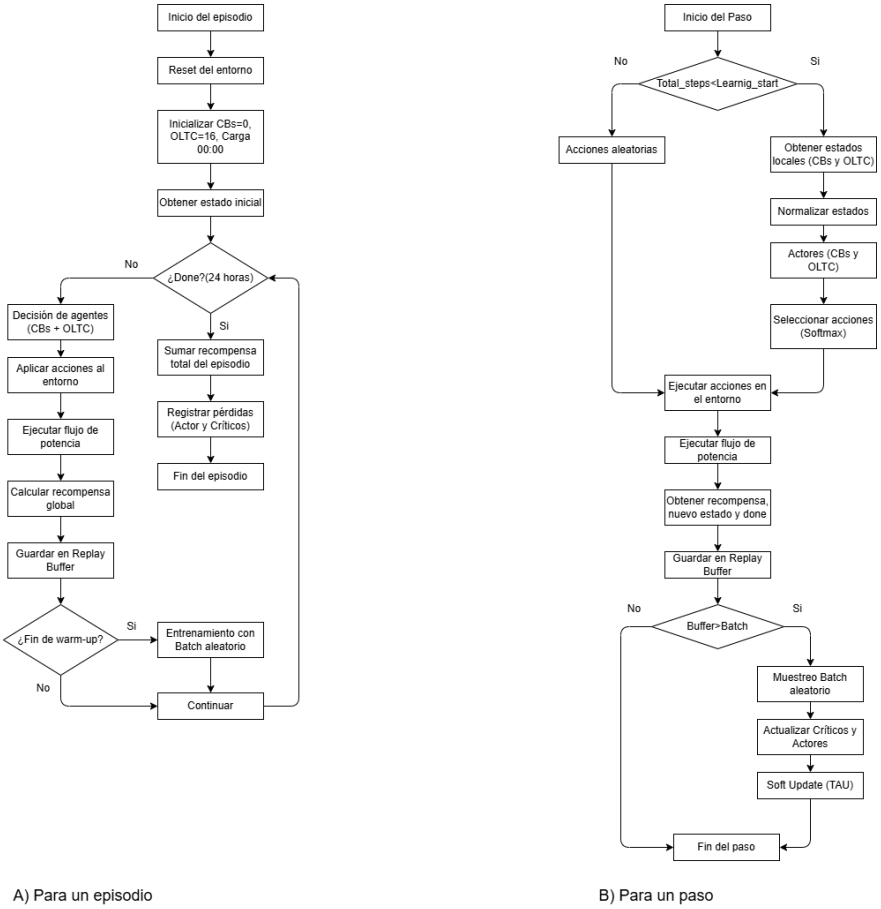
Síntesis de trabajos representativos revisados (continuación)

Artículo	**Objetivo principal**	**Estrategia de control**	**Dispositivos**	**Método**
Safe Off-Policy Deep Reinforcement Learning Algorithm for Volt-VAR Control in Power Distribution Systems	Reducir pérdidas y costo operativo bajo restricciones de tensión	Control centralizado sobre dispositivos discretos	OLTC, CBs y reguladores	CSAC
Sensitivity-Based Discrete Coordinate-Descent for Volt/Var Control in Distribution Networks	Reducir pérdidas y violaciones de tensión	Optimización iterativa basada en sensibilidad	OLTC, CBs y generación distribuida	DCD
Stabilizing Voltage in Power Distribution Networks via Multi-Agent Reinforcement Learning with Transformer	Mantener tensión y reducir pérdidas	Entrenamiento centralizado con ejecución local por agentes	Inversores inteligentes	T-MAAC

Nota. Síntesis elaborada a partir de los trabajos priorizados durante la revisión bibliográfica.

Figura 14

Flujo de entrenamiento para un paso y para un episodio

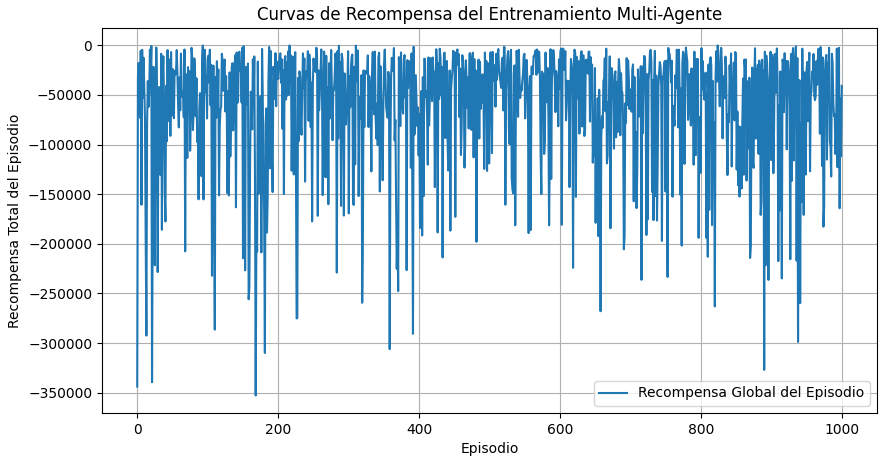


Nota. Elaboración propia.

Apéndice B.Resultado de entrenamiento

Figura 15

Resultado de la recompensa al finalizar el entrenamiento

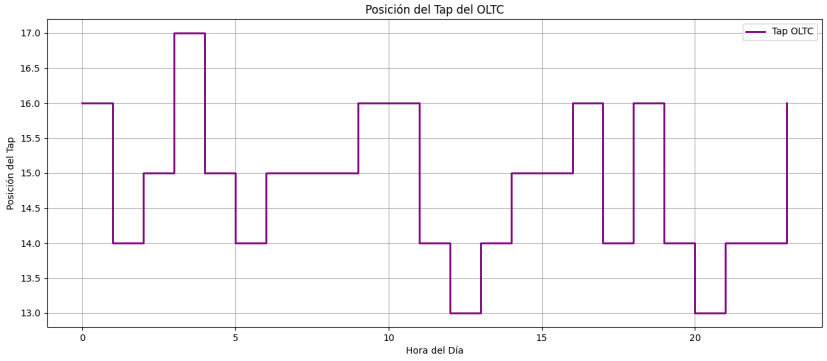


Nota. Datos recopilados durante la fase de pruebas.

Apéndice C.Resultado de evaluación

Figura 16

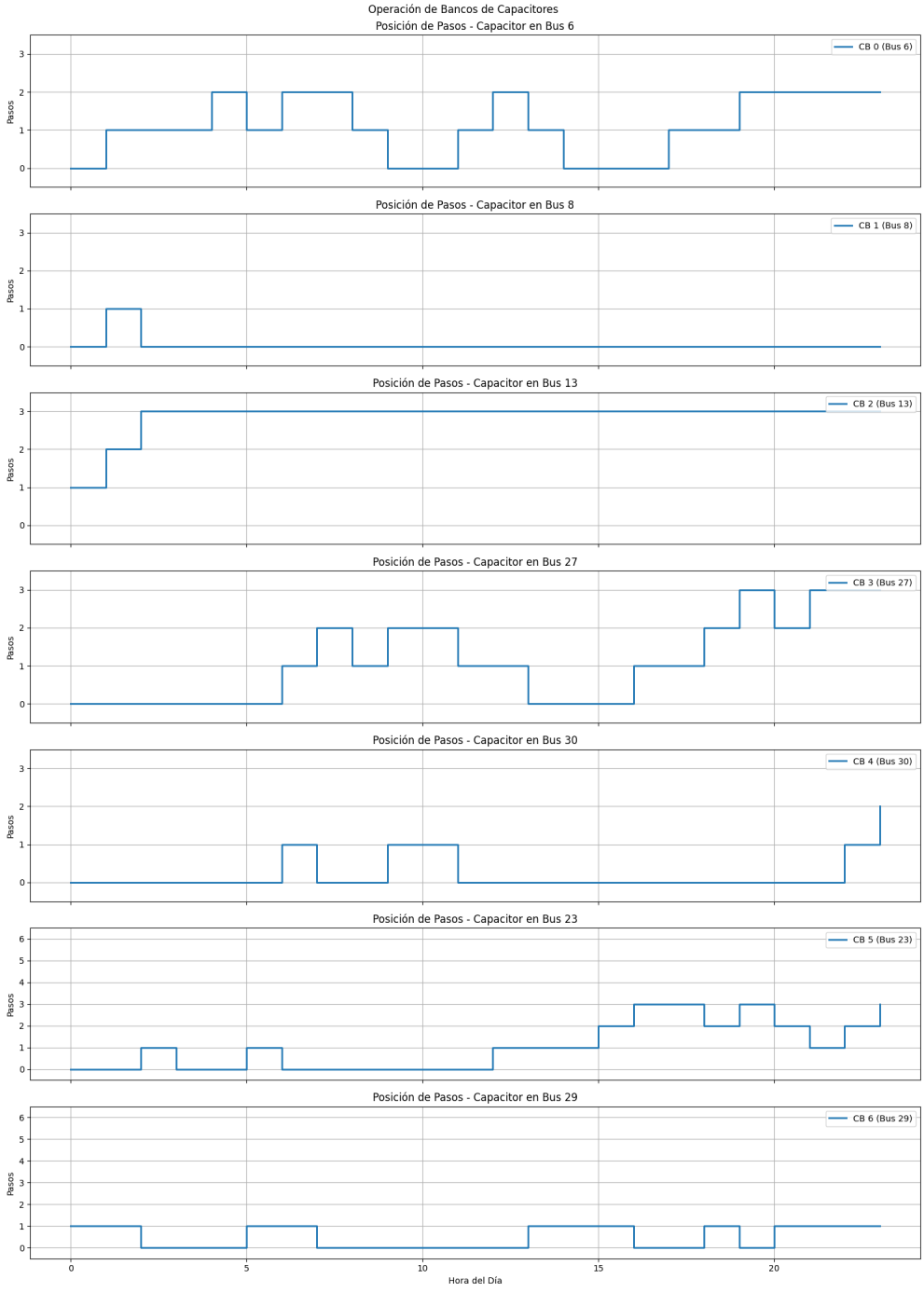
Acciones realizadas por el OLTC durante el día de operación



Nota. Datos recopilados durante la fase de pruebas.

Figura 17

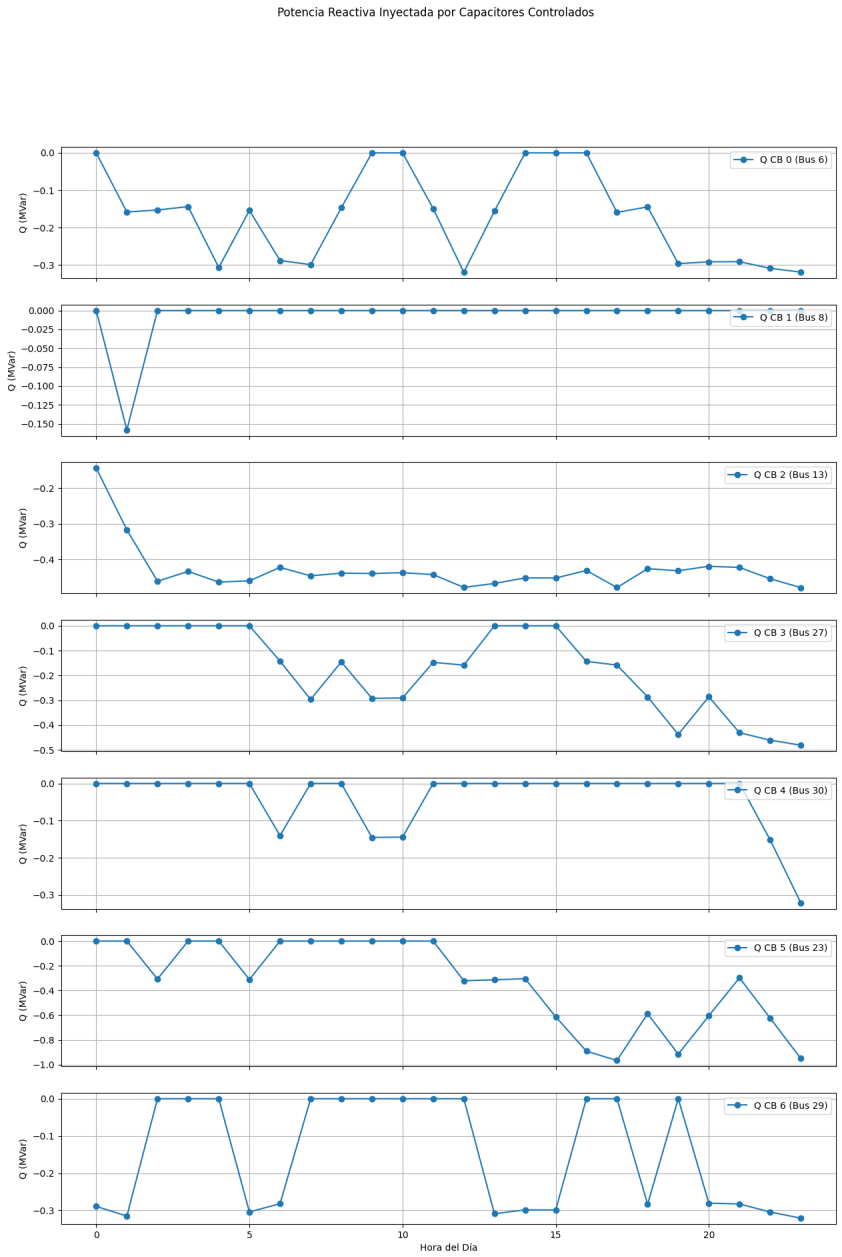
Operaciones de los bancos de capacitores



Nota. Datos recopilados durante la fase de pruebas.

Figura 18

Potencia reactiva entregada por los bancos de capacitores

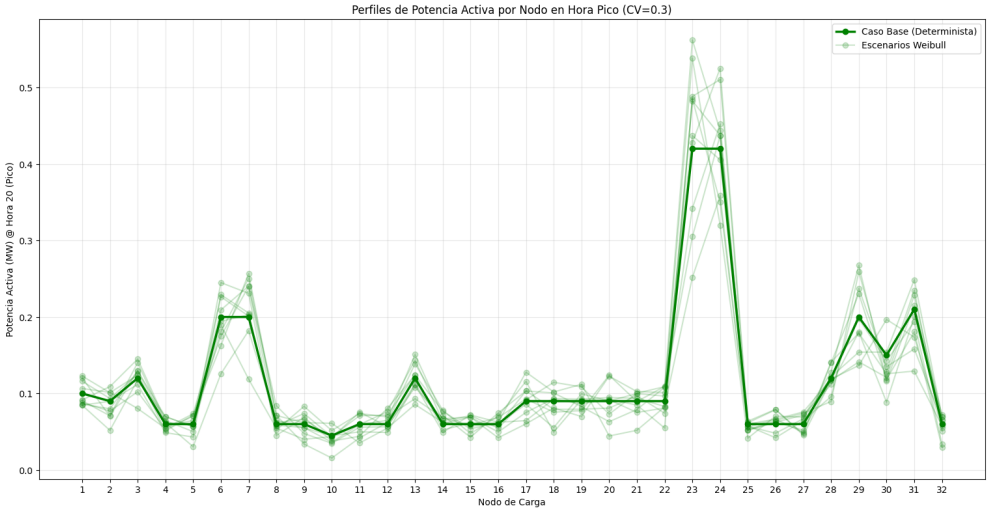


Nota. Datos recopilados durante la fase de pruebas.

Apéndice D.Escenarios de evaluación con Weibull

Figura 19

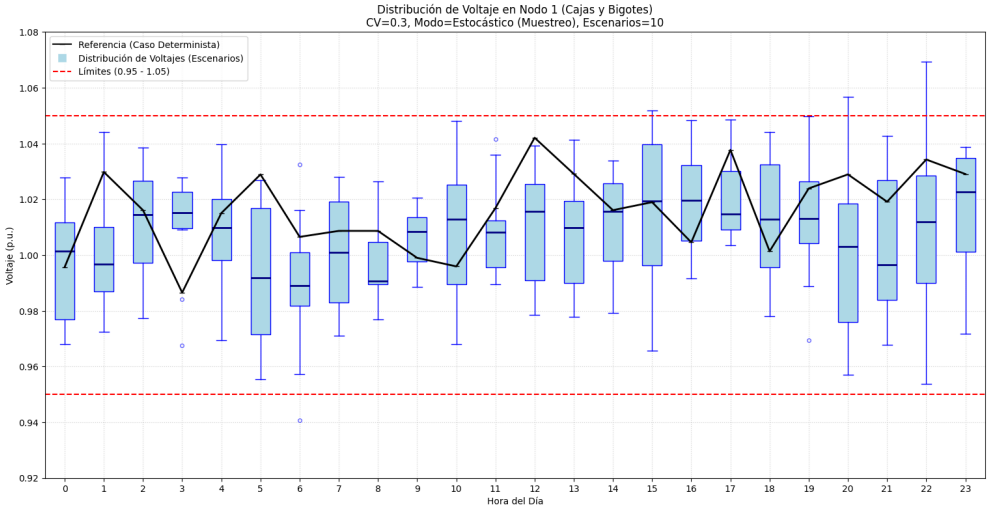
Escenario aleatorio generado con Weibull para CV de 30 %



Nota. Datos recopilados durante la fase de pruebas.

Figura 20

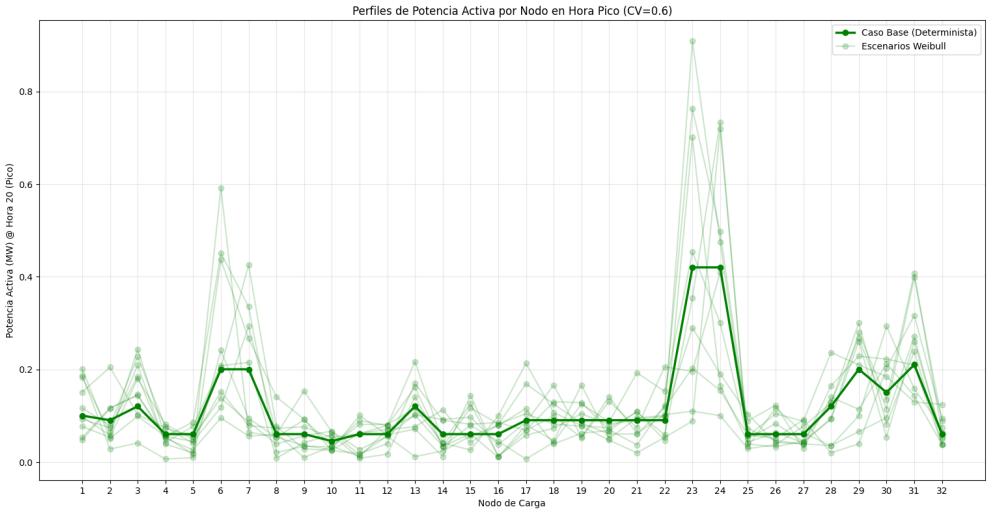
Resultado de la evaluación con Weibull para CV de 30 %



Nota. Datos recopilados durante la fase de pruebas.

Figura 21

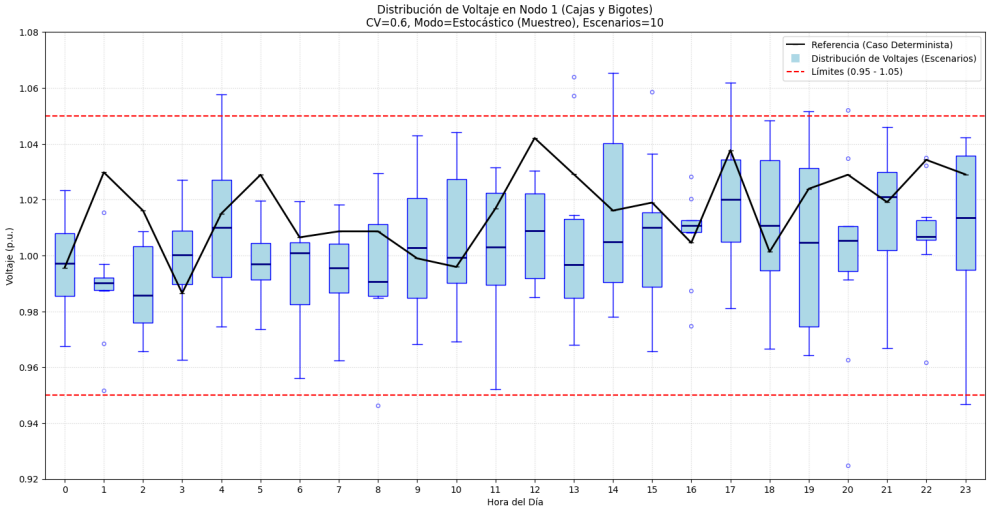
Escenario aleatorio generado con Weibull para CV de 60 %



Nota. Datos recopilados durante la fase de pruebas.

Figura 22

Resultado de la evaluación con Weibull para CV de 60 %



Nota. Datos recopilados durante la fase de pruebas.