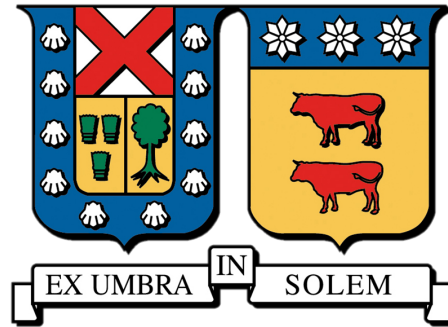




UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INDUSTRIAS

PAIRS TRADING CON REINFORCEMENT LEARNING Y
PREDICCIÓN LSTM: UN ENFOQUE DE ESCALAMIENTO DINÁMICO
APLICADO A SECTORES DEL S&P 500

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

AUTOR

FRANCISCO FARÍAS CHÁVEZ

PROFESOR GUÍA

DR. WERNER KRISTJANPOLLER

PROFESOR CO-REFERENTE

DR. BERNARDO PINCHEIRA

VALPARAÍSO, JULIO 2026



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: PAIRS TRADING CON REINFORCEMENT LEARNING Y PREDICCIÓN LSTM: UN ENFOQUE DE ESCALAMIENTO DINÁMICO APLICADO A SECTORES DEL S&P 500

Nombre del candidato(a): Francisco Ignacio Farías Chávez

Carrera / Grado: Ingeniería Civil Industrial

Campus: Casa Central Valparaíso

Departamento: Industrias

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, _____ Werner Kristjanpoller Rodríguez _____, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por (**marcar una opción**):

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 30-07-2026

Firma:

Estudiante o Candidato(a):

Fecha: 23-07-2026

Firma:

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Tabla de Contenidos

1	Agradecimientos	4
2	Resumen	5
3	Introducción	7
4	Problema de Investigación	8
5	Objetivos	11
5.1	Objetivo General	11
5.2	Objetivos Específicos	11
6	Alcance	12
7	Marco Teórico	14
7.1	Arbitraje estadístico	14
7.2	Pairs Trading	16
7.3	Formación de pares	18
7.4	Estrategia clásica	22
7.5	Reinforcement Learning (RL)	23
7.6	Algoritmos RL relevantes	25
7.7	RL aplicado a Pairs Trading	28
7.8	LSTM como componente predictivo	32
8	Metodología	34

8.1	Datos	35
8.2	Diseño experimental	36
8.3	Selección de pares	38
8.4	Construcción del spread	39
8.5	Oráculo LSTM	41
8.6	Entorno RL	43
8.6.1	Máquina de estados	43
8.6.2	Espacio de observación	44
8.6.3	Espacio de acción	46
8.6.4	Mecanismos de control	46
8.6.5	Función de recompensa	47
8.7	Optimización de hiperparámetros	49
8.8	Entrenamiento del agente	51
8.9	Selección de pares post entrenamiento	52
8.10	Backtest OOS y construcción de portafolio	52
8.11	Benchmark y comparaciones experimentales	54
8.12	Métricas de evaluación	55
9	Resultados	57
10	Conclusiones	64
11	Discusión	66
12	Limitaciones	68



13 Anexos

69

1 Agradecimientos

Aprovecho este espacio para agradecer el apoyo que me han entregado desde casa, mi mamá, papá y mis dos hermanos por darme ánimos constantemente durante el proceso universitario; a mi polola que me ha animado en todo momento y a mis amigos con los cuales pude disfrutar todos estos años. También a mi profesor guía por darme directrices para poder realizar mi memoria.

2 Resumen

La presente memoria aborda el tema pairs trading en mercados financieros, una estrategia de arbitraje estadístico que busca aprovechar desviaciones temporales entre activos relacionados. En su formulación clásica, esta estrategia suele emplear reglas fijas de entrada y salida; sin embargo, su desempeño puede verse afectado por cambios en la volatilidad, quiebres en la relación entre activos y falta de adaptación ante nuevas condiciones de mercado. En este contexto, surge la necesidad de evaluar enfoques más flexibles que permitan decidir no solo cuándo operar, sino también cuánto capital asignar y cuándo limitar pérdidas.

El objetivo de esta investigación es desarrollar y evaluar un modelo de pairs trading basado en Reinforcement Learning, aplicado a sectores del S&P 500, con el fin de comparar su rentabilidad y riesgo frente a una estrategia clásica. Además, se analiza cómo distintas frecuencias de reentrenamiento mensual, bimestral y trimestral afectan el desempeño del modelo, buscando determinar si existe una periodicidad óptima común para todos los sectores o si esta depende de las características propias de cada uno.

La metodología considera precios diarios ajustados de acciones del S&P 500, agrupadas por sectores GICS. Para la formación de pares se aplica cointegración mediante los test de Engle-Granger y Johansen, construyendo un spread y su respectivo z-score. El modelo propuesto integra un oráculo LSTM para estimar la probabilidad de divergencia del par, optimización de hiperparámetros mediante Optuna y un agente PPO con espacio de acción continuo, lo que permite aplicar dynamic scaling sobre la exposición al par. La evaluación se realiza mediante un esquema walk-forward, donde en cada rolling se rehace la selección de pares, el entrenamiento del LSTM, la optimización y el entrenamiento del agente.

Los resultados muestran que el agente basado en Reinforcement Learning supera al benchmark clásico de Gatev en los once sectores evaluados, considerando la mejor frecuencia de reentrenamiento de cada sector. Mientras el modelo propuesto obtiene retornos positivos en todos los sectores, la estrategia clásica presenta retornos negativos en la mayoría de los casos. Asimismo, se observa que la frecuencia de reentrenamiento tiene un efecto relevante sobre el desempeño: la frecuencia trimestral obtiene el mayor retorno en siete sectores, la bimestral en tres sectores y la mensual solo en uno. Esto evidencia que no existe una única periodicidad óptima aplicable a todo el mercado.

El principal aporte del estudio consiste en integrar selección de pares por cointegración, predicción LSTM, control de riesgo mediante stop-loss y escalamiento dinámico con PPO dentro de un mismo sistema aplicado a nivel sectorial. Los resultados indican que la política aprendida por el agente, junto con la calibración adaptativa de hiperparámetros, puede mejorar el desempeño del pairs trading clásico. Como recomendación, en futuras investigaciones se podría evaluar el modelo incorporando fricciones de mercado más realistas, como slippage y costos de préstamo para posiciones cortas, además de extender la evaluación a otros mercados e instrumentos financieros.

Palabras clave: *Pairs Trading; Reinforcement Learning; PPO; LSTM; Dynamic Scaling; Cointegración; S&P 500; Walk-forward.*

3 Introducción

El pairs trading es una estrategia de arbitraje estadístico que busca aprovechar desviaciones temporales entre dos activos relacionados. En su formulación clásica, utiliza reglas fijas de entrada y salida; sin embargo, dichos enfoques pueden perder efectividad cuando cambian las condiciones de mercado, la volatilidad o la relación histórica entre los activos. En este contexto, el uso de Reinforcement Learning permite abordar el problema como una decisión secuencial, donde un agente aprende cuándo y cuánto capital asignar.

La presente memoria desarrolla y evalúa un modelo de pairs trading aplicado a sectores del S&P 500, integrando selección de pares por cointegración, un componente predictivo LSTM, optimización de hiperparámetros y un agente PPO con escalamiento dinámico de posiciones. Además, se analiza el efecto de distintas frecuencias de reentrenamiento mensual, bimestral y trimestral, con el fin de determinar si la periodicidad de actualización del modelo influye en su desempeño.

El documento se organiza de la siguiente manera: se presenta el problema de investigación, los objetivos y el alcance del estudio. Luego, se desarrolla el marco teórico, abordando los fundamentos del arbitraje estadístico, pairs trading, cointegración, Reinforcement Learning y LSTM. Posteriormente, se presenta la metodología utilizada, incluyendo el diseño, selección de pares, entrenamiento del agente y el backtest fuera de muestra. Finalmente, se presentan los resultados, conclusiones, discusión, limitaciones y anexos correspondientes.

4 Problema de Investigación

El trading de pares, también conocido como *Pairs Trading* en inglés, es una estrategia de arbitraje estadístico neutral al mercado que busca conseguir rentabilidad aprovechando brechas temporales en los precios de dos activos que presentan una relación histórica de comovimiento cuyo diferencial (*spread*), permite tomar posición larga y posición corta en cada activo del par, asumiendo que este revertirá a su media (Gatev et al., 2006; Vidyamurthy, 2004). En enfoques basados en la cointegración, esta expectativa se fundamenta en que el *spread* cointegrado presente un comportamiento estacionario, por lo que sus variaciones respecto a la media pueden interpretarse como oportunidades de arbitraje estadístico (Engle y Granger, 1987; Johansen, 1991; Flori y Regoli, 2021).

Tradicionalmente el trading de pares se ha ejecutado aplicando reglas predefinidas y umbrales fijos (estáticos) para abrir y cerrar posiciones. En la literatura se ha documentado que el rendimiento de este enfoque ha ido disminuyendo significativamente debido a su poca capacidad de adaptación frente a cambios en la volatilidad, cambios en la relación de los activos o inestabilidad del mercado. Como resultado, una estrategia basada en parámetros fijos puede generar operaciones excesivas, cerrar posiciones de forma ineficiente o mantener exposición cuando el *spread* deja de revertir a su media (Do y Faff, 2010).

Esta ineficiencia se vuelve importante en los mercados financieros modernos, donde las relaciones históricas entre activos pueden cambiar repentinamente debido a cambios de régimen, shocks de información o quiebres estructurales. Frente a estos escenarios no basta con identificar pares que se hayan movido de forma similar en el pasado, ya que también es necesario saber cuándo operar, cuándo detener una pérdida y cuánto capital asignar a cada oportunidad. Por ello,

el problema central del trading de pares no se limita a la formación de pares, sino que involucra decisiones secuenciales bajo incertidumbre (Lu et al., 2022).

Para abordar estas limitaciones, la literatura reciente ha incorporado técnicas de aprendizaje por refuerzo profundo (*Deep Reinforcement Learning*) para optimizar las decisiones comerciales. Modelos basados en Redes Q Profundas (*DQN*) han demostrado ser superiores a los métodos tradicionales al establecer umbrales dinámicos. Asimismo, la incorporación de redes de Memoria a Corto y Largo Plazo (*LSTM*) ha permitido extraer información temporal desde series financieras y complementar las señales de convergencia o divergencia del spread.

A pesar de estos avances todavía existe una brecha respecto a la forma en que estos modelos deben actualizarse a lo largo del tiempo. Muchos enfoques entrenan agentes sobre ventanas históricas y luego evalúan su desempeño fuera de muestra, pero no analizan de manera sistemática cómo la frecuencia de reentrenamiento afecta la rentabilidad, el riesgo y la estabilidad de la estrategia. Esto es relevante ya que reentrenar con demasiada frecuencia puede llevar a que el agente reaccione al ruido de corto plazo propio de cada ventana, generando inestabilidad en las decisiones y una operativa excesiva que merma las ganancias mediante costos de transacción. En el extremo opuesto, un reentrenamiento poco frecuente puede dejar al agente operando con una configuración desactualizada durante períodos prolongados, siendo incapaz de adaptarse a cambios de régimen y arrastrando pérdidas.

Dicho esto, esta investigación aborda la interrogante de cómo la frecuencia de reentrenamiento incide en el desempeño de una estrategia pairs trading basada en reinforcement learning para el mercado accionario estadounidense S&P 500. Para ello, se plantea evaluar un sistema que integra formación y selección de pares, predicción mediante LSTM y un agente PPO con capacidad de ajuste dinámico de posiciones, bajo un esquema walk forward. A partir de lo anterior se

plantean las siguientes preguntas de investigación: ¿Puede un agente de Reinforcement Learning con ajuste dinámico, complementado con un oráculo LSTM y gestión de riesgo por stop-loss, superar en rentabilidad a la estrategia clásica de pairs trading en el mercado accionario estadounidense evaluado por sectores?, ¿Presenta dicho agente un perfil de riesgo, medido a través de la desviación estándar y el máximo drawdown, más favorable que el de la estrategia clásica de pairs trading en condiciones fuera de muestra?, ¿Cómo influye la frecuencia de reentrenamiento del sistema (mensual, bimestral, trimestral) en su rentabilidad y riesgo, y existe una granularidad que optimice el desempeño?

5 Objetivos

5.1 Objetivo General

Desarrollar y evaluar un modelo de Reinforcement Learning para pairs trading en el mercado accionario estadounidense, aplicado a nivel sectorial, con el fin de determinar su rentabilidad y riesgo frente a la estrategia tradicional de pairs trading.

5.2 Objetivos Específicos

Construir un sistema de pairs trading que integre selección de pares por cointegración, un oráculo LSTM y un agente de Reinforcement Learning con dynamic scaling y gestión de riesgo por stop-loss, aplicado a sectores del S&P 500.

Evaluar el desempeño del sistema propuesto fuera de muestra en términos de rentabilidad y riesgo, mediante métricas como el retorno total, la desviación estándar y el máximo drawdown, frente a la estrategia clásica de pairs trading.

Analizar el efecto de la frecuencia de reentrenamiento mensual, bimestral y trimestral sobre la rentabilidad y el riesgo de la estrategia mediante validación walk-forward, determinando si la granularidad óptima es consistente entre sectores o depende de las características de cada uno.

6 Alcance

La presente investigación tiene un alcance aplicado y cuantitativo, ya que propone, implementa y evalúa una estrategia de pairs trading basada en Reinforcement Learning, predicción LSTM y escalamiento dinámico de posiciones. Se busca evaluar el desempeño de una estrategia algorítmica bajo condiciones fuera de muestra, comparando su rentabilidad y riesgo frente al planteamiento clásico de pairs trading.

Desde la perspectiva del alcance metodológico, la investigación posee un componente descriptivo, comparativo y correlacional. Es descriptiva en cuanto caracteriza el desempeño de la estrategia mediante métricas como retorno, volatilidad, Maximum Drawdown, número de operaciones y activaciones de stop-loss. Es comparativa porque contrasta el agente PPO con un benchmark basado en la estrategia clásica de Gatev et al. (2006), además de comparar distintas frecuencias de reentrenamiento mensual, bimestral y trimestral. El componente correlacional se debe a que se busca analizar cómo la frecuencia de actualización del modelo incide en la rentabilidad, el riesgo y la estabilidad de la estrategia.

El estudio se desarrolla sobre acciones pertenecientes al índice S&P 500, agrupadas según la clasificación sectorial GICS. Por lo tanto, el universo de inversión se restringe a un mercado accionario desarrollado, líquido y ampliamente utilizado en la literatura financiera. La evaluación se realiza a nivel sectorial, evitando la formación de pares entre activos de distintos sectores, con el objetivo de trabajar sobre relaciones económicas más comparables y reducir la exposición a dinámicas intersectoriales heterogéneas.

En cuanto a sus delimitaciones, la investigación considera precios diarios ajustados de acciones estadounidenses, por lo que no incluye otros mercados ni instrumentos financieros como

criptomonedas, divisas, commodities o futuros. Además, la estrategia se evalúa bajo esquema de backtest, por lo que no incorpora ejecución en tiempo real, restricciones operativas de mercado ni impacto de liquidez. Los costos de transacción se aplican según la tarifa de un broker de referencia (Fintual).

Respecto de la formación de pares, solo se considera cointegración evaluada mediante Engle-Granger y Johansen. Si bien en la literatura se mencionan criterios como correlación o distancia mínima entre precios normalizados, estos se utilizan principalmente como referencia teórica. Así, el análisis se enfoca en pares cuyo spread presenta una relación de equilibrio con reversión a la media.

Los resultados obtenidos se evalúan dentro del horizonte temporal, frecuencia de datos, universo de activos y supuestos definidos. Debido a esto, las conclusiones no buscan establecer una regla universal para todos los mercados o instrumentos financieros, sino evaluar si, bajo las condiciones consideradas, un agente PPO con asignación de capital y señales LSTM puede mejorar el desempeño del pairs trading clásico, y si la frecuencia de reentrenamiento afecta dicho desempeño de manera relevante.

7 Marco Teórico

7.1 Arbitraje estadístico

El arbitraje estadístico corresponde a un conjunto de estrategias cuantitativas que buscan aprovechar las brechas temporales de precios entre activos financieros mediante herramientas estadísticas. A diferencia del arbitraje puro cuyo objetivo es capturar discrepancias de precios para obtener ganancias prácticamente libres de riesgo, el arbitraje estadístico se basa en relaciones probabilísticas. Debido a esto, las oportunidades no están libres de riesgo ya que la rentabilidad esperada depende de la formación de patrones históricos como la reversión a la media o el movimiento conjunto de activos, que se mantengan durante el periodo de inversión. En consecuencia, su aplicación requiere una adecuada calibración de reglas, control de riesgo y actualización de los modelos (Pole, 2007).

Desde la perspectiva general, estas estrategias intentan construir posiciones entre activos comparables, cuando el precio de un activo se aleja temporalmente del valor sugerido por otro activo o por un portafolio que mantiene características similares, el inversionista puede tomar una posición larga en el activo que esté subvalorado y una posición corta en el activo sobrevalorado esperando que el diferencial revierta a su nivel de equilibrio, la estrategia genera beneficios a partir de la convergencia de precios, independiente de la dirección general del mercado (Gatev et al., 2006).

Una característica importante del arbitraje estadístico es su relación con la neutralidad al mercado. Al combinar posiciones largas y cortas, estas estrategias buscan reducir la exposición a movimientos generales del mercado y concentrar la fuente de retorno en los desfases relativos entre activos (Pole, 2007). En esta misma línea, Avellaneda y Lee (2010) señalan que una forma

de lograr esta neutralidad es descontar el efecto de factores comunes, como el mercado o el sector, para operar sobre desviaciones propias de los activos. De esta manera, la rentabilidad esperada no depende principalmente de anticipar la dirección del mercado, sino de que dichas desviaciones relativas se corrijan en el tiempo.

Pese a lo antes mencionado, el arbitraje estadístico no está libre de riesgo, la relación histórica entre los activos puede debilitarse, el diferencial de precios puede continuar ampliándose en vez de revertir, o los costos de transacción pueden consumir una parte considerable de la rentabilidad esperada. Además, dado que estas estrategias se apoyan de evidencia estadística obtenida de datos históricos, existe el riesgo de sobreajuste, inestabilidad de parámetros y pérdida de validez ante cambios de régimen. Por esta razón, una estrategia de arbitraje estadístico debe resolver tres elementos fundamentales: la selección de activos o portafolios comparables, la obtención de señales que indiquen desviaciones temporales y la definición de una regla de asignación que determine cómo operar esas señales considerando restricciones, riesgo y costos de mercado (Guijarro-Ordonez et al., 2022).

Dentro de este marco, el pairs trading puede entenderse como una forma particular de arbitraje estadístico. En lugar de construir carteras complejas de referencia, esta se enfoca en dos activos que presentan una relación histórica de comovimiento. Cuando el diferencial entre ambos se amplía, se abre una posición relativa esperando que dicho diferencial vuelva a su equilibrio. Debido a esto, el pairs trading representa una de las aplicaciones más conocidas y estudiadas del arbitraje estadístico.

7.2 Pairs Trading

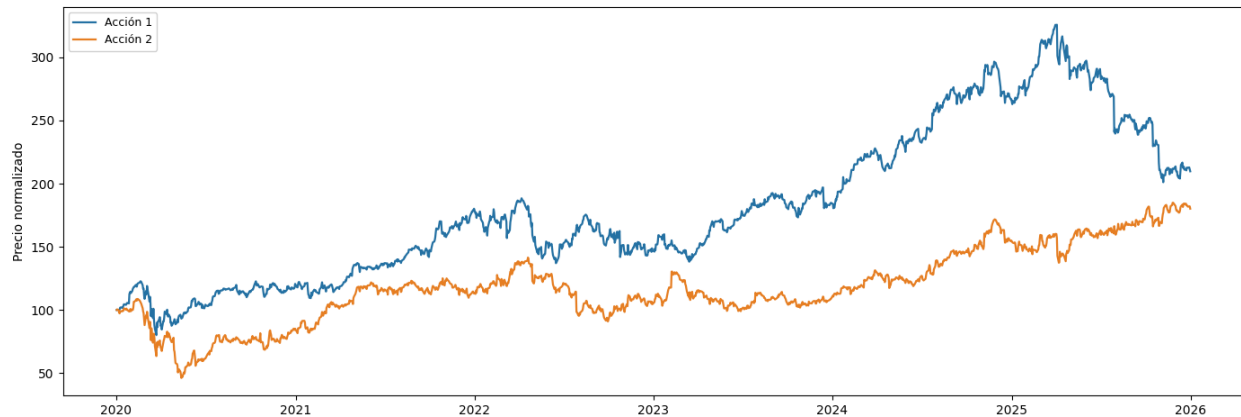
Como se mencionó previamente el pairs trading es una forma de arbitraje estadístico, la cual se basa en identificar dos activos financieros, tales como acciones, ETFs, futuros, divisas o criptomonedas, que han mostrado una relación histórica de comovimiento. Cuando esta relación se altera temporalmente y el diferencial (*spread*) de precios entre ambos activos se amplía, es decir diverge, la estrategia consiste en tomar una posición larga en el activo que se encuentra subvalorado, con la expectativa de que suba de precio y una posición corta en el activo sobrevalorado, vendiendo el activo prestado con la expectativa de que su precio disminuya. Posteriormente, si el spread vuelve a su relación histórica, a su media, las posiciones se cierran y se obtiene ganancia debido a la convergencia del par (Gatev et al., 2006).

Gatev et al. (2006) plantean que este tipo de arbitraje cumple un rol relevante dentro del funcionamiento del mercado, ya que los agentes que explotan desviaciones temporales también ayudan a mantener la eficiencia de primer orden del mercado. Esta idea conecta con la *Law of One Price*, la cual dice que con activos de pagos similares deberían tener precios equivalentes o cercanos. En otras palabras, las ganancias del arbitrajista provienen de identificar diferencias relativas entre activos similares, pero al mismo tiempo su actividad ayuda a que esas diferencias tiendan a desaparecer. Desde esta perspectiva, se puede decir que el pairs trading no solo puede entenderse como una estrategia de inversión, sino también como un mecanismo que favorece la eficiencia en los precios.

En términos operativos, la estrategia requiere definir al menos tres elementos. Primero, se debe seleccionar un par de activos con una relación histórica suficientemente estable. Luego, se debe construir una medida del spread entre ambos mediante z-score. Finalmente, se deben estable-

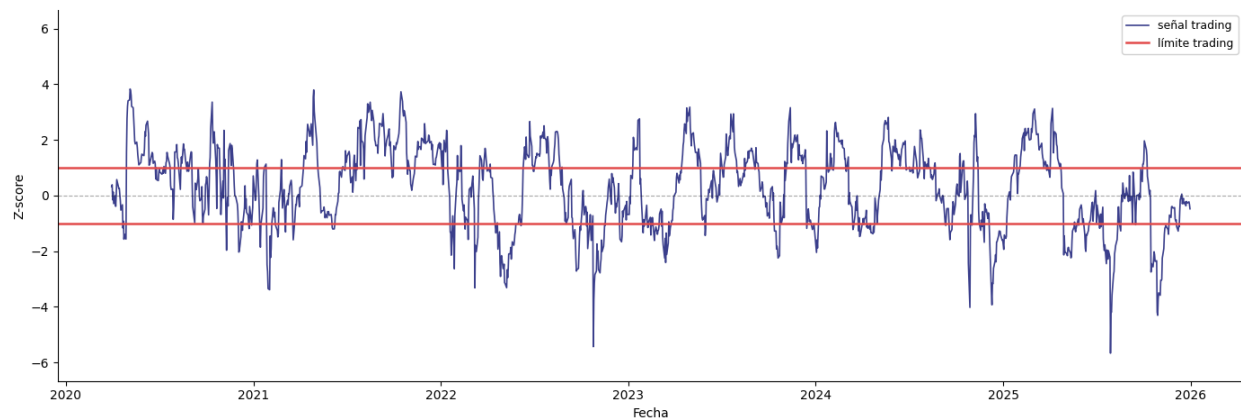
cer reglas para abrir y cerrar posiciones, considerando que la convergencia no está garantizada y que la relación histórica entre los activos puede debilitarse en el tiempo.

Figura 1: Movimiento de precios normalizados de un par de acciones



Nota: Elaboración propia

Figura 2: Señal de z-score en pairs trading



Nota: Elaboración propia

La Figura 1 muestra el comportamiento de precios de dos activos relacionados en el tiempo, mientras que la Figura 2 muestra la lógica operativa del pairs trading mediante el z-score del spread de dicho par. La línea azul representa la señal de trading, mientras que las líneas rojas indican los umbrales de apertura de posiciones. Cuando la señal supera el límite superior o inferior, se

interpreta que existe una divergencia relevante entre los activos y se abre una posición apostando a la convergencia del par.

7.3 Formación de pares

La formación de pares corresponde a una etapa fundamental del pairs trading, ya que determina qué activos serán considerados como candidatos para operar. El objetivo es identificar instrumentos financieros que hayan mostrado una relación histórica suficientemente estable, de modo que una desviación temporal entre ellos pueda interpretarse como una posible oportunidad de arbitraje estadístico. La literatura ha propuesto distintos criterios para esta tarea, entre los que destacan la correlación, la distancia entre precios normalizados y la cointegración (Vidyamurthy, 2004; Lu et al., 2022).

El primer criterio consiste en analizar la correlación entre los activos, para esto se usa la correlación de Pearson que mide el grado de asociación lineal entre las dos series usando retornos, permite identificar aquellos pares que tienden a moverse en la misma dirección. Para dos activos i y j , este criterio se expresa como:

$$\rho_{ij} = \frac{\text{Cov}(r_i, r_j)}{\sigma_{r_i} \sigma_{r_j}}, \quad (1)$$

donde r_i y r_j representan los retornos de los activos, $\text{Cov}(r_i, r_j)$ corresponde a la covarianza entre ambos, y σ_{r_i} y σ_{r_j} son sus desviaciones estándar. Un valor cercano a 1 indica que ambos tienden a moverse en la misma dirección, mientras que un valor cercano a -1 indica movimientos opuestos. Sin embargo, una alta correlación no garantiza que exista una relación en el largo plazo, ya que dos activos pueden presentar retornos altamente correlacionados durante un periodo y

posteriormente sus precios pueden divergir por un largo tiempo. Debido a esto la correlación se ha usado en la literatura como filtro inicial pero no como criterio suficiente para definir los pares.

Otro enfoque clásico utilizado para la formación de pares corresponde al método de distancia utilizado por Gatev et al. (2006). En este, los pares se seleccionan buscando aquellos cuyas trayectorias de precios normalizados sean lo más similares durante un periodo de formación. La idea es crear un índice de rendimiento total acumulado para las acciones durante el periodo de formación, en base a este comparar el recorrido histórico de los precios de los activos y escoger aquellos emparejamientos que presenten la menor separación acumulada, es decir, aquellos que minimicen la suma de las desviaciones al cuadrado entre las dos series de precios normalizados:

$$D_{ij} = \sum_{t=1}^T (\bar{P}_{i,t} - \bar{P}_{j,t})^2, \quad (2)$$

donde $\bar{P}_{i,t}$ y $\bar{P}_{j,t}$ corresponden a los precios normalizados de los activos i y j en el periodo t . Bajo este criterio mientras menor sea D_{ij} , mayor será la similitud histórica entre ambos activos; por lo tanto, el par seleccionado será aquel que minimice esta distancia durante el periodo de formación. Este criterio es simple y permite identificar activos que se hayan comportado de forma parecida, pero no necesariamente asegura que el diferencial (spread) entre ellos sea estacionario.

Debido a lo antes mencionado, la literatura ha utilizado la cointegración como criterio para formar pares. Esta permite evaluar si dos series de precios no estacionarias mantienen una relación de equilibrio en el largo plazo. En términos formales, si dos series de precios $P_{i,t}$ y $P_{j,t}$ son integradas de orden uno, es decir, $I(1)$, pero existe una combinación lineal entre ellas que es estacionaria, $I(0)$, entonces ambas series se consideran cointegradas (Engle y Granger, 1987; Johansen, 1991). Esta propiedad es altamente relevante para el pairs trading, ya que esto implica

que el spread es estacionario, y esto significa que las desviaciones respecto a su media tienden a ser temporales.

La representación de esta serie resultante tiende a realizarse con el spread:

$$s_t = P_{i,t} - \beta P_{j,t}, \quad (3)$$

donde β corresponde al coeficiente de cobertura que tiene $P_{j,t}$ sobre $P_{i,t}$. Si el spread es estacionario, entonces el diferencial entre ambos activos fluctúa alrededor de un nivel de equilibrio; esto puede interpretarse como una relación de ajuste, es decir, cuando un activo se aleja demasiado del otro, el spread tiende a corregirse en el tiempo, lo que es coherente con la lógica de reversión a la media del arbitraje estadístico.

Para la aplicación de cointegración se tienen dos métodos bastante conocidos:

- El test de Engle y Granger (1987) es uno de los más utilizados, su aplicación se basa en dos etapas. Primero se estima una regresión entre los precios de ambos activos:

$$P_{i,t} = \alpha + \beta P_{j,t} + \varepsilon_t. \quad (4)$$

Luego, se analiza si los residuos ε_t de la regresión son estacionarios; si estos no presentan raíz unitaria, se concluye que la combinación lineal entre ambos precios es estacionaria, por lo tanto, existe cointegración entre los activos. En este contexto, los residuos representan la desviación del par respecto al punto de equilibrio, es decir, este residuo constituye el spread sobre el cual se construyen las señales de trading

- De forma complementaria se tiene el test presentado por Johansen (1991) el cual permite

evaluar un sistema multivariado. Contrario a Engle-Granger, este se basa en transformar el vector autorregresivo (VAR) a un modelo de corrección de errores vectoriales (VECM) y permite identificar en un rango cuántas relaciones de largo plazo existen en el sistema que estén cointegradas. Para el caso de un par de activos, si hay al menos un vector de cointegración, significa que existe una relación de equilibrio entre las series. Cabe destacar que este método de Johansen trata a las variables de manera conjunta y no exige definir una variable como dependiente de la otra.

El test se apoya en el estadístico de traza y estadístico de máximo eigenvalue:

$$\lambda_{\text{traza}}(r) = -T \sum_{i=r+1}^n \ln(1 - \hat{\lambda}_i) \quad (5)$$

$$\lambda_{\text{max}}(r, r + 1) = -T \ln(1 - \hat{\lambda}_{r+1}) \quad (6)$$

En estas expresiones, T corresponde al número de observaciones, $\hat{\lambda}_i$ son los eigenvalues estimados ordenados de mayor a menor, y r representa el número de relaciones de cointegración bajo la hipótesis nula. El estadístico de traza evalúa la hipótesis nula de que existen como máximo r relaciones de cointegración frente a la alternativa de que existe un número mayor, acumulando la información de todos los eigenvalues desde $\hat{\lambda}_{r+1}$ en adelante. Por su lado, el estadístico de máximo eigenvalue contrasta la hipótesis de r relaciones de cointegración frente a la alternativa específica de $r + 1$, evaluando únicamente el eigenvalue siguiente $\hat{\lambda}_{r+1}$. En ambos casos, un valor del estadístico superior al valor crítico correspondiente conduce a rechazar la hipótesis nula, indicando la presencia de al menos una relación de cointegración

adicional. En este trabajo se utiliza el estadístico de traza para el caso bivariado, con el objetivo de evaluar si existe una relación de cointegración entre los dos activos que conforman cada par.

Por lo tanto, para la formación de pares se busca identificar activos que presenten una relación histórica suficientemente estable para justificar una estrategia de convergencia; para el caso de esta investigación se aplica el criterio de cointegración ya que entrega una base econométrica más directa para sostener la existencia de un spread estacionario.

7.4 Estrategia clásica

La estrategia clásica de pairs trading se asocia principalmente al trabajo de Gatev et al. (2006), quienes proponen la formación de pares mediante distancia mínima entre precios normalizados. Bajo este enfoque, la estrategia busca capturar desviaciones temporales entre activos históricamente similares, bajo la expectativa de que sus precios vuelvan a converger.

En esta literatura, los pares se forman durante una ventana de doce meses y se negocian durante los seis meses siguientes. La posición se abre cuando los precios normalizados divergen en más de dos desviaciones estándar históricas, si el diferencial vuelve a cerrarse la posición se liquida, en caso opuesto, esta permanece abierta hasta el final del periodo de trading, el cual es el momento en que todas las posiciones cierran (Gatev et al., 2006).

El estudio evalúa portafolios formados por distintos grupos de pares, desde los 5 a 20 pares con menor distancia histórica, tras pasar el periodo de test sus resultados indicaron que generó cerca de 11 % en retornos anualizados para portafolios autofinanciados de pares, esto considerando que los retornos superan estimaciones conservadoras de costos de transacción. Esto posicionó a Gatev et al. (2006) como uno de los principales referentes empíricos de la literatura de pairs trading.

Sin embargo, esta formulación clásica presenta limitaciones, ya que utiliza umbrales fijos y no incorpora mecanismos para detener pérdidas en caso de que el par rompa su relación, lo cual puede resultar en un desempeño deficiente en los contextos que la relación se quiebre.

Gatev et al. (2006) concluyen que las ganancias del pairs trading provienen del mis-pricing temporal de sustitutos cercanos identificados por el criterio de mínima distancia. Asimismo, relacionan la estrategia con la presencia de un factor común en los retornos de los pares, que es distinto de las medidas de riesgo convencionales. Finalmente, muestran que las ganancias son robustas ante estimaciones conservadoras de los costos de transacción.

7.5 Reinforcement Learning (RL)

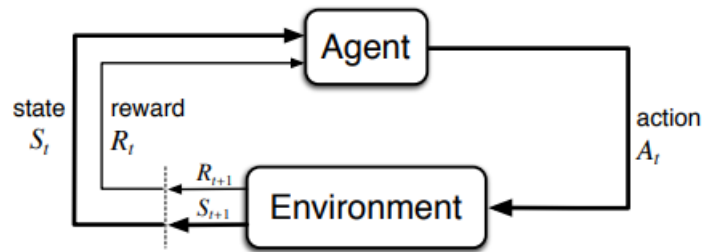
El Reinforcement Learning es una rama del aprendizaje automático (*Machine Learning*) que está enfocada a problemas de decisión secuencial. A diferencia de los enfoques supervisados en donde el modelo aprende a partir de ejemplos etiquetados proporcionados por un externo, en RL un agente aprende mediante la interacción con un entorno. Para cada periodo el agente interactúa con el entorno seleccionando una acción y recibe una recompensa asociada a la consecuencia de dicha acción, a partir de esta retroalimentación el objetivo es aprender una política que maximice la recompensa acumulada en el tiempo (Sutton y Barto, 2018).

Esta estructura es bastante útil en contextos financieros, en especial al tratar con series de tiempo, ya que las estrategias de inversión actúan bajo la incertidumbre del mercado. En una estrategia de pairs trading, el agente debe decidir si abrir, mantener o cerrar la posición considerando su consecuencia en el resultado de la estrategia. Debido a esto el problema no es solo de anticipar el movimiento de los precios, sino también cómo actuar frente a la información disponible.

Un problema de RL suele representarse como un proceso de decisión de Markov en donde

se define un conjunto de estados S , un conjunto de acciones A , una función de transición entre estados que en el contexto de mercados es desconocida y estocástica, una función de recompensa y un factor de descuento γ ; con esto el agente busca aprender una política ($\pi(a|s)$) que determine la probabilidad de seleccionar una acción a dado un estado s (Sutton y Barto, 2018).

Figura 3: Proceso interacción agente-entorno



Nota: Gráfico tomado desde (Sutton y Barto, 2018)

La recompensa es la que cumple el rol central, ya que define el criterio mediante el cual el agente evalúa sus decisiones. En aplicaciones financieras esta puede construirse a partir del retorno de la estrategia, considerando penalizaciones por costos de transacción, pérdidas o exceso de operaciones. De esta manera la función de recompensa condiciona el comportamiento aprendido por el agente.

El objetivo del agente es maximizar la recompensa, cuyo planteamiento general es:

$$G_t = \sum_{k=0}^T \gamma^k R_{t+k+1}, \quad (7)$$

donde G_t corresponde al retorno acumulado desde el periodo t , R_{t+k+1} representa la recompensa futura y $\gamma \in [0, 1]$ es el factor de descuento. Un valor alto de γ otorga mayor importancia a recompensas futuras, mientras que un valor bajo prioriza resultados más inmediatos. La forma en que el agente aprende la política que maximiza G_t depende del algoritmo de Reinforcement Lear-

ning empleado; en este trabajo se utiliza Proximal Policy Optimization (PPO), cuyas características se detallarán más adelante.

7.6 Algoritmos RL relevantes

En el Reinforcement Learning se agrupan familias de algoritmos que difieren en su estrategia para aprender una política de decisión, entre las más relevantes se tiene: los métodos que estiman el valor esperado de cada acción y los que optimizan directamente la regla de comportamiento (Arulkumaran et al., 2017). La elección algorítmica depende directamente de las decisiones que el agente deba tomar, es decir, si el agente se enfrenta a un espacio de acciones discreto como comprar o vender, o a uno continuo como definir la fracción de capital a invertir en una posición.

Una de las arquitecturas representativas dentro de los métodos basados en valor es Deep Q-Network (DQN), la cual usa las redes neuronales para aproximar la función de valor acción $Q(s, a)$. Esta función evalúa qué tan conveniente resulta ejecutar una acción a en un estado s , asumiendo que se seguirá una política determinada. En pairs trading el DQN ha sido aplicado en la literatura por Kim y Kim (2019) en donde demostraron su utilidad al optimizar entre niveles de entrada de posición, mostrando que es efectivo y es capaz de asimilar oportunidades de convergencia, permitiendo obtener ganancias.

Sin embargo, esta formulación discreta presenta una limitación en el contexto financiero. Como señalan Yang y Malik (2024), la mayoría de las aplicaciones de pairs trading invierten un monto fijo o la totalidad del portafolio y omiten el componente de la asignación de capital según la calidad de la oportunidad. Por lo tanto, restringir las decisiones a entrar o salir impide determinar cuánto invertir, por lo que los métodos basados en política resultan más adecuados para incorporar la dimensión de asignación de capital ya que operan en espacios de acción continuos.

Para suplir esta limitación destaca el método Proximal Policy Optimization (PPO) propuesto por Schulman et al. (2017). PPO pertenece a la familia de métodos de gradiente de política, en donde ajusta iterativamente los parámetros de la red neuronal para maximizar la recompensa esperada. La razón principal de su alta estabilidad radica en el diseño de su función objetivo, la cual restringe matemáticamente la magnitud de las actualizaciones de la política para evitar cambios destructivos en la política entre las iteraciones.

Tal como propone Schulman et al. (2017), PPO establece un ratio de probabilidad $r_t(\theta)$ que mide la divergencia entre la política actual (π_θ) y la anterior ($\pi_{\theta_{old}}$) al tomar una acción a_t en el estado s_t

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \quad (8)$$

Para evitar actualizaciones desproporcionadas, el algoritmo optimiza una función objetivo recortada (*clipped surrogate objective*) que limita la variación del ratio de probabilidad mediante un margen de recorte ϵ . Su lógica se desglosa en tres operadores matemáticos que la componen:

1. *Esperanza empírica* ($\hat{\mathbb{E}}_t$): Calcula el promedio esperado sobre un lote de trayectorias en el instante de tiempo t , permitiendo que el agente aprenda de un conjunto representativo de experiencias y no de una sola operación aislada.
2. *Función de recorte* (*clip*): Acota estrictamente el ratio de probabilidad $r_t(\theta)$ dentro del intervalo cerrado $[1 - \epsilon, 1 + \epsilon]$, neutralizando matemáticamente los saltos bruscos.
3. *Operador mínimo* (*mín*): Compara el objetivo original (sin recortar) con el objetivo recortado, seleccionando siempre el menor valor entre ambos, garantizando una evaluación “pesi-

mista” del rendimiento de la política.

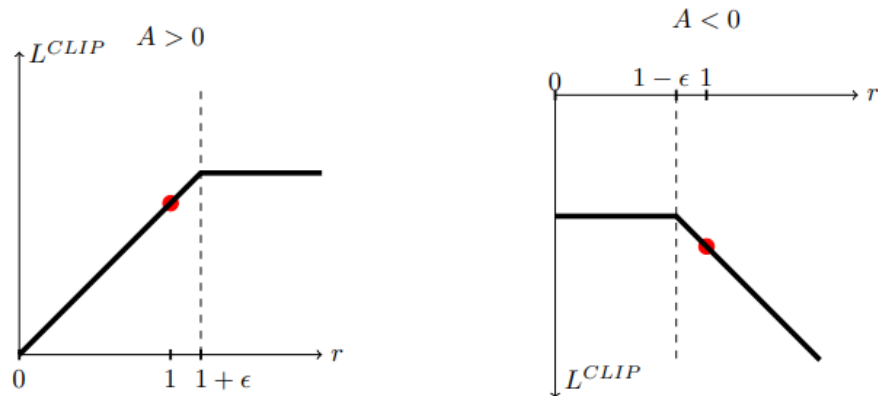
Integrando estos componentes, la función objetivo formal se define como:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t) \right] \quad (9)$$

En donde $\hat{A}_t$ representa la estimación de la función de ventaja, la cual evalúa si una acción resulta mejor o peor de lo esperado; ϵ define el margen de recorte. Por lo que la función actúa de forma asimétrica dependiendo de la calidad de la acción (Schulman et al., 2017):

- Si la ventaja es positiva ($\hat{A}_t > 0$): La acción fue mejor de lo esperado y el ratio $r_t(\theta)$ se limita a un máximo de $1 + \epsilon$. Esto evita que el modelo se vuelva excesivamente optimista y asuma posturas extremas ante una sola buena señal.
- Si la ventaja es negativa ($\hat{A}_t < 0$): La acción fue peor de lo esperado y el ratio $r_t(\theta)$ se limita a un mínimo de $1 - \epsilon$. Esto impide que la política penalice y destruya por completo un comportamiento aprendido debido a una mala experiencia aislada.

Figura 4: Comportamiento según ventaja A_t



Nota: Gráfico tomado desde (Schulman et al., 2017)

Al tomar el mínimo entre la versión original y la recortada, PPO plantea un límite inferior pesimista, el cual asegura que la política no se aleje abruptamente de su versión previa, incluso si las señales de recompensa son extremas.

Esta propiedad es la que vuelve al PPO el adecuado para el modelo propuesto en esta investigación, ya que al operar con series de tiempo financieras que presentan mucho ruido, la restricción de actualizaciones bruscas reduce la posibilidad de ajustes excesivos ante eventos aislados como lo pueden ser falsos quiebres de mercado. Además, al emplear el PPO, la estrategia pasa de reglas estáticas a un esquema de escalado dinámico, por lo que el agente puede ajustar dinámicamente la exposición al par basándose en el estado del spread. Finalmente, dado que su arquitectura permite operar recopilando trayectorias en ventanas de tiempo, PPO es compatible con esquemas que incorporan información secuencial, lo cual es funcional para este trabajo que se complementa con LSTM.

7.7 RL aplicado a Pairs Trading

La aplicación de Reinforcement Learning en pairs trading surge como una respuesta a algunas limitaciones de la estrategia clásica. Particularmente, las reglas tradicionales dependen de umbrales fijos de entrada y salida, además de asumir que la relación histórica de los activos se mantendrá durante el periodo de operación; sin embargo, ante situaciones en las que el spread no converge o cuando la relación entre los activos se debilita, una regla fija puede generar pérdidas de oportunidades y/o pérdidas de capital. En este contexto, el RL permite plantear el pairs trading como un problema de decisión secuencial, en donde el agente decide cuándo operar, no operar y cómo reaccionar ante cambios en el estado del par.

Algunos aportes de RL en pairs trading son:

- Uno de los aportes relevantes en este punto corresponde a Kim y Kim (2019), quienes proponen una estrategia de pairs trading basada en Deep Reinforcement Learning. Los autores utilizan Deep Q-Network (DQN) para optimizar las decisiones del agente a partir del comportamiento del spread entre dos activos. Su propuesta, denominada Pairs Trading Deep Q-Network (PTDQN), busca maximizar las recompensas acumuladas en el largo plazo, en vez de depender únicamente de reglas fijas de entrada y salida.

Un aspecto distintivo de este enfoque es la incorporación explícita de trading y de stop-loss, ya que mientras los límites de entrada permiten identificar cuándo el spread se ha alejado lo suficiente de su media para abrir una posición, el *stop-loss boundary* define un nivel máximo de divergencia tolerado antes de cerrar la operación, esto debido a que el modelo reconoce que la convergencia del par no está garantizada y prefiere liquidar para evitar que la pérdida siga aumentando. Este aporte es relevante para la presente investigación debido a que introduce el control de riesgo dentro de la lógica de aprendizaje.

En términos experimentales, evalúan su enfoque sobre pares seleccionados del S&P 500 y consideran distintas configuraciones de límites de trading y stop-loss. Sus resultados muestran que el uso de DQN permite ajustar la estrategia de forma más flexible que una regla tradicional, ya que el agente aprende a asociar estados del spread con decisiones de operación y recompensas futuras.

- Otro aporte relevante corresponde a la incorporación de quiebres estructurales en el pairs trading por Lu et al. (2022), quienes plantean que una relación de cointegración aparentemente estable puede romperse ante cambios abruptos del mercado, afectando la validez de una estrategia basada en convergencia.

Los autores desarrollan un marco de trabajo de dos fases denominado Structural Break-Aware Pairs Trading (*SAPT*), diseñado para operar pares de acciones en un contexto intradiario en el mercado de Taiwán (TAIEX). En la fase 1, el modelo combina dos redes neuronales, CNN para obtener las características de las frecuencias y LSTM para obtener patrones, esto con el fin de detectar posibles rupturas estructurales en la relación de precios. En la fase 2, los autores incorporan un agente basado en DQN que toma decisiones de trading considerando no solo el comportamiento del spread, sino también la probabilidad de quiebre estructural, los costos de transacción y el riesgo asociado a mantener posiciones en un mercado intradiario. De esta forma, el agente puede evitar operar o cerrar posiciones en situaciones donde la divergencia del par no represente una oportunidad de convergencia, y sea un posible cambio en la relación estructural de los activos.

- Han et al. (2023) proponen un enfoque de RL recurrente denominado CREDIT (reCurrent Reinforcement lEarning methoD for paIrs Trading), en donde modelan el pairs trading como un entorno parcialmente observable. Este planteamiento busca representar de manera más realista el funcionamiento del mercado financiero, donde el agente no observa toda la información relevante del entorno y debe aprender a partir del historial disponible de precios y su memoria. Para esto los autores incorporan una arquitectura recurrente basada en Bi-GRU junto con un mecanismo de atención temporal, lo que permite capturar patrones de largo plazo en los movimientos de precios de los activos.

Un aspecto relevante de CREDIT es que presenta una recompensa consciente del riesgo, que no solo considera la rentabilidad, sino también el riesgo asumido durante el periodo de trading. En base a esto, el agente aprende una preferencia más robusta, evitando operacio-

nes riesgosas. Los autores muestran que este enfoque permite reducir operaciones excesivas asociadas a altos costos de transacción y mejora el desempeño frente a otros métodos de RL aplicados al pairs trading.

- Por su parte, Yang y Malik (2024) construyen entornos de Reinforcement Learning para determinar cuándo y cómo operar pares en el mercado de las criptomonedas. Su propuesta se basa en que el agente aprenda a tomar decisiones de operación a partir del estado observado del par, considerando la oportunidad de convergencia y la asignación de capital.

Los autores proponen dos agentes: El primero, denominado RL1, está enfocado en optimizar el momento de la operación, es decir, el *timing* y la dirección de la posición, mediante un espacio de acciones discretas como comprar, vender o cerrar. Por otro lado, el segundo agente RL2, presenta el aporte más relevante del estudio, ya que incorpora la lógica de la escala dinámica (*dynamic scaling*). Bajo este enfoque, el agente opera en un espacio de acción continuo, por lo que no solo decide cuándo entrar o salir del mercado, sino también qué proporción de capital asignar a la operación según la oportunidad observada y el nivel de exposición, nótese que indirectamente también calcula el timing del RL1 pero como algo secundario.

Para evaluar ambos enfoques, Yang y Malik (2024) consideran algoritmos de RL como PPO, DQN, A2C y SAC. Sus resultados muestran que el enfoque de dynamic scaling obtiene un desempeño superior al RL1 basado únicamente en acciones discretas. En concreto, el agente RL2 alcanza un retorno de 31,53 % en un mes, frente al 9,94 % obtenido por RL1. Además, los autores muestran que RL2 realiza menos operaciones pero con mayor efectividad, lo que sugiere que la asignación dinámica de capital permite priorizar oportunidades de mayor

calidad en lugar de aumentar simplemente la frecuencia de trading.

Este aporte es relevante para la presente investigación, ya que respalda el uso de acciones continuas para plantear un esquema de exposición dinámica, permitiendo ajustar el tamaño de la posición en función del estado del spread, en vez de operar siempre con una asignación fija de capital.

7.8 LSTM como componente predictivo

Las redes neuronales recurrentes (RNN) fueron diseñadas para procesar datos secuenciales, incorporando la noción de que el estado depende de la información observada previamente. Sin embargo, presentaban una limitación conocida como el problema del gradiente desvaneciente, ya que durante el entrenamiento al mover el gradiente hacia atrás en los pasos temporales, reduce su importancia significativamente y la red pierde la capacidad de retener información de los primeros periodos de la secuencia (Hochreiter y Schmidhuber, 1997).

Para superar esta limitación Hochreiter y Schmidhuber (1997) propusieron la arquitectura Long Short Term Memory (LSTM), la cual consiste en incorporar una celda de memoria que mantiene un resumen del estado pasado y que se conecta a sí misma mediante una arista recurrente de peso fijo, permitiendo que el gradiente atravesase numerosos pasos temporales sin desvanecerse.

En su formulación moderna, el flujo de información hacia y desde la celda de memoria se regula mediante tres compuertas:

- *Compuerta de entrada (input gate)*: Planteada en la arquitectura original (Hochreiter y Schmidhuber, 1997), determina en qué medida la información del período actual se incorpora a la celda de memoria.

- *Compuerta de salida (output gate)*: Determina cuánto del contenido de la celda influye en el estado oculto que se transmite al período siguiente.
- *Compuerta de olvido (forget gate)*: Propuesta posteriormente por Gers et al. (2000), determina qué proporción de la información almacenada previamente se descarta para evitar la saturación del estado.

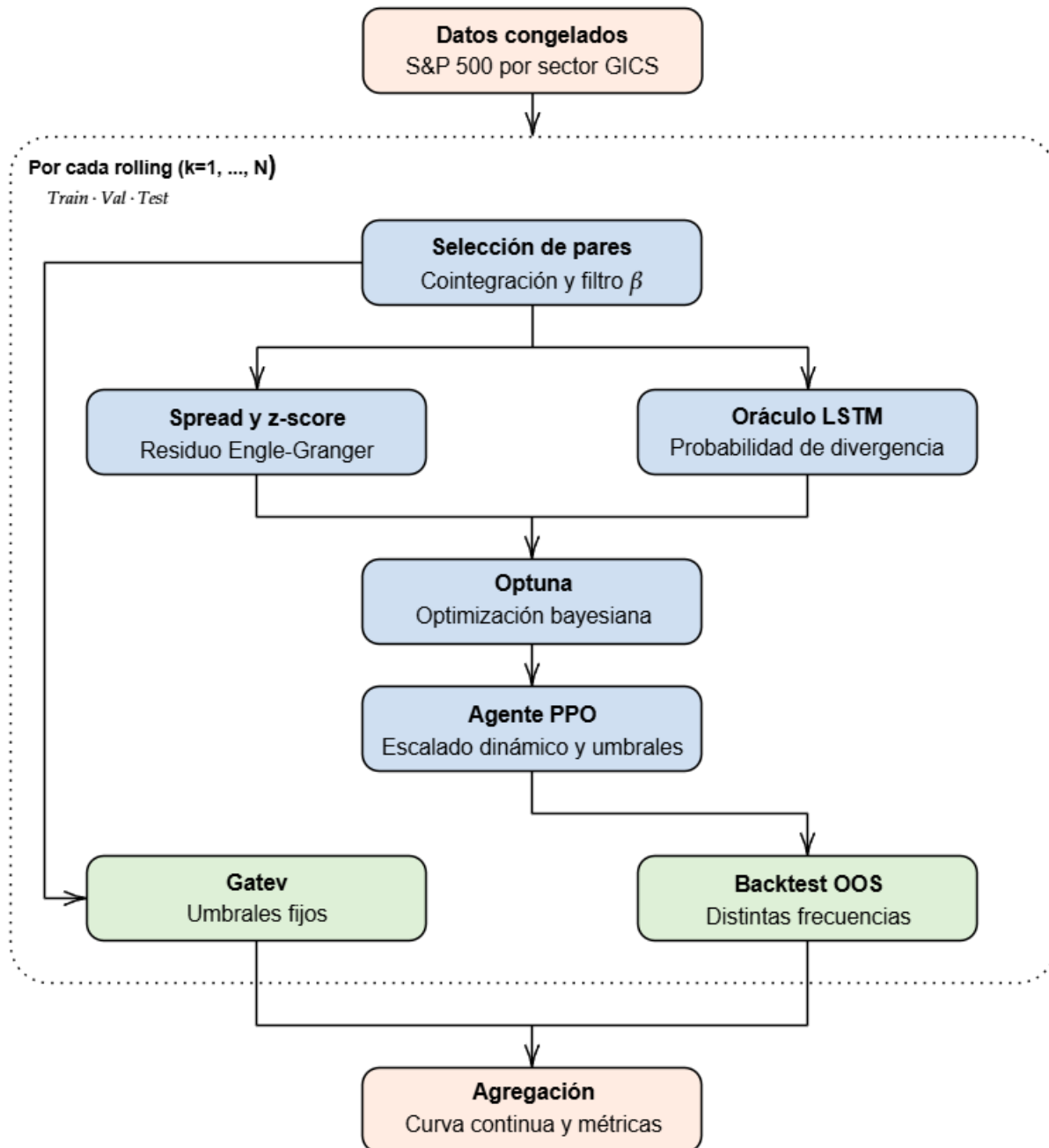
Este planteamiento de compuertas permite que la red aprenda de manera selectiva la información a conservar y cuál descartar a lo largo del tiempo, permitiendo capturar dependencias en el corto y largo plazo.

Esta capacidad del LSTM fue aprovechada por Flori y Regoli (2021), quienes no buscaban predecir precios ni los momentos exactos para abrir posiciones largas o cortas, sino estimar la probabilidad de que la brecha de rendimiento entre un par de activos aumente en el futuro cercano (en un horizonte de 1 a 10 días). El estudio demuestra que la predicción del LSTM no sustituye los umbrales tradicionales, sino que los complementa. Al cruzar la información, el inversor puede distinguir si se encuentra en un escenario de reversión a la media para comprar/vender, o en un escenario de momentum, lo cual indicaría que el spread seguirá divergiendo.

En este trabajo el LSTM se emplea como un componente predictivo externo al agente, es decir, la red se entrena de forma independiente para estimar la probabilidad de que el spread diverja o converja en el periodo siguiente, para luego incorporarlo dentro del espacio de observación del agente.

8 Metodología

Figura 5: Diagrama metodológico del modelo propuesto



Nota: Elaboración propia

8.1 Datos

Para el desarrollo de este trabajo se utilizaron precios diarios ajustados de las acciones pertenecientes al índice bursátil S&P 500, extraídos desde *Yahoo Finance*. La clasificación sectorial de cada activo se obtuvo a partir de un repositorio en GitHub que organiza las acciones según el estándar “Global Industry Classification Standard” (GICS). Los sectores son:

Tabla 1: Sectores considerados según la clasificación GICS.

Sector	
Information Technology	<i>(Tecnología de la Información)</i>
Financials	<i>(Finanzas)</i>
Health Care	<i>(Salud)</i>
Consumer Discretionary	<i>(Consumo Discrecional)</i>
Communication Services	<i>(Servicios de Comunicación)</i>
Industrials	<i>(Industrial)</i>
Consumer Staples	<i>(Consumo Básico)</i>
Energy	<i>(Energía)</i>
Utilities	<i>(Servicios Públicos)</i>
Real Estate	<i>(Bienes Raíces)</i>
Materials	<i>(Materiales)</i>

El universo de inversión se define a nivel sectorial, es decir, cada agente opera únicamente con acciones pertenecientes a un mismo sector GICS, por lo que no se consideran pares cruzados entre sectores. Esta decisión se justifica en que los pares intersectoriales pueden estar expuestos a dinámicas económicas distintas, lo que incrementa la probabilidad de quiebres estructurales y puede debilitar la cointegración.

En cuanto al tratamiento de datos, al momento de la descarga se descartan aquellas acciones que presentan más de un 10 % de datos faltantes sobre el rango histórico completo; este criterio asegura una calidad mínima de las series y evita que la formación de pares se vea afectada por

información incompleta. Los precios corresponden únicamente a días hábiles bursátiles, debido a que la fuente no reporta observaciones para fines de semana ni feriados. Ante datos faltantes puntuales en un activo dentro del histórico, estos no se imputan ya que la construcción de cada par se realiza sobre la intersección de fechas en que ambos activos registran precios, evitando valores artificiales dentro de las series. Una vez descargadas y procesadas las series, almacenan en archivos *pickle*, tal que todas las corridas posteriores utilicen la misma base, lo que mantiene la reproducibilidad de los experimentos y evita variaciones debido a revisiones retroactivas en la fuente de datos.

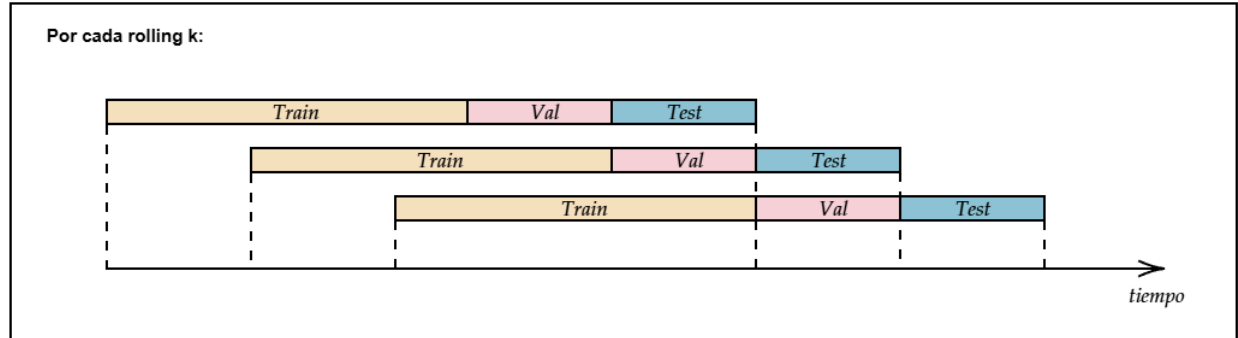
La descarga de datos históricos comienza en 2017, con el objetivo de disponer de información suficiente para las ventanas iniciales de entrenamiento y validación. Sin embargo, la evaluación realizada mediante *walk-forward* se desarrolla desde 2022 hasta 2025, considerando el primer periodo de entrenamiento en 2022 y el último periodo de prueba fuera de muestra (OOS) en 2025.

8.2 Diseño experimental

El diseño experimental aplicado en este trabajo es de tipo “walk-forward”, o validación por avance, en donde se plantea una estructura de rollings mediante ventanas móviles. Cada rolling considera un periodo de 12 meses de entrenamiento (*train*), 3 meses de validación (*validation*) y un periodo posterior de evaluación fuera de muestra (*out-of-sample*, OOS). Estos bloques son contiguos y sin solapes entre las etapas de entrenamiento, validación y prueba, lo cual evita la filtración de información (*data leakage*) entre etapas. En cambio, entre rollings sucesivos las ventanas de entrenamiento y validación se desplazan y comparten observaciones, mientras que los periodos de prueba son disjuntos y cubren de forma continua el horizonte fuera de muestra sin repetir observaciones.

La Figura 6 muestra cómo las ventanas se desplazan en el tiempo; cada rolling avanza una longitud igual a su período de prueba, de modo que los test sucesivos cubren el horizonte fuera de muestra sin superponerse.

Figura 6: Movimiento de rolling en el tiempo



Nota: Elaboración propia. Esquema referencial, no a escala

Añadido a esto, se evalúan tres frecuencias de reentrenamiento. En el primer caso se considera una frecuencia mensual, lo que implica 24 rollings para cubrir dos años de evaluación OOS; en el segundo caso se utiliza una frecuencia bimestral, equivalente a 12 rollings; mientras que en el tercer caso se emplea una frecuencia trimestral, correspondiente a 8 rollings. En todos los casos se cubre el mismo horizonte fuera de muestra (enero de 2024 a diciembre de 2025), lo que permite comparar directamente el retorno total y la desviación estándar entre frecuencias, al estar medidos sobre el mismo período calendario.

Tabla 2: Configuraciones de reentrenamiento evaluadas

Frecuencia	Train	Validación	Test	Rollings OOS
Mensual	12 meses	3 meses	1 mes	24
Bimestral	12 meses	3 meses	2 meses	12
Trimestral	12 meses	3 meses	3 meses	8

En cada rolling se rehace completamente el proceso de modelamiento, incluyendo la se-

lección de pares, el entrenamiento del modelo LSTM, la optimización de hiperparámetros y el entrenamiento del agente RL. De esta forma, la comparación no solo evalúa el desempeño de una estrategia fija, sino también el efecto que tiene la frecuencia de reentrenamiento sobre el desempeño.

Este diseño busca analizar cómo afecta la frecuencia de reentrenamiento al desempeño de cada sector, evitando asumir de manera previa que todos los sectores deben actualizarse con la misma periodicidad. Así, la metodología permite estudiar si existen sectores que se benefician de una actualización más frecuente u otros en los que se obtiene mayor retorno con una frecuencia menor.

8.3 Selección de pares

Para la formación de pares se utiliza el criterio de cointegración, aplicando primero el test de Engle-Granger seguido de Johansen. El objetivo de esta etapa es seleccionar pares de activos que mantengan una relación de equilibrio relativamente estable en el tiempo, de modo que las desviaciones del spread puedan implicar oportunidades de convergencia.

Inicialmente, dentro de cada sector GICS se evalúan todas las combinaciones posibles de pares mediante el test de Engle-Granger sobre las series de precios ajustados. Se considera como primer filtro una significancia del 5 %, por lo que solo se mantienen los pares cuyo resultado permite rechazar la hipótesis nula de no cointegración, es decir, aquellos con $p < 0,05$.

Posteriormente, los pares que superan este primer filtro son evaluados mediante el test de Johansen, utilizando el estadístico de traza con el mismo nivel de significancia de 5 %. Esta segunda prueba se incorpora como una confirmación adicional de la relación de cointegración, ya que permite evaluar la existencia de al menos un vector de cointegración entre las series de precios

consideradas. Este segundo test se aplica debido a que Engle-Granger es un procedimiento que plantea un activo como una variable dependiente y su resultado puede variar según el orden, mientras que Johansen trata ambas series de forma simétrica dentro de un sistema VECM, reduciendo así la posibilidad de falsos positivos.

Una vez confirmada la cointegración, se aplica un filtro sobre el coeficiente de cobertura ($|\beta|$), restringiendo su valor al intervalo $[0,3, 3,0]$. Este criterio busca evitar relaciones de cobertura excesivamente desbalanceadas, ya que un $|\beta|$ elevado implica que una de las patas domina el spread, amplificando su volatilidad y produciendo retornos no realizables durante la operación del agente.

Adicionalmente, se impone una restricción de no reutilización de tickers, lo cual significa que cada activo puede pertenecer como máximo a un solo par dentro de cada rolling, evitando dependencias de varios pares sobre un mismo ticker.

Finalmente, los pares que cumplen con los criterios anteriores son ordenados de forma ascendente según el p valor del test de Engle-Granger seleccionando hasta un máximo de 8 pares por sector y por rolling. Estos pares representan el universo sobre el cual se desarrollará el proceso de entrenamiento, validación y evaluación de la estrategia de pairs trading.

8.4 Construcción del spread

Una vez seleccionados los pares cointegrados se construye el spread, el cual adopta la misma forma funcional que el residuo de la regresión de Engle-Granger. Sin embargo, ambos cumplen roles distintos, mientras el test de cointegración se estima sobre la totalidad del período de entrenamiento y opera como criterio de selección de pares, el spread operativo se recalcula sobre una ventana móvil de M días. El tamaño de dicha ventana se define como hiperparámetro

dentro del proceso de optimización bayesiana realizado con Optuna; cabe aclarar que Optuna no estima los coeficientes de la regresión, sino que selecciona la configuración de ventana con mejor desempeño en la etapa de validación.

Los parámetros α_t y β_t no se estiman una única vez, sino que se recalculan en cada período t mediante mínimos cuadrados sobre los M días previos. De esta manera la relación de equilibrio se adapta a las condiciones recientes del mercado y debido a que tanto los coeficientes como los estadísticos de la ventana se construyen exclusivamente con información anterior a t , no se utiliza información futura en su cálculo.

La relación entre ambos activos se expresa como:

$$P_{1,t} = \alpha_t + \beta_t P_{2,t} + \varepsilon_t \quad (10)$$

Donde $P_{1,t}$ y $P_{2,t}$ corresponden a los precios ajustados de los activos que conforman el par en el periodo t , α_t y β_t corresponden respectivamente al intercepto y al coeficiente de cobertura estimados en la ventana $[t - M, t)$ y ε_t es el residuo de la relación estimada. A partir de esto, el spread se define como el residuo:

$$s_t = P_{1,t} - (\alpha_t + \beta_t P_{2,t}) \quad (11)$$

Luego, para obtener una señal adimensional y comparable entre pares con distintas escalas de precios, el spread se estandariza mediante el z-score. Esta medida permite evaluar qué tan alejado se encuentra el diferencial respecto de su comportamiento reciente expresado en unidades de desviación estándar de la ventana, cuyo cálculo es:

$$z_t = \frac{s_t - \mu_{\text{ventana}}}{\sigma_{\text{ventana}}} \quad (12)$$

Donde μ_{ventana} y σ_{ventana} representan, respectivamente, la media y la desviación estándar del spread dentro de la ventana móvil. De esta forma, el spread representa la desviación económica entre ambos activos y determina el resultado monetario de las posiciones, mientras que el z-score transforma dicha desviación en una señal normalizada que puede ser utilizada para identificar oportunidades de entrada, salida o ajuste de posición dentro de la estrategia.

8.5 Oráculo LSTM

El agente de RL observa el mercado a través de variables que describen principalmente el estado presente del par, como el z-score o la zona vigente, incorporando únicamente un período de memoria mediante la variación del z-score respecto al período anterior. Dicha variación informa si la desviación se amplía o se reduce, pero no permite reconocer patrones que se despliegan a lo largo de varias semanas. Para complementar dicha información se incorpora un componente predictivo basado en una red LSTM, cuya función es comprimir el comportamiento de los últimos treinta períodos en una única señal que el agente recibe como observación adicional.

La arquitectura se compone de una capa LSTM de 35 unidades, seguida de una capa de normalización por lotes (*batch normalization*) y una capa densa de salida con dos neuronas y función de activación softmax, que entrega una distribución de probabilidad sobre dos clases, definidas según si el diferencial de retornos logarítmicos aumenta (clase 1) o no (clase 0) en el período siguiente. El entrenamiento se realiza con el optimizador RMSprop que es adecuado para redes recurrentes por su capacidad de adaptar la tasa de aprendizaje durante la optimización.

Las variables de entrada corresponden a tres características construidas a partir de la información histórica del par: el diferencial de retornos logarítmicos entre ambos activos, el volumen transado del primer activo y el retorno logarítmico del primer activo. Todas se normalizan restando la media y dividiendo por la desviación estándar calculadas exclusivamente sobre el período de entrenamiento del rolling, de modo que el escalamiento no incorpore información posterior a dicha ventana.

El modelo emplea una ventana retrospectiva (*lookback*) de 30 días: para cada período, la red recibe la secuencia de las 30 observaciones previas y estima la clase del período siguiente. La clase positiva se define como un incremento del diferencial de retornos logarítmicos respecto del período actual, lo que se interpreta como una señal de divergencia entre ambos activos. En los primeros 30 días de cada serie, en que no existe historia suficiente para construir la ventana, la salida se completa con el valor 0.5, que corresponde a una probabilidad neutra, lo cual no introduce sesgo en las observaciones del agente.

Cabe aclarar que se entrena un único modelo por rolling, construido a partir de las secuencias de todos los pares seleccionados del sector. De esta forma la red aprende un patrón de comportamiento a nivel sectorial y no una regla específica de un par, y ese mismo modelo se aplica sobre cada par de manera individual para generar sus probabilidades.

Dado que las clases pueden presentarse desbalanceadas, se aplica una ponderación mediante *class weights*, calculada con la función *compute_class_weight*. Con esto se evita que el modelo concentre su aprendizaje en la clase más frecuente. Para controlar el sobreajuste se emplea *early stopping*, reservando el 10 % final del entrenamiento como validación interna, sin mezclar aleatoriamente las observaciones. Esta validación se utiliza únicamente como criterio de detención del entrenamiento del LSTM y es independiente del esquema de validación de la estrategia, que

corresponde al bloque de validación definido en el diseño walk-forward.

Finalmente, la salida del modelo es la probabilidad asociada a la clase positiva, denominada $Prob_{LSTM}$, que se integra como una dimensión en el espacio de observación del agente RL, es decir, no reemplaza al z-score ni a las variables del entorno, sino que es una señal que aporta una lectura basada en patrones secuenciales.

8.6 Entorno RL

El entorno del agente RL consta de distintas componentes que corresponden a:

8.6.1. Máquina de estados

La operación de cada par se organiza mediante una máquina de estados que determina qué transiciones son admisibles en cada período, incorporando el mecanismo de *stop-loss boundary* propuesto por Kim y Kim (2019). Su propósito es reducir el riesgo de que el agente sostenga una posición perdedora cuando la relación de equilibrio deja de verificarse. Se definen tres estados: *flat*, cuando el par no mantiene posición y el capital permanece en efectivo; *in trade*, cuando mantiene una posición abierta cuya magnitud y dirección determina el agente; y *stopped*, cuando el par fue detenido por el stop-loss y permanece inhabilitado hasta que surja una oportunidad válida.

Desde *flat*, el par pasa a *in trade* cuando el z-score supera el umbral de cierre. Desde *in trade* existen tres caminos: el retorno a *flat* por reversión del spread dentro del umbral de cierre, el retorno a *flat* por tiempo máximo de mantención de la misma posición, y la transición a *stopped* cuando el z-score excede el umbral de stop-loss, lo que se interpreta como un posible quiebre de la relación del par. Esta última fuerza el cierre de la posición y aplica una penalización en la función de recompensa.

Cabe aclarar que la posición tomada por el agente no se condiciona al cruce del umbral de apertura, sino al del umbral de cierre. Esta decisión se debe a que, en implementaciones preliminares en que la entrada se restringía solo al cruce del umbral de apertura, el agente desaprovechaba oportunidades de convergencia que se presentaban en niveles intermedios del z-score. Al ampliar el rango admisible, el umbral de apertura deja de operar como condición fija de entrada y pasa a constituir información que el agente recibe a través de la zona, quedando a su criterio determinar el momento y la magnitud de la entrada. De esta forma, los umbrales delimitan el espacio de decisión sin sustituir al agente en ella, lo que resulta coherente con el objetivo de evaluar si el aprendizaje por refuerzo mejora sobre reglas predefinidas.

8.6.2. Espacio de observación

El espacio de observación corresponde a la información que el agente recibe del entorno en cada período para tomar su decisión. Este se compone de seis dimensiones:

1. **Posición vigente** ($A_{t-1} \in [-1, 1]$): Es la fracción del capital comprometida en el par, heredada de la decisión del período anterior. El signo indica la dirección de la exposición, positivo para el *long* y negativo para el *short*, mientras que un valor nulo indica que no se mantiene posición.
2. **Z-score** ($z_t \in \mathbb{R}$): Es la señal de trading del par, que indica cuán alejado se encuentra el spread de su media en unidades de desviación estándar de la ventana.
3. **Zona**: Es una variable categórica que sintetiza la ubicación de $|z_t|$ respecto de los umbrales de la estrategia. Toma el valor 0 cuando el z-score se encuentra dentro del umbral de cierre, 1 entre los umbrales de cierre y apertura, 2 entre el umbral de apertura y el de stop-loss, y 3

más allá de este último.

4. **Probabilidad del oráculo** ($\text{Prob}_{\text{LSTM}} \in [0, 1]$): Corresponde a la señal predictiva entregada por la red LSTM descrita en el punto anterior, que resume el comportamiento de los treinta períodos previos.
5. **Estado**: Codifica el estado vigente del par, con valores 0 para *flat* cuando el par no mantiene una posición, 0.5 para *in trade* cuando el par toma una posición y 1 para *stopped* si el par fue detenido y permanece inhabilitado para operar debido al stop-loss hasta que se cumpla la condición de reingreso. Permite que el agente distinga situaciones que el z-score por sí solo no diferencia, dado que un mismo valor de z_t tiene implicancias distintas según si el par se encuentra sin posición, operando o detenido.
6. **Variación del z-score** ($\Delta z_t = |z_t| - |z_{t-1}|$): Esta dimensión informa si la desviación se amplía o se reduce respecto del período anterior. Representa la única memoria de mercado del vector de observación, y resulta necesaria porque un mismo valor de z_t admite lecturas opuestas, es decir, si proviene de un valor menor, la desviación se encuentra en expansión; si proviene de uno mayor, el spread está revirtiendo hacia su media.

Cabe precisar que las dimensiones primera y quinta no describen el mercado sino el estado del propio agente, mientras que las restantes caracterizan la condición del par. Asimismo, todas se construyen exclusivamente con información disponible hasta el período t , sin incorporar datos posteriores.

8.6.3. Espacio de acción

El espacio de acción es continuo y corresponde a la decisión $A_t \in [-1, 1]$ que el agente emite en cada período. El signo determina la dirección de la exposición, positivo para el *long* y negativo para el *short*, el valor absoluto determina la fracción del capital comprometida en el par. Por ejemplo, $A_t = 0,5$ representa una posición larga con el 50 % del capital y $A_t = -0,2$ una posición corta con el 20 %, en tanto que $A_t = 0$ implica no mantener posición (recordar que por cobertura toma la posición del segundo activo, Ecuación 11).

Esta es la implementación del *dynamic scaling* propuesto por Yang y Malik (2024), que otorga al agente no solo la decisión de cuándo operar sino también la de cuánto capital comprometer, a diferencia de las implementaciones que invierten un monto fijo o la totalidad del portafolio.

8.6.4. Mecanismos de control

El entorno incorpora tres mecanismos que acotan el comportamiento del agente. El primero es una banda muerta (*deadband*) sobre la magnitud de la posición, si la acción es inferior a 0.15 en valor absoluto, la posición se lleva a cero. Con esto se evitan cambios de posiciones muy bajos cuyo resultado esperado no justifica el costo de transacción asociado a abrirlas y cerrarlas.

El segundo es el stop-loss, que fuerza el cierre de la posición cuando el z-score excede el umbral θ_{SL} , definido como el umbral de apertura más un margen optimizado, de modo que se cumpla siempre $SL > \theta_{open}$ conforme a la restricción planteada por Kim y Kim (2019). Este mecanismo evita la acumulación de grandes pérdidas cuando el spread diverge demasiado, lo cual sugiere quiebres estructurales del par.

El tercero es un cierre por tiempo máximo de mantención, si una posición permanece abier-

ta durante más períodos que el límite *max_hold*, se cierra de forma forzada. Este límite se define como hiperparámetro dentro del proceso de optimización. Su propósito es evitar que el capital quede inmovilizado en posiciones que no rewerten ni alcanzan el stop-loss, liberándolo para otras oportunidades.

8.6.5. Función de recompensa

La función de recompensa define el criterio con el que el agente evalúa sus decisiones durante el entrenamiento, por tanto, condiciona la política que aprende. En este trabajo no se emplea el retorno puro de la estrategia, sino un esquema de *reward shaping* que incorpora términos adicionales orientados a guiar el aprendizaje hacia el comportamiento buscado. La recompensa en el período t se define como:

$$r_t = \underbrace{\frac{A_{t-1}(s_t - s_{t-1})}{\nu_t}}_{\text{PnL normalizado}} + \underbrace{c_a \cdot \phi(A_t, z_t)}_{\text{shaping por zona}} - \underbrace{\frac{|A_t - A_{t-1}| \nu_{t-1} \kappa}{\nu_t}}_{\text{costo de transacción}} - \underbrace{\frac{M_t}{\nu_t}}_{\text{penalización stop-loss}} \quad (13)$$

donde A_{t-1} y A_t son la posición vigente y la posición objetivo, s_t es el spread, $\nu_t = P_{1,t} + |\beta_t|P_{2,t}$ corresponde al valor nocional del par, es decir, el monto total comprometido al tomar una unidad completa de la operación (comprar una acción del primer activo y vender β_t acciones del segundo), κ es la comisión por transacción (se considera 0.1 % en base a broker Fintual) y c_a el coeficiente de shaping.

El primer término mide el resultado económico de la posición mantenida durante el período. La ganancia o pérdida (PnL) depende de la posición del período anterior A_{t-1} y del cambio del spread, y se normaliza por el nocional ν_t para que sea comparable entre pares con distinta escala de

precios. El tercer término descuenta el costo de transacción, proporcional a la magnitud del cambio de posición. El cuarto penaliza la activación del stop-loss, definida como:

$$M_t = \begin{cases} 0,05 \cdot |A_{t-1}| \cdot \nu_{t-1} & \text{si se activa el stop-loss en } t \\ 0 & \text{en caso contrario} \end{cases} \quad (14)$$

El segundo término corresponde al *shaping* por zona, que orienta al agente respecto de cuándo resulta razonable mantener exposición:

$$\phi(A_t, z_t) = \begin{cases} |A_t| & \text{si } |z_t| > \theta_{\text{close}} \\ -|A_t| & \text{en caso contrario} \end{cases} \quad (15)$$

Cuando el z-score se encuentra fuera del umbral de cierre, es decir, en una zona donde existe una posibilidad de revertir, el término premia mantener posición de forma proporcional a su magnitud; en caso contrario, cuando el spread se encuentra próximo a su media y no hay oportunidad que explotar, penaliza sostener exposición. Su propósito es acelerar el aprendizaje de la lógica de reversión, evitando que el agente deba inferirla únicamente a partir del PnL, cuya señal podría ser ruidosa en series financieras.

Cabe aclarar que este esquema de reward shaping introduce un sesgo de diseño, lo que el agente maximiza durante el entrenamiento no coincide exactamente con el retorno financiero que se mide en la evaluación.

8.7 Optimización de hiperparámetros

Para definir los hiperparámetros de la estrategia se utiliza la librería Optuna, haciendo uso del muestreador *Tree-structured Parzen Estimator* (TPE). Este método corresponde a un enfoque de optimización bayesiana, el cual a partir de los resultados obtenidos en ensayos anteriores, construye modelos probabilísticos para identificar configuraciones que prometen un mejor desempeño. De esta manera, en lugar de explorar aleatoriamente todo el espacio de búsqueda, Optuna propone nuevas combinaciones de hiperparámetros concentrándose progresivamente en regiones con mayor probabilidad de mejorar el criterio objetivo. En este caso, el proceso de optimización considera 100 ensayos para definir los hiperparámetros.

Los hiperparámetros de la estrategia de pairs trading optimizados por Optuna son los siguientes:

- Ventana móvil (*window_size*), con valores posibles $\{40, 60, 80, 120, 160\}$.
- Umbral de apertura θ_{open} (*open_threshold*), definido en el intervalo $[1,0, 2,5]$ con paso de 0.1.
- Umbral de cierre θ_{close} (*close_threshold*), definido en el intervalo $[0,1, 1,0]$ con paso de 0.1.
- Margen de stop-loss (*sl_buffer*), definido en el intervalo $[1,5, 3,0]$ con paso de 0.1.
- Tiempo máximo de mantención de una posición (*max_hold_steps*), con valores posibles $\{30, 45, 60, 90\}$.

Además, se establecen restricciones sobre los umbrales para evitar configuraciones degeneradas. Se exige que el umbral de apertura sea suficientemente mayor que el umbral de cierre, de

modo que se cumpla:

$$open_threshold - close_threshold \geq 0,5. \quad (16)$$

Asimismo, siguiendo la lógica de control de riesgo planteada por Kim y Kim (2019), el nivel de stop-loss debe ubicarse por encima del umbral de apertura. Debido a esto se define el límite de pérdida como:

$$SL = open_threshold + sl_buffer. \quad (17)$$

De esta forma, el parámetro *sl_buffer* no representa directamente el nivel de stop-loss, sino la distancia adicional permitida entre el umbral de apertura y el límite máximo de divergencia tolerado antes de cerrar la posición.

Por otro lado, los hiperparámetros propios del entrenamiento del agente PPO se mantienen fijos, utilizando una tasa de aprendizaje $learning_rate = 3 \times 10^{-4}$ y un coeficiente de entropía $ent_coef = 0,01$. Esta decisión se debe a que durante las pruebas preliminares, permitir la variación de estos generó inestabilidad en los resultados y afectó la reproducibilidad.

El criterio objetivo utilizado por Optuna para definir los hiperparámetros corresponde a la maximización del índice de Sharpe durante el periodo de validación. Así, para cada ensayo se entrena y evalúa la estrategia con una combinación específica de hiperparámetros, seleccionando aquella que logra el mejor desempeño ajustado por riesgo en la ventana de validación.

El criterio objetivo utilizado por Optuna para la definición de los hiperparámetros es la maximización del índice de Sharpe calculado sobre la serie de recompensas obtenidas en el periodo de validación. Cabe aclarar que esta serie corresponde a la función de recompensa definida

previamente en la Ecuación 13 y no al retorno puro.

8.8 Entrenamiento del agente

El agente se entrena mediante Proximal Policy Optimization (PPO), decisión que se justifica en tres razones. En primer lugar, permite trabajar con un espacio de acciones continuo, lo que es necesario para implementar el esquema de dynamic scaling, donde la decisión del agente no se limita a entrar o salir de una operación, sino que también define la proporción de capital asignada al par. En segundo lugar, corresponde a un enfoque *model-free*, por lo que no requiere conocer explícitamente la función de transición del mercado, la cual en la práctica es desconocida y estocástica. Finalmente, PPO entrega mayor estabilidad durante el entrenamiento mediante el mecanismo de clipping, que restringe cambios bruscos en la política entre actualizaciones sucesivas.

En cada sector, el agente se entrena utilizando un subconjunto de los top 4 pares seleccionados. A partir de estos pares, se busca aprender una política compartida a nivel sectorial, que permita capturar patrones generales del comportamiento entre activos pertenecientes a una misma industria. De esta manera, la política no se ajusta a solo un par, sino que aprende a operar bajo distintas configuraciones de spread, niveles de exposición y estados de posición dentro del mismo sector. Además, el entrenamiento se realiza de forma secuencial sobre cada par, acumulando las actualizaciones de la política sin reiniciar el proceso, tal que lo aprendido sobre un par se transfiere al siguiente.

El entrenamiento del agente se realiza en la ventana de train de cada rolling, utilizando los hiperparámetros seleccionados previamente mediante Optuna. Durante este proceso, el agente observa el estado del entorno, ejecuta acciones continuas asociadas a la exposición del par y recibe recompensas según su desempeño, considerando tanto el resultado económico como las penaliza-

ciones definidas por costos, riesgo y control de pérdidas. Posteriormente, la política entrenada se evalúa en la ventana de validación para seleccionar la configuración con mejor desempeño y luego realizar la prueba fuera de muestra (OOS) en el periodo correspondiente.

8.9 Selección de pares post entrenamiento

Una vez entrenado, el agente se evalúa sobre todos los pares cointegrados del rolling en el período de validación y se seleccionan aquellos cuyo índice de Sharpe del reward supera el valor 1, bajo el supuesto de que un Sharpe superior indica mayor probabilidad de desempeño favorable durante la evaluación; en caso de que ningún par supere este filtro, se consideran los dos con mejor desempeño. Esta selección puede incluir pares que no fueron utilizados durante el entrenamiento, decisión que permite verificar si la política aprendida generaliza a otros pares del mismo sector durante el periodo de evaluación, en vez de medir únicamente su desempeño sobre los pares observados durante el entrenamiento. De esta manera, la metodología propuesta se diferencia de enfoques vistos en la literatura en donde el agente se entrena y evalúa sobre un mismo par específico, ya que aquí se busca construir una política a nivel sectorial.

8.10 Backtest OOS y construcción de portafolio

El backtest fuera de muestra se realiza sobre la ventana de evaluación definida en cada rolling, para evitar que los primeros días del periodo de evaluación queden vacíos, la ventana móvil utilizada para calcular el z-score se inicia con observaciones inmediatamente anteriores al inicio del periodo de evaluación ($TEST_START$). De esta forma, el cálculo del spread se mantiene contiguo entre la ventana histórica y el periodo OOS, sin incorporar información futura.

La ganancia o pérdida diaria de cada par se calcula de manera causal, utilizando la posición

mantenida al cierre del día anterior y la variación observada del spread en el día actual. Esto evita que se utilice información del mismo día para decidir una posición que afecta ese mismo retorno.

Para cada par, el PnL diario se define como:

$$PnL_{i,t} = pos_{i,t-1} \cdot \frac{\Delta s_{i,t}}{\nu_{i,t}} - |pos_{i,t} - pos_{i,t-1}| \cdot \kappa \quad (18)$$

donde $pos_{i,t-1}$ corresponde a la posición en el par i el día anterior, $\Delta s_{i,t}$ la variación diaria del spread, κ corresponde a la comisión por transacción y $\nu_{i,t}$ es el nocional del par.

Al dividir por $\nu_{i,t}$, el PnL queda expresado como fracción del capital necesario para construir la posición. Esto permite sumar los retornos de pares con distintos niveles de precios sin que aquellos de mayor precio dominen el resultado del portafolio.

El segundo término incorpora los costos de transacción asociados al cambio de posición. De esta forma, la comisión se aplica únicamente cuando existe variación en la exposición del par, capturando el costo generado por abrir, cerrar o ajustar una posición.

Para la construcción del portafolio, se agregan los PnL de los pares activos en cada día. Dado que el agente puede asignar exposición simultáneamente a más de un par, se controla la exposición bruta diaria mediante una regla de escalamiento. Si la suma de exposiciones absolutas supera la unidad, es decir, si $\sum_i |pos_{i,t}| > 1,0$, todas las posiciones se reescalan por $\eta = \frac{1}{\sum_i |pos_{i,t}|}$

$$r_{\text{portafolio},t} = \sum_i PnL_{i,t} \cdot \eta, \quad (19)$$

Esta regla permite distribuir el capital entre los pares activos cada día, evitando que el portafolio asuma una exposición total superior al 100 % del capital asignado.

8.11 Benchmark y comparaciones experimentales

Para evaluar el desempeño del agente propuesto se consideran dos comparaciones. La primera corresponde a un benchmark basado en la estrategia clásica de Gatev et al. (2006), que utiliza precios normalizados, reglas de apertura definidas por umbrales fijos de dos desviaciones estándar y cierre cuando los precios normalizados se cruzan. Esto permite contrastar el desempeño del agente RL frente a una regla tradicional de pairs trading, ampliamente utilizada como referencia en la literatura.

Sin embargo, en este trabajo se aplica una modificación respecto de la metodología original, ya que los pares utilizados por el benchmark no son seleccionados mediante el criterio de distancia mínima, sino que provienen del mismo proceso de selección por cointegración. Esta decisión busca que ambas estrategias operen sobre el mismo conjunto de pares, de manera que la comparación aíse el efecto de la regla de decisión; por un lado, la estrategia clásica basada en umbrales fijos y por otro política aprendida por el agente PPO.

La segunda comparación corresponde al análisis de distintas frecuencias de reentrenamiento del modelo propuesto, en donde se evalúa reentrenamiento mensual, bimestral y trimestral, manteniendo el mismo horizonte fuera de muestra para todos los casos. Esta comparación permite analizar cómo la frecuencia de actualización del sistema afecta la rentabilidad y la estabilidad de los resultados, considerando que cada sector puede enfrentar dinámicas distintas, por lo que no necesariamente requieren la misma periodicidad de reentrenamiento.

8.12 Métricas de evaluación

Para evaluar el desempeño de las estrategias se consideran métricas de rentabilidad, riesgo y actividad dentro del periodo fuera de muestra. En primer lugar, se calcula el retorno total acumulado durante el horizonte OOS, el cual permite medir la ganancia o pérdida obtenida por la estrategia en el periodo evaluado, este se calcula mediante composición geométrica de los retornos diarios:

$$R = \prod_{t=1}^T (1 + r_t) - 1 \quad (20)$$

donde R corresponde al retorno total acumulado, r_t el retorno diario en el periodo t , y T es el número total de días evaluados. Junto con esto, se obtiene también su desviación estándar, utilizada como medida de volatilidad de la estrategia.

También se calcula el índice de Sharpe como la relación entre el retorno sobre el riesgo anualizado para valores diarios, sin embargo, al considerar distintas frecuencias no es un indicador adecuado para comparar entre los distintos esquemas por lo que se reporta únicamente como referencia.

Adicionalmente, se reportan métricas operativas que permiten caracterizar el comportamiento del agente. Entre ellas se considera el número total de operaciones ejecutadas, el ratio win/loss que expresa la relación entre operaciones ganadoras y perdedoras, y la cantidad de stop-loss activados durante el periodo de evaluación.

Finalmente, se construye una curva acumulada de retornos para cada estrategia y frecuencia de reentrenamiento, la cual permite evaluar visualmente la estabilidad del desempeño en el

tiempo, identificando periodos de crecimiento, caídas persistentes, recuperación y posibles episodios de alta volatilidad. También se reporta el drawdown máximo (MDD) que mide la mayor caída porcentual desde un máximo acumulado hasta el punto más bajo, el cual permite cuantificar el peor escenario de pérdida experimentado. De esta manera, la evaluación no se limita al resultado final del periodo, sino que también considera la trayectoria seguida por la estrategia durante la ventana OOS.

9 Resultados

Tras la evaluación en el periodo fuera de muestra se obtuvieron los siguientes resultados:

Tabla 3: Retorno total OOS del agente RL frente al benchmark de Gatev, por sector

Sector	Mejor frec.	RL	Gatev	Dif. (pp)
Communication Services	Trimestral	47.75 %	−29,19 %	76.94
Consumer Discretionary	Trimestral	74.61 %	−1,59 %	76.20
Consumer Staples	Trimestral	38.22 %	−5,17 %	43.39
Energy	Bimestral	71.18 %	−11,57 %	82.75
Financials	Trimestral	40.96 %	−3,89 %	44.85
Health Care	Trimestral	83.66 %	−3,53 %	87.19
Industrials	Bimestral	53.00 %	−17,42 %	70.42
Information Technology	Bimestral	81.05 %	−26,27 %	107.32
Materials	Trimestral	61.42 %	11.76 %	49.66
Real Estate	Mensual	22.57 %	−13,81 %	36.38
Utilities	Trimestral	13.89 %	−3,99 %	17.88

Nota: Se reporta la mejor frecuencia de reentrenamiento de cada sector

Desde la Tabla 3, se observa que el agente basado en RL, al considerar la mejor frecuencia de reentrenamiento de cada uno, supera al benchmark clásico en todos los sectores evaluados. El modelo propuesto obtiene retornos positivos en los once sectores, mientras que la estrategia clásica presenta retornos negativos en diez de ellos, siendo Materials el único caso en que logra un resultado positivo. Esta diferencia apunta a que la mejora no se debe a la selección de pares, ya que ambos enfoques operan sobre los mismos, sino a la política de decisión aprendida por el agente, calibración de umbrales, ventana móvil, ratio de cobertura, la incorporación de control de riesgo y la posibilidad de ajustar dinámicamente la exposición versus los umbrales fijos.

Las mayores diferencias ocurren en los sectores Information Technology, Health Care y Energy, donde el desempeño del agente supera los 80 puntos porcentuales respecto al benchmark.

Para estos casos, se puede interpretar que el modelo RL capturó oportunidades de convergencia de forma más eficiente que los umbrales fijos. Por otro lado, el sector Utilities es el que presenta la menor diferencia porcentual, aunque igualmente el agente logra un mayor retorno frente al modelo clásico.

Tabla 4: Retorno total OOS del agente RL (%) por sector y frecuencia de reentrenamiento

Sector	Mensual	Bimestral	Trimestral
Communication Services	25.63	3.46	47.75
Consumer Discretionary	62.21	37.39	74.61
Consumer Staples	9.65	26.73	38.22
Energy	6.19	71.18	15.58
Financials	26.23	13.02	40.96
Health Care	29.14	74.03	83.66
Industrials	16.14	53.00	44.28
Information Technology	15.03	81.05	57.03
Materials	53.93	37.10	61.42
Real Estate	22.57	7.00	3.97
Utilities	-6.66	9.42	13.89

Nota: En negrita se indica la mejor frecuencia de cada sector

Los resultados en la Tabla 4 muestran que la frecuencia de reentrenamiento tiene un efecto relevante en el desempeño del agente y que no hay una que domine en todos los sectores. Se observa que la frecuencia trimestral es la que obtiene mayor retorno en siete sectores (Communication Services, Consumer Discretionary, Consumer Staples, Financials, Health Care, Materials y Utilities); la bimestral tiene un mejor desempeño en tres sectores (Energy, Industrials e Information Technology) y la mensual presenta un mejor rendimiento en el sector Real Estate.

Este resultado es consistente con la hipótesis planteada, en donde cada sector presenta dinámicas distintas, por lo que puede requerir distintas actualizaciones en el modelo. En algunos sectores la frecuencia trimestral parece entregar mayor estabilidad, posiblemente se debe a que

evita que el agente reaccione en exceso al ruido del corto plazo. En cuanto a los sectores que presentan mejor desempeño con bimestral, puede deberse a que necesitan mayor adaptación frente a cambios rápidos en las relaciones de los pares. En cuanto al caso mensual, el único caso que resulta ganador es en Real Estate aunque con retornos más moderados respecto a los otros sectores.

Los desempeños menores en algunos sectores podrían deberse a quiebres estructurales en las relaciones de los pares, los cuales deterioran la cointegración y provocan cierres tempranos de posición.

Tabla 5: Métricas OOS del agente RL por sector y frecuencia de reentrenamiento

Sector	Frecuencia	Retorno	Desv.	Sharpe	MDD	Stops
Communication Services	Mensual	25.63 %	1.05 %	0.55	-15,83 %	5
	Bimestral	3.46 %	0.73 %	-0.15	-17,81 %	2
	Trimestral	47.75 %	1.29 %	0.87	-17,42 %	2
Consumer Discretionary	Mensual	62.21 %	0.84 %	1.67	-7,34 %	5
	Bimestral	37.39 %	0.77 %	1.06	-13,86 %	9
	Trimestral	74.61 %	0.86 %	1.85	-11,06 %	7
Consumer Staples	Mensual	9.65 %	0.66 %	0.12	-13,55 %	8
	Bimestral	26.73 %	0.61 %	0.88	-5,48 %	5
	Trimestral	38.22 %	0.53 %	1.49	-5,28 %	11
Energy	Mensual	6.19 %	0.81 %	-0.02	-10,85 %	7
	Bimestral	71.18 %	1.01 %	1.52	-10,49 %	2
	Trimestral	15.58 %	0.77 %	0.32	-10,24 %	3
Financials	Mensual	26.23 %	0.40 %	1.29	-2,67 %	6
	Bimestral	13.02 %	0.40 %	0.36	-3,99 %	7
	Trimestral	40.96 %	0.45 %	1.90	-2,40 %	5
Health Care	Mensual	29.14 %	0.62 %	0.99	-4,67 %	10
	Bimestral	74.03 %	0.74 %	2.12	-6,34 %	7
	Trimestral	83.66 %	0.63 %	2.74	-4,94 %	10
Industrials	Mensual	16.14 %	0.56 %	0.46	-5,82 %	10
	Bimestral	53.00 %	0.58 %	1.96	-7,61 %	10
	Trimestral	44.28 %	0.55 %	1.69	-5,82 %	10
Information Technology	Mensual	15.03 %	0.68 %	0.35	-10,81 %	9
	Bimestral	81.05 %	0.67 %	2.53	-11,19 %	24
	Trimestral	57.03 %	0.76 %	1.62	-13,04 %	7
Materials	Mensual	53.93 %	0.67 %	1.79	-8,44 %	5
	Bimestral	37.10 %	0.79 %	1.02	-8,53 %	7
	Trimestral	61.42 %	0.86 %	1.54	-20,22 %	7
Real Estate	Mensual	22.57 %	0.55 %	0.79	-6,96 %	2
	Bimestral	7.00 %	0.57 %	-0.03	-8,14 %	2
	Trimestral	3.97 %	0.59 %	-0.19	-16,86 %	0
Utilities	Mensual	-6,66 %	0.46 %	-1.05	-9,99 %	3
	Bimestral	9.42 %	0.50 %	0.09	-9,85 %	3
	Trimestral	13.89 %	0.58 %	0.31	-6,76 %	5

Nota: La desviación estándar corresponde a la volatilidad observada en OOS. En negrita se indica la frecuencia de mayor retorno total en cada sector. El Sharpe se reporta como referencia descriptiva. Los stops corresponden al total de activaciones de stop-loss durante el horizonte OOS.

La Tabla 5 resume el desempeño del agente RL para todas las evaluaciones que se realizaron. A continuación, se presenta una lectura comparativa por sector:

- **Communication Services:** La frecuencia trimestral obtiene el mayor retorno total, aunque también presenta la desviación estándar más elevada. El caso bimestral muestra el MDD más severo, lo que indica una caída acumulada más profunda durante el horizonte OOS. En cuanto a activaciones de stop-loss, la frecuencia mensual presenta la mayor cantidad, aunque esta métrica debe interpretarse con cuidado debido a que esa frecuencia posee una mayor cantidad de rollings.
- **Consumer Discretionary:** La frecuencia trimestral alcanza el mayor retorno, con una desviación estándar ligeramente mayor a las demás. Sin embargo, el caso bimestral presenta el MDD más severo, lo que sugiere una trayectoria menos estable. Las activaciones de stop-loss son mayores en frecuencia bimestral, lo que puede indicar mayor exposición a episodios de divergencia.
- **Consumer Staples:** La frecuencia trimestral domina tanto en retorno como en estabilidad, ya que obtiene el mayor retorno, la menor desviación estándar y el MDD menos severo. Esto sugiere que, para este sector, una actualización menos frecuente permitió capturar más oportunidades de convergencia. No obstante, también registra la mayor cantidad de stop-loss, por lo que el resultado positivo no significa que no ocurran episodios de riesgo.
- **Energy:** La frecuencia bimestral presenta un retorno que supera en gran medida al resto de configuraciones, debido a oportunidades puntuales de alta rentabilidad durante el periodo evaluado, junto con la mayor desviación estándar, lo que indica mayor variabilidad en los retornos. El MDD es similar entre las tres frecuencias, aunque el caso mensual presenta la

caída más severa.

- **Financials:** La frecuencia trimestral obtiene el mayor retorno y el MDD menos severo. Aunque su desviación estándar es levemente superior. Los niveles de MDD son bajos en las tres configuraciones, lo que indica un comportamiento relativamente estable en el sector.
- **Health Care:** La frecuencia trimestral alcanza el mayor retorno total, mientras que la mayor desviación estándar se observa en la frecuencia bimestral. En cuanto al MDD, el caso mensual presenta la caída menos severa, aunque las tres mantienen niveles de MDD leves, lo cual sugiere que el modelo es estable en este sector.
- **Industrials:** La frecuencia bimestral obtiene el mejor retorno y una desviación estándar ligeramente mayor al resto. Los MDD se mantienen cercano, aunque el caso bimestral presenta la mayor caída acumulada. Los tres casos muestran la misma cantidad de activaciones de stop-loss.
- **Information Technology:** La frecuencia bimestral presenta el mayor retorno y la menor desviación estándar. Sin embargo, también muestra una alta cantidad de activaciones de stop-loss, lo que indica que el agente enfrentó varias situaciones de divergencia. Aun así, el desempeño agregado muestra que fueron compensadas por operaciones de mayor rentabilidad.
- **Materials:** La frecuencia trimestral obtiene el mayor retorno, pero también la mayor desviación estándar y el MDD más severo. Esto indica que el resultado presentó una trayectoria más riesgosa. En este sector, la frecuencia mensual muestra una alternativa más estable, aunque con menor retorno total.

- **Real Estate:** La frecuencia mensual es la que entrega el mayor retorno, la menor desviación estándar y el MDD menos severo. A diferencia de la mayoría de los sectores, la actualización más frecuente resulta más beneficiosa para el agente, posiblemente porque permite adaptarse mejor a cambios en las relaciones entre pares dentro del sector.
- **Utilities:** Este sector presenta el desempeño más débil, siendo el único que muestra un caso con retorno total negativo. La frecuencia trimestral presenta el mayor retorno y el MDD menos severo, aunque con una desviación estándar mayor. En términos generales, el agente logra mejorar frente a la frecuencia mensual, pero con una capacidad de generación de retorno más moderada que en otros sectores.

Cabe destacar que las activaciones de stop-loss no presentan una relación directa con el retorno obtenido, dado que la función de este es limitar las pérdidas ante divergencias del spread y no generar rentabilidad. Esto se observa en sectores donde, a pesar de un mayor número de activaciones, se obtienen retornos superiores.

En cuanto a los resultados, mostraron de forma consistente que el agente supera al benchmark clásico en todos los sectores evaluados y que la frecuencia de reentrenamiento de mayor retorno varía según el sector.

10 Conclusiones

La investigación tuvo como objetivo desarrollar y evaluar un modelo de pairs trading basado en Reinforcement Learning aplicado a sectores del mercado accionario estadounidense, con el fin de determinar su desempeño en retorno y riesgo frente a la estrategia clásica. Tras los resultados se concluye lo siguiente.

En cuanto al retorno superó al benchmark clásico de Gatev et al. (2006), el modelo propuesto superó en los once sectores evaluados, considerando la mejor frecuencia de reentrenamiento de cada uno. Mientras el agente presentó retornos positivos, la estrategia clásica presentó retornos negativos en su mayoría. Por lo tanto, se concluye que la política aprendida junto con la calibración adaptativa de hiperparámetros y el ajuste dinámico de capital permite mejorar el desempeño de la estrategia de pairs trading.

Sobre el riesgo, el agente mostró niveles de máximo drawdown acotados en la mayoría de los sectores, sin una relación directa que mayor retorno implicase mayores pérdidas. Asimismo con las activaciones del stop-loss, las cuales no mostraron relación con el retorno obtenido, lo cual permite decir que el mecanismo cumple una función de contención de pérdidas y no de generación de rentabilidad.

Respecto a las frecuencias de reentrenamiento, se concluye que no existe una que optimice el desempeño en todos los sectores por sobre las otras; la que predominó fue la trimestral, seguida de la bimestral y en un caso la mensual. Este resultado indica que la periodicidad de actualización del modelo podría considerarse como un hiperparámetro dentro de cada sector y no como una constante del sistema en otras investigaciones.

En conclusión los resultados permiten afirmar las preguntas de investigación planteadas, ya

que el agente basado en Reinforcement Learning supera a la estrategia clásica de pairs trading en rentabilidad, presenta un perfil de riesgo favorable, y la frecuencia óptima de reentrenamiento sí depende del sector.

11 Discusión

Como primer punto de discusión, los trabajos previos tienden a proponer componentes de forma aislada, Gatev et al. (2006) proponen la formación de pares mediante distancia mínima y umbrales fijos, Kim y Kim (2019) plantean el mecanismo de stop-loss sobre un agente de acción discreta, Yang y Malik (2024) proponen el escalamiento dinámico de capital y Flori y Regoli (2021) consideran LSTM para obtener probabilidades de divergencia; en cada caso demuestran beneficios que aportan a la estrategia, pero de forma separada. El aporte de este trabajo consiste en la integración de estos elementos, cointegración, LSTM, escalamiento dinámico y control de riesgo dentro de un mismo modelo que opera a nivel sectorial y construye portafolios de múltiples pares en lugar de evaluar un par de forma aislada.

Un segundo punto consiste en que parte de la literatura atribuye las mejoras de desempeño a la capacidad del agente para ajustar dinámicamente la exposición. Sin embargo, la evidencia de este trabajo varía levemente dicha atribución, al comparar con la formulación clásica, se observó que la calibración de umbrales, ventana móvil y coeficiente de cobertura forman parte de la mejora y no exclusivamente de la asignación dinámica de capital. En otras palabras, bajo esta propuesta, el aprendizaje por refuerzo aporta una capacidad adicional de adaptación, pero este aporte se construye sobre una calibración de parámetros que resulta determinante en el desempeño.

Un tercer punto de discusión corresponde a la frecuencia de reentrenamiento del modelo, la literatura en aprendizaje por refuerzo aplicado a pairs trading suele entrenar el agente y evaluarlo sobre un período posterior, sin considerar la periodicidad con que el modelo debería actualizarse. En este trabajo, la frecuencia de reentrenamiento se incorpora como una variable de diseño, evaluándose los esquemas mensual, bimestral y trimestral bajo un mismo horizonte fuera de muestra.

Los resultados muestran que no existe una frecuencia que optimice el desempeño en todos los sectores, sino que la frecuencia de mayor retorno varía según el sector evaluado. Esto sugiere que la periodicidad de actualización no debe tratarse como una constante del sistema, sino como un parámetro dependiente de las características de cada sector.

12 Limitaciones

Algunas limitaciones presentadas en esta investigación, tanto metodológicas como de alcance son:

La ausencia de fricciones de ejecución reales en la evaluación, si bien se considera una comisión por transacción, el backtest no modela la diferencia de precios entre el esperado al enviar una orden y el precio real (*slippage*) ni los costos asociados al préstamo de acciones para posiciones cortas. En un entorno realista, dichas consideraciones reducirían los retornos, por lo que, los resultados observados se consideran una cota superior optimista del desempeño.

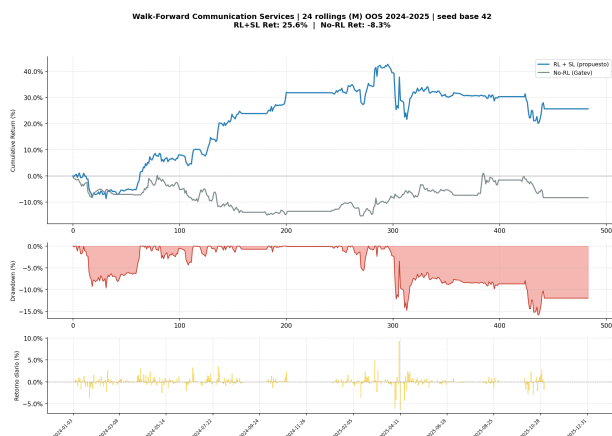
Una segunda limitación se relaciona con la variabilidad propia del aprendizaje por refuerzo, aunque se tomaron medidas para garantizar la reproducibilidad, el proceso tiene una componente estocástica. Los resultados son reproducibles bajo condiciones idénticas de hardware y versiones de librerías, por lo que cambios en estas pueden significar variaciones en el desempeño.

La tercera limitación está ligada al alcance, debido a decisiones de diseño, en donde la evaluación se restringe solo a acciones del índice S&P 500, por lo que las conclusiones no son directamente extrapolables a otros mercados o instrumentos.

Algunas limitaciones de carácter técnico, con el fin de preservar la reproducibilidad, se optó por no emplear mecanismos de poda en la optimización de hiperparámetros y por ejecutar el sistema en modo determinista, decisiones que incrementan el costo computacional del proceso.

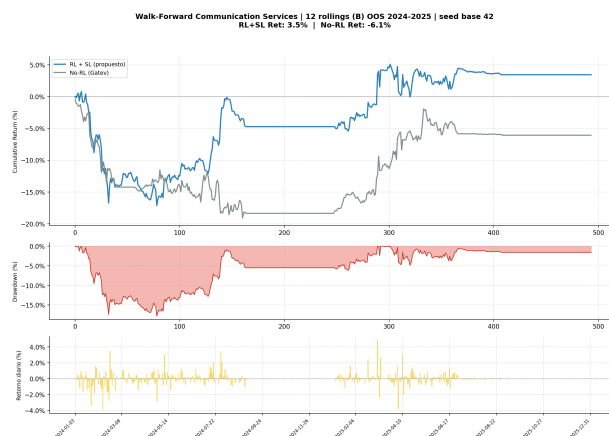
13 Anexos

Figura 7: Gráficos Communication Services.



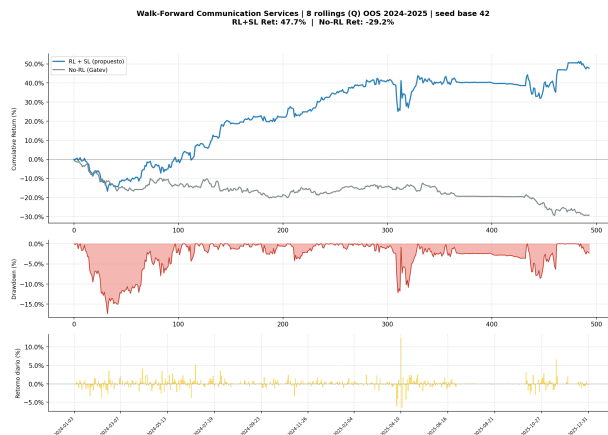
Nota: Frecuencia Mensual

Figura 8: Gráficos Communication Services.



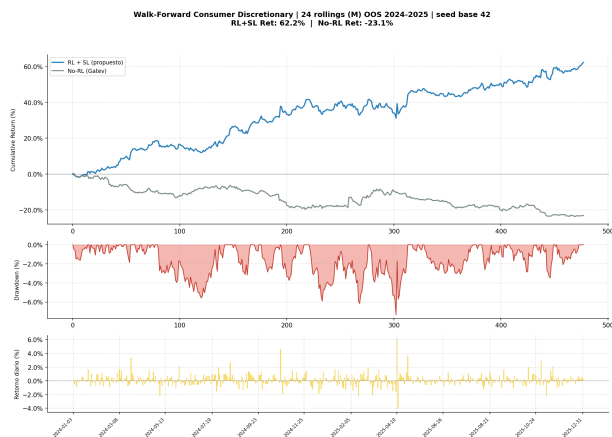
Nota: Frecuencia Bimestral

Figura 9: Gráficos Communication Services.



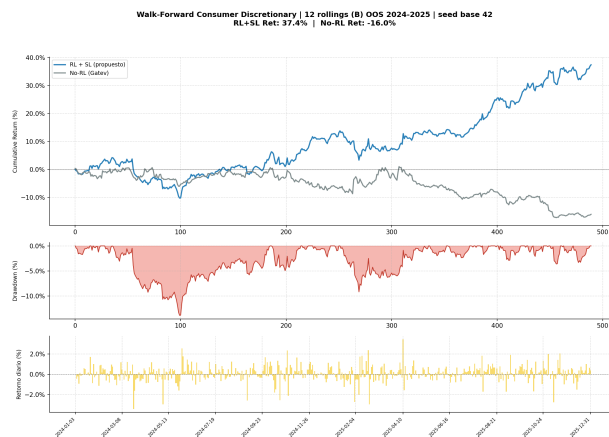
Nota: Frecuencia Trimestral

Figura 10: Gráficos Consumer Discretionary.



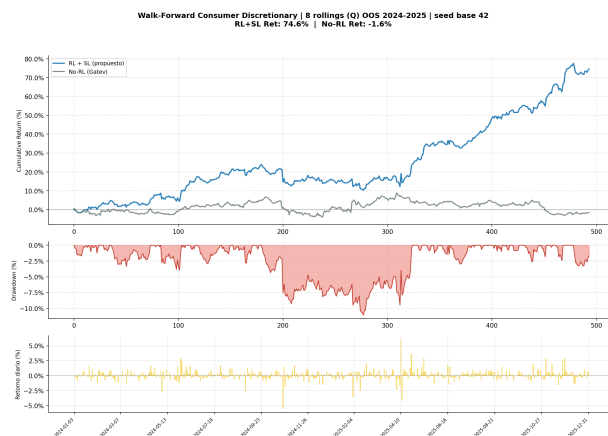
Nota: Frecuencia Mensual

Figura 11: Gráficos Consumer Discretionary.



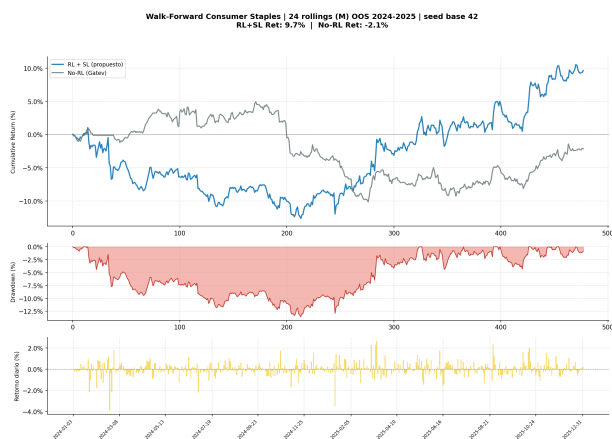
Nota: Frecuencia Bimestral

Figura 12: Gráficos Consumer Discretionary.



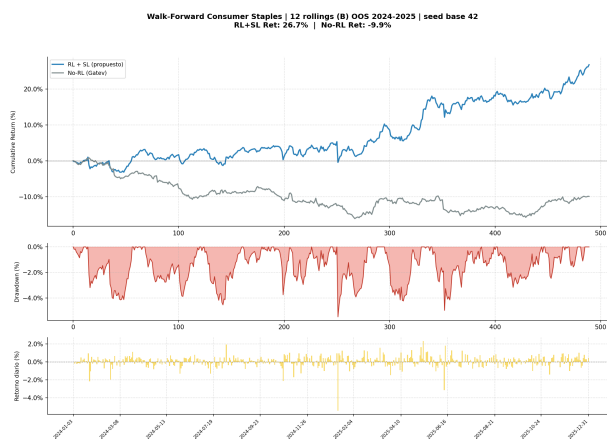
Nota: Frecuencia Trimestral

Figura 13: Gráficos Consumer Staples.



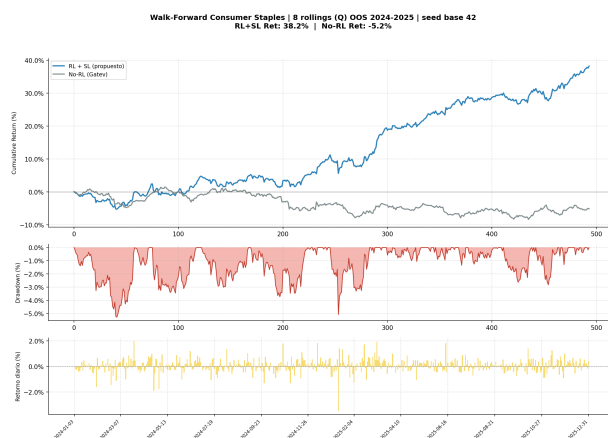
Nota: Frecuencia Mensual

Figura 14: Gráficos Consumer Staples.



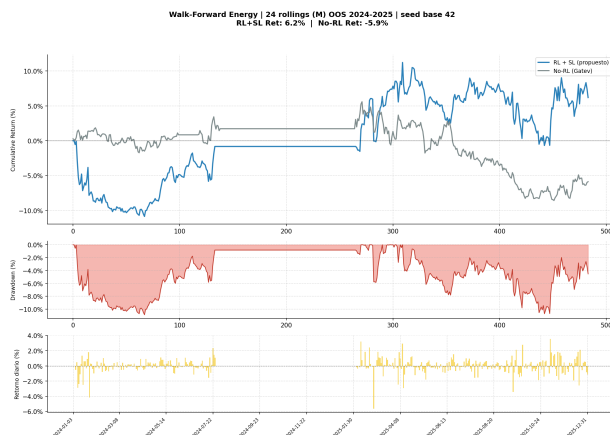
Nota: Frecuencia Bimestral

Figura 15: Gráficos Consumer Staples.



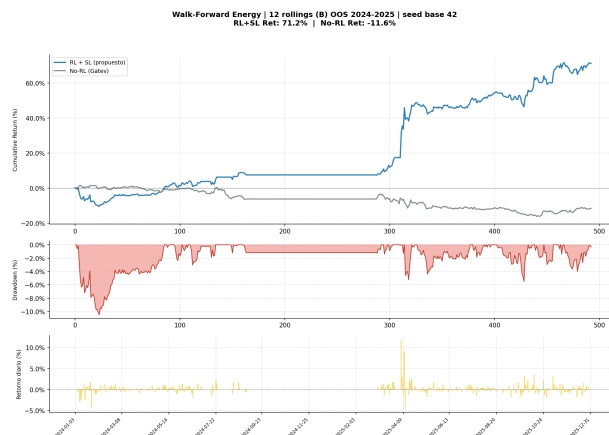
Nota: Frecuencia Trimestral

Figura 16: Gráficos Energy.



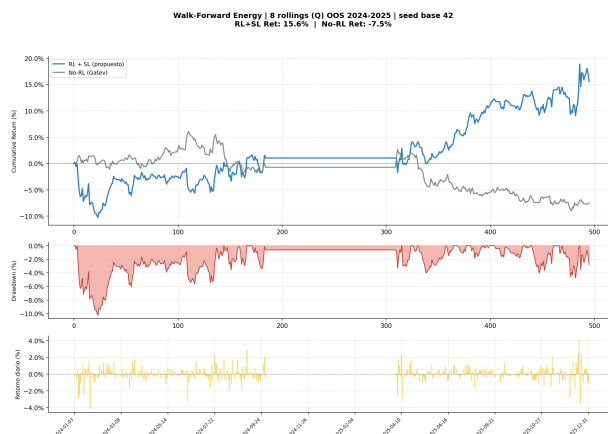
Nota: Frecuencia Mensual

Figura 17: Gráficos Energy.



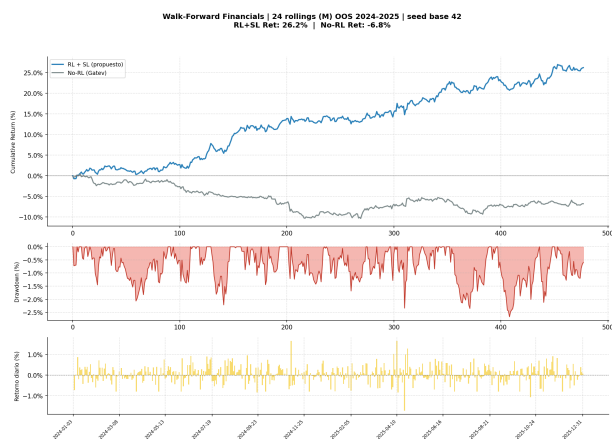
Nota: Frecuencia Bimestral

Figura 18: Gráficos Energy.



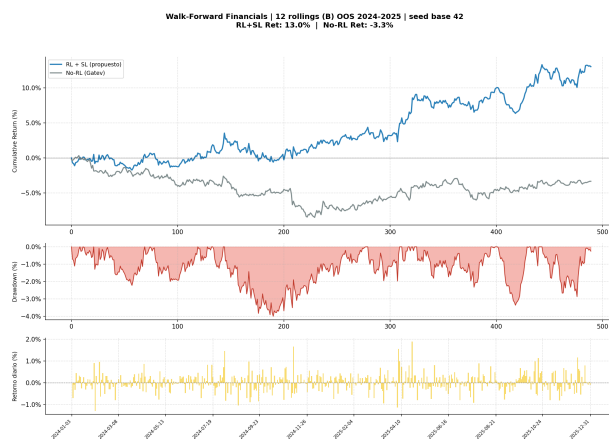
Nota: Frecuencia Trimestral

Figura 19: Gráficos Financieros.



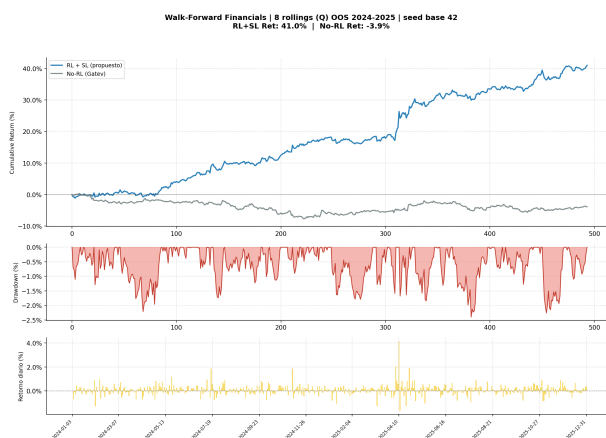
Nota: Frecuencia Mensual

Figura 20: Gráficos Financieros.



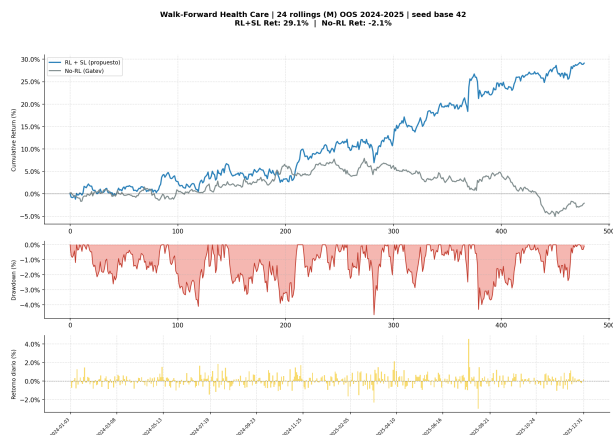
Nota: Frecuencia Bimestral

Figura 21: Gráficos Financieros.



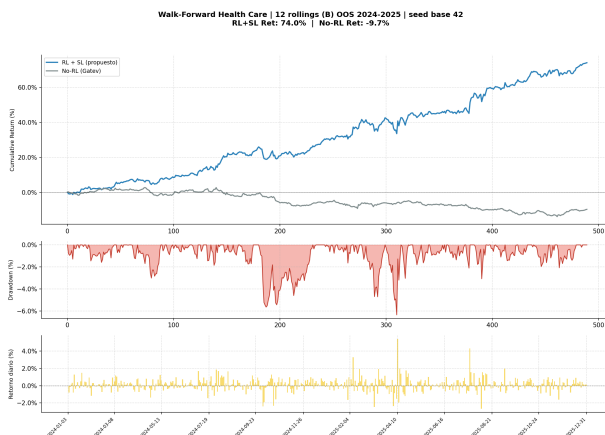
Nota: Frecuencia Trimestral

Figura 22: Gráficos Health Care.



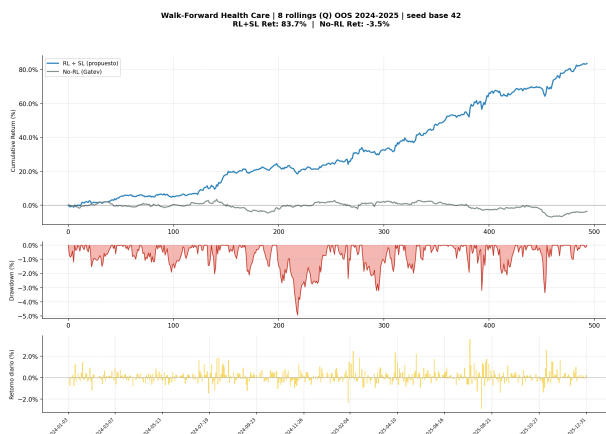
Nota: Frecuencia Mensual

Figura 23: Gráficos Health Care.



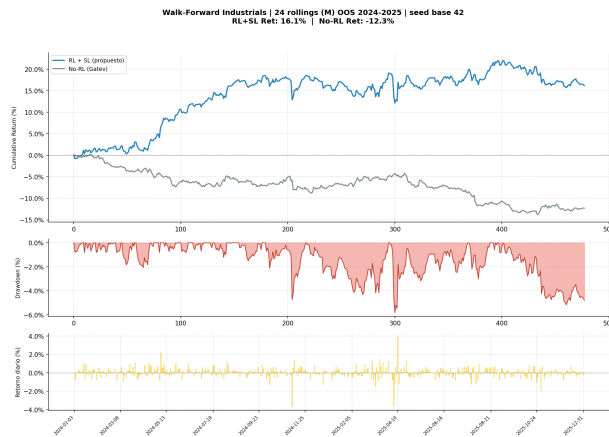
Nota: Frecuencia Bimestral

Figura 24: Gráficos Health Care.



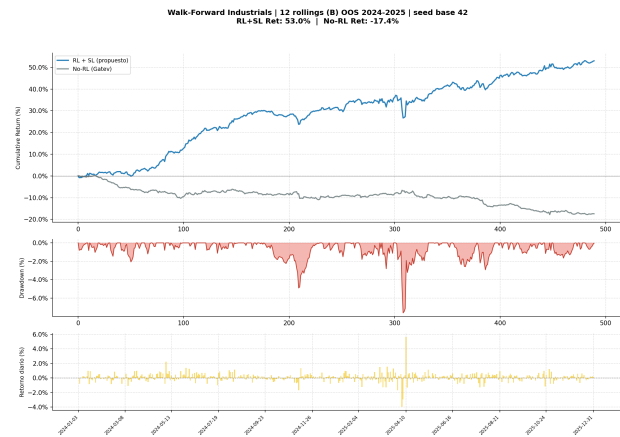
Nota: Frecuencia Trimestral

Figura 25: Gráficos Industrials.



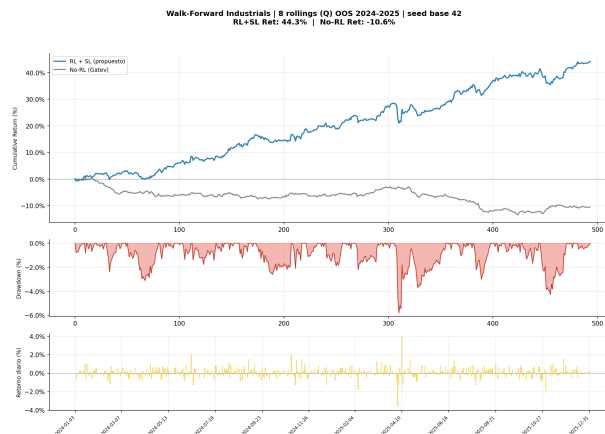
Nota: Frecuencia Mensual

Figura 26: Gráficos Industrials.



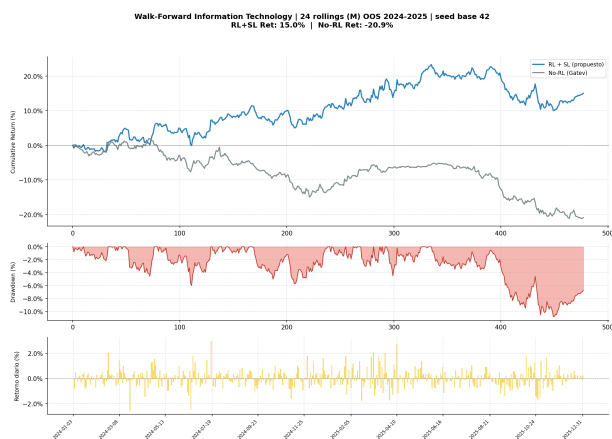
Nota: Frecuencia Bimestral

Figura 27: Gráficos Industrials.



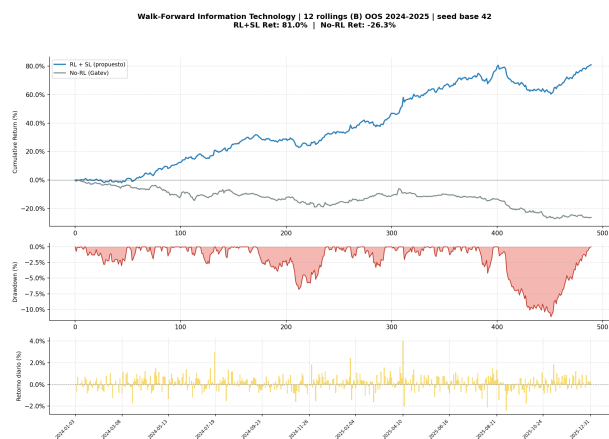
Nota: Frecuencia Trimestral

Figura 28: Gráficos Information Technology.



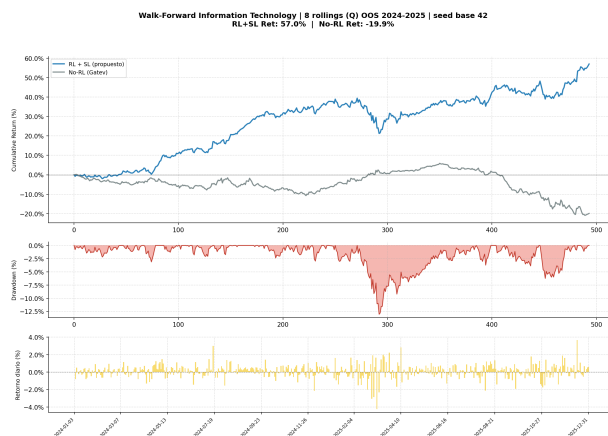
Nota: Frecuencia Mensual

Figura 29: Gráficos Information Technology.



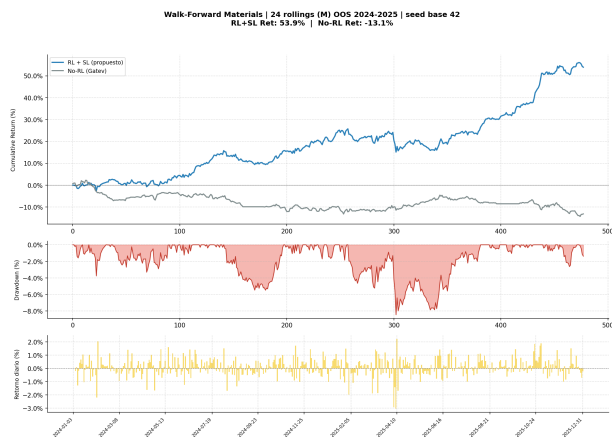
Nota: Frecuencia Bimestral

Figura 30: Gráficos Information Technology.



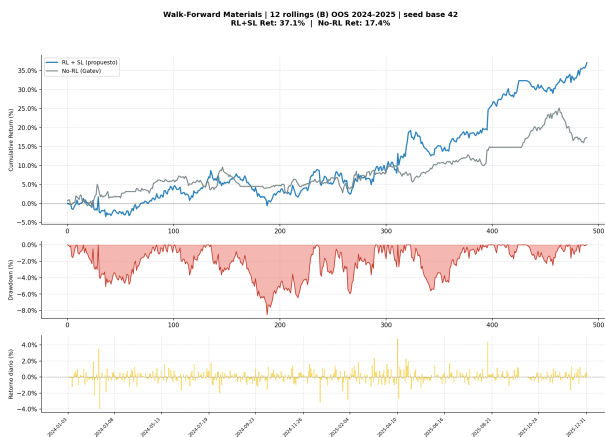
Nota: Frecuencia Trimestral

Figura 31: Gráficos Materials.



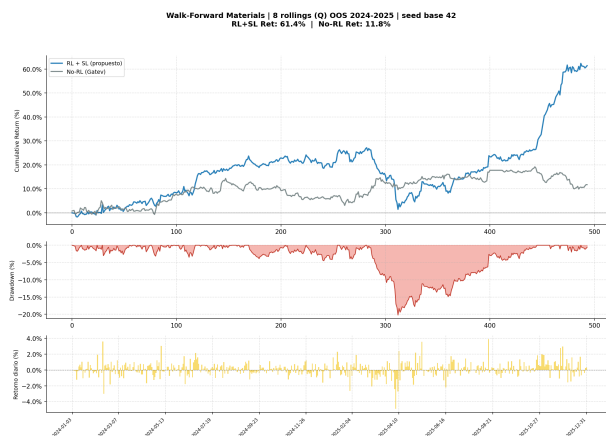
Nota: Frecuencia Mensual

Figura 32: Gráficos Materials.



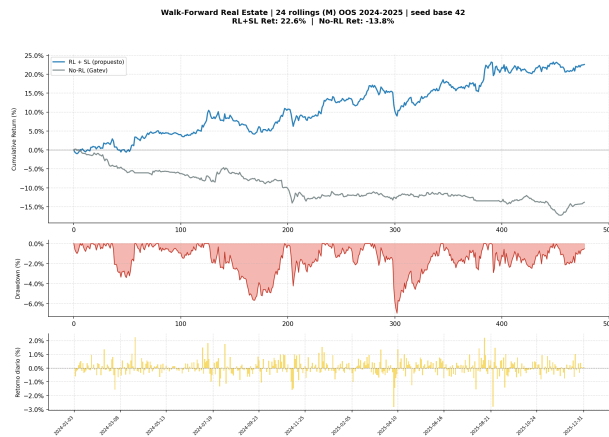
Nota: Frecuencia Bimestral

Figura 33: Gráficos Materials.



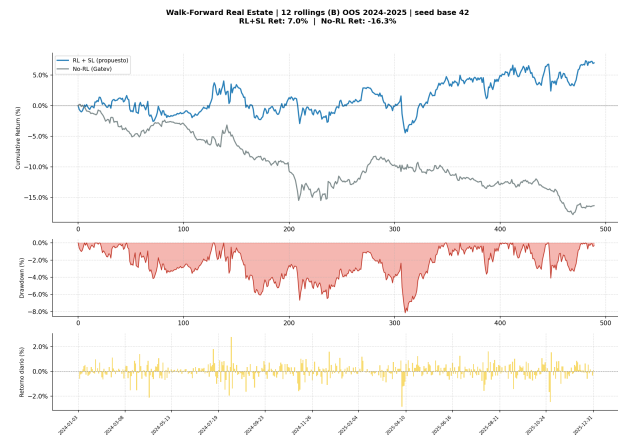
Nota: Frecuencia Trimestral

Figura 34: Gráficos Real Estate.



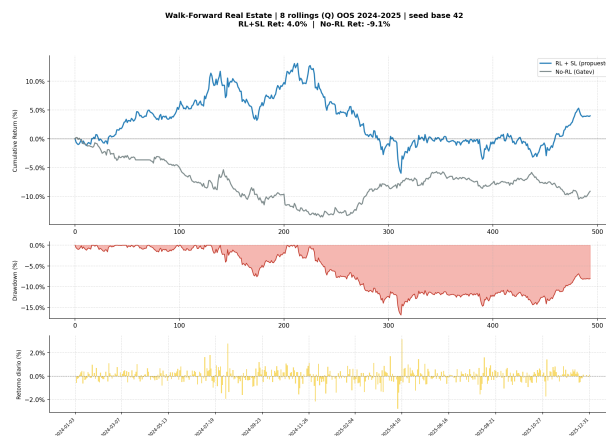
Nota: Frecuencia Mensual

Figura 35: Gráficos Real Estate.



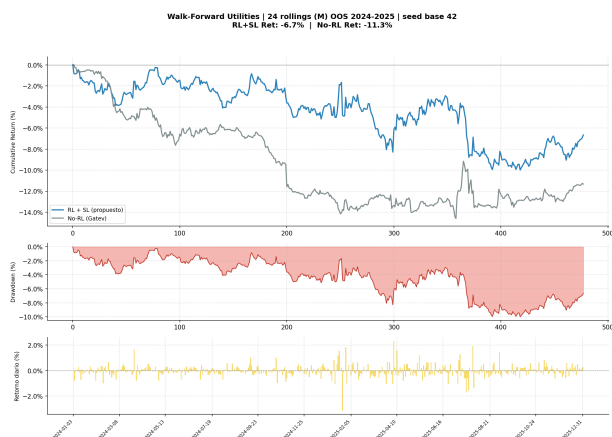
Nota: Frecuencia Bimestral

Figura 36: Gráficos Real Estate.



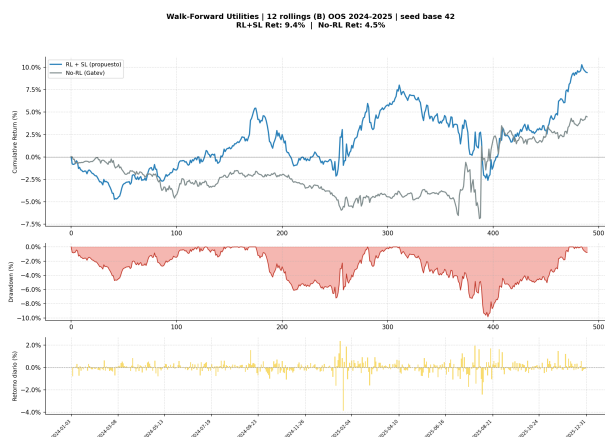
Nota: Frecuencia Trimestral

Figura 37: Gráficos Utilities.



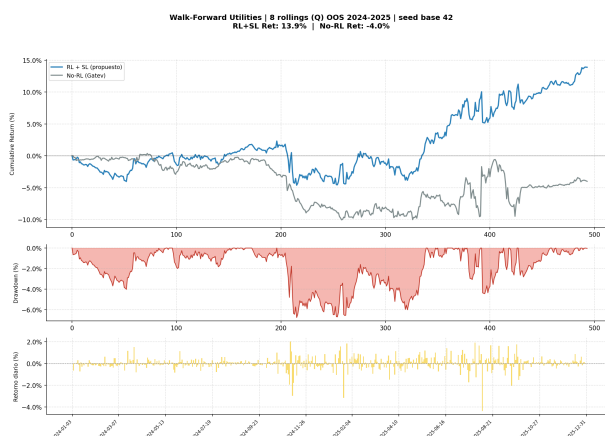
Nota: Frecuencia Mensual

Figura 38: Gráficos Utilities.



Nota: Frecuencia Bimestral

Figura 39: Gráficos Utilities.



Nota: Frecuencia Trimestral

Referencias

- Arulkumaran K., Deisenroth M. P., Brundage M., y Bharath A. A. (2017). Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Processing Magazine*, 34(6):26–38. <https://doi.org/10.1109/MSP.2017.2743240>.
- Avellaneda M. y Lee J.-H. (2010). Statistical Arbitrage in the US Equities Market. *Quantitative Finance*, 10(7):761–782. <https://doi.org/10.1080/14697680903124632>.
- Do B. y Faff R. (2010). Does Simple Pairs Trading Still Work? *Financial Analysts Journal*, 66(4):83–95. <https://doi.org/10.2469/faj.v66.n4.1>.
- Engle R. F. y Granger C. W. J. (1987). Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, 55(2):251–276. <https://doi.org/10.2307/1913236>.
- Flori A. y Regoli D. (2021). Revealing Pairs-Trading Opportunities with Long Short-Term Memory Networks. *European Journal of Operational Research*, 295(2):772–791. <https://doi.org/10.1016/j.ejor.2021.03.009>.
- Gatev E., Goetzmann W. N., y Rouwenhorst K. G. (2006). Pairs Trading: Performance of a Relative-Value Arbitrage Rule. *The Review of Financial Studies*, 19(3):797–827. <https://doi.org/10.1093/rfs/hhj020>.
- Gers F. A., Schmidhuber J., y Cummins F. (2000). Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471.

- Guijarro-Ordóñez J., Pelger M., y Zanotti G. (2022). Deep Learning Statistical Arbitrage. arXiv:2106.04028. <https://arxiv.org/abs/2106.04028>.
- Han W., Huang J., Xie Q., Zhang B., Lai Y., y Peng M. (2023). Mastering Pair Trading with Risk-Aware Recurrent Reinforcement Learning. arXiv:2304.00364. <https://arxiv.org/abs/2304.00364>.
- Hochreiter S. y Schmidhuber J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Johansen S. (1991). Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica*, 59(6):1551–1580. <https://doi.org/10.2307/2938278>.
- Kim T. y Kim H. Y. (2019). Optimizing the Pairs-Trading Strategy Using Deep Reinforcement Learning with Trading and Stop-Loss Boundaries. *Complexity*, 2019:3582516. <https://doi.org/10.1155/2019/3582516>.
- Lu J.-Y., Lai H.-C., Shih W.-Y., Chen Y.-F., Huang S.-H., Chang H.-H., Wang J.-Z., Huang J.-L., y Dai T.-S. (2022). Structural Break-Aware Pairs Trading Strategy Using Deep Reinforcement Learning. *The Journal of Supercomputing*, 78:3843–3882. <https://doi.org/10.1007/s11227-021-04013-x>.
- Pole A. (2007). *Statistical Arbitrage: Algorithmic Trading Insights and Techniques*. John Wiley & Sons.
- Schulman J., Wolski F., Dhariwal P., Radford A., y Klimov O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sutton R. S. y Barto A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press, 2 edición.

Vidyamurthy G. (2004). *Pairs Trading: Quantitative Methods and Analysis*. John Wiley & Sons.

Yang H. y Malik A. (2024). Reinforcement Learning Pair Trading: A Dynamic Scaling Approach.

Journal of Risk and Financial Management, 17(12):555. [https://doi.org/10.3390/](https://doi.org/10.3390/jrfm17120555)

[jrfm17120555](https://doi.org/10.3390/jrfm17120555).