Detección no supervisada de anomalías en telemetría satelital mediante caracterización estructural y umbralización adaptativa basada en la teoría de valores extremos

Óscar Alcarria[1*], Rafael Sánchez[1**], Javier Sempere[2], Pablo Torrijos[1], Juan C. Alfaro[1], Juan M. Auñón[3], José A. Gámez[1] y José M. Puerta[1]

[1] Universidad de Castilla-La Mancha, Albacete, Spain
[2] GMV GmbH, Darmstadt, Germany
[3] GMV, Madrid, Spain

Resumen La detección operacional de anomalías en telemetría satelital está limitada por la escasez de anomalías históricas etiquetadas, la naturaleza no estacionaria de las señales y la necesidad de iniciar la monitorización sin largos periodos de acumulación de datos. En este trabajo se propone un marco completamente no supervisado y orientado al despliegue, capaz de generar predicciones a partir del segundo mes de operación sin utilizar etiquetas, información semántica de la misión, selección supervisada de canales ni ajuste supervisado de hiperparámetros.
El enfoque combina reentrenamiento incremental mensual, caracterización estadística de cada misión y umbralización adaptativa basada en la teoría de valores extremos (*Extreme Value Theory*, EVT) para controlar las falsas alarmas en escenarios no estacionarios. La caracterización estadística entre misiones, basada en la distancia de energía, el coeficiente de variación y la estructura de agrupamiento, guía la selección del paradigma de detección más adecuado: un *autoencoder Long Short-Term Memory* (LSTM) ligero para misiones con dinámicas estables y anomalías temporales sutiles, y un esquema basado en análisis de componentes principales (PCA), *k-means* y *Local Outlier Factor* (LOF) para misiones con alta variabilidad y regímenes operacionales diferenciados.
La evaluación se realiza sobre el *ESA Anomalies Dataset Benchmark* (ESA-ADB) bajo una configuración cronológica estricta que preserva la causalidad temporal y evita la fuga de información, utilizando la métrica corregida $F_{0,5}$ a nivel de evento, que prioriza la precisión y penaliza la sobredetección en periodos nominales. El método propuesto alcanza $F_{0,5} = 0{,}700$ en la Misión 1 y $F_{0,5} = 0{,}698$ en la Misión 2, igualando o superando soluciones industriales no supervisadas, sin recurrir a etiquetas, selección supervisada de canales ni hardware GPU dedicado.

Palabras clave: Detección de anomalías · Telemetría satelital · Aprendizaje no supervisado · Aprendizaje incremental · Clustering · Teoría de valores extremos · Series temporales multivariantes · ESA-ADB

* Autor de correspondencia: oscar.alcarria@uclm.es
** Autor de correspondencia: rafael.sanchez@uclm.es

1. Introducción

La detección de anomalías en telemetría satelital consiste en identificar comportamientos fuera de lo nominal en series temporales multivariantes y no estacionarias bajo estrictas restricciones operacionales. En la práctica, los entornos de control de misión rara vez disponen de historiales completos de fallos etiquetados, hardware dedicado con GPU o largos periodos de telemetría acumulada antes de que deba comenzar la monitorización.

El conjunto de datos *ESA Anomalies Dataset Benchmark* (ESA-ADB) [5] constituye el banco de pruebas público más completo disponible para la detección de anomalías en telemetría satelital. Proporciona datos anotados por expertos procedentes de tres misiones completas de la ESA, junto con un procedimiento de evaluación jerárquico basado en eventos. Dos de estas misiones se utilizan para la evaluación: la Misión 1 (M1), con 76 canales, de los cuales 58 son canales objetivo, 118 anomalías y 78 eventos nominales raros; y la Misión 2 (M2), con 100 canales, 47 canales objetivo, 31 anomalías y 613 eventos nominales raros.

Un análisis exploratorio de ESA-ADB revela diferencias estadísticas marcadas entre misiones, lo que sugiere que una única estrategia de modelado puede no generalizar de forma consistente. En este trabajo se propone un marco de detección no supervisado y consciente de la estructura de cada misión que: (1) selecciona el paradigma de detección en función de las propiedades estadísticas a nivel de misión; (2) integra reentrenamiento incremental mensual, permitiendo generar predicciones a partir del segundo mes de operación; y (3) aplica umbralización adaptativa basada en la teoría de valores extremos (EVT) para una detección de anomalías robusta y orientada a la precisión a nivel de evento.

El resto del artículo se organiza de la siguiente manera: la Sección 2 describe la metodología propuesta y su configuración incremental; la Sección 3 presenta los resultados experimentales obtenidos sobre ESA-ADB y el coste computacional; la Sección 4 discute las implicaciones y limitaciones del enfoque; finalmente, la Sección 5 resume las conclusiones y plantea las líneas de trabajo futuro.

2. Metodología

2.1. Definición del problema y restricciones

Abordamos la detección de eventos anómalos en telemetría satelital multivariante bajo condiciones completamente anonimizadas, lo que impide utilizar semántica específica de la misión o conocimiento previo basado en la física del sistema. Siguiendo la definición del *ESA Benchmark for Anomaly Detection in Satellite Telemetry* (ESA-ADB) [5], las anomalías se modelan como segmentos temporales contiguos fuera de lo nominal, en lugar de valores atípicos aislados.

El *benchmark* ESA-ADB distingue entre anomalías y eventos nominales raros desde una perspectiva operacional. Las anomalías son cambios atípicos, raros, no planificados y no deseados en la telemetría, mientras que los eventos nominales raros son cambios también atípicos o poco frecuentes, pero esperados o planificados, como maniobras, reinicios, calibraciones u otras operaciones especiales

de la misión. En este trabajo, ambas categorías se consideran eventos anotados de interés a predecir, manteniendo su distinción únicamente para el análisis e interpretación de los resultados.

Para reflejar restricciones realistas de despliegue [11], el marco propuesto opera sin etiquetas ni hardware GPU dedicado, y se evalúa mediante una configuración incremental y cronológica que preserva la causalidad temporal y evita la fuga de información. El rendimiento se mide utilizando la métrica corregida $F_{0,5}$ a nivel de evento [5], donde $\beta = 0{,}5$ pondera la precisión por encima del *recall*. Además, la precisión a nivel de evento se corrige mediante la tasa de verdaderos negativos a nivel de muestra, penalizando así los métodos que sobredetectan periodos nominales.

2.2. Preprocesamiento de datos

La telemetría se preprocesa siguiendo el protocolo operativo del sistema PitIA [11]: las señales se remuestrean en los intervalos propios de cada misión (30 s para M1 y 18 s para M2), y los valores faltantes se imputan mediante interpolación causal de orden cero (*zero-order hold*). Además, los pasos de alineación temporal y preservación de anomalías garantizan que los eventos etiquetados se conserven tras el remuestreo.

2.3. Configuración de aprendizaje incremental

En lugar de depender de la partición estática de ESA-ADB, se implementa una estrategia de aprendizaje incremental basada en una ventana mensual. En esta configuración, el modelo se entrena exclusivamente con los datos del mes anterior y se evalúa sobre el mes siguiente, lo que permite generar predicciones a partir del segundo mes de operación sin necesidad de grandes volúmenes de datos históricos. Esta metodología resulta especialmente adecuada para entornos operacionales, ya que proporciona una adaptación continua al *concept drift* [13], habitual en misiones espaciales de larga duración, y se alinea con los ciclos de actualización propios de los sistemas de monitorización en producción.

2.4. Caracterización de las misiones

Para caracterizar las misiones y guiar la selección de familias de modelos de detección de anomalías, se realiza un análisis estadístico comparativo sobre la telemetría disponible de M1 y M2 en ESA-ADB, incluyendo periodos nominales, eventos raros y anomalías.

En primer lugar, la distancia de energía (*Energy Distance*) [10] entre M1 y M2 es 4,33, lo que indica diferencias sustanciales entre sus distribuciones multivariantes. La significación estadística se evaluó mediante una prueba de permutación con 200 permutaciones [2], en la que las etiquetas de misión se reasignaron aleatoriamente manteniendo el tamaño de las muestras. Ningún estadístico permutado superó el valor observado ($p < 0{,}005$), lo que respalda la existencia de

diferencias estadísticas significativas entre ambas misiones. Para cada misión, el coeficiente de variación (CV) se calculó de forma independiente para cada sensor como la razón entre su desviación estándar y su media absoluta, y posteriormente se promedió sobre todas las variables numéricas. Este análisis revela una baja variabilidad en M1 (CV= 0,13), consistente con dinámicas relativamente estables, mientras que M2 presenta una variabilidad notablemente mayor (CV= 1,27), reflejando una mayor dispersión global entre sensores.

La estructura de agrupamiento se evaluó utilizando *k-means* [6] con distancia euclídea sobre variables estandarizadas, explorando múltiples valores de k. El coeficiente de *Silhouette* [8], mostrado en el recuadro de la Figura 1 para distintos valores de k, presenta puntuaciones más altas para M2, lo que sugiere clústeres más compactos y mejor separados, mientras que M1 presenta una segmentación intrínseca más débil.

Las proyecciones de análisis de componentes principales (PCA) [3] (Figura 1), obtenidas ajustando PCA sobre M1 con el subconjunto de variables común a ambas misiones y proyectando posteriormente ambas misiones en ese subespacio, refuerzan esta distinción. Los dos primeros componentes principales explican el 33,48 % y el 18,27 % de la varianza total, respectivamente. En este subespacio compartido, M1 presenta una distribución en gran medida continua y solapada, estructurada en torno a un número limitado de direcciones lineales principales. En cambio, M2 muestra regiones densas más separadas y áreas aisladas de baja densidad, en consonancia con una estructura de agrupamiento local más marcada y una varianza distribuida a lo largo de múltiples dimensiones.

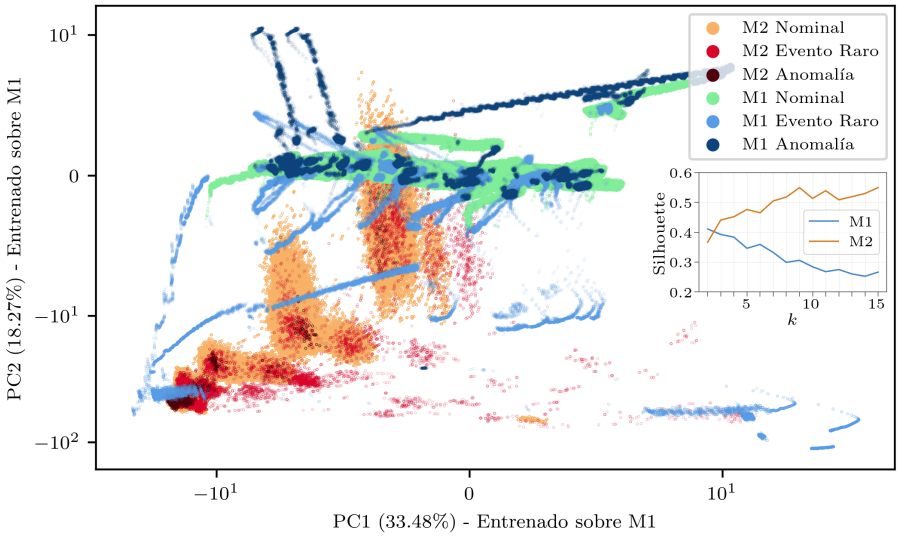


Figura 1. Proyección PCA de M1 y M2.

En conjunto, estas observaciones motivan la hipótesis de que en M1 el comportamiento anómalo puede manifestarse como desviaciones sutiles incrustadas en dinámicas estables, mientras que en M2 las anomalías pueden alinearse con separaciones estructurales. El análisis estadístico desaconseja un enfoque basado en un único modelo y motiva una selección guiada por la estructura. Modelos genéricos independientes (por ejemplo, un único *Local Outlier Factor* (LOF) o un *autoencoder*) mostraron un rendimiento deficiente en experimentos preliminares, lo que confirma que se requieren estrategias adaptadas para capturar la variabilidad entre misiones.

2.5. Selección y entrenamiento de modelos

La estrategia de selección de modelos se basa en la caracterización estadística presentada en la sección anterior. Dado que las dos misiones presentan estructuras estadísticas claramente diferenciadas, se adoptan enfoques de modelado distintos dentro de un mismo marco metodológico.

Modelo de reconstrucción temporal. M1 se caracteriza por una baja variabilidad y una débil estructura de agrupamiento intrínseca. Estas propiedades sugieren que las anomalías pueden manifestarse como desviaciones temporales sutiles incrustadas en señales por lo demás estables.

En consecuencia, se emplea un *autoencoder* no supervisado basado en redes *Long Short-Term Memory* (LSTM) [7]. Esta arquitectura permite capturar tanto las dinámicas temporales como las relaciones no lineales entre variables, lo cual resulta adecuado en escenarios caracterizados por señales estables y anomalías de baja magnitud. En cada ventana mensual w, la telemetría preprocesada se representa como una serie multivariante con N_w muestras y F variables. A partir de esta serie se construyen secuencias deslizantes $\mathbf{x}_t \in \mathbb{R}^{L \times F}$ de longitud L, desplazadas una muestra respecto a la secuencia anterior e indexadas por su instante final t, con $t = L, \ldots, N_w$. Cada secuencia $\mathbf{x}_t$ es reconstruida por el modelo como $\hat{\mathbf{x}}_t$.

Para cumplir con las restricciones de uso exclusivo de CPU y el reentrenamiento incremental mensual, se adopta un *autoencoder* LSTM ligero. Se procesan secuencias de longitud $L = 64$ mediante una arquitectura de una sola capa con 64 unidades ocultas y un espacio latente de 16 dimensiones, limitando la capacidad del modelo para reducir el sobreajuste y garantizar un reentrenamiento eficiente. El modelo se optimiza con Adam, con tasa de aprendizaje 10^{-3}, durante un máximo de 30 épocas y con un tamaño de lote de 128. El *early stopping*, con paciencia igual a 3, monitoriza la pérdida de reconstrucción durante el entrenamiento y se activa en la mayoría de las ventanas mensuales, lo que indica que la convergencia suele alcanzarse antes del límite de épocas. Antes de cada ciclo de entrenamiento, se aplica un filtrado robusto a nivel de muestra basado en la desviación absoluta mediana (MAD) para reducir la influencia de valores atípicos en los patrones de reconstrucción aprendidos.

El error de reconstrucción para cada secuencia se cuantifica mediante el error cuadrático medio (MSE):

$$s_t = \frac{1}{LF} \sum_{i=1}^{L} \sum_{j=1}^{F} \left(x_{t-L+i,\,j} - \hat{x}_{t-L+i,\,j} \right)^2, \qquad t = L, \ldots, N_w. \tag{1}$$

Por tanto, para cada ventana mensual w se obtiene una puntuación de anomalía por cada secuencia deslizante, es decir, $\mathcal{S}_w = \{s_t\}_{t=L}^{N_w}$. Estas puntuaciones pueden variar entre ventanas temporales debido a la dinámica no estacionaria del sistema, lo que hace que los umbrales fijos sean poco fiables. Por ello, se establece un umbral inicial en el percentil 98 del error de reconstrucción en entrenamiento, que posteriormente se refina mediante el procedimiento adaptativo basado en EVT descrito en la Sección 2.6.

Marco basado en *clustering*. M2 presenta alta variabilidad, una estructura de agrupamiento pronunciada y una varianza distribuida a lo largo de múltiples dimensiones. Estas características indican la presencia de regímenes operacionales bien definidos y espacios de estado localmente estructurados. Para explotar esta estructura, se adopta un marco de detección de anomalías basado en *clustering*. El proceso consta de: (1) escalado robusto y filtrado preliminar no supervisado de valores atípicos extremos; (2) PCA para aprender una representación compacta del comportamiento nominal; (3) *clustering* mediante *k-means* para modelar los regímenes operacionales en el subespacio reducido; (4) LOF [1] para cuantificar desviaciones de densidad local; y (5) una puntuación de anomalía combinada que integra la proximidad a los regímenes y la información de densidad local.

PCA reduce la dimensionalidad preservando la estructura dominante del funcionamiento normal. *k-means* captura modos operacionales recurrentes, mientras que LOF evalúa el grado de desviación local dentro de cada régimen. La puntuación de anomalía se define como la combinación ponderada

$$s_t = 0{,}7\, s_{\mathrm{LOF}}(t) + 0{,}3\, d_{\mathrm{centroid}}(t), \tag{2}$$

donde $s_{\mathrm{LOF}}(t)$ y $d_{\mathrm{centroid}}(t)$ son, respectivamente, la puntuación LOF y la distancia al centroide más cercano en el subespacio PCA, ambas ya normalizadas. De este modo, se priorizan las distorsiones de densidad local sin perder sensibilidad ante cambios globales de régimen.

Los hiperparámetros se infieren automáticamente a partir de cada ventana de entrenamiento sin ajuste supervisado. La dimensionalidad de PCA se establece para explicar el 95 % de la varianza acumulada. El número de clústeres se estima utilizando el *gap statistic* [12] dentro de un rango dependiente de los datos, y el tamaño del vecindario de LOF escala con $\sqrt{n}$ y la dimensionalidad reducida para garantizar una estimación de densidad estable. Las puntuaciones de anomalía se calibran inicialmente utilizando el percentil 98 antes de aplicar el procedimiento de umbralización dinámica descrito en la Sección 2.6.

2.6. Umbralización dinámica

En escenarios de monitorización dinámica, la distribución de las puntuaciones de anomalía puede fluctuar significativamente entre periodos de evaluación debido a cambios operacionales, variaciones en los regímenes de funcionamiento y efectos propios de la no estacionariedad de la telemetría. En M1, estas puntuaciones corresponden al error de reconstrucción del *autoencoder*, mientras que en M2 proceden de la combinación entre desviación de densidad local y proximidad a los regímenes operacionales estimados. Aunque su origen es distinto en cada misión, en ambos casos valores elevados de la puntuación indican un mayor grado de desviación respecto al comportamiento aprendido como nominal.

Estas variaciones hacen que el uso de umbrales estáticos definidos *a priori* sea inadecuado, especialmente cuando se requiere controlar las falsas alarmas bajo distribuciones no estacionarias. Un umbral fijo puede resultar demasiado permisivo en periodos de baja variabilidad, generando sobredetección, o demasiado restrictivo en periodos de mayor dispersión, reduciendo la capacidad de detectar anomalías reales. Para abordar este problema, se utiliza EVT [9] con el fin de estimar, en cada ventana de evaluación, un umbral adaptativo basado en la cola superior de la distribución de puntuaciones.

Sea $\mathcal{S}_r = \{s_i\}_{i=1}^{n_r}$ el conjunto de n_r puntuaciones de anomalía obtenidas en un periodo de evaluación r. El operador $Q_q(\mathcal{S}_r)$ denota el cuantil empírico de orden q, es decir, el valor por debajo del cual se sitúa aproximadamente una fracción q de las puntuaciones. Siguiendo el enfoque *peaks-over-threshold*, se selecciona un umbral preliminar $u = Q_q(\mathcal{S}_r)$ y se modela únicamente la cola superior de la distribución. Para ello, se define el conjunto de excesos sobre dicho umbral como

$$\mathcal{X}_r = \{s_i - u \mid s_i > u,\ i = 1, \ldots, n_r\}.$$

Si el número de excesos es suficiente, esto es, si $|\mathcal{X}_r| \geq N_{\min}$, se ajusta una distribución de Pareto generalizada $\mathrm{GPD}(\xi, \sigma)$ a $\mathcal{X}_r$ mediante máxima verosimilitud, donde ξ y σ representan los parámetros de forma y escala, respectivamente. A partir de este ajuste, y para una tasa de falsos positivos objetivo α, el nuevo umbral del periodo r se define como:

$$T_{\mathrm{new},r} = \begin{cases} u + \frac{\sigma}{\xi}\left(\left(\frac{\alpha}{1-q}\right)^{-\xi} - 1\right) & \text{si } |\xi| > \varepsilon, \\ u + \sigma \log\left(\frac{1-q}{\alpha}\right) & \text{en otro caso,} \end{cases} \tag{3}$$

donde ε es una pequeña tolerancia numérica. Si $|\mathcal{X}_r| < N_{\min}$, se establece $T_{\mathrm{new},r} = Q_{1-\alpha}(\mathcal{S}_r)$.

Para garantizar la estabilidad temporal, sea T_{r-1} el umbral utilizado en el periodo anterior. En el primer periodo, este valor se inicializa con el umbral calculado sobre las puntuaciones de entrenamiento. Primero se aplica suavizado exponencial:

$$\tilde{T}_r = \alpha_{\mathrm{smooth}} \cdot T_{\mathrm{new},r} + (1 - \alpha_{\mathrm{smooth}}) \cdot T_{r-1}, \tag{4}$$

y posteriormente se limita la variación relativa del umbral respecto a T_{r-1}:

$$T_r = \mathrm{mín}\left(\mathrm{máx}\left(\tilde{T}_r,\, T_{r-1}(1-\delta)\right), T_{r-1}(1+\delta)\right), \quad \delta > 0. \tag{5}$$

Finalmente, cada puntuación $s_i \in \mathcal{S}_r$ se considera anómala si $s_i > T_r$. En la evaluación experimental se fijan $\alpha = 0{,}01$, $q = 0{,}98$, $\alpha_{\text{smooth}} = 0{,}05$, $\delta = 0{,}05$ y $N_{\min} = 30$, siguiendo prácticas estándar en modelos *peaks-over-threshold* y monitorización orientada a la precisión.

2.7. Agrupación de anomalías tras la detección

Siguiendo la práctica operativa del sistema PitIA [11], las detecciones separadas por menos de 6 horas se consolidan en un único evento, reduciendo alarmas redundantes y mejorando la claridad diagnóstica. Esta agrupación evita que activaciones próximas en el tiempo, asociadas a una misma incidencia, se interpreten como eventos independientes. Este umbral refleja recomendaciones de expertos de centros de operaciones de satélites.

3. Resultados

La Tabla 1 resume el rendimiento global a nivel de evento para ambas misiones. En esta evaluación, un evento corresponde a un intervalo anotado, ya sea anomalía o evento raro, mientras que una instancia representa una observación individual de la serie temporal. Por ello, precisión, *recall* y $F_{0,5}$ se calculan a nivel de evento, mientras que la tasa de verdaderos negativos se calcula a nivel de instancia para penalizar la sobredetección en periodos nominales.

En M1, el marco alcanza una alta precisión con un *recall* moderado, obteniendo $F_{0,5} = 0{,}700$. Este comportamiento conservador es operativamente deseable: el marco detecta menos eventos, pero con alta confianza, minimizando intervenciones innecesarias por parte de los operadores. En M2, el compromiso es más equilibrado ($F_{0,5} = 0{,}698$), detectando una mayor proporción de anomalías a costa de una menor precisión.

Tabla 1. Rendimiento a nivel de evento para M1 y M2 bajo la configuración incremental propuesta.

Métrica	**Misión 1 (M1)**	**Misión 2 (M2)**
Precisión a nivel de evento	0,852	0,721
Recall a nivel de evento	0,408	0,621
$F_{0,5}$ a nivel de evento	0,700	0,698
Tasa de verdaderos negativos	0,977	0,801

Dentro del ecosistema ESA-ADB, soluciones industriales no supervisadas como el sistema PitIA [11] han reportado valores de $F_{0,5}$ a nivel de evento de 0,424 para la Misión 1 y 0,882 para la Misión 2 bajo configuraciones estáticas de entrenamiento/prueba. En comparación, el marco propuesto obtiene resultados sustancialmente mejores en la Misión 1, manteniendo un rendimiento competitivo en la Misión 2. Estos resultados se obtienen bajo un esquema estrictamente

incremental con reentrenamiento mensual desde el segundo mes de operación, sin depender de particiones históricas extensas ni de un ajuste específico de hiperparámetros por misión, y dentro de una configuración orientada al despliegue.

Además de los valores agregados, resulta útil analizar la evolución temporal del rendimiento para comprobar la estabilidad del marco a lo largo de las ventanas mensuales. Las Figuras 2 y 3 muestran la evolución temporal del $F_{0,5}$ corregido acumulado. En M1, los eventos anómalos y raros se concentran en el año 2000, predominantemente en la clase de evento 22 [5, Fig. 1], y posteriormente se vuelven más dispersos. La curva de $F_{0,5}$ refleja este comportamiento: valores superiores a 0,9 durante el periodo inicial denso, estabilizándose gradualmente entre 0,7 y 0,8 a medida que las anomalías se distribuyen entre distintas clases. En M2, las clases de eventos son más constantes a lo largo de la misión. El $F_{0,5}$ acumulado desciende brevemente durante los primeros meses, pero se recupera rápidamente, estabilizándose en torno a 0,7 con oscilaciones más reducidas.

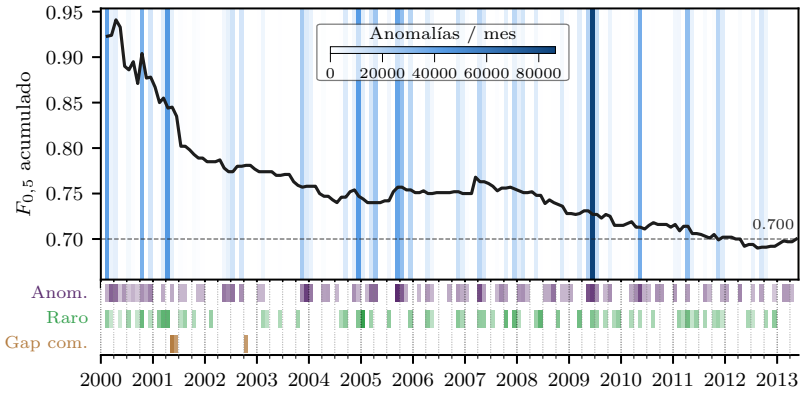


Figura 2. $F_{0,5}$ corregido acumulado a lo largo del tiempo para la Misión 1.

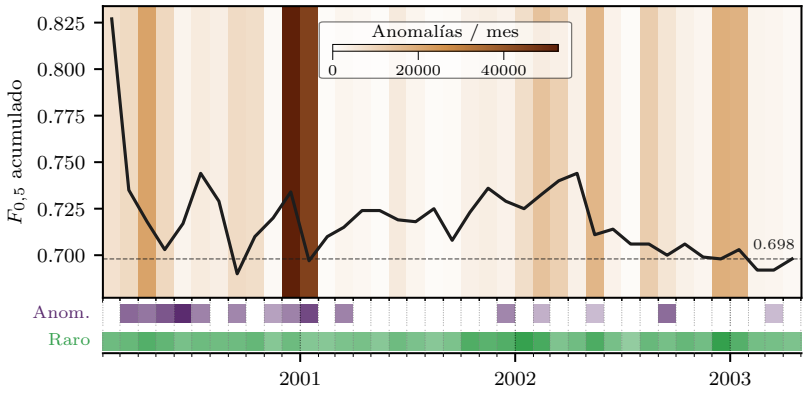


Figura 3. $F_{0,5}$ corregido acumulado a lo largo del tiempo para la Misión 2.

Coste computacional. Bajo las mismas restricciones de hardware (solo CPU, 16 GB de RAM), la evaluación en M1 a lo largo de 161 ventanas mensuales requirió un promedio de 27,5 épocas por ventana, con un tiempo medio de reentrenamiento de 1041,71 s/ventana (total: 46,64 h) y un uso de memoria máximo medio de 7,72 GB (máx.: 8,11 GB). La evaluación en M2, a lo largo de 40 ventanas mensuales, requirió 214,65 s/ventana (total: 2,5 h) y un uso de memoria máximo medio de 0,54 GB (máx.: 1,09 GB). No se observó acumulación de memoria entre ventanas en ninguna de las misiones. Estos valores son compatibles con un esquema de reentrenamiento mensual fuera de línea bajo restricciones de hardware moderadas.

4. Discusión

Aunque el marco se evalúa bajo un régimen estrictamente incremental, el análisis exploratorio de la Sección 2.4 se realiza sobre los conjuntos de datos completos de las misiones únicamente con fines descriptivos, con el objetivo de evidenciar diferencias estadísticas globales y justificar la elección de distintas familias de modelos. No se utiliza para el ajuste de hiperparámetros, la calibración de umbrales ni ninguna decisión dependiente de etiquetas. En escenarios operacionales, puede emplearse una breve fase inicial de *warm-up* con detectores genéricos para acumular datos históricos, tras la cual una evaluación estadística restringida al periodo observado puede guiar la selección del modelo sin introducir fuga de información.

Para un contexto más amplio, ESA-ADB [5] reporta resultados bajo una partición estática 50/50 de entrenamiento/prueba. Utilizando todos los canales, los mejores modelos alcanzan $F_{0,5} = 0{,}061$ para M1 y $F_{0,5} = 0{,}241$ para M2. Los enfoques supervisados con selección de subconjuntos de canales alcanzan $F_{0,5} = 0{,}786$ y $F_{0,5} = 0{,}949$, respectivamente. Aunque no son directamente comparables, el marco propuesto alcanza $F_{0,5} = 0{,}700$ y $F_{0,5} = 0{,}698$ sin utilizar etiquetas ni selección de canales, reduciendo sustancialmente la brecha con los métodos supervisados. Esto es particularmente significativo para M1, donde los *baselines* no supervisados existentes presentan mayores dificultades.

Los resultados también reflejan el compromiso entre precisión y *recall* asociado a una monitorización orientada a reducir falsas alarmas. En aplicaciones operacionales, este balance puede ajustarse en función de la criticidad del escenario y del coste asociado a no detectar determinados eventos.

Los hiperparámetros se seleccionaron siguiendo prácticas estándar en lugar de optimización supervisada, en coherencia con el escenario de despliegue completamente no supervisado. Cuando se dispone de datos históricos etiquetados, el marco podría ampliarse mediante estrategias de AutoML [4] para optimizar conjuntamente las arquitecturas de los modelos, los hiperparámetros y los umbrales mediante procedimientos sistemáticos de búsqueda guiados por el $F_{0,5}$ a nivel de evento. Esto permitiría una selección de modelos adaptativa y reproducible entre misiones, reduciendo el sesgo asociado al ajuste manual.

5. Conclusiones y trabajo futuro

En este trabajo se ha presentado un marco no supervisado y orientado al despliegue para la detección de eventos anómalos en telemetría satelital multivariante. La propuesta combina una caracterización estadística previa de cada misión, reentrenamiento incremental mensual y umbralización adaptativa basada en EVT, permitiendo generar predicciones desde el segundo mes de operación sin utilizar etiquetas, información semántica de la misión, selección supervisada de canales ni hardware GPU dedicado.

Los resultados sobre ESA-ADB muestran que la adaptación del paradigma de detección a la estructura estadística de cada misión permite obtener un rendimiento competitivo bajo una configuración estrictamente cronológica. En M1, el modelo de reconstrucción temporal basado en LSTM alcanza una alta precisión a nivel de evento, mientras que en M2 el enfoque basado en PCA, *k-means* y LOF proporciona un compromiso más equilibrado entre precisión y *recall*. Estos resultados refuerzan la idea de que no existe una única estrategia de modelado adecuada para todas las misiones, y que la variabilidad entre regímenes operacionales debe considerarse explícitamente en el diseño de detectores no supervisados.

Como trabajo futuro, se plantea extender el marco hacia una selección de modelos más automática y reproducible, incorporando una caracterización estadística periódica, por ejemplo mensual, que permita revisar si el paradigma de detección sigue siendo adecuado para la dinámica observada de la misión. Esta extensión permitiría adaptar no solo los parámetros del modelo, sino también la propia familia de detectores ante cambios persistentes en los regímenes operacionales. Asimismo, sería interesante integrar mecanismos de aprendizaje activo o *one-shot learning* para memorizar eventos nominales raros y reducir alarmas repetidas asociadas a operaciones especiales de la misión. Finalmente, la evaluación podría ampliarse con métricas de diagnóstico a nivel de canal o subsistema, así como con nuevas misiones y escenarios de monitorización en línea, con el objetivo de acercar el marco a sistemas operacionales de soporte a operadores.

Agradecimientos

Este trabajo se ha desarrollado en el marco del proyecto INCIBE Módulo de detección de anomalías de comportamiento basado en IA y ML, ref. 250120UCTR-INCIBE (Laboratorio SIMD-I3A), y ha contado con financiación asociada a contratos de asistencia técnica financiados por GMV Soluciones Globales Internet, S.A.

También ha sido parcialmente financiado por la Universidad de Castilla-La Mancha y el «FEDER Una manera de hacer Europa» en el marco del proyecto 2025-GRIN-38476.

Referencias

1. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: Identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. pp. 93–104 (2000). https://doi.org/10.1145/342009.335388
2. Good, P.I.: Permutation, Parametric and Bootstrap Tests of Hypotheses. Springer, 3 edn. (2005)
3. Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology **24**(6), 417–441 (1933). https://doi.org/10.1037/h0071325
4. Hutter, F., Kotthoff, L., Vanschoren, J.: Automated Machine Learning: Methods, Systems, Challenges. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-05318-5
5. Kotowski, K., Haskamp, C., Andrzejewski, J., Ruszczak, B., Nalepa, J., Lakey, D., Collins, P., Kolmas, A., Bartesaghi, M., Martinez-Heras, J., Canio, G.D.: European space agency benchmark for anomaly detection in satellite telemetry. arXiv preprint arXiv:2406.17826 (2025), https://arxiv.org/abs/2406.17826
6. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. vol. 1, pp. 281–297. University of California Press (1967)
7. Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., Shroff, G.: LSTM-based encoder-decoder for multi-sensor anomaly detection. arXiv preprint arXiv:1607.00148 (2016), https://arxiv.org/abs/1607.00148
8. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (1987). https://doi.org/10.1016/0377-0427(87)90125-7
9. Siffer, A., Fouque, P.A., Termier, A., Largouet, C.: Anomaly detection in streams with extreme value theory. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1067–1075 (2017). https://doi.org/10.1145/3097983.3098144
10. Székely, G.J., Rizzo, M.L.: Energy statistics: A class of statistics based on distances. Journal of Statistical Planning and Inference **143**(8), 1249–1272 (2013). https://doi.org/10.1016/j.jspi.2013.03.018
11. Tejedor, M., Jiménez, H., Pozo, J.A., Perea, I., Auñón, J.M.: Enhancing space operations with unsupervised anomaly detection: The PitIA system. In: Proceedings of the 2025 Conference on Big Data from Space (BiDS'25). pp. 57–60. Publications Office of the European Union (2025). https://doi.org/10.2760/2119408
12. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society Series B: Statistical Methodology **63**(2), 411–423 (2002). https://doi.org/10.1111/1467-9868.00293
13. Widmer, G., Kubat, M.: Learning in the presence of concept drift and hidden contexts. Machine Learning **23** (1996). https://doi.org/10.1007/BF00116900