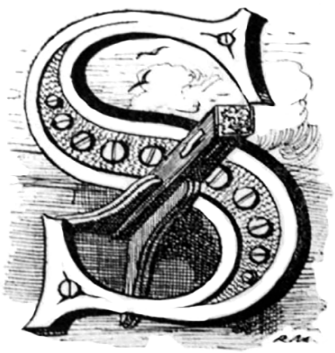
BÚSQUEDA DE ANOMALÍAS CON INTELIGENCIA ARTIFICIAL EN DATOS DE INTERÉS NAVAL CON MÉTODOS COMPUTACIONALMENTE LIGEROS (II)

Francisco LAMAS LÓPEZ
Doctor ingeniero ENPC Paris Tech

Rodrigo SANCHO MOYA
Magíster en Ciencia de Datos

Aprendizaje máquina aplicado al análisis de datos de posicionamiento en series temporales



I bien en el anterior artículo se describía una metodología para detectar anomalías en conjuntos de datos en series temporales ligadas a activos de plataforma naval, en este se trata una temática distinta, de un entorno potencialmente más operacional y con un enfoque diferente.

En entornos operacionales, por ejemplo de vigilancia de tráfico durante la navegación (1) aérea (2) o marítima (3), se registran series temporales de datos de posicionamiento (figura 1) de los nodos vigilados (aviones o buques).

---

(1) LI, M.; LI, B.; QI, Z.; LI, J.; WU, J. (2024): «Enhancing Maritime Navigational Safety: Ship Trajectory Prediction Using ACoAtt-LSTM and AIS Data». *ISPRS International Journal of Geo-Information*, 13(3), p. 85, *https://doi.org/10.3390/ijgi13030085*

(2) STANDFUSS, T.; HIRTE, G.; SCHULTZ, M.; FRICKE, H. (2024): «Efficiency assessment in European air traffic management. A fundamental analysis of data, models, and methods». *Journal of Air Transport Management,* 115, 102523,*https://doi.org/10.1016/j.jairtraman.2023.102523*

(3) DARANDA, A.; DZEMYDA, G. (2020): «Navigation Decision Support: Discover of Vessel Traffic Anomaly According to the Historic Marine Data». *International Journal of*

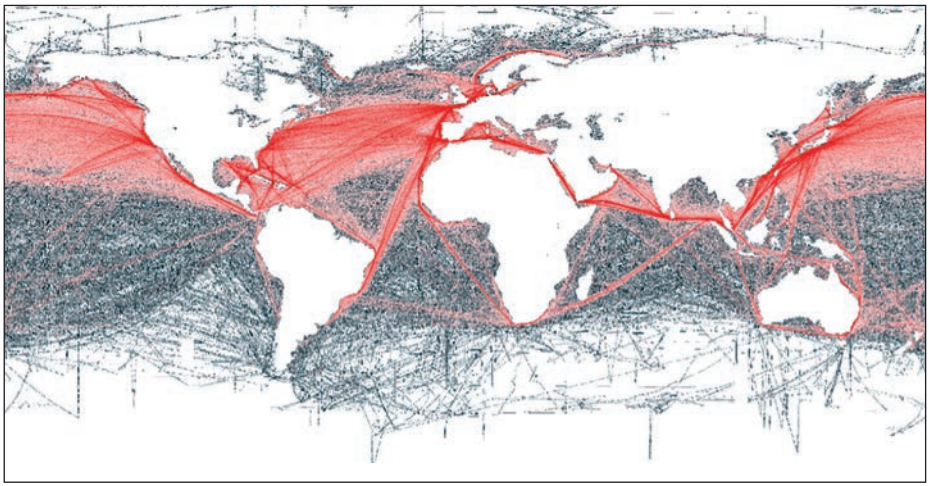


Figura 1. Las rutas de tráfico marítimo global han crecido un 300 por 100 desde 1992. (Fuente: TOURNADRE, 2014) (4)

Esto es típico en centros de vigilancia de tráfico aéreo civiles (5) y militares, como los de control de tráfico marítimo (6). En este artículo se discute la necesidad de evaluar las anomalías de comportamiento de nodos en navegación a través de las series temporales de las trazas registradas, y se propone una metodología adecuada para control con *machine learning* (aprendizaje máquina) de aquellas trazas que presentan anomalías respecto a la normalidad entrenada para cada ruta seguida. Siempre que se trate el tema de navegación en este artículo se referirá a ambos dominios, aéreo y marítimo, indistintamente, por la capacidad de aplicar las técnicas descritas en sendos campos.

En los últimos años, el tráfico aéreo y marítimo ha experimentado un incremento significativo (figura 2), lo que plantea desafíos considerables en cuanto a la seguridad de la navegación. La seguridad en el tráfico de navegación aérea y marítima depende en gran medida de las decisiones tomadas por sus tripulaciones y de las situaciones específicas que enfrenten (meteorológicas y otras). La

---

Computers Communications & Control, 15(3), 3864, *https://doi.org/10.15837/ijccc.2020.3.3864*

(4) TOURNADRE, J. (2014): «Anthropogenic pressure on the open ocean: The growth of ship traffic revealed by altimeter data analysis». *Geophysical Research Letters*, 41(22), 7924-7932, *https://doi.org/10.1002/2014GL061786*

(5) EUROCONTROL (2022): «Analysis Paper: Performance 2022. The year European aviation bounced back, despite war & Omicron/COVID». *Aviation Intelligence Unit*.

(6) MADARIAGA, E.; ORTEGA, A.; ORIA, J. M.; DÍAZ, E.; SALAMA, R. (2015): «Maritime Security with website of the Spanish Armada». *ResearchGate*.

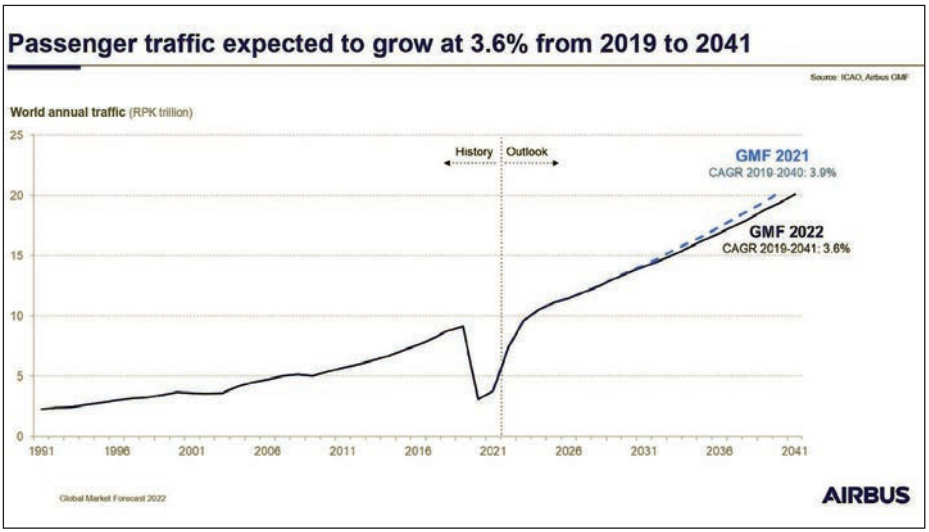


Figura 2. Evolución de la demanda del tráfico aéreo de pasajeros.
(Fuente: SHPARBERG & LANGE, 2022) (7)

detección de anomalías en el tráfico controlado es crucial para prevenir situaciones peligrosas y tomar decisiones oportunas que garanticen una navegación segura.

En este contexto, este artículo presenta un método para detectar anomalías en el tráfico aéreo (con datos obtenidos de EUROCONTROL), que es directamente aplicable al control del tráfico marítimo, combinando el algoritmo de agrupamiento DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)* (8) con un análisis de los k-vecinos más cercanos entre los clústeres. Este enfoque se aplica a datos históricos del tráfico, concretamente a puntos de giro de las aeronaves, proporcionando un criterio numérico formal para distinguir entre casos de tráfico normal y situaciones anómalas.

El propósito de realizar este tipo de análisis es doble. Por un lado, busca mejorar la toma de decisiones en materia de navegación mediante la identificación temprana de comportamientos inusuales que puedan indicar potenciales riesgos o actividades ilegales, como por ejemplo sería en tráfico marítimo la pesca

(7) SHPARBERG, S.; LANGE, B. (2022): «Global Market Forecast 2022», *https://www.airbus.com/sites/g/files/jlcbta136/files/2022-07/GMFPresentation-2022-2041.pdf*

(8) ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. (1996): «A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise». *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. Institute for Computer Science, University of Munich.

ilegal, el contrabando u otros. Por otra parte, este análisis responde a la necesidad de adaptar y desarrollar métodos que puedan manejar la complejidad y la especificidad de los datos de navegación aérea, que son dinámicos y multidimensionales. Dado que el error humano se identifica como una causa principal de accidentes, el desarrollo de herramientas automatizadas de apoyo a la decisión que puedan alertar a las autoridades competentes sobre anomalías en tiempo real es esencial para mejorar la seguridad y la eficiencia del tráfico, tanto aéreo como marítimo.

En la literatura existente en el estado del arte se comprueba que éste es un tema de alto interés y se exploran diversas técnicas de aprendizaje profundo para la predicción de trayectorias de navegación, con énfasis en la mejora de la seguridad. Un enfoque es el uso de redes neuronales LSTM (*Long Short-Term Memory*) para predecir trayectorias. Este método se distingue por su habilidad para procesar secuencias temporales de datos AIS (*Automatic Identification System*) y capturar relaciones complejas entre atributos dinámicos de los barcos, como velocidad y dirección.

Además del LSTM, existen otras técnicas avanzadas en el campo del aprendizaje automático y profundo que podrían aplicarse a la predicción de trayectorias en series temporales. Entre ellas, se encuentran las redes neuronales convolucionales (CNN), que pueden ser útiles para capturar patrones espaciales en datos geolocalizados, y los modelos generativos adversarios (GAN), que podrían emplearse para generar trayectorias futuras potenciales basadas en patrones aprendidos de datos históricos de navegación. Estas técnicas, al ser exploradas y adaptadas al contexto específico de la navegación marítima, podrían ofrecer nuevas perspectivas y mejoras significativas en la predicción de trayectorias, contribuyendo así a una navegación más segura y eficiente. Sin embargo, el entrenamiento de estos modelos suele ser más costoso que otros de *machine learning* más sencillos y que pueden dar resultados satisfactorios para los objetivos planteados.

Justificación del modelo DBSCAN para detección de anomalías en navegación

El modelo DBSCAN para la clasificación de normalidad en trazas de vuelos comerciales se fundamenta en varias particularidades y ventajas que se alinean con los objetivos y necesidades de la investigación, debido a su capacidad de análisis de datos espaciales complejos. Primero, DBSCAN es capaz de identificar clústeres de formas arbitrarias, lo que lo distingue de otros algoritmos de agrupamiento de datos, como K-means, que se limita a identificar clústeres de formas esféricas. Esta característica es crucial para el análisis de trazas de vuelos comerciales, ya que las rutas pueden adoptar diversas configuraciones espaciales en función de múltiples factores, como restricciones geográficas y

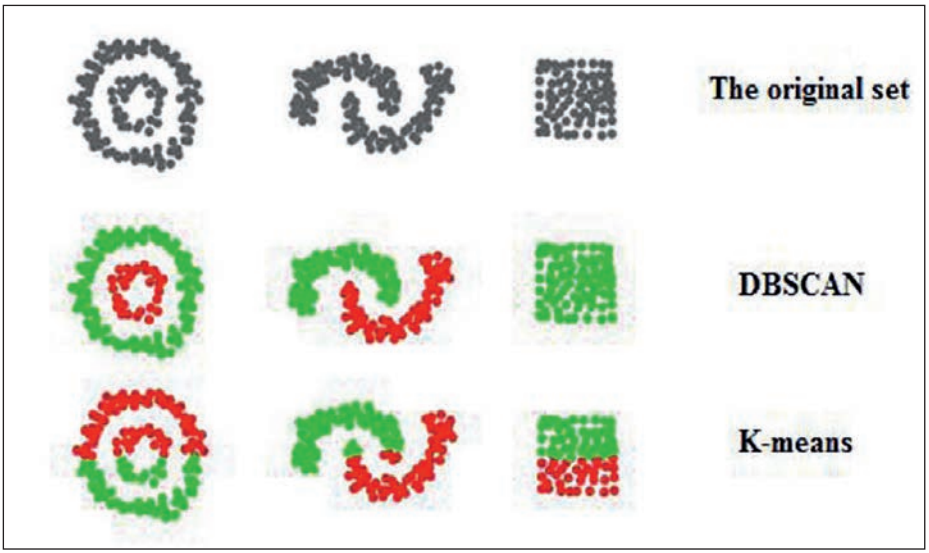


Figura 3. Una comparación esquemática de la naturaleza de los clústeres formados por los algoritmos DBSCAN y K-means. (Fuente: MOCHURAD, L., *et al*., 2023) (9)

regulaciones aéreas. Además, DBSCAN minimiza la influencia de puntos atípicos o ruido al no forzar la inclusión de estos puntos en clústeres, a diferencia de algoritmos como K-means (figura 3), donde cada punto debe ser asignado a un clúster. Esto resulta especialmente útil en el contexto de navegación que atañe a este artículo, donde es común encontrar datos anómalos o comportamientos atípicos debidos a emergencias, condiciones meteorológicas adversas o desviaciones de ruta por razones de seguridad.

El algoritmo se basa en la densidad de los puntos, agrupando los que están próximos entre sí y marcando como ruido aquéllos que se hallan en regiones de baja densidad. Este enfoque permite una representación más fiel de las concentraciones naturales de las trazas, facilitando la identificación de patrones normales y anómalos en los vuelos. DBSCAN requiere dos parámetros principalmente: ε (10), que define la máxima distancia entre dos puntos para que sean considerados vecinos, y minPts, el número mínimo de puntos

(9) MOCHURAD, L., SYDOR, A.; RATINSKIY, O. (2023): «A fast parallelized DBSCAN algorithm based on OpenMP for detection of criminals on streaming services». *Frontiers in Big Data*, 6, 1292923, *https://doi.org/10.3389/fdata.2023.1292923*

(10) RAHMAH, N.; SUKAESIH SITANGGANG, I. (2016): «Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra». IOP Conference Series: Earth and Environmental Science, 31.

requeridos para formar un clúster. Estos parámetros permiten una gran flexibilidad en la definición de qué constituye un clúster, adaptándose a las variaciones en la densidad de los datos.

Las ecuaciones que describen el modelo DBSCAN son fundamentales para entender su funcionamiento. Un punto p es considerado como punto núcleo si el número de puntos dentro de un radio ε (incluido él mismo) es al menos minPts. Un punto q es directamente alcanzable desde p si q está a no más distancia de ε de p, y p es un punto núcleo. Un clúster se forma entonces por un punto núcleo p y todos los puntos que son alcanzables directa o indirectamente desde p. Los puntos que no son alcanzables desde ningún punto núcleo son considerados ruido. Estas definiciones permiten a DBSCAN adaptarse a la densidad variable de los datos, identificando clústeres basados en la proximidad espacial y la densidad de puntos.

Las ecuaciones fundamentales y condiciones que describen el modelo DBSCAN, centradas en los conceptos de densidad y alcanzabilidad, son las siguientes:

— **Punto núcleo (*core point*):** un punto p de la traza de navegación es considerado un punto núcleo si el número de puntos dentro de su vecindario ε (incluyéndose a sí mismo) alcanza o supera un umbral mínimo minPts. Esto se formaliza como:

$$N_\epsilon(p) = \{q \in D | dist(p,q) \leq \epsilon\}$$

donde $N\varepsilon(p)$ es el ε-vecindario de p, incluyendo todos los puntos q, tal que la distancia entre p y q no excede ε, y D es el conjunto de datos. Si $/N\varepsilon(p)/ \geq minPts$, entonces p es un punto núcleo.

— **Punto directamente alcanzable:** un punto q es directamente alcanzable desde un punto p si q está dentro del ε-vecindario de p y p es un punto núcleo. Esto implica una relación directa que depende del valor de ε y minPts.

— **Punto alcanzable:** un punto q es alcanzable desde p si existe una secuencia de puntos $p1, p2 \ldots pn$ tal que $p1 = p$, $pn = q$, y $pi + 1$ es directamente alcanzable desde pi para $1 \leq i < n$. Esto establece una cadena de puntos directamente alcanzables que conecta p con q.

— **Ruido:** un punto que no es alcanzable desde ningún punto núcleo es considerado ruido. En otras palabras, si un punto p no cumple las condiciones para ser un punto núcleo o no es alcanzable desde algún punto núcleo, entonces p es clasificado como ruido.

Del mismo modo que el anteriormente descrito DBSCAN, el modelo HDBSCAN, o Hierarchical DBSCAN, extiende DBSCAN al incorporar una

perspectiva jerárquica, lo que lo hace más flexible en la detección de clústeres con diferentes densidades. No necesita el parámetro ε; en cambio, se basa en un parámetro de persistencia de clústeres para determinar qué clústeres son significativos. Este enfoque jerárquico permite que HDBSCAN se ajuste a variaciones locales de densidad, haciéndolo más adecuado para datos complejos o con estructuras de clústeres anidados. Además, HDBSCAN es más robusto en la selección de parámetros y a menudo produce resultados de agrupamiento más consistentes.

Para la detección de anomalías en series temporales de tráfico de navegación, la elección entre DBSCAN y HDBSCAN depende de la naturaleza de los datos. DBSCAN puede ser útil si el tráfico aéreo tiene patrones de densidad relativamente uniformes y si los parámetros pueden ser determinados de manera confiable a través de la experiencia o la experimentación. Por otro lado, HDBSCAN es preferible cuando se enfrenta a series temporales más complejas, donde los patrones de tráfico pueden variar ampliamente en densidad debido a factores como los horarios de vuelo o las condiciones meteorológicas. La habilidad de HDBSCAN para adaptarse a variaciones de densidad y su menor sensibilidad a la selección de parámetros lo hacen un candidato fuerte para detectar anomalías en escenarios de tráfico aéreo donde los patrones pueden no ser uniformemente densos o bien definidos.

Los objetivos de este artículo son, pues, comparar los resultados de los algoritmos DBSCAN y HDBSCAN para detección de anomalías en trazas de vuelos comerciales, analizar el impacto de la geometría del conjunto de datos, evaluar el papel del preprocesado de datos en las trazas de series temporales de navegación (aérea en este caso) y finalmente comparar el coste computacional de ambas soluciones.

Datos utilizados para experimentación con el modelo

Para este estudio se utilizó un conjunto de datos de EUROCONTROL (figura 4), organización paneuropea que se encarga, entre otras cosas, de recopilar datos estadísticos y predictivos sobre la aviación y los pone a disposición de la comunidad investigadora. La información obtenida para la investigación incluye mediciones realizadas durante los meses de marzo, junio, septiembre y diciembre, desde marzo de 2015 hasta marzo de 2020.

Los datos están estructurados originalmente, antes del preprocesamiento, en conjuntos separados por mes y año, resultando en un total de 21 conjuntos de datos que requerirían ser unificados para la investigación.

El objeto principal de este *dataset* son los vuelos, y las diferentes dimensiones disponibles son puntos de vuelo (*flight_points*), regiones de información de vuelo por las que se pasa a lo largo del vuelo (*flight*_FIR) y unidades de control aeroespacial del vuelo (*flight*_AUA). Cada vuelo tiene un número

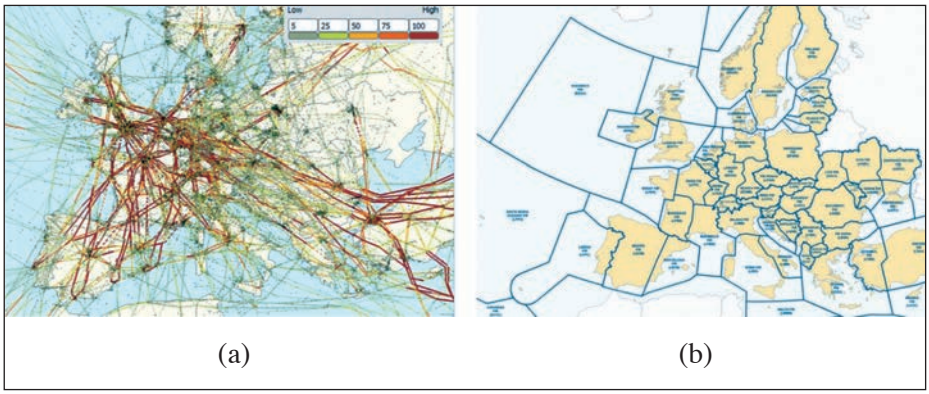


Figura 4. (a) Densidad de rutas de vuelos en tráfico aéreo europeo; (b) Regiones de información de vuelos (FIR) en Europa. (Fuente: STANDFUSS & SCHULTZ, 2018) (11)

de identificación interno denominado ECTRL ID, que será referenciado en las tablas de cada dimensión. Las tablas que modelan la dimensión de los puntos de vuelos disponen de una referencia al identificador interno del vuelo (ECTRL ID), un número en la secuencia de vuelo de dicho punto (*sequence number*), hora (*time over*), altitud, latitud y longitud (*flight level, latitude* y *longitude* respectivamente), entre otros datos presentes en los datos originales.

El conjunto de datos incluye tanto información planificada como real capturada por radares y otras fuentes y abarca vuelos de todos los aeropuertos europeos y algunos internacionales, pero se centra el estudio en 1.710 con origen en Madrid y destino en A Coruña, Oviedo, Bilbao, Barcelona, Córdoba y Jerez entre los años 2015 a 2020, sólo para los cuatro meses de marzo, junio, septiembre y diciembre.

Metodología de análisis de normalidad planteada, preprocesado de los datos y descripción de la experimentación

El objetivo es realizar modelos de normalidad con DBSCAN y HDBSCAN detectando los vuelos que sean categorizados como *outliers* (anómalos) y no se puedan hacer corresponder con los clústeres creados. Estos clústeres, o agrupamientos de trazas de vuelos, son calculados a partir de las distancias (similitud)

(11) STANDFUSS, T.; SCHULTZ, M. (2018): «Performance Assessment of European Air Navigation Service Providers». *Conference Paper*, September 2018, *https://doi.org/10.1109/DASC.2018.8569839*

de las trazas. Es decir, que se crean a partir del histórico de datos de navegación total conjuntos de vuelos asociados por similitud entre ellos a rutas navegadas, y con el modelo entrenado se puede asociar una nueva traza a alguno de los clústeres. En función de la similitud de la traza se asignará a un clúster o se clasificará como *outlier*. Esto es de interés debido a que, con un coste computacionalmente bajo, se puede determinar la normalidad de la trayectoria seguida por una traza de navegación respecto a todas las existentes.

Aunque el objetivo puede parecer trivial, la forma en que se resuelva el problema (el cual tiene infinitas soluciones) determina la bondad de los resultados. El objetivo, como en cualquier problema de *machine learning* (aprendizaje automático), es la correcta generalización del modelo, sin cometer *overfitting* (considerar todos los modelos normales o dentro de un clúster) o hacer *underfitting* (obtener excesivas anomalías y criterios de clúster complejos).

Lo primero que hay que considerar en los datos de trazas de vuelos será filtrar, entre todos los descargados, los necesarios para la experimentación. El conjunto de datos aporta información para cada trayectoria de los aeropuertos de entrada, variables ADEP (aeropuerto de salida) y ADES (aeropuerto de destino). Se usarán esos campos para filtrar la información. También, en algunos experimentos, se propone un paso previo en el que se eliminarán de las trayectorias ya filtradas, los puntos de taxi (figura 5).

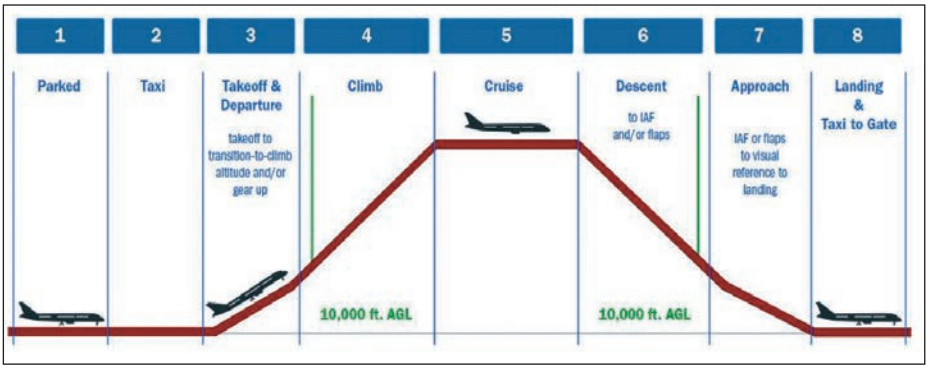


Figura 5. Fases de vuelo de una aeronave. (Fuente: adaptado de HANIFA, *et al*., 2018) (12)

(12) HANIFA, N., *et al*. (2018): «Detection of unstable approaches in flight track with recurrent neural network». In Proceedings of the 2018 International Conference on Information and Communications Technology (ICOIACT) (pp. 735-740). *IEEE, doi:10.1109/ICOIACT. 2018. 8350754*

Se denomina período de taxi a la fase en la que las aeronaves se desplazan desde la posición de estacionamiento hasta la pista de despegue y desde que aterrizan hasta que llegan a la posición de estacionamiento. Para cada trayectoria, los puntos de taxi tienen las mismas coordenadas, las del aeropuerto de salida o de llegada, dependiendo de si se trata del taxi de salida o de llegada. Es por esto por lo que se propone eliminar estos puntos y dejar uno único coincidente con las coordenadas de los aeropuertos.

Por último, y más importante en cuanto al preprocesado de los datos, se procede a interpolar puntos en las trayectorias para poder compararlas, igualando el número de puntos de las trayectorias de cada experimento al número de puntos de la trayectoria con mayor número de puntos. La interpolación se hace de forma lineal en tres dimensiones y no sólo en las coordenadas de longitud y latitud, sino que también se interpola la altura (figura 6).

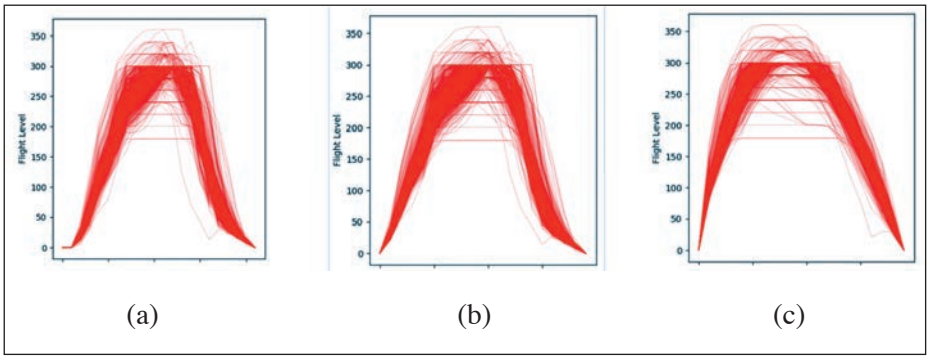


Figura 6. Ejemplo de altimetrías respecto a la distancia de vuelo normalizada en los 757 efectuados entre Madrid-Barcelona durante marzo de 2015: (a) Datos originales, (b) Datos eliminando puntos taxi y (c) Datos con interpolación de puntos respecto al vuelo con mayor número de puntos. (Elaboración propia)

En este artículo se presentan dos experimentos con los vuelos descritos con salida desde Madrid y llegada a los aeropuertos de Alicante, Barcelona, Bilbao, A Coruña, Córdoba, Granada o Jerez durante cuatro meses de 2015. La diferencia entre ambos radica en el preprocesado. En el primero se aplican todos los pasos de preprocesado, pero se dejan intactos los puntos de taxi.

El primer objetivo es crear la matriz de distancias *M* de los 1.710 vuelos; para cada uno se calcula una distancia de traza respecto a todos los demás, obteniendo una matriz cuadrada de dimensiones *n x n* para un conjunto de *n* vuelos.

$$\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1N} \\ m_{21} & m_{22} & \cdots & m_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ m_{N1} & m_{N2} & \cdots & m_{NN} \end{bmatrix}$$

Se calculan estas distancias entre vuelos siguiendo dos metodologías diferentes: la ERP (*Edit Distance with Real Penalty*) y la SSPD (*Symmetrized Segment-Path Distance*).

El algoritmo de ERP (Chen & Ng, 2004) (13) es de la familia del *edit framework*, que nos devuelve el número de acciones necesarias para transformar una trayectoria y añade una penalización al coste de cada acción en base a la distancia que existe entre los segmentos que se están comparando. Este último detalle, el coste en base a la distancia, es la principal diferencia con respecto a EDR (*Endpoint Detection and Response*). En esta definición dos puntos coinciden si se encuentran a menos de una distancia ε, siendo esta distancia mínima mayor que 0.

La SSPD (Besse, Guillouet, Loubes & Royer, 2016) (14) es una medida de distancia propuesta especialmente para ser utilizada en tareas de agrupamiento. Como peculiaridad, discretiza las trayectorias en base a segmentos en vez de puntos (figura 7).

$$SSPD(A,B) = \frac{SPD(A,B)+SPD(B,A)}{2}$$

donde:

$$SPD(A,B) = \frac{1}{m}\sum_{i=1}^{m} D_{pt}(a_i, B)$$

con:

$$D_{pt}(a_i, B) = \min_{j\in[0,m-1]} D_{ps}(a_i, b_j)$$

A diferencia de las anteriores metodologías ERP y SSPD, en la familia de las semejanzas espaciales, la comúnmente utilizada distancia euclídea (ED) no

(13) Chen, L.; Ng, R. (2004): «On the marriage of lp-norms and edit distance». Proceedings of the Thirtieth International Conference on Very Large Data Bases, volume 30 (pp. 792-803). *VLDB Endowment*.

(14) Besse, P.; Guillouet, B.; Loubes, J.-M.; Royer, F. (2016): «Review and perspective for distance-based clustering of vehicle trajectories». *IEEE Transactions on Intelligent Transportation Systems*, 17(11), pp. 3.306-3.317.

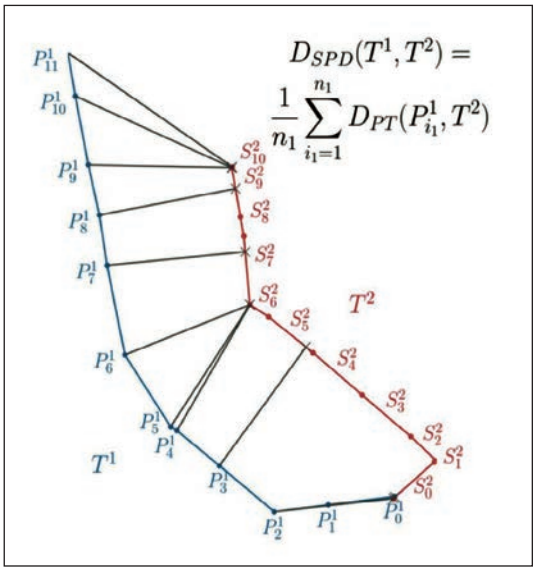


Figura 7. Distancia SSPD entre la trayectoria T^1 y T^2. (Fuente: BESSE, GUILLOUET, LOUBES & ROYER, 2016)

tiene en cuenta el tiempo, solamente la geometría espacial de la trayectoria. Existe además una versión de esta medida en la que se asignan pesos a los puntos en base a la importancia que tienen con respecto al resto (WED, Distancia Euclídea Ponderada) y que también ha sido utilizada para explorar este problema (Corrado, Puranik, Pinon & Mavris, 2020) (15).

$$ED(A,B) = \sqrt{\sum_{i=1}^{n} dist^2(a_i, b_i)}$$

Con la matriz *M* calculada con los métodos ERP y SSPD, se entrenan los modelos DBSCAN y HDBSCAN, utilizando el módulo clúster de la librería *scikit-learn*. Se itera cada uno de ellos calculando el valor óptimo de distancia ε que minimiza la varianza interna de los clústeres (uno para SSPD y otro para ERP, que son utilizados en el modelo DBSCAN). Esto se hace con el «método de la rodilla», representando el gráfico del k-ésimo más cercano. En el eje *X* del gráfico se muestran todos los puntos del conjunto de datos ordenados de acuerdo con la distancia al k-ésimo vecino más cercano, generalmente comenzando por el punto con la menor distancia. Mientras que en el eje *Y* se representa la distancia al k-ésimo vecino más cercano para cada punto. La distancia se calcula normalmente con una métrica de distancia como la euclidiana (figura 8).

El punto de inflexión o «rodilla» suele ser una buena elección para épsilon (ε) porque representa un aumento en la distancia del k-ésimo vecino que no se debe a la variación natural de las distancias dentro de los clústeres, sino al espacio entre éstos, y se cal-culan los minPts (y *min_cluster_size* en el caso de HDBSCAN) a tener en cuenta durante el entrenamiento de los modelos. Para validar la elección de minPts y *min_cluster_size* se evalúan los modelos con

(15) CORRADO, S. J.; PURANIK, T. G.; PINON, O. J.; MAVRIS, D. N. (2020): «Trajectory Clustering within the Terminal Airspace Utilizing a Weighted Distance Function». Eighth OpenSky Symposium, *https://doi.org/10.3390/proceedings2020059007*

distintos hiperparámetros basándonos en las métricas de dos índices: el de Davies-Bouldin (DB) y el de Silueta. El Índice DB se centra en medir la cohesión dentro de cada clúster y la separación entre los clústeres. Se calcula como la media del cociente de la distancia intraclúster y la interclúster. Formalmente, se expresa como:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

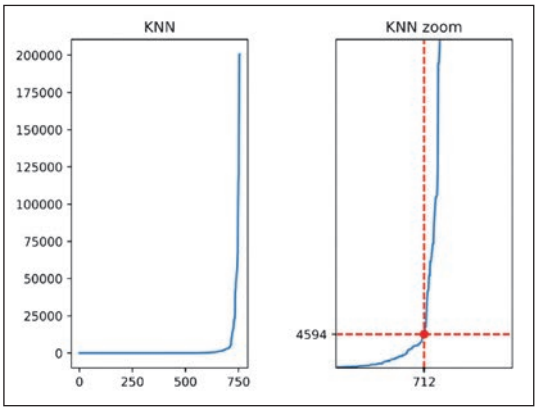


Figura 8. Ejemplo de cálculo analítico del parámetro ε. (Elaboración propia)

donde k es el número de clústeres, σ_i la dispersión intraclúster del clúster i, σ_j la dispersión intraclúster del clúster j, y $d(c_i, c_j)$ es la distancia entre los centroides de los clústeres i y j. Una ventaja del índice de Davies-Bouldin es su interpretación intuitiva: valores más bajos indican una mejor partición de los datos. Sin embargo, puede verse afectado por la elección de la métrica de distancia y su sensibilidad a la dimensionalidad de los datos.

Por otro lado, el Índice de Silueta mide la cohesión intraclúster y la separación interclúster, proporcionando una puntuación para cada punto de datos en función de cuán similar es a su propio clúster en comparación con otros clústeres. La fórmula resultante es:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

donde $a(i)$ es la distancia media del punto i a los demás puntos dentro de su clúster, y $b(i)$ es la distancia media más pequeña del punto i a los puntos en un clúster diferente. El coeficiente de silueta $s(i)$ varía entre -1 y 1, donde valores más altos indican una mejor calidad de agrupamiento. El índice de silueta es ventajoso debido a su capacidad para manejar diferentes formas y densidades de clúster y no requiere el conocimiento previo del número de clústeres. Sin embargo, puede ser computacionalmente costoso para grandes conjuntos de datos y su interpretación puede resultar más difícil que la del índice de Davies-Bouldin debido a la evaluación de cada punto individualmente.

Una vez con los hiperparámetros de cada modelo calculados en base a los resultados de estas dos métricas, pueden analizarse los resultados de vuelos anómalos en las trazas consideradas. Hay entonces cuatro resultados para cada

experimento, con ERP, con SSPD, y para cada uno de ellos resultados basándose en la combinación de hiperparámetros de la que se obtienen mejores valores de las métricas de los índices de Davies-Bouldin y de Silueta.

Resultados de la experimentación realizada y discusión

En la primera experimentación se mantienen todos los pasos del preprocesado, incluyendo la eliminación de los puntos de taxi. En ella se obtienen:

	ÍNDICE DAVIES-BOULDIN		**ÍNDICE SILUETA**	
	DBSCAN	**HDBSCAN**	**DBSCAN**	**HDBSCAN**
SSPD	1.11	0.55	0.67	0.91
ERP	0.99	0.37	0.72	0.93

Tabla 1. Métricas obtenidas para los modelos entrenados sobre la experimentación de 1.710 trazas de vuelo sin datos taxi. (Elaboración propia)

A primera vista, los resultados, teniendo en cuenta la forma de medir de cada métrica como se ha indicado previamente, parece mejorar notablemente con el modelo HDBSCAN. Además, el SSPD, pese a ser una metodología más sencilla de computar para calcular distancias entre trazas de vuelos, obtiene unos resultados similares en rendimiento a metodologías de cálculo computacionalmente más costosas como el ERP.

En cuanto a las trazas determinadas como anómalas para cada caso respecto al total de las 1.710 trazas de navegación tenidas en cuenta:

	ÍNDICE DAVIES-BOULDIN		**ÍNDICE SILUETA**	
	DBSCAN	**HDBSCAN**	**DBSCAN**	**HDBSCAN**
SSPD	190	57	116	18
ERP	335	47	142	47

Tabla 2. Trazas de navegación anómalas obtenidas en los modelos entrenados sobre la experimentación de 1.710 trazas de vuelo sin datos taxi. (Elaboración propia)

En el tratamiento de vuelos anómalos, el algoritmo HDBSCAN suele resultar en mayores ajustes de las rutas y menores tasas de anomalías. Se pueden ver las

trazas anómalas graficadas en rojo en las siguientes imágenes (figuras 9 y 10) para los modelos entrenados. En cada ruta más del 90 por 100 de trazas de media son consideradas normales y son indistinguibles generalmente bajo líneas de color azul superpuestas.

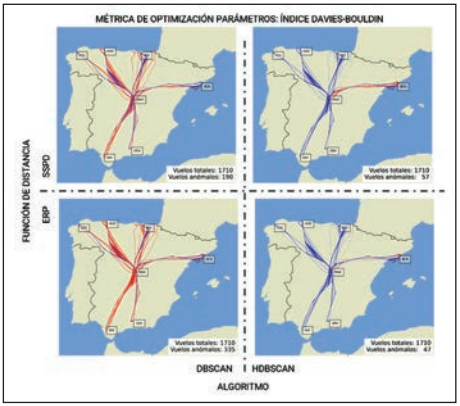


Figura 9. Comparativa visual de los agrupamientos de trazas normales (azules) y anómalas (rojas) para la métrica Índice de Davies-Bouldin, habiendo eliminado tiempos de taxi. (Elaboración propia)

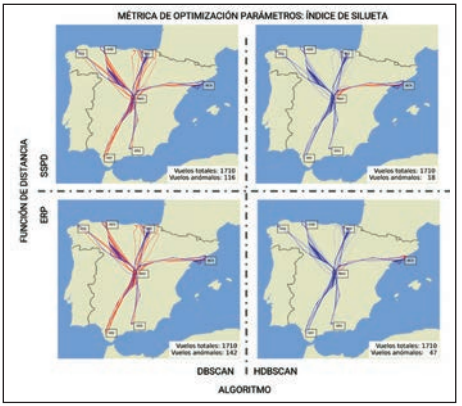


Figura 10. Comparativa visual de los agrupamientos de trazas normales (azules) y anómalas (rojas) para la métrica Índice de Silueta, habiendo eliminado tiempos de taxi. (Elaboración propia)

En la segunda experimentación no se eliminan los puntos de taxi de las trazas. En ella se obtienen las siguientes métricas:

	ÍNDICE DAVIES-BOULDIN		**ÍNDICE SILUETA**	
	DBSCAN	**HDBSCAN**	**DBSCAN**	**HDBSCAN**
SSPD	1.11	0.55	0.68	0.91
ERP	0.95	0.38	0.72	0.93

Tabla 3. Métricas obtenidas para los modelos entrenados sobre la experimentación de 1.710 trazas de vuelo sin datos taxi. (Elaboración propia)

Las métricas son muy similares a las obtenidas tras eliminar los puntos de taxi de las trazas de navegación. En cuanto a las trazas determinadas como anómalas para cada caso, respecto al total de 1.710 trazas de navegación tenidas en cuenta, siguen siendo muy similares a las resultantes tras eliminar puntos de taxi (con el costo computacional y el esfuerzo que podrían conllevar):

	ÍNDICE DAVIES-BOULDIN		ÍNDICE SILUETA	
	DBSCAN	**HDBSCAN**	**DBSCAN**	**HDBSCAN**
SSPD	192	54	106	15
ERP	357	53	146	13

Tabla 4. Trazas de navegación anómalas obtenidas en los modelos entrenados sobre experimentación de 1.710 trazas de vuelo sin datos de taxi. (Elaboración propia)

En el tratamiento de vuelos anómalos, el algoritmo HDBSCAN sigue resultando en mayores ajustes de las rutas y menores tasas de anomalías, aumentando las trazas de vuelos anómalos muy ligeramente (< 2 por 100 de media) en la mayoría de los casos. Se pueden ver las trazas anómalas graficadas en rojo en las siguientes imágenes (figuras 11 y 12), para los modelos entrenados incluyendo datos de taxi.

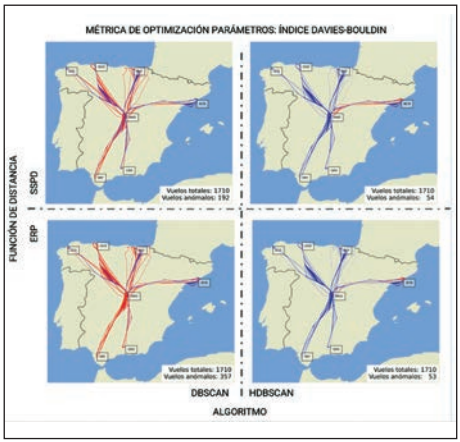


Figura 11. Comparativa visual de los agrupamientos de trazas normales (azules) y anómalas (rojas) para la métrica Índice de Davies-Bouldin sin eliminar tiempos de taxi. (Elaboración propia)

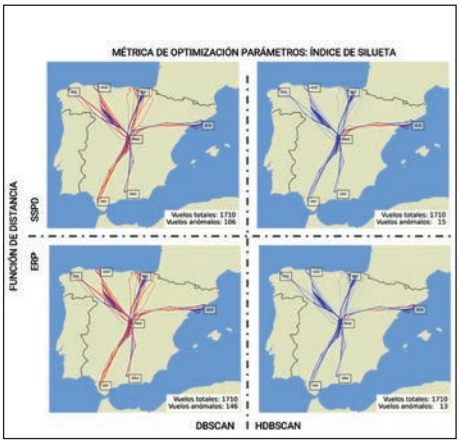


Figura 12. Comparativa visual de los agrupamientos de trazas normales (azules) y anómalas (rojas) para la métrica Índice de Silueta sin eliminar tiempos de taxi. (Elaboración propia)

En cuanto al rendimiento de los distintos modelos procesados y sus entrenamientos, resulta computacionalmente menos costoso (en torno a un 75 por 100 más eficiente) calcular distancias entre trazas de navegación siguiendo la metodología SSPD (figura 13). Sin embargo, la métrica de optimización

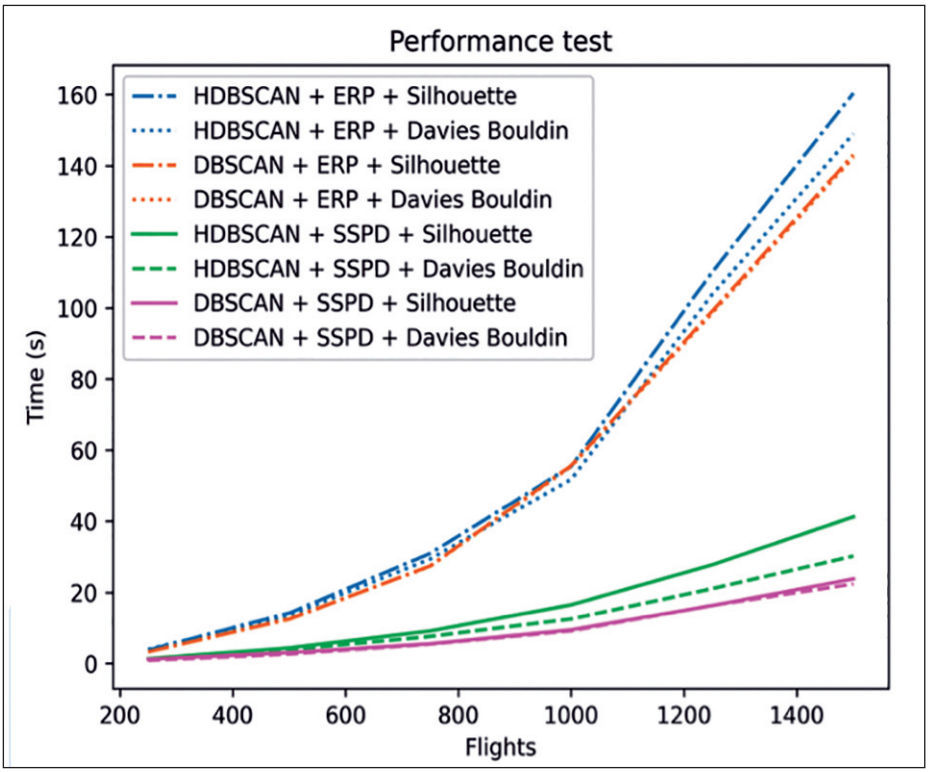


Figura 13. Rendimiento de los modelos entrenados en tiempo de computación *vs*. número de trazas de navegación consideradas. (Elaboración propia)

escogida, para optimización de los hiperparámetros de cada modelo, no tiene apenas impacto en el tiempo total de cómputo.

Conclusiones y perspectivas

A la vista de los resultados obtenidos en la experimentación se puede concluir que la metodología de cálculo de trazas de navegación anómalas presentada generaliza correctamente los datos introducidos, incluyendo múltiples rutas en el mismo conjunto de datos. Además, es directamente aplicable a datos de navegación marítima, como los registrados en el Centro de Operaciones y Vigilancia de Acción Marítima (COVAM) de la Armada. Es de utilidad para automatizar la detección de trazas que sigan rutas no agrupables a las típicas (por las razones que sean, por ejemplo meteorológicas) sin tener

que establecer reglas expertas para la detección (definición de canales de ruta o similar).

En este caso, el algoritmo HDBSCAN no se ve apenas influenciado por la matriz de distancias o, lo que es lo mismo, es capaz de generalizar mejor, indistintamente de las trayectorias del conjunto de datos de trazas de navegación. Sin embargo, para obtener un buen resultado con HDBSCAN es necesario estudiar cada uno de los escenarios para conseguir buenos resultados, y resulta más difícil generalizar los modelos. La generalización correcta de un modelo DBSCAN se debe en gran medida al cálculo analítico del valor ε, que permite obtener el punto a partir del cual el aumento de clústeres no hace más que favorecer un sobreajuste. En cuanto a las metodologías de cálculo de distancia, SSPD y ERP, ambas han tenido un comportamiento parecido respecto al ruido detectado, pero ERP ha demostrado un peor rendimiento computacional. Esto hace que, para usos como el agrupamiento de datos de trazas en tiempo real y a coste computacional bajo, la metodología de cálculo de distancias SSPD sea una mejor opción.

Para futuros trabajos se propone seguir trabajando en datos que contengan el mayor número de rutas posibles, detectando los agrupamientos en torno a cada ruta, y estudiar el desempeño de estos modelos con flujos de navegación marítima, como, por ejemplo, los datos registrados por el COVAM.