VOLUMEN 37
XXX Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

Preprocesamiento e identificación de patrones espacio-temporales en la base de datos Intel Lab Data

Preprocessing and Spatio-Temporal Pattern Discovery in the Intel Lab Dataset

E. Muñoz-Rivera[1], M. J. Peñaran-Prieto[1], M.A. Vázquez-Olguín[3]*
[1] Licenciatura en Ingeniería de Datos e Inteligencia Artificial, División de Ingenierías, Campus Irapuato Salamanca.
[2] Departamento de Electrónica, División de Ingenierías, Campus Irapuato Salamanca.
e.munozrivera@ugto.mx, mj.penaranprieto@ugto.mx, vazquez.ma@ugto.mx

*Autor de correspondencia

Resumen

En este trabajo se presenta el preprocesamiento y análisis de la base de datos Intel Lab Data, un registro público de observaciones de variables ambientales recolectadas por cincuenta y cuatro sensores distribuidos en el laboratorio de Intel en Berkeley, el cual tiene por objetivo su uso en aplicaciones de redes inalámbricas de sensores. Se realizó una limpieza profunda de los datos, incluyendo la eliminación de valores faltantes y atípicos, corrección de inconsistencias temporales y tratamiento del jitter mediante interpolación inteligente. Además, se aplicó un análisis de clustering con técnicas estadísticas y aprendizaje no supervisado, destacando la relevancia de los métodos del codo y Silhouette para determinar el número adecuado de grupos. Los resultados permitieron identificar claramente patrones espaciales y temporales en el comportamiento de los sensores, facilitando la interpretación de las condiciones ambientales dentro del laboratorio.

Palabras clave: limpieza de datos; series temporales, data clustering.

Introducción

En la actualidad, las redes inalámbricas de sensores (Wireless Sensor Networks, WSNs, por sus siglas en inglés) han adquirido una importancia considerable en la modernización y tecnificación de diversos campos de la ingeniería, tales como la agroindustria, la manufactura y los espacios de vivienda inteligentes, entre otros (López-Ramírez & Aragon-Zavala, 2023). Sin embargo, el uso de esta tecnología presenta desafíos particulares, como mediciones ruidosas, cambios inesperados en la topología de red y pérdida de datos (Kandris & Anastasiadis, 2024; Patel & Kumar, 2018; Akyildiz *et al*., 2002).

Entre los distintos problemas que afectan la implementación eficiente de las WSNs, la estimación precisa de cantidades físicas —como temperatura, humedad o luminosidad— se considera un requerimiento fundamental en el diseño de redes modernas (Asad *et al*. 2024; Chen *et al*., 2014; Toh, 2001). Esta necesidad se ve intensificada por el uso frecuente de hardware de bajo costo, el cual tiende a generar datos poco fiables, ya sea por la presencia de ruido o por pérdidas de información.

Para mitigar estas deficiencias, se ha investigado el desarrollo de diversos algoritmos de estimación distribuida (Chen *et al*. 2022; Modalavalasa *et al*. 2021). Estos algoritmos aprovechan la redundancia en la información de la variable física de interés para alcanzar un consenso en la medición (Tian, 2025; Carli *et al*., 2008; Rashvand & Calero, 2012).

En la Universidad de Guanajuato, el Cuerpo Académico de Procesamiento Digital de Señales ha realizado varias contribuciones al desarrollo de algoritmos de estimación distribuida (Vazquez-Olguin *et al*., 2025; Vazquez-Olguin, Shmaliy, & Ibarra-Manzano, 2018; Vazquez-Olguin, Shmaliy, & Ibarra-Manzano, 2019; Vazquez-Olguin *et al*., 2019). No obstante, para validar y verificar la confiabilidad de los algoritmos desarrollados, es necesario contar con bases de datos adecuadas.

Para este proyecto, se seleccionó la base de datos pública disponible en MIT CSAIL: Lab Data. Esta base de datos contiene aproximadamente 2.3 millones de registros recolectados de 54 sensores distribuidos (ver Figura 1) en el laboratorio de investigación de Intel, ubicado en Berkeley, California. Al realizar una inspección detallada de la base de datos, se identificaron diversas problemáticas que serán descritas a continuación.

- **Falta de encabezados**. La base de datos no cuenta con los encabezados necesarios para su correcto uso.
- **Nodos anómalos**. De acuerdo con la descripción de la base de datos, solo deberían existir 54 nodos, sin embargo, se encuentran una cantidad mayor.
- **Presencia de datos atípicos**. Las mediciones presentan valores anómalos, además de que existe una gran cantidad de datos faltantes.
- **Diferentes escalas de tiempo**. Los nodos presentan variaciones entre el tiempo de muestreo debido a que la falta de sincronía entre ellos.
- **Presencia de jitter en las mediciones**. Los relojes internos de los nodos son de bajo costo, lo que provoca que el periodo de muestreo en cada nodo no sea constante.

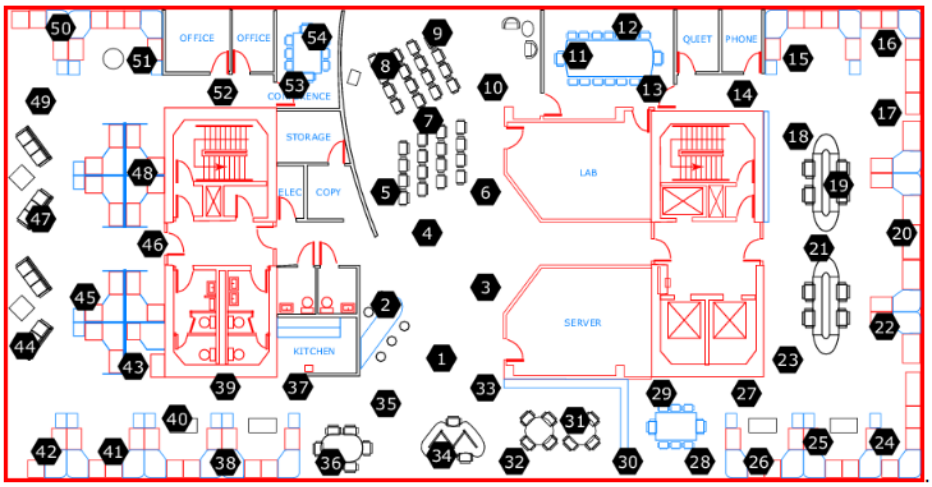

Figura 1. Esquema que indica la posición de los sensores colocados en el laboratorio de Intel en Berkeley. Fuente: Adaptado de Intel Lab Data, MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), 2004. https://db.csail.mit.edu/labdata/labdata.html

Para que la base de datos pueda ser utilizada en el desarrollo y prueba de nuevos algoritmos de consenso, es necesario que las problemáticas previamente mencionadas sean solucionadas, además de que es indispensable identificar agrupamientos de nodos que efectivamente realicen mediciones redundantes de las variables físicas de interés, por ejemplo, no podemos considerar en el mismo grupo a aquellos nodos que midan temperatura cerca de la cocina a aquellos que realicen mediciones cerca de los servidores. Por este motivo, la principal motivación para este proyecto de investigación es la limpieza de la base de datos, así como la identificación de grupos de nodos para garantizar mediciones redundantes de las variables físicas de interés.

Metodología

Análisis y Limpieza de Datos

El dataset del Intel Berkeley Research Lab contiene lecturas de 54 sensores (motes) recolectadas entre el 28 de febrero y el 5 de abril de 2004 con variables como temperatura, humedad, luminosidad y voltaje. En la Tabla 1 se muestran distintas medidas estadísticas obtenidas de los datos. Se puede apreciar como el número de registros es diferente para cada variable física, lo que indica la presencia de datos faltantes. Con respecto a los valores máximos y mínimos se observan valores que sugieren la presencia de datos atípicos (Outliers).

Tabla 1. *Estadísticas descriptivas de las variables medidas por los sensores. Fuente: Elaboración propia con base en Intel Lab Data (MIT CSAIL, 2004).*

	CONTEO	MEDIA	STD	MIN	25%	50%	75%	MAX
TEMPERATURA	2313156	39.2	37.42	-38.4	20.4	22.44	27.02	385.57
HUMEDAD	2312783	33.91	17.32	-8983.13	31.9	39.28	43.59	137.51
LUMINOSIDAD	2312779	390.94	534.4	0	18.4	143.5	507.8	1847.4
VOLTAJE	2219803	2.49	0.18	0.01	2.39	2.53	2.63	3.16

Dado que los datos crudos presentaban problemas de calidad, se implementó un proceso estructurado de limpieza:

1. **Identificación y eliminación de valores faltantes**. Se detectaron registros incompletos, especialmente en las variables *humidity* (899 NaN), *light* (903 NaN) y *voltage* (93,879 NaN). Todos los registros con valores faltantes fueron eliminados, reduciendo el dataset de aproximadamente 2.3 millones a 2.2 millones de registros válidos.
2. **Filtrado de outliers basado en dominio**. Se descartaron mediciones de humedad fuera del rango 0-100%, voltajes menores a 2V o mayores a 3V, y lecturas de luminosidad negativas o superiores a 100,000 Lux (equivalente a la luz directa del sol). La temperatura se ajustó para excluir valores extremos inconsistentes con las condiciones ambientales de un laboratorio.
3. **Detección y manejo de sensores anómalos**. Un análisis de los identificadores de sensores reveló la presencia de tres motes no documentados (55, 56 y 58). Adicionalmente, se confirmó la ausencia esperada de los sensores 5 y 28, documentada en la metadata original. Los sensores anómalos contienen en promedio 3,000 mediciones, las cuales son a lo largo del mes de marzo e inicios de abril. Esto representa pocos datos durante ese lapso.

Corrección de Inconsistencias Temporales y de Epoch

El campo epoch, diseñado como identificador de sincronización, presentaba anomalías que requerían atención especial. Se detectaron casos donde múltiples registros compartían el mismo epoch pero mostraban marcas de tiempo diferentes, violando el principio fundamental de sincronización. En la tabla 2 se muestran ejemplos concretos de estos registros conflictivos. Se puede observar como para el mismo nodo y en el mismo valor de epoch se presentan mediciones diferentes para las mismas variables físicas.

Para resolver estas inconsistencias, se implementó un algoritmo que evalúa la coherencia física de las mediciones conflictivas. Cada registro problemático se comparó con sus vecinos temporales inmediatos, calculando la discrepancia relativa en las variables medidas. La medición que presentó menor divergencia con respecto al comportamiento esperado, basado en los registros adyacentes, fue seleccionada como válida, mientras que las demás se descartaron. Este enfoque preservó la continuidad temporal y física de las señales.

La tabla 3. Ejemplos de registros corregidos tras el tratamiento de inconsistencias en epoch muestra cómo los registros se corrigen tras la aplicación del tratamiento.

Por último, para facilitar el análisis temporal posterior, las columnas de fecha y hora se fusionaron en un único campo timestamp estandarizado.

Tabla 2. *Ejemplo de registros con inconsistencias temporales detectadas en el campo epoch.*

epoch	moteid	temperature	humidity	light	voltage	timestamp
4764	1	19.9394	41.48	82.8	2.68	29/02/2004 16:39
4764	1	121.918	4.774	9.2	2.01	31/03/2004 04:38
4806	1	19.8414	41.58	79.12	2.69	29/02/2004 17:00
4806	1	121.997	14.95	8.28	2.01	31/03/2004 04:59
29066	1	21.8112	36.47	1.38	2.59	09/03/2004 03:11

Fuente: Elaboración propia con base en Intel Lab Data (MIT CSAIL, 2004).

Tabla 3. *Ejemplos de registros corregidos tras el tratamiento de inconsistencias en epoch.*

epoch	moteid	temperature	humidity	light	voltage	timestamp
4764	1	19.9394	41.48	82.8	2.68	29/02/2004 16:39
4765	1	19.9492	41.44	79.12	2.68	29/02/2004 16:40
4770	1	19.91	41.54	79.12	2.69	29/02/2004 16:40
4774	1	19.9198	41.51	79.12	2.69	29/02/2004 16:44
4776	1	19.91	41.51	79.12	2.69	29/02/2004 16:45

Fuente: Elaboración propia con base en Intel Lab Data (MIT CSAIL, 2004).

Tratamiento del Jitter en los Datos

El análisis de los intervalos entre mediciones reveló una significativa variabilidad temporal en la frecuencia de muestreo. Aunque la documentación original especifica un intervalo teórico de 31 segundos, la distribución real mostró que la mayoría de las mediciones (aproximadamente 18.1% del total) se concentraron entre 29.5 y 31.5 segundos, con una marcada dispersión hacia intervalos tanto menores como mayores. Este comportamiento irregular, conocido como jitter, puede introducir artefactos en análisis temporales posteriores. En la Figura 2 se presenta un histograma que ilustra esta distribución de intervalos entre mediciones consecutivas. Para mitigar el jitter, se implementaron dos estrategias complementarias de re-muestreo, descritas a continuación.

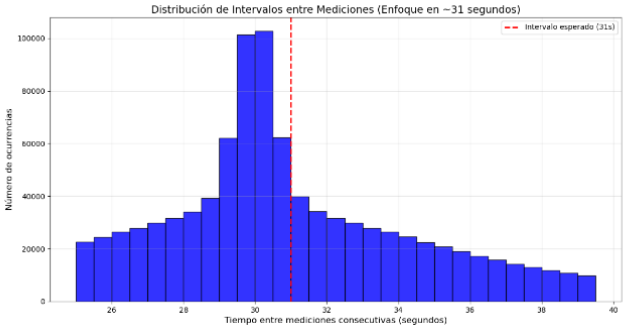

Figura 2. *Histograma de la distribución de intervalos entre mediciones consecutivas.*
Fuente: Elaboración propia con datos de Intel Lab Data (MIT CSAIL, 2004).

Re-muestreo a Intervalos Regulares sin Interpolación

El primer enfoque consistió en redefinir la estructura temporal del dataset mediante un re-muestreo (downsampling) a intervalos fijos de 30 segundos, sin aplicar técnicas de interpolación. Las acciones que implican este método son:

- Agrupación de las mediciones de cada sensor en ventanas temporales de 30 segundos.
- Cálculo del valor promedio de cada variable física dentro de cada ventana.
- Generar valores perdidos (missing values) en los casos donde no existían mediciones originales para una ventana determinada

Figura 3. Mapa de disponibilidad temporal por sensor tras el re-muestreo a 30 segundos sin interpolación.
Fuente: Elaboración propia con datos de Intel Lab Data (MIT CSAIL, 2004).

La Figura 3 presenta el mapa de disponibilidad temporal por sensor tras aplicar este método, donde se observan segmentos sin datos debido a la ausencia de mediciones en ciertas ventanas; este enfoque preservó la integridad de los datos originales al no introducir valores estimados, aunque como consecuencia produjo un dataset con discontinuidades temporales.

Re-muestreo con Interpolación Inteligente

Para abordar el problema de los huecos temporales, se desarrolló un segundo enfoque que combinó el re-muestreo a 30 segundos con una interpolación lineal condicional. Para este método se realizaron las siguientes acciones:

- Se aplicó interpolación lineal únicamente en huecos temporales menores o iguales a 3 minutos (6 intervalos consecutivos), umbral determinado para tratar de balancear completitud de datos y confiabilidad.
- Los huecos mayores a 3 minutos se mantuvieron como valores faltantes, evitando extrapolaciones potencialmente erróneas.
- La interpolación se realizó de forma independiente para cada variable física (temperatura, humedad, etc.), considerando sus características específicas.

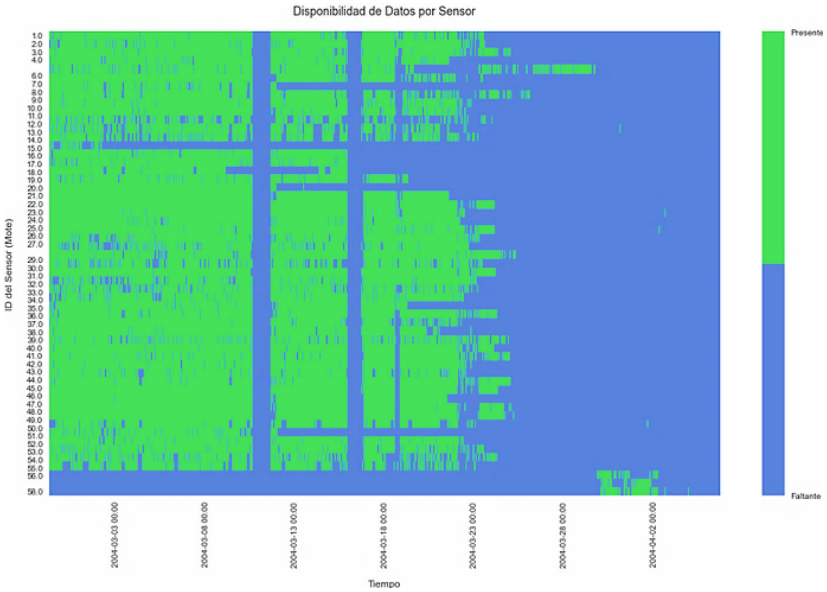

Figura 4. Mapa de disponibilidad temporal por sensor tras interpolación condicional con umbral de 3 minutos
Fuente: Elaboración propia con datos de Intel Lab Data *(MIT CSAIL, 2004).*

El resultado puede observarse en la Figura 4, donde se muestra el mapa de disponibilidad temporal por sensor después de la interpolación. El incremento en las zonas verdes refleja la recuperación efectiva de información faltante en segmentos confiables. Este método produjo una serie temporal más continua. La restricción en la duración de los huecos interpolados garantizó que solo se recuperara información con una alta probabilidad de coherencia física.

La Figura 5 muestra una comparación puntual entre los datos re-muestreados sin interpolación (líneas rojas discontinuas) y los datos re-muestreados con interpolación inteligente (líneas azules continuas), destacando cómo se preserva la estructura original sin introducir valores artificiales en intervalos mayores a tres minutos.

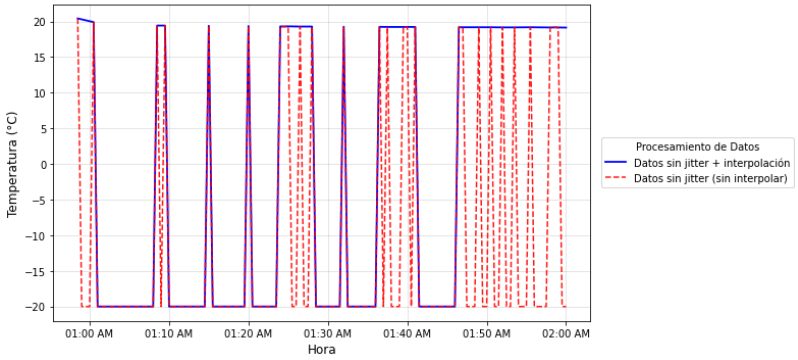

Figura 5. Efecto de la interpolación inteligente para el Mote 6 durante el intervalo de las 12:58 A.M. a 2:00 A.M para el día 2004-02-28.
Fuente: Elaboración propia con datos de Intel Lab Data *(MIT CSAIL, 2004).*

Análisis de Clustering Espacio-Temporal

El proceso de clustering se diseñó para identificar patrones en el comportamiento de los sensores, considerando tanto su ubicación física como las variaciones temporales de las mediciones. La metodología combinó técnicas de aprendizaje no supervisado con análisis espacial para revelar agrupaciones naturales en los datos.

Se utilizó el dataset preprocesado con tiempo de muestreo a 30 segundos e interpolación inteligente. Cada sensor se caracterizó mediante un vector de features que incluyó la media, mediana, desviación estándar y asimetría, además de usarse las coordenadas espaciales (x,y) obtenidas del archivo de posiciones.

Se implementó un pipeline de análisis que ejecutó las siguientes etapas para cada hora y fecha seleccionada:

1. Extracción de Features Temporales. Para cada sensor en la hora analizada, se calcularon métricas estadísticas que capturan el comportamiento dinámico de las variables medidas. Estas features se estandarizaron usando StandardScaler para asegurar comparabilidad entre variables con distintas unidades.
2. Elección de Numero de Clusters. Se utilizo el método del codo y el coeficiente de Solhouette para encontrar el valor k con el que se probara k-Means en la hora y fecha seleccionada.
3. Aplicación de K-Means. El algoritmo de K-means se ejecutó sobre los datos normalizados, utilizando como número de clusters el valor antes encontrado.
4. Visualización Espacial. Los resultados del clustering se representaron gráficamente mediante diagramas de dispersión que muestran la posición física de cada sensor junto con un color asignado por cluster y un mapa de densidad para visualizar la distribución espacial de los clusters.

Resultados

El día 4 de marzo de 2004 fue seleccionado para el análisis detallado debido a su alta completitud de datos, como se evidencia en la Figura 4, que muestra una ausencia mínima de interrupciones en las mediciones. Este criterio aseguró que los patrones identificados no estuvieran distorsionados por valores faltantes.

Para determinar el número óptimo de clusters se aplicaron los dos métodos antes mencionados, sus resultados se muestran en la Figura 6 y Figura 7.

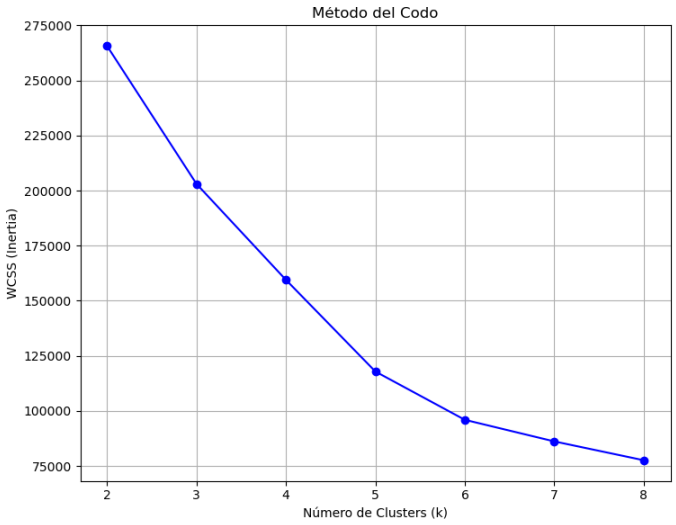

Figura 6. Determinación del Número Óptimo de Clusters mediante el Método del Codo.
Fuente: Elaboración propia con datos de Intel Lab Data *(MIT CSAIL, 2004).*

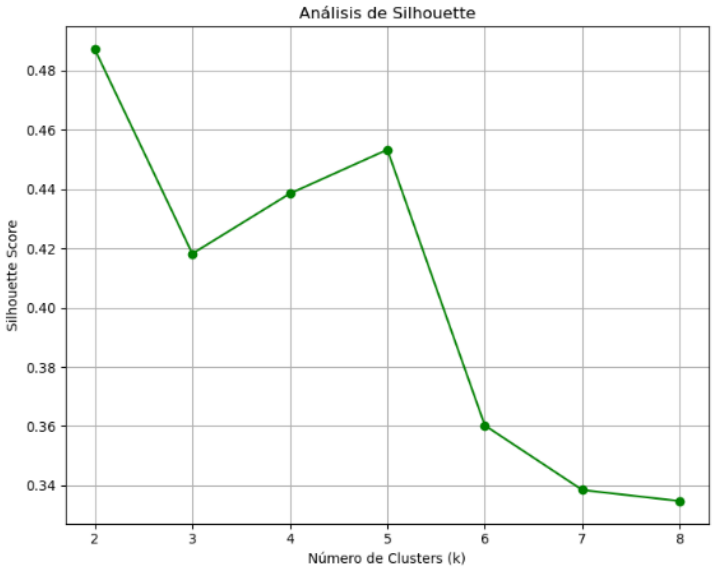

Figura 7. Determinación del Número Óptimo de Clusters mediante Análisis de Silhouette.
Fuente: Elaboración propia con datos de Intel Lab Data *(MIT CSAIL, 2004).*

Se logra observar en el método del codo mostrado en la **Figura 6** que k=5 es un punto de inflexión, por lo que es posible que se pueda utilizar en el análisis. Sin embargo, en el análisis de Silhouette presentado en la Figura 7, k=2 muestra el valor más alto, aunque k=5 es el segundo mejor. Por esto mismo se opta por probar con ambas opciones y visualizar las distribuciones.

Al utilizar dos clusters, la Figura 8 permite observar claramente lo siguiente:

- Cluster 0: La mayor parte de los sensores están en este cluster, abarcando la parte sur del laboratorio, la gran parte este y todo el centro.
- Cluster 1: Predomina en la zona oeste del laboratorio y en la zona noreste. Cabe resaltar que en la zona suroeste del laboratorio hay un nodo que pertenece al cluster 1 y puede parecer un valor atípico.

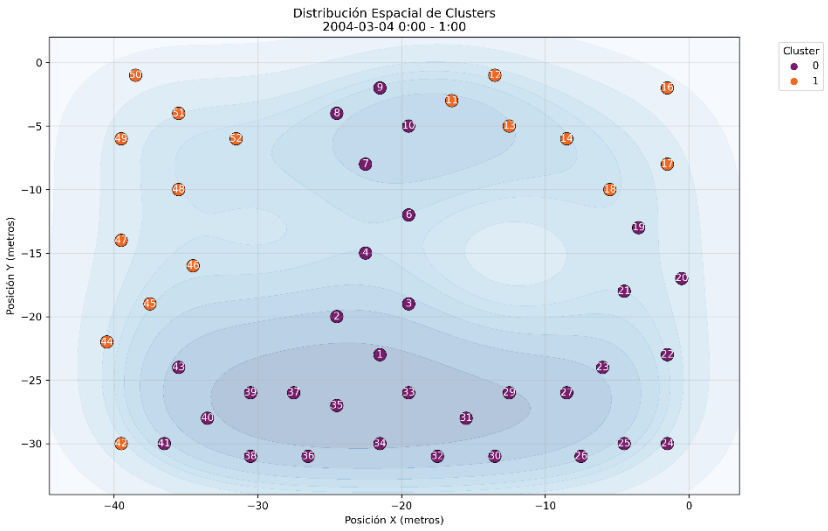

Figura 8. Distribución Espacial del Laboratorio con Dos Clusters (k=2).
Fuente: Elaboración propia con datos de Intel Lab Data *(MIT CSAIL, 2004).*

Cuando se analizan cinco clusters en la Figura 9, se observan las siguientes distribuciones:

- Cluster 0: Los sensores pertenecientes a este cluster están dispersos por el laboratorio, aun así, se observa que la mayoría están en la zona sureste.
- Cluster 1: Los sensores que se encuentran en el centro del laboratorio pertenecen a este cluster.
- Cluster 2: La mayor parte se encuentra en la zona noroeste del laboratorio.
- Cluster 3: Solo 3 sensores pertenecen a este cluster, dos de ellos se encuentran en la zona suroeste y el otro en contraparte.
- Cluster 4: Solo pertenece un solo nodo.

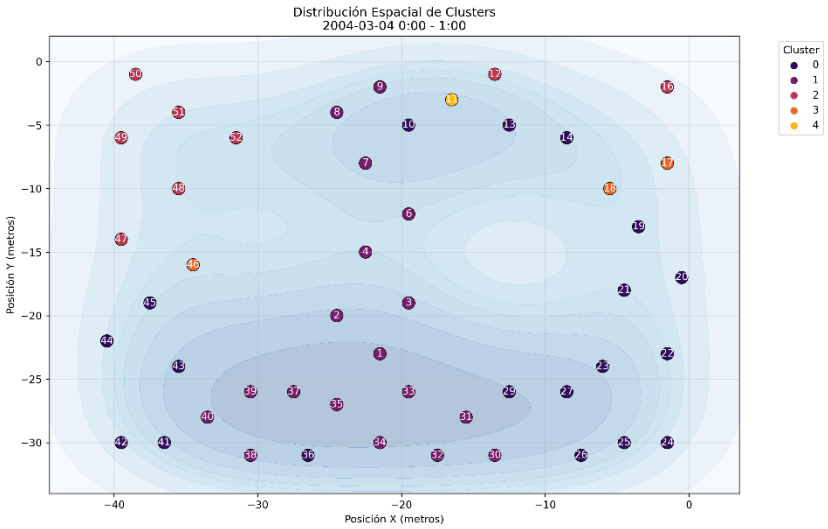

Figura 9. Distribución Espacial del Laboratorio con Cinco Clusters (k=5).
Fuente: Elaboración propia con datos de Intel Lab Data *(MIT CSAIL, 2004).*

En el caso de dos clusters, a lo largo del día se observan determinados patrones, por ejemplo, los sensores en la zona central regularmente pertenecen al mismo cluster. Este comportamiento se observa tanto en la parte sur y norte del laboratorio. En cinco clusters ocurre lo mismo, pero con grupos más pequeños tal y como se muestra en la Figura 10.

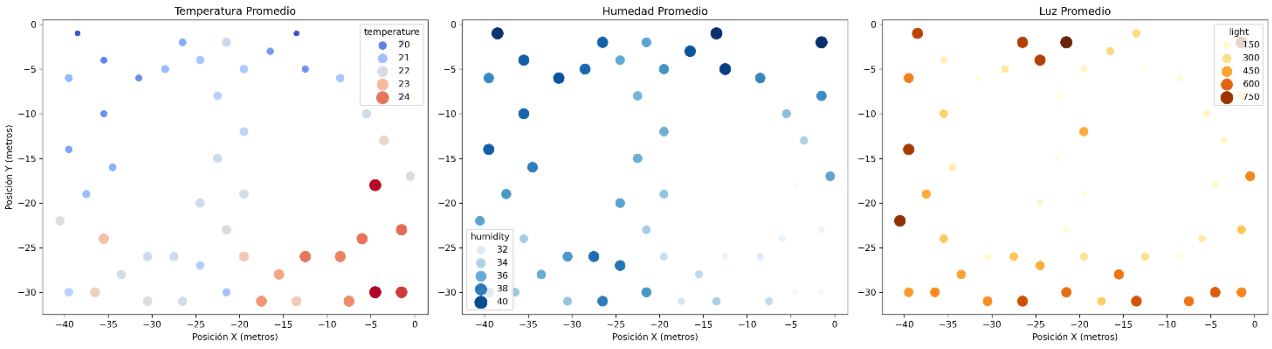

Figura 10. Promedios Diarios de Temperatura, Humedad y Luz.
Fuente: Elaboración propia con datos de Intel Lab Data *(MIT CSAIL, 2004).*

La Figura 10 muestra las gráficas en las que se realizó el promedio de los valores registrados por el sensor a lo largo del día. Teniendo como resultado que:

- La parte sureste es la más caliente del laboratorio y la más fría es la noroeste.
- La zona norte, por lo general, es la más húmeda.
- Y los costados del laboratorio son los más iluminados.

Conclusiones

El proceso integral de limpieza, acondicionamiento y análisis del conjunto de datos Intel Lab Data permitió solucionar de manera correcta múltiples problemas iniciales como la presencia de valores faltantes, datos atípicos, inconsistencias en las escalas de tiempo y jitter en las mediciones. El uso de técnicas como la interpolación inteligente facilitó la recuperación efectiva de información temporal, mejorando notablemente la continuidad del dataset. Asimismo, la aplicación de métodos estadísticos y clustering permitió identificar claramente agrupamientos espaciales y temporales significativos de los sensores. Estos resultados proporcionan una base sólida para futuras investigaciones y desarrollo de algoritmos de consenso en redes inalámbricas de sensores.

Referencias

Asad, M., Rehan, M., Ahn, C. K., Tufail, M., & Basit, A. (2024). Distributed $ H_\infty $ State and Parameter Estimation Over Wireless Sensor Networks Under Energy Constraints. *IEEE Transactions on Network Science and Engineering,* 11(3), 2976-2988.

Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., & Cayirci, E. (2002). Wireless sensor networks: A survey. *Computer Networks*, 38(4), 393–422. https://doi.org/10.1016/S1389-1286(01)00302-4

Carli, R., Chiuso, A., Schenato, L., & Zampieri, S. (2008). Distributed Kalman filtering based on consensus strategies. *IEEE Journal on Selected Areas in Communications*, 26(4), 622–633. https://doi.org/10.1109/JSAC.2008.080503

Chen, C., Zhu, S., Guan, X., & Shen, X. S. (2014). Wireless sensor networks: Distributed consensus estimation. Springer. https://doi.org/10.1007/978-3-319-02780-0

Chen, W., Wang, Z., Ding, D., Yi, X., & Han, Q. L. (2022). Distributed state estimation over wireless sensor networks with energy harvesting sensors. *IEEE transactions on cybernetics,* 53(5), 3311-3324.

Kandris, D., & Anastasiadis, E. (2024). Advanced wireless sensor networks: Applications, challenges and research trends. *Electronics,* 13(12), 2268.

López-Ramírez, G. A., & Aragon-Zavala, A. (2023). Wireless sensor networks for water quality monitoring: A comprehensive review. *IEEE Access,* 11, 95120–95142. https://doi.org/10.1109/ACCESS.2023.3303163

MIT Computer Science and Artificial Intelligence Laboratory. (n.d.). Intel Lab Data. https://db.csail.mit.edu/labdata/labdata.html

Modalavalasa, S., Sahoo, U. K., Sahoo, A. K., & Baraha, S. (2021). A review of robust distributed estimation strategies over wireless sensor networks. *Signal Processing,* 188, 108150.

Patel, N. R., & Kumar, S. (2018, November). Wireless sensor networks' challenges and future prospects. In 2018 International Conference on System Modeling & Advancement in Research Trends (SMART) (pp. 60–65). IEEE. https://doi.org/10.1109/SYSMART.2018.8746922

Rashvand, H. F., & Calero, J. M. A. (2012). Distributed sensor systems: Practice and applications. Wiley.

Tian, H. (2025). Enhancing the effectiveness of wireless sensor networks through consensus estimation and universal coverage. *Scientific Reports,* 15(1), 24930.

Toh, C. K. (2001). Ad hoc mobile wireless networks: Protocols and systems. Pearson Education.

Vazquez-Olguin, M., Shmaliy, Y. S., & Ibarra-Manzano, O. (2018). Developing UFIR filtering with consensus on estimates for distributed wireless sensor networks. *WSEAS Transactions on Circuits and Systems*, 17, 30–37.

Vazquez-Olguin, M., Shmaliy, Y. S., & Ibarra-Manzano, O. G. (2019). Distributed UFIR filtering over WSNs with consensus on estimates. *IEEE Transactions on Industrial Informatics*, 16(3), 1645–1654. https://doi.org/10.1109/TII.2019.2934164

Vazquez-Olguin, M., Shmaliy, Y. S., Ibarra-Manzano, O., Munoz-Minjares, J., & Lastre-Dominguez, C. (2019). Object tracking over distributed WSNs with consensus on estimates and missing data. *IEEE Access*, 7, 39448–39458. https://doi.org/10.1109/ACCESS.2019.2906596

Vazquez-Olguin, M., Shmaliy, Y. S., Ibarra-Manzano, O. G., & Munoz-Minjares, J. (2025). Enhancing Estimates Over WSN Using Robust Distributed Unbiased FIR Filtering With Consensus on Noise Power Gain. *IEEE Sensors Journal.*