UACON21
Un espacio de Divulgación Científica

Análisis de los Perfiles de Consumo Eléctrico Paraguayo Utilizando Machine Learning

Ing. Félix Morales

CONSEJO NACIONAL DE CIENCIA Y TECNOLOGÍA
CONACYT
PROCIENCIA
PROGRAMA PARAGUAYO PARA EL DESARROLLO DE LA CIENCIA Y TECNOLOGÍA
UNIVERSIDAD AMERICANA
Con el apoyo de:
Feei
Fondo para la Excelencia de la Educación y la Investigación

UACON21
Un espacio de Divulgación Científica

Resumen

- Introducción
- Materiales y Métodos
  - Datos
  - Técnicas Utilizadas
  - Medidas de Validación
- Resultados Obtenidos
- Discusión y Conclusión

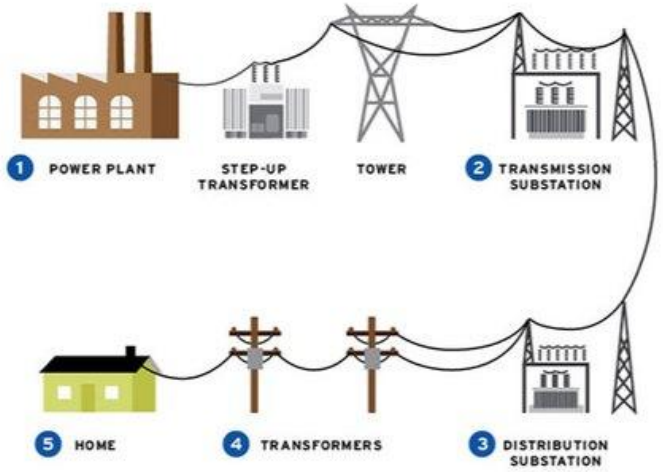
Resumen

- Introducción
- Materiales y Métodos
  - Datos
  - Técnicas Utilizadas
  - Medidas de Validación
- Resultados Obtenidos
- Discusión y Conclusión

Introducción

La identificación de un conjunto de alimentadores representativos es de gran interés para los planificadores y operadores de redes de distribución de energía eléctrica.

El conjunto seleccionado de alimentadores podría utilizarse en varias tareas, como la realización de un número limitado de simulaciones para evaluar el impacto de las nuevas tecnologías de red, el estudio del impacto de una nueva tarifa o la reconfiguración de la red.



Introducción

Las aplicaciones importantes de la agrupación del consumo eléctrico incluyen:

- La caracterización de las curvas de carga en un sistema de distribución real.
- La elaboración de perfiles de carga para el diseño de tarifas.
- La previsión de la carga o la planificación de la distribución.
- La determinación de la ubicación óptima de las fuentes de generación distribuida en los sistemas de distribución eléctrica

Introducción

En el presente trabajo se aborda este problema aplicando dos estrategias de Clustering en un conjunto de datos que contiene datos de consumo eléctrico generados en Paraguay, y proporcionados por la compañía eléctrica paraguaya.

Que es Clustering?

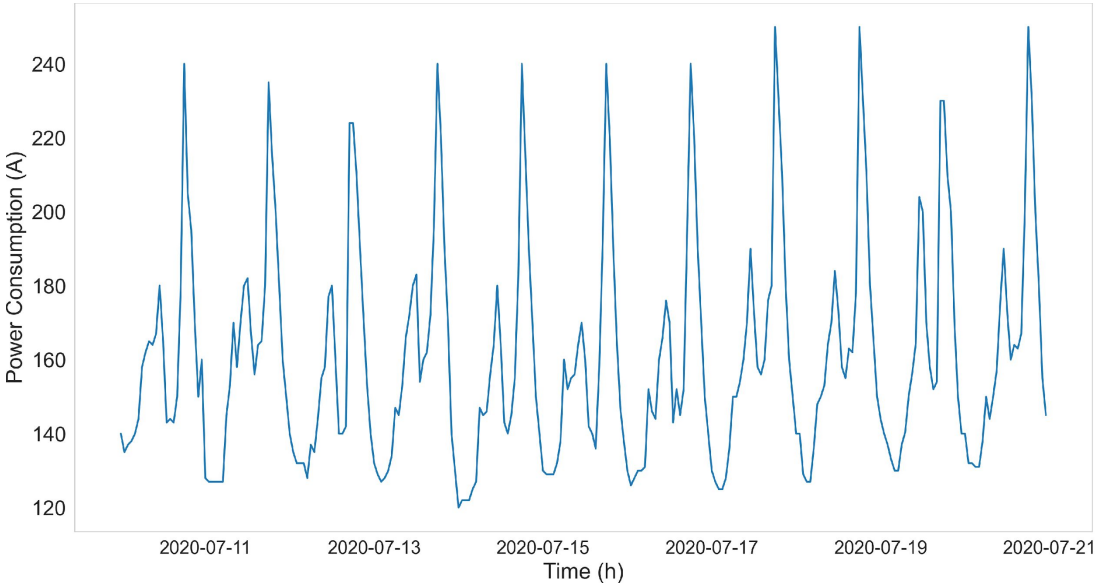
Es una tarea que tiene como finalidad principal lograr el agrupamiento de conjuntos de objetos no etiquetados, para lograr construir subconjuntos de datos conocidos como Clusters.

Resumen

- Introducción
- Materiales y Métodos
  - Datos
  - Técnicas Utilizadas
  - Medidas de Validación
- Resultados Obtenidos
- Discusión y Conclusión

Datos

El consumo de energía eléctrica es usualmente representado como una serie temporal a través de una secuencia discreta de valores medidos con un mismo intervalo de tiempo.



Matemáticamente:

$$X = \{x_t\}_{t=1}^{T}$$

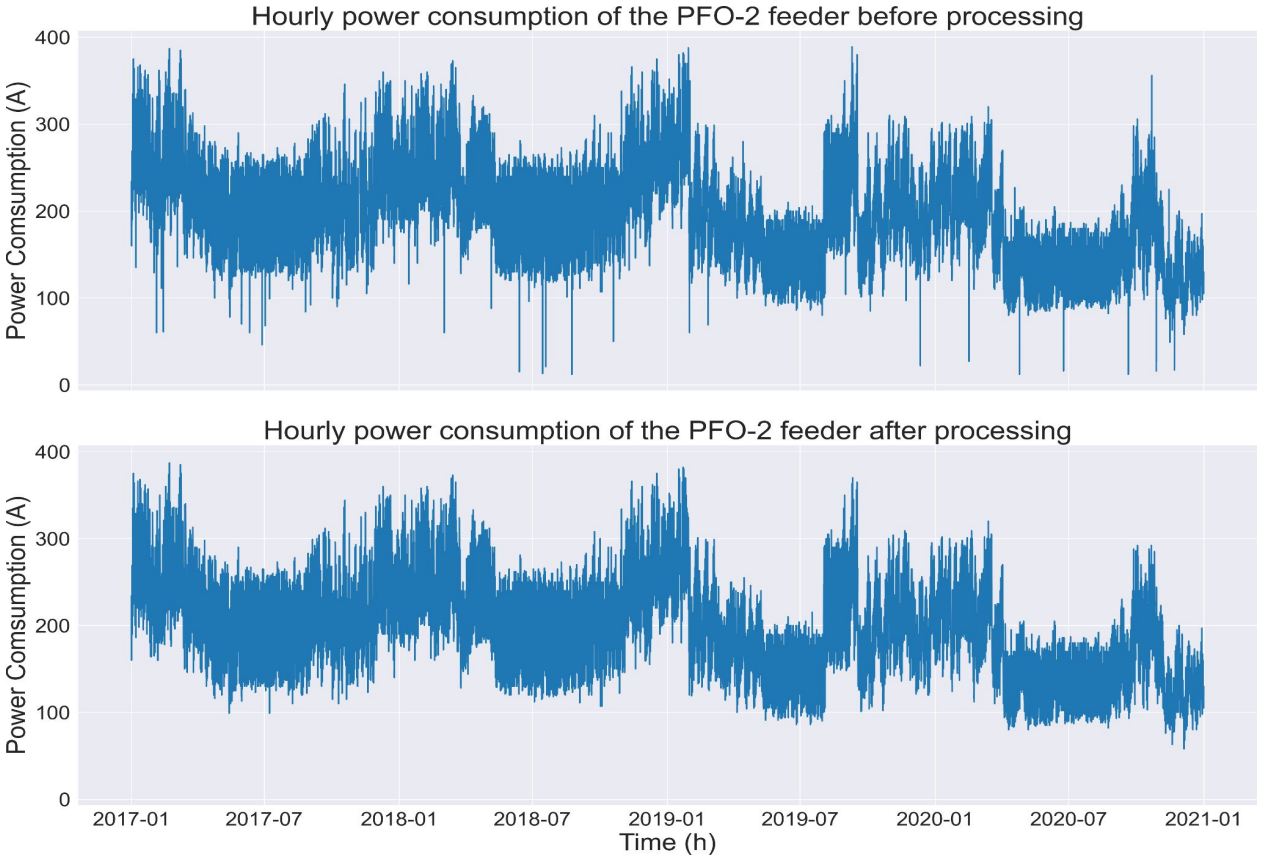
Datos

El conjunto de datos utilizado en este trabajo contiene 2.967.224 registros de consumo eléctrico medidos en Ampere desde enero de 2017 hasta diciembre de 2020 (4 años), de 115 alimentadores distribuidos en 17 subestaciones de la región Este, Paraguay.

Sin embargo, debido a la existencia de valores faltantes y atípicos, se ha procedido a aplicar una técnica para la imputación de estos datos. Para ello, se hizo uso del método propuesto por Vallis et. al.[1]

[1]Owen Vallis, Jordan Hochenbaum, and Arun Kejariwal. “A novel technique for long-term anomaly detection in the cloud”. In:6th {USENIX} workshop on hot topics in cloud computing (HotCloud 14). 2014.

Datos


Hourly power consumption of the PFO-2 feeder before processing
Power Comsumption (A)
400
300
200
100
0
Hourly power consumption of the PFO-2 feeder after processing
Power Comsumption (A)
400
300
200
100
0
2017-01
2017-07
2018-01
2018-07
2019-01
2019-07
2020-01
2020-07
2021-01
Time (h)

Datos

Una vez imputadas las series temporales correspondientes al consumo de energía eléctrica, se procedió a normalizarlas en base a cada alimentador según la ecuación:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Seguidamente, se han considerado los siguientes conjuntos de datos como series temporales:

- Serie temporal del consumo eléctrico semanal
- Serie temporal del consumo eléctrico mensual

Datos

Otro enfoque para representar el consumo de energía eléctrica es calcular un conjunto de características que representen cada secuencia de consumo eléctrico en lugar de considerarlo como una serie temporal.

Ventajas:

- Reducción de la dimensionalidad.
- Más robusta a los valores faltantes.
- Puede manejar distintas longitudes de series temporales.

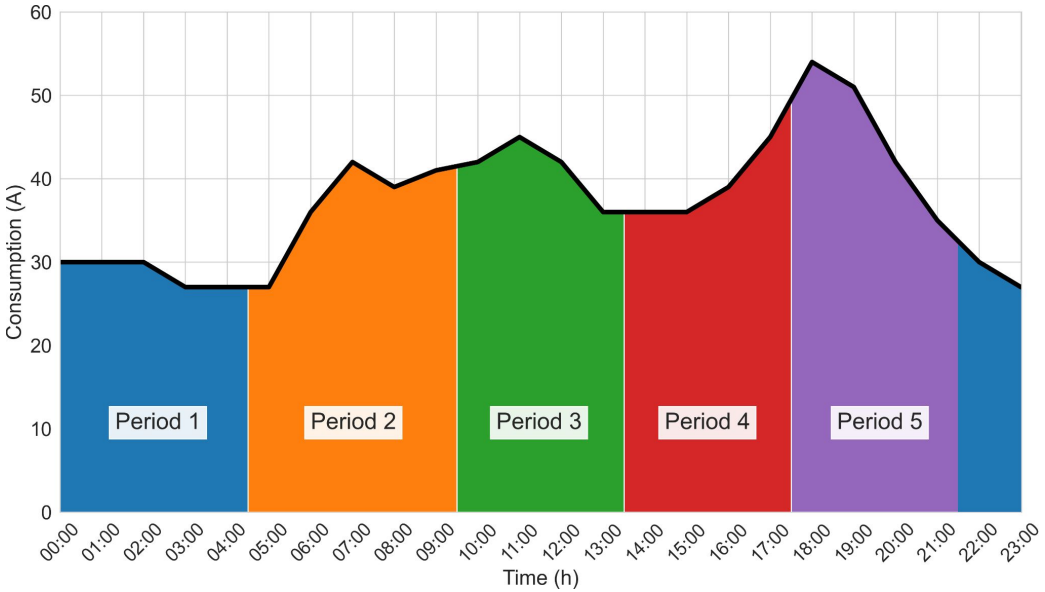
Datos

El primer conjunto de datos basado en características se obtuvo en base a características estadísticas por cada semana.
Se tuvieron en cuenta 7 variables estadísticas en un total de 207 semanas, resultando en 1449 características por cada alimentador.

- Media $\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{i,j}$
- Desviación estándar $\sigma_i = \sqrt{\frac{1}{N_i} \sum_{j=1}^{N_i} (x_{i,j} - \mu_i)^2}$
- Asimetría estadística $\mathcal{S}_i = \frac{1}{N_i(\sigma_i)^3} \sum_{j=1}^{N_i} (x_{i,j} - \mu_i)^3$
- Curtosis $\mathcal{K}_i = \frac{1}{N_i(\sigma_i)^4} \sum_{j=1}^{N_i} (x_{i,j} - \mu_i)^4$
- Exponente máximo de Lyapunov $\lambda_i = \frac{1}{N_i \times \Delta t} \ln \frac{|\delta \mathbf{Z}_t|}{|\delta \mathbf{Z}_0|}$
- Energía $\mathcal{E}_i = \frac{\sum_{\mathbb{k}=1}^{N_i} |\mathcal{X}[\mathbb{k}]|^2}{N_i}$
- Periodicidad $\mathcal{P}_i = \underset{\mathcal{T}}{\operatorname{argmax}}\, \mathcal{P}_{xx,i}(\omega)$

Datos

El segundo conjunto de datos basado en características se obtuvo en base a características estacionales y de consumo.



Time period	Interval
1	10:00 pm - 04:00 am
2	05:00 am - 09:00 am
3	10:00 am - 01:00 pm
4	02:00 pm - 05:00 pm
5	06:00 pm - 09:00 pm

Datos

Se han seleccionado las siguientes variables:

- P_i : Consumo eléctrico medio en el periodo i.
- σ_i : Desviación estándar del consumo eléctrico en el periodo i.
- $\hat{P}$: Consumo eléctrico medio en toda la serie temporal.
- P_i^S : Consumo medio de Verano en el periodo i.
- P_i^W : Consumo medio de Invierno en el periodo i.
- P_i^{WD}: Consumo medio de los días entre semana en el periodo i.
- P_i^{WE}: Consumo medio de los fines de semana en el periodo i.

Datos

Dado lo anterior las características obtenidas fueron 8, las cuales son:

- Potencia media relativa $P_i^R = \frac{P_i}{\hat{P}} \quad \text{for} \quad i = 1, \ldots, 5$
- Desviación estándar media relativa $\hat{\sigma} = \frac{1}{5} \sum_{i=1}^{5} \frac{\sigma_i}{P_i}$
- Puntuación estacional $\mathcal{S} = \sum_{i=1}^{5} \frac{|P_i^W - P_i^S|}{P_i}$
- Puntuación de diferencia de días entre semana vs fines de semana

$$\mathcal{W} = \sum_{i=1}^{5} \frac{|P_i^{WD} - P_i^{WE}|}{P_i}$$

Datos

En resumen, se han considerado 4 conjuntos de datos para analizar los perfiles de consumo energético.

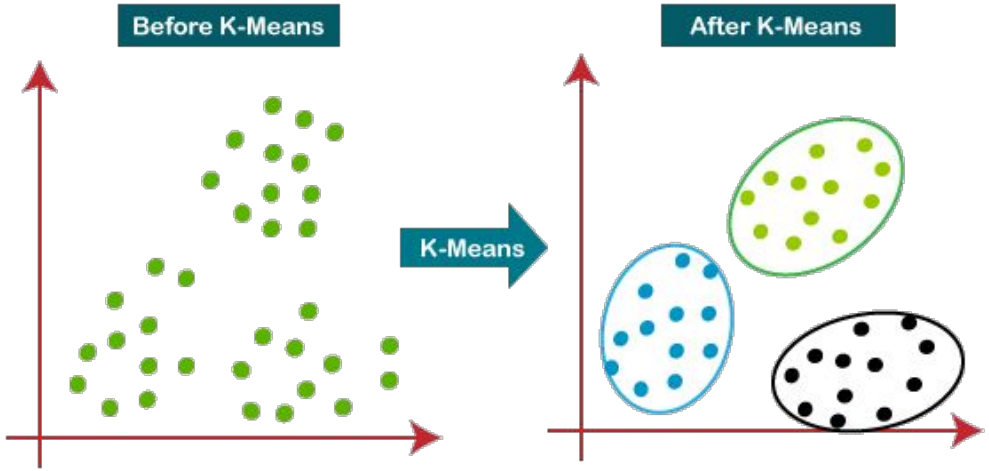
- Serie temporal del consumo eléctrico semanal
- Serie temporal del consumo eléctrico mensual
- Conjunto basado en características estadísticas
- Conjunto de datos basado en características estacionales y de consumo

Resumen

- Introducción
- Materiales y Métodos
  - Datos
  - Técnicas Utilizadas
  - Medidas de Validación
- Resultados Obtenidos
- Discusión y Conclusión

Técnicas Utilizadas

K-Means



El algoritmo K-Means es una de las técnicas de clustering más sencillas y utilizadas. Determina los centroides de los clusters pertenecientes a un conjunto de datos, en función de un valor K que representa el número de clusters en los que se dividirán.

Técnicas Utilizadas

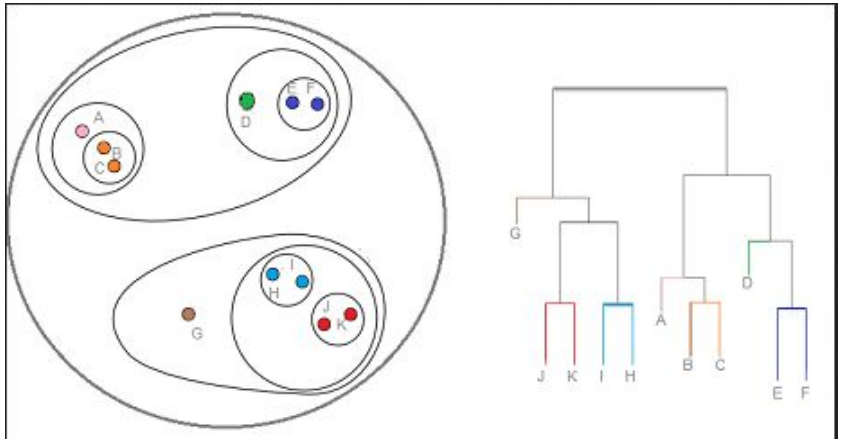
K-Means (Clustering)

- En primer lugar, asigna el centroide más cercano a cada dato para minimizar la suma de la distancia al cuadrado.
- Luego, vuelve a calcular los centroides basándose en la media de los datos que le fueron asignados.

$$E = \sum_{k=1}^{K} \sum_{o \in C_k} \|o - c_k\|^2$$

Técnicas Utilizadas

Agrupación Jerárquica (Clustering)



- La agrupación jerárquica permite construir una estructura jerárquica o de enlace entre los clusters formados, que puede ser aglomerativa o divisoria.
- En el método aglomerativo, cada objeto se considera inicialmente como un grupo. A continuación, los grupos se combinan de forma iterativa para formar una jerarquía ascendente de grupos hasta llegar a un único grupo raíz.

Técnicas Utilizadas

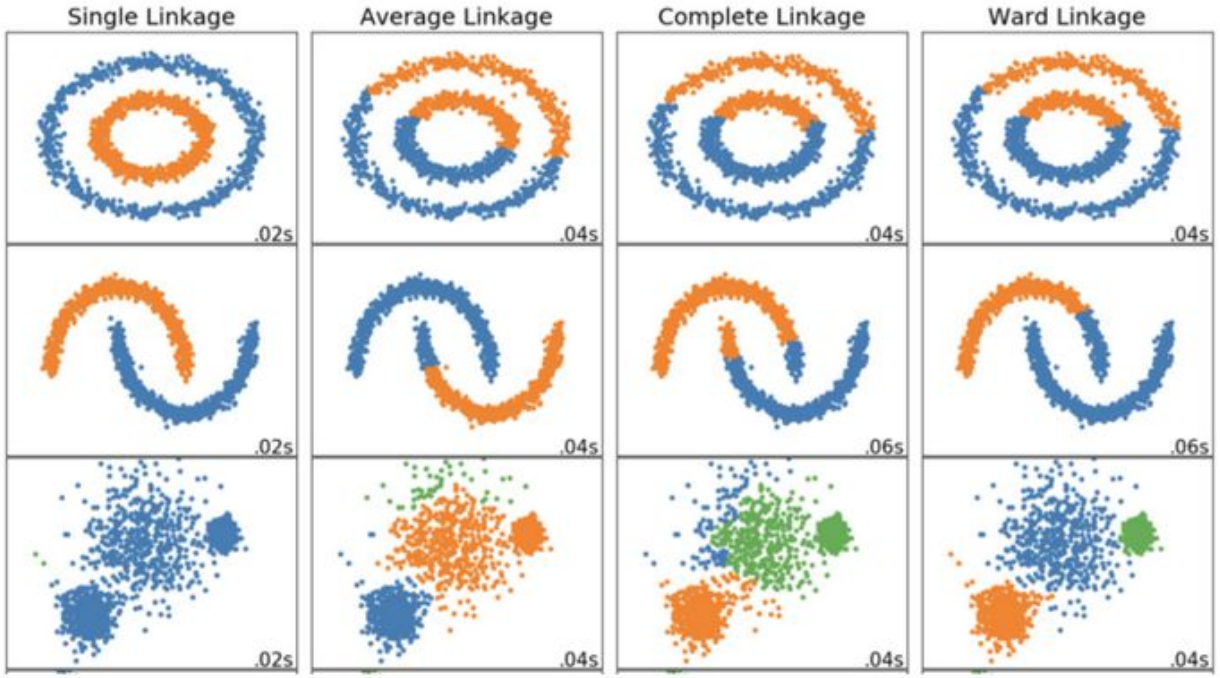
Agrupación Jerárquica (Clustering)

Para determinar lo agrupacion a realizarse, se deben considerar ciertos criterios de enlace:

Criterion	Formula
Single	$D(C_i, C_j) = \min_{o \in C_i, o' \in C_j} d(o, o')$
Complete	$D(C_i, C_j) = \max_{o \in C_i, o' \in C_j} d(o, o')$
Average	$D(C_i, C_j) = \frac{1}{\lvert C_i \rvert} \frac{1}{\lvert C_j \rvert} \sum_{o \in C_i} \sum_{o' \in C_j} d(o, o')$
Centroid	$D(C_i, C_j) = d(c_i, c_j)$
Ward	$D(C_i, C_j) = \sum_{o \in C_i \cup C_j} d(o, c_{i,j})^2$

Técnicas Utilizadas

Agrupación Jerárquica



Técnicas Utilizadas

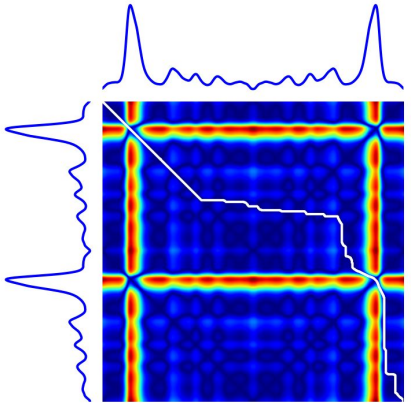
Medidas de Distancia

Distancia Euclidiana

$$d_e(x,y) = \sqrt{\sum_{i=1}^{N} \|x_i - y_i\|^2}$$

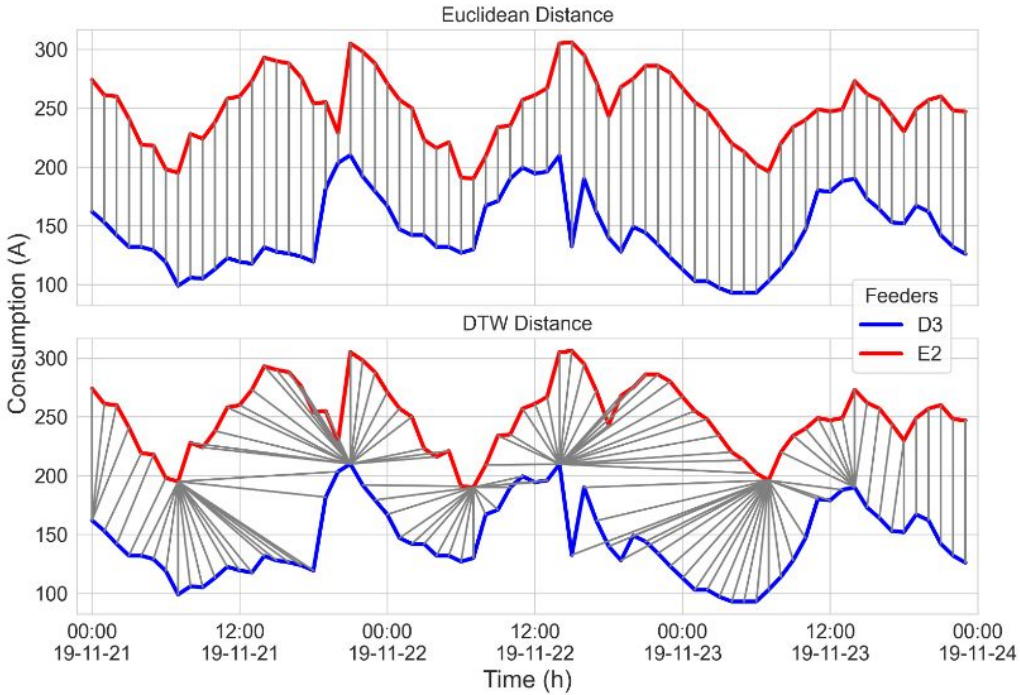
Deformación temporal dinámica (DTW)

$$d_{DTW}(x,y) = m_{wp}(x,y)$$



Técnicas Utilizadas

Medidas de Distancia



Resumen

- Introducción
- Materiales y Métodos
  - Datos
  - Técnicas Utilizadas
  - Medidas de Validación
- Resultados Obtenidos
- Discusión y Conclusión

Técnicas Utilizadas

Medidas de Validación

Índice de Silhouette $SIL = \frac{1}{N}\sum_{i=1}^{N} \int_i$ ↑ Mayor es mejor

Índice de Davies-Bouldin $DB = \frac{1}{K}\sum_{i=1}^{K} \max_{i \neq j} R_{ij}$ ↓ Menor es mejor

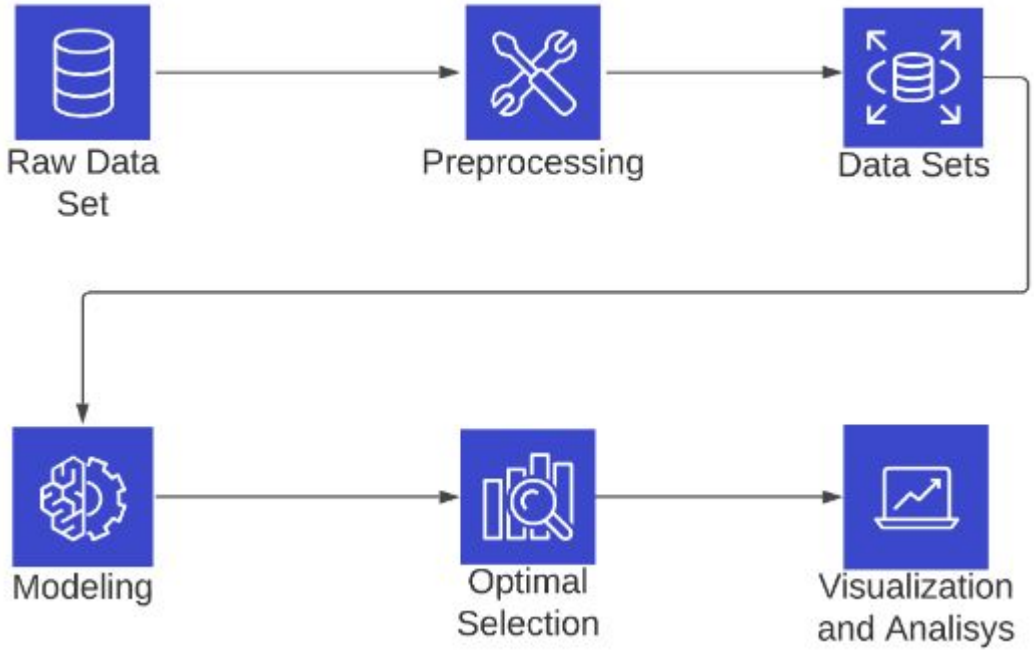
Índice de Calinski-Harabasz $CH = \frac{tr(B)}{tr(w)} \times \frac{n_o - K}{K - 1}$ ↑ Mayor es mejor

UACON21
Un espacio de Divulgación Científica

Resumen

- Introducción
- Materiales y Métodos
  - Datos
  - Técnicas Utilizadas
  - Medidas de Validación
- Resultados Obtenidos
- Discusión y Conclusión

Resultados Obtenidos



Resultados Obtenidos

Data Sets	Algorithm	Distance	Linkage Criteria	Model ID
Weekly time series	K-Means	Euclidean	-	week_k-means_euclid
		DTW	-	week_k-means_dtw
	Hierarchical	Euclidean	Single	week_hier_euclid_single
			Complete	week_hier_euclid_complete
			Average	week_hier_euclid_average
			Centroid	week_hier_euclid_centroid
			Ward	week_hier_euclid_ward
		DTW	Single	week_hier_dtw_single
			Complete	week_hier_dtw_complete
			Average	week_hier_dtw_average
			Centroid	week_hier_dtw_centroid
			Ward	week_hier_dtw_ward
Monthly time series	K-Means	Euclidean	-	month_k-means_euclid
		DTW	-	month_k-means_dtw
	Hierarchical	Euclidean	Single	month_hier_euclid_single
			Complete	month_hier_euclid_comple
			Average	month_hier_euclid_average
			Centroid	month_hier_euclid_centroid
			Ward	month_hier_euclid_ward
		DTW	Single	month_hier_dtw_single
			Complete	month_hier_dtw_complete
			Average	month_hier_dtw_average
			Centroid	month_hier_dtw_centroid
			Ward	month_hier_dtw_ward

UACON21
Un espacio de Divulgación Científica

Resultados Obtenidos

Data Sets	Algorithm	Distance	Linkage Criteria	Model ID
Statistical Based	K-Means	Euclidean	-	stats_k-means
	Hierarchical	Euclidean	Single	stats_hier_single
			Complete	stats_hier_complete
			Average	stats_hier_average
			Centroid	stats_hier_centroid
			Ward	stats_hier_ward
Seasonal Based	K-Means	Euclidean	-	seas_k-means
	Hierarchical	Euclidean	Single	seas_hier_single
			Complete	seas_hier_complete
			Average	seas_hier_average
			Centroid	seas_hier_centroid
			Ward	seas_hier_ward

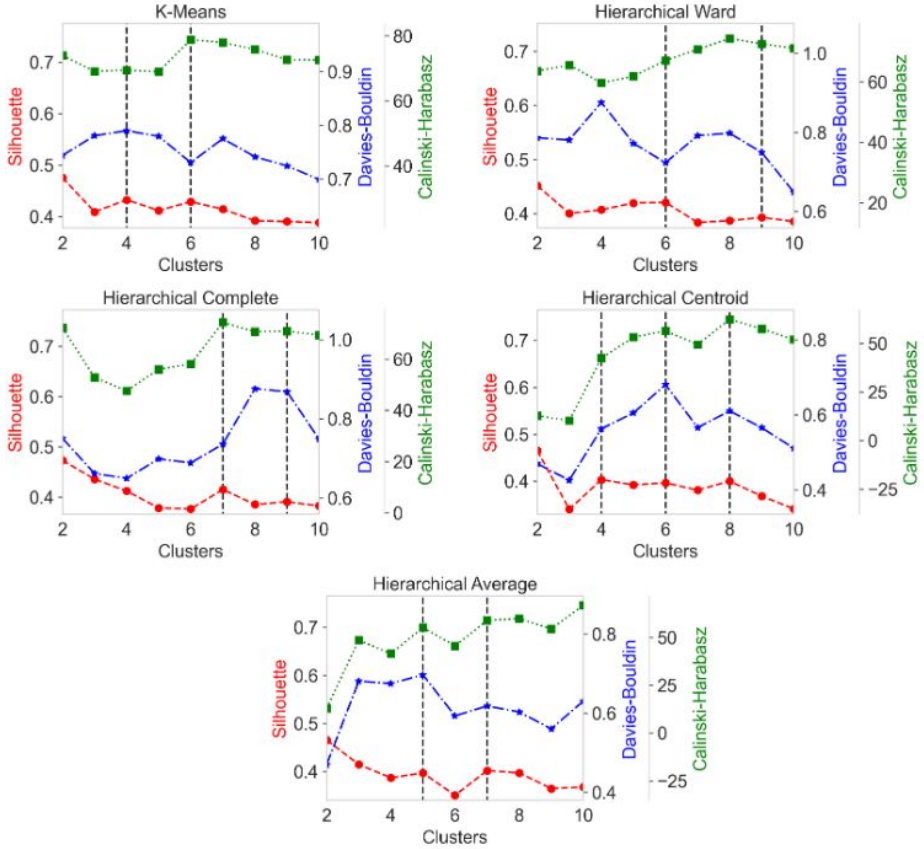
UACON21
Un espacio de Divulgación Científica

Resultados Obtenidos

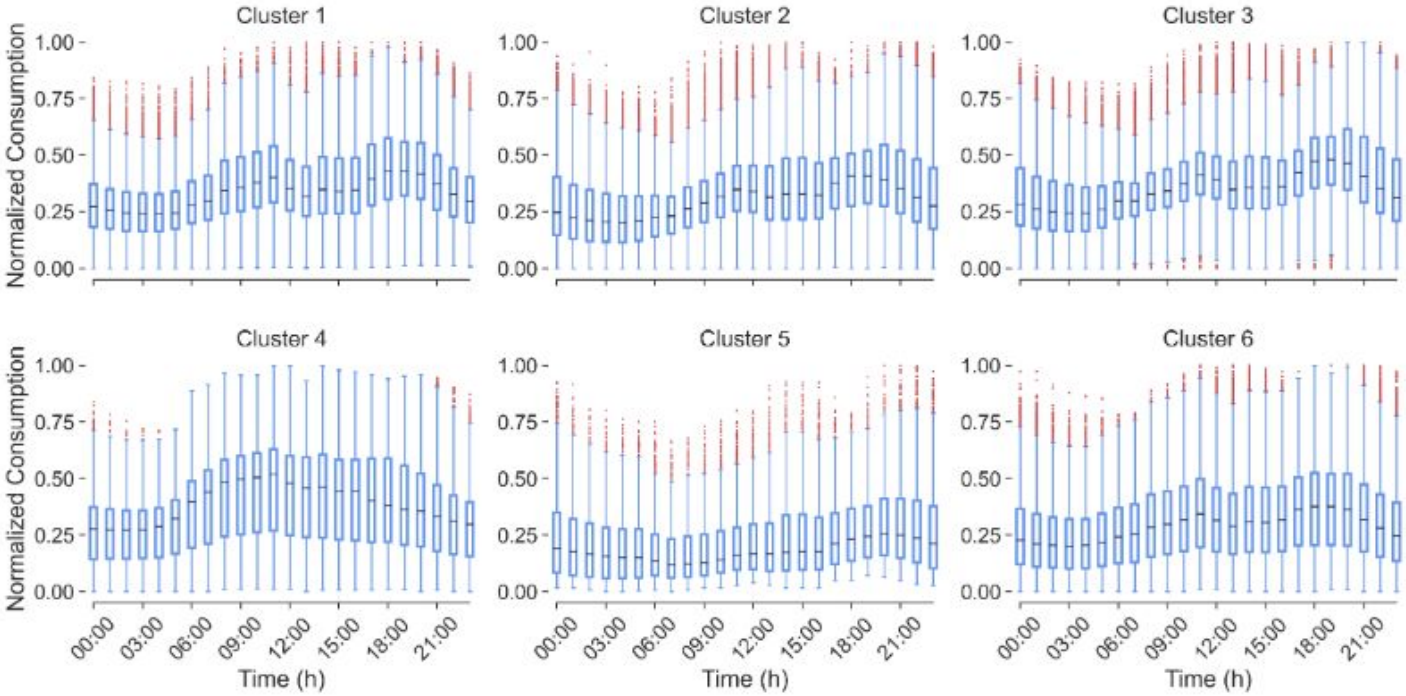
Rank	Model ID	Silhouette Score	Clusters
1	seas_k-means	0.432	4
2	seas_k-means	0.428	6
3	seas_hier_ward	0.421	6
4	seas_hier_complete	0.415	7
5	seas_hier_centroid	0.403	4
6	seas_hier_average	0.402	7
7	seas_hier_centroid	0.400	8
8	seas_hier_average	0.397	5
9	seas_hier_centroid	0.397	6
10	seas_hier_ward	0.393	9
11	seas_hier_complete	0.391	9
12	week_k-means_dtw	0.239	3
13	week_hier_dtw_complete	0.224	4
14	week_hier_euclid_complete	0.216	5
15	month_hier_euclid_ward	0.213	6

UACON21
Un espacio de Divulgación Científica

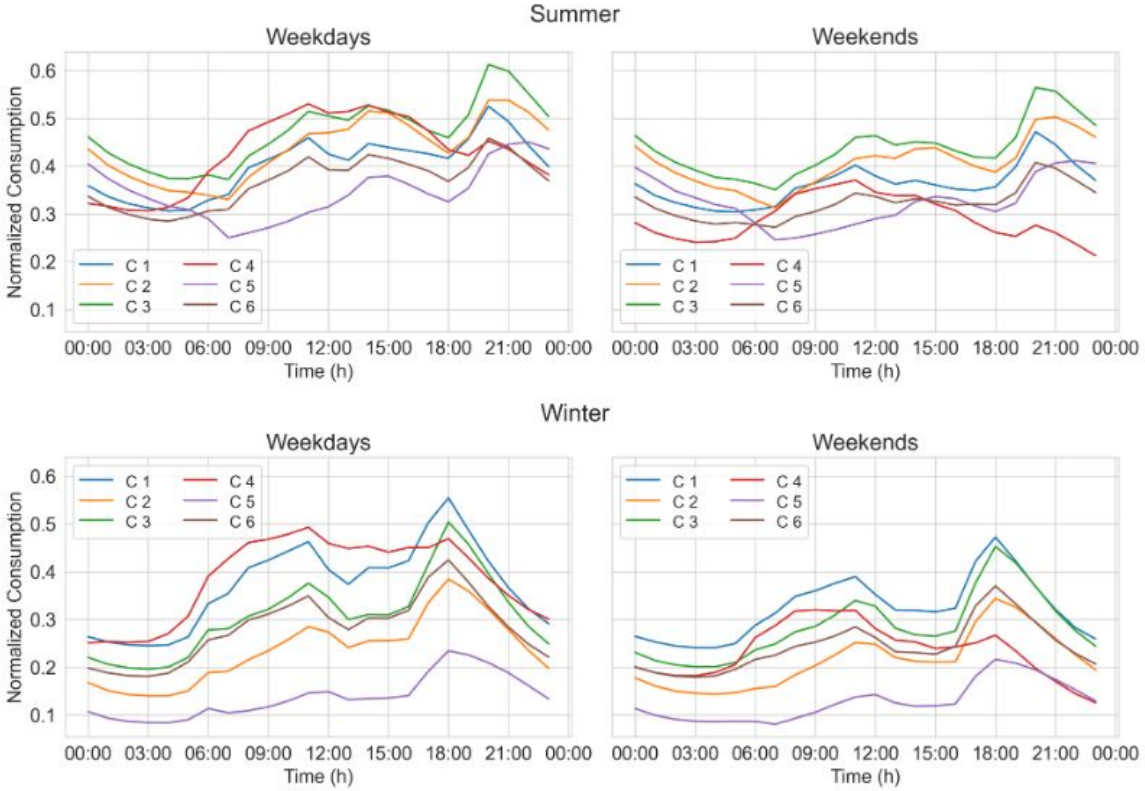
Resultados Obtenidos



Resultados Obtenidos



Resultados Obtenidos



Resultados Obtenidos

La distribución de los cluster es tal como:

- Cluster 1: A1, N1, M5, L3, K3, I1, I2, I5, D1, N4 (10 alimentadores)
- Cluster 2: H3, M6, G3, L1, I3, E4, G1, H1, E7, F1, I4 (11 alimentadores)
- Cluster 3: C1, K2, M4, A2, K1,J1, B5, H2, G4, G2, E6, E1, E2, D3 (14 alimentadores)
- Cluster 4: B1, B3, B4 (3 alimentadores)
- Cluster 5: E3, L4 (2 alimentadores)
- Cluster 6: M7, N2, D2, M3, H4, M1, B2, N3, E5, F2, H5, M2, L2 (13 alimentadores)

Resumen

- Introducción
- Materiales y Métodos
  - Datos
  - Técnicas Utilizadas
  - Medidas de Validación
- Resultados Obtenidos
- Discusión y Conclusión

Discusión y Conclusiones

- Se presenta por primera vez un análisis de cluster de datos reales del sistema eléctrico de la región este de Paraguay. Los datos contienen cuatro años de consumo eléctrico horario de 115 alimentadores distribuidos en 17 subestaciones.
- Los datos se pre procesaron para generar cuatro conjuntos de datos útiles para los algoritmos de agrupación según lo siguiente: i) una demanda semanal, ii) una demanda mensual, iii) un conjunto de características estadísticas, y iv) un conjunto de características de consumo estacional y diario.
- Se utilizaron los algoritmos K-means y de aglomeración jerárquica con las medidas euclidianas y de deformación temporal dinámica (DTW) como métricas de distancia

Discusión y Conclusiones

- El conjunto de características estacionales obtuvo los mejores resultados.
- K-means con 6 cluster presenta el mejor rendimiento en general.
- En futuros trabajos, otros algoritmos de clustering como Kernel DBScan, Fuzzy C-Means modificado, o K-Medoids Based Genetic Clustering. También se propone un enfoque de Biclustering como una alternativa interesante para futuros trabajos de esta investigación.

UACON21
Un espacio de Divulgación Científica

MUCHAS GRACIAS!