

CINVESTAV

ANÁLISIS DE DATOS

# PROYECTO

**Integrantes:**

AUTOR – JOSÉ CARLOS MORÍN GARCÍA

Fecha de realización: 1 de marzo de 2021

Fecha de entrega: 21 de enero de 2021

Firma del docente: .....



## **Problema o Resumen**

Encontrar la mejor manera de correlacionar topoformas de México con los k clusters identificados con algoritmos de clustering, en esta ocasión se presentan dos tipos de algoritmos el Gaussian Mixture Model y DBSCAN para poder obtener la mayor cantidad de formas de topoformas a partir de datos climatológicos..

# **Índice**

<b>1. Introducción</b>	<b>2</b>
<b>2. Marco conceptual</b>	<b>2</b>
2.1. Agrupamiento / Clustering . . . . .	2
2.2. Clasificación . . . . .	2
2.3. GMM - Gaussian Mixture Model . . . . .	3
2.3.1. Densidad de probabilidad . . . . .	3
2.4. DBSCAN - Density Based Spatial Clustering of Applications with Noise . . . . .	3
<b>3. Dataset</b>	<b>4</b>
3.1. Muestra . . . . .	5
<b>4. Preprocesamiento</b>	<b>6</b>
4.1. Eliminación de datos nulos y remplazo de datos . . . . .	6
4.2. Rellenado de valores faltantes . . . . .	7
<b>5. Implementación</b>	<b>8</b>
5.1. Visualización de los Datos . . . . .	8
5.2. GMM . . . . .	9
5.2.1. MERRA . . . . .	9
5.2.2. EMAS . . . . .	13
5.3. DBSCAN . . . . .	15
5.3.1. MERRA . . . . .	15
5.4. EMAS . . . . .	18
<b>6. Resultados</b>	<b>20</b>
6.1. Resultados en Mapa . . . . .	20
6.1.1. GMM . . . . .	20
6.1.2. DBSCAN . . . . .	22
6.2. Validación . . . . .	23
6.2.1. GMM . . . . .	23
6.2.2. DBSCAN . . . . .	30
<b>7. Conclusión</b>	<b>30</b>

# **1. Introducción**

En el presente año, la información es fundamental para los seres humanos, se dice que es indispensable porque a partir de ella se realizar estudios científicos, que pueden obtener predicciones, incluso tomar decisiones a partir de ellas. Nos percatamos de que no se trata de pocos datos almacenados, si no que son millones de datos de los cuales a la hora de transformarlos podemos crear información valiosa. Ya que hoy en día las herramientas que nos ofrece la tecnología tienen el inconveniente de que el proceso de transformacional de estos datos en información es demasiado tardado y no es conveniente para las personas que estudian antecedentes del día a día como los datos meteorológicos, incluso estos datos vienen dados en fracciones mas pequeñas como 10 minutos de registros sobre espacio-temporal.

Este proyecto se realiza con el fin de encontrar una correlación entre las topoformas y las fuentes de Emas y Merra, mediante el agrupamiento realizado mediante temperaturas máximas y mínimas, los resultados van a ser comparados mediante un clustering de K-means con datos provenientes de un dataset. Un objetivo secundario sería la implementación de este estudio en el GeoPortal propio para tener variantes de algoritmos de agrupamiento y así los usuarios que utilicen el sistema puedan realizar diferentes experimentos con las variantes que se puedan ofrecer.

## **2. Marco conceptual**

En esta sección se abordaran los conceptos y definiciones importantes para el mejor entendimiento de este proyecto, tendrá terminologías como clustering, algoritmos de agrupamientos, y los tipos de validación que se utilizan, esto con el fin de que sea mejor comprensión al lector.

### **2.1. Agrupamiento / Clustering**

Clustering o Agrupamiento es el proceso de agrupar datos en clases o clusters de tal forma que los objetos de un cluster tengan una similaridad alta entre ellos, y baja (sean muy diferentes) con objetos de otros clusters. Clustering puede ser aplicado, por ejemplo, para caracterizar clientes, formar taxonomías, clasificar documentos, etc. (1)

### **2.2. Clasificación**

Según la forma en que los clusters se relacionan entre sí y con los objetos del dataset, podemos establecer una primera división entre los diversos algoritmos existentes:(2)

- Clustering Duro: donde cada objeto pertenece a un solo cluster (por lo que los clusters pasarían a ser algo así como una partición del dataset).
- Clustering Blando (o difuso): donde los objetos pertenecen a los clusters según un grado de confianza (o grado de pertenencia).

Pero a veces también podemos encontrar una clasificación más fina atendiendo a cómo se relacionan con detalle:

- Partición estricta: cada objeto pertenece exactamente a un cluster.
- Partición estricta con outliers: puede haber objetos que no pertenecen a ningún cluster (los outliers).
- Clustering con superposiciones: un objeto puede pertenecer a más de un cluster.
- Clustering Jerárquico: Los clusters se ordenan jerárquicamente de forma que los objetos que pertenecen a un cluster también pertenecen a su cluster padre.

## 2.3. GMM - Gaussian Mixture Model

Gaussian Mixture Model (GMM) es un modelo probabilístico en el que se considera que las observaciones siguen una distribución probabilística formada por la combinación de múltiples distribuciones normales (componentes). Puede entenderse como una generalización de K-means con la que, en lugar de asignar cada observación a un único cluster, se obtiene una distribución probabilidad de pertenencia a cada uno.

Ajustar un modelo GMM consiste en estimar los parámetros que definen la función de distribución de cada componente: la media y la matriz de covarianza. Por ejemplo, si el modelo tiene dos componentes, hay que encontrar la media y la matriz de covarianzas de cada una. Si es un problema multidimensional, por ejemplo de 3 variables, la media de cada componente es un vector de 3 valores y la matriz de covarianza una matriz de 3x3. El algoritmo más empleado para realizar el ajuste es Expectation-Maximization (EM).

Una vez aprendidos los parámetros, se puede calcular la densidad de probabilidad que tiene cada observación de pertenecer a cada componente y al conjunto de la distribución. Observaciones con muy poca densidad de probabilidad pueden considerarse como anomalías.(3)

### 2.3.1. Densidad de probabilidad

Al tratarse de un modelo probabilístico, el ajuste de un modelo GMM genera es en realidad una función de densidad de probabilidad. El concepto de densidad puede entenderse como un análogo al de probabilidad en distribuciones discretas, pero, en lugar de acotada en el rango [0,1], puede tomar valores [0, +inf]. El valor de densidad de probabilidad es una medida relativa de verosimilitud (likelihood), cuanto mayor es el valor de densidad de una observación, mayor evidencia de que la observación pertenece a una determinada distribución.(3)

## 2.4. DBSCAN - Density Based Spatial Clustering of Applications with Noise

La agrupación en clústeres basada en la densidad se refiere a métodos de aprendizaje no supervisados que identifican grupos / clústeres distintivos en los datos, basándose en la idea de que un clúster en el espacio de datos es una región contigua de alta densidad de puntos, separada de otros clústeres por regiones contiguas de baja densidad de puntos.

El agrupamiento espacial de aplicaciones con ruido basado en densidad (DBSCAN) es un algoritmo básico para el agrupamiento en clúster basado en densidad. Puede descubrir grupos de diferentes formas y tamaños a partir de una gran cantidad de datos que contienen ruido y valores atípicos. El algoritmo DBSCAN utiliza dos parámetros:

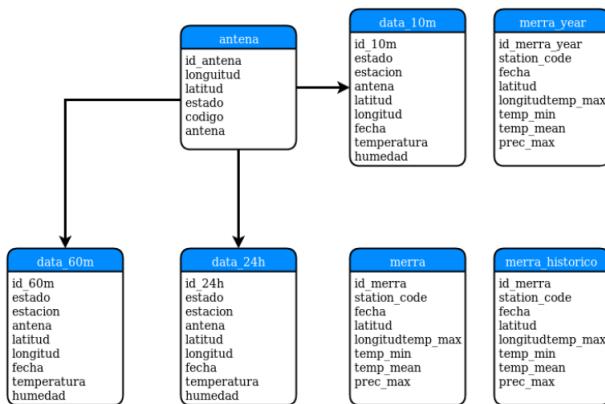
- minPts: el número mínimo de puntos (un umbral) agrupados para que una región se considere densa.
- eps ( $\epsilon$ ): una medida de distancia que se utilizará para ubicar los puntos en la vecindad de cualquier punto.

Hay tres tipos de puntos después de que se completa la agrupación en clústeres de DBSCAN:

- Core: este es un punto que tiene al menos m puntos dentro de una distancia n de sí mismo.
- Border: este es un punto que tiene al menos un punto central a una distancia n.
- Noise: este es un punto que no es ni un núcleo ni un borde. Y tiene menos de m puntos dentro de una distancia n de sí mismo.

### 3. Dataset

En esta sección explicaremos el dataset utilizado para realizar el clustering, este dataset contiene todos los registros de dos fuentes de datos. En la Figura 1 podemos observar como se compone la base de datos en donde se toman la muestras a procesar.



**Figura 1: Arquitectura de la Base de Datos**

La base de datos se encuentra ubicada en un servidor ya que esta es utilizada por un Geoportal que realiza estudios de estos datos dentro de un tiempo y espacio determinado por el usuario. En esta ocasión se cuentan con des fuentes de datos las cuales son:

1. EMAS: Estación Meteorológica Automática, es un sistema propuesto por la CONAGUA que obtiene sus datos mediante antenas, es decir, son datos capturados desde el suelo de la tierra.
2. MERRA: Modern-Era Retrospective analysis for Research and Applications, este sistema mucho más completo que EMAS. Fue realizado este sistema por la NASA y recolecta los datos mediante satélites.

### 3.1. Muestra

Todos los datos capturados dentro de estas dos fuentes fueron a partir de 1980 hasta la actualidad, en varias proporciones como lo son por 10 minutos, días, y años por mencionar algunas. Debido a esto y como se sigue trabajando en el Geoportal se decidió tomar valores del mes de abril del año 2019, porque es uno de los estados donde EMAS es más estable en cuestión de sus registros.

Para la generación del archivo CSV utilizando en este proyecto, se tomo la muestra directamente desde el Geoportal, ya que este hace una integración de los datos entre EMAS y MERRA con los siguientes atributos (ver Figura 2):

#	Column	Non-Null Count	Dtype
0	Antena	63434	non-null
1	Fecha	63434	non-null
2	Latitud	63434	non-null
3	Longitud	63434	non-null
4	Codigo	63434	non-null
5	Temp_max_emas	63126	non-null
6	Temp_min_emas	63126	non-null
7	Hidroregion	63042	non-null
8	Topoforma	62328	non-null
9	Temp_max_merra	63434	non-null
10	Temp_min_merra	63434	non-null
11	Differential_max	63126	non-null
12	Differential_min	63126	non-null
13	Temp_mean_merra	63434	non-null
14	Temp_mean_emas	4051	non-null
15	Humedad	4051	non-null
16	Presion_barometrica	4051	non-null
17	Precipitacion	4051	non-null
18	Radiacion_solar	4051	non-null
19	Etiqueta_clase	63434	non-null
			int64

**Figura 2:** Atributos de los registros Meteorológicos

Es importante mencionar que las Antenas, no necesariamente son antenas físicas si no que se tomo como un nombre general para guardar tanto la antena como el nombre del punto que genera resultados de MERRA.

Se puede ver en la Figura 2 que la cantidad de datos ronda los 63 000 registros, pero dentro de las columnas Temp\_mean\_emas, Humedad, Presion\_barometrica y Radiacion\_solar, solo aparecen 4000 registros, esto quiere decir que tiene datos nulos en sus casillas correspondientes al resto de las Antenas, incluso es algo engañoso, ya que los registros de Temperatura máxima y mínima de EMAS la mayor parte están en un número determinado por el Geoportal para rechazar los valores nulos, este número es él -99.0, en pocas palabras los registros reales de EMAS, están alrededor de los 4000 registros completos en la muestra generada.

## 4. Preprocesamiento

En esta sección explicaremos la preparación de los datos para su posterior estudio. Para poder generar estos cambios en el dataset se hizo uso de la librería numpy que pertenece a Python.

### 4.1. Eliminación de datos nulos y remplazo de datos

Como se observó en la Figura 2 hay un porcentaje algo de valores nulos en algunas columnas y otras que no se ocupaban para el clustering a realizar, entonces se opto por eliminar estas columnas. Las columnas eliminadas fueron las siguientes:

- Fecha: no es necesaria para el agrupamiento.
- Codigo: no es necesaria para el agrupamiento .
- Hidroregion: no es necesaria para el agrupamiento.
- Topoforma: no es necesaria para el agrupamiento.
- Differential\_max: no es necesaria para el agrupamiento.
- Differential\_min: no es necesaria para el agrupamiento.
- Humedad: valores nulos en mas del 90 %.
- Presion\_barometrica: valores nulos en mas del 90 %.
- Precipitacion: valores nulos en mas del 90 %.
- Radiacion\_solar: valores nulos en mas del 90 %.
- Etiqueta\_clase': no es necesaria para el agrupamiento.

Otro problema encontrado dentro de los datos de cada casilla en los que son nulos, no están 100 % marcados como nulos, sino que están en un tipo de dato String y no se pueden convertir a un tipo flotante, esto depende de la fuente, ya que en unos datos viene la palabra 'Null' y en otros la palabra 'NaN'. El remedio para este inconveniente mediante la librería numpy se utilizó el método de replace, para cambiar estas dos presentaciones de datos, cambiándolas por None, ya que python toma esta forma como un valor nulo.

```
data_clima_clus = data_clima_clus.replace('Null',None)
data_clima_clus = data_clima_clus.replace('NaN',None)
```

Los resultados obtenidos se pueden ver en la Figura 3, como vemos el interpretador los datos de None son desplegados como NaN solo para su visualización y de los 20 atributos el dataset se cambio a 9 atributos.

	Antena	Latitud	Longitud	Temp_max_emas	Temp_min_emas	Temp_max_merra	Temp_min_merra	Temp_mean_merra	Temp_mean_emas
0	1003	21.880000	-102.720000	-99.0	-99.0	28.54462	8.47449	17.70834	NaN
1	1003	21.880000	-102.720000	-99.0	-99.0	28.94238	8.73407	18.08676	NaN
2	1003	21.880000	-102.720000	-99.0	-99.0	30.52585	9.16574	18.78622	NaN
3	1003	21.880000	-102.720000	-99.0	-99.0	29.86624	9.59552	19.84659	NaN
4	1003	21.880000	-102.720000	-99.0	-99.0	29.81949	8.31488	18.53400	NaN
...	...	...	...	...	...	...	...	...	...
63429	QO05	20.590000	-100.490000	-99.0	-99.0	29.85660	10.20856	19.36948	NaN
63430	SI09	23.220000	-106.410000	-99.0	-99.0	26.93439	18.46060	22.52142	NaN
63431	MC18	19.720000	-101.180000	-99.0	-99.0	29.45911	10.46365	19.04816	NaN
63432	VILLAGRAN	24.470556	-99.488611	NaN	NaN	31.85730	17.40424	24.21817	NaN
63433	TEPEATLOXTOC	19.569167	-98.824722	26.1	11.4	27.76379	8.74628	17.23352	NaN

63434 rows x 9 columns

**Figura 3:** Eliminación de atributos y cambio de datos nulos

## 4.2. Rellenado de valores faltantes

Como se puede observar en las columnas que le corresponden a EMAS, existen valores de -99.0, estoy por cuestiones de procesamiento del Geoportal corresponden a datos nulos.

La solución empleada en esta ocasión fue llenar los valores faltantes con los datos de EMAS, pero colocando algunas restricciones. Los valores de EMAs tiene mucha variabilidad, en la mayoría de las ocasiones son menores a su semejante de MERRA y el resto son mayores. A continuación se estipulan las condiciones efectuadas dentro del relleno de estos datos:

- Si es mayor al dato de MERRA por más de 5 grados.
- Si es menor al dato de MERRA por más de 5 grados.
- Si el valor de EMAS es -99.0.
- Si el valor de EMAS es nulo, en dado caso de que este uno por alguna condición especial.

Los resultados obtenidos fueron se pueden ver en la Figura , las columnas afectadas a estos cambios fueron Temp\_max\_emas, Temp\_min\_emas y Temp\_mean\_emas.

	Antena	Latitud	Longitud	Temp_max_emas	Temp_min_emas	Temp_max_merra	Temp_min_merra	Temp_mean_merra	Temp_mean_emas
0	1003	21.880000	-102.720000	28.54462	8.47449	28.54462	8.47449	17.70834	17.7083
1	1003	21.880000	-102.720000	28.94238	8.73407	28.94238	8.73407	18.08676	18.0868
2	1003	21.880000	-102.720000	30.52585	9.16574	30.52585	9.16574	18.78622	18.7862
3	1003	21.880000	-102.720000	29.86624	9.59552	29.86624	9.59552	19.84659	19.8466
4	1003	21.880000	-102.720000	29.81949	8.31488	29.81949	8.31488	18.53400	18.534
...	...	...	...	...	...	...	...	...	...
63429	QO05	20.590000	-100.490000	29.85660	10.20856	29.85660	10.20856	19.36948	19.3695
63430	SI09	23.220000	-106.410000	26.93439	18.46060	26.93439	18.46060	22.52142	22.5214
63431	MC18	19.720000	-101.180000	29.45911	10.46365	29.45911	10.46365	19.04816	19.0482
63432	VILLAGRAN	24.470556	-99.488611	31.85730	17.40424	31.85730	17.40424	24.21817	24.2182
63433	TEPEATLOXTOC	19.569167	-98.824722	26.10000	11.40000	27.76379	8.74628	17.23352	17.2335

63434 rows x 9 columns

**Figura 4:** Relleno de valores faltantes

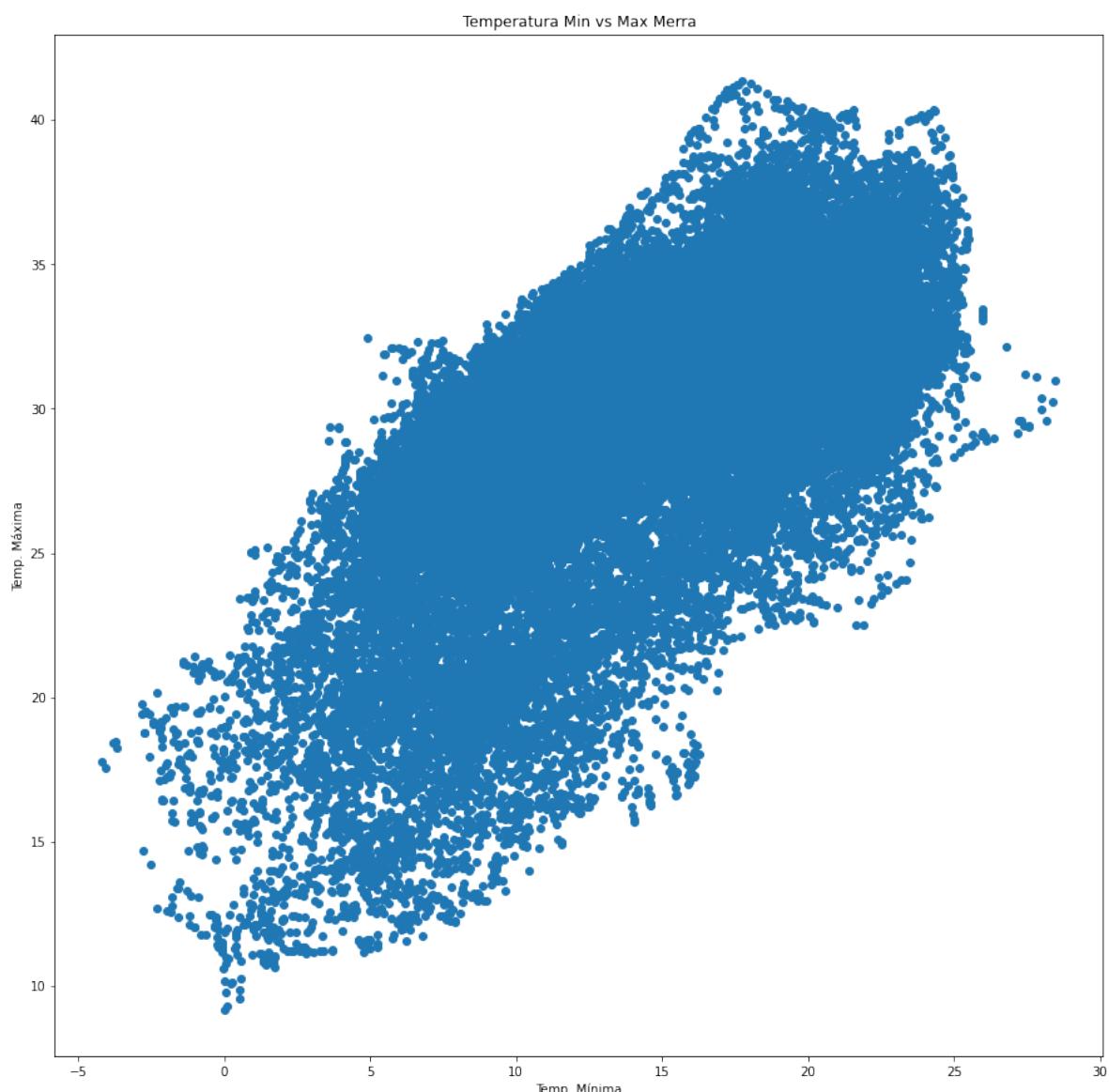
## 5. Implementación

En esta sección se expresará lo realizado en cuestión de los algoritmos de agrupamiento, en primera instancia se explica el Gaussian Mixture Model y al finalizar con el menos representativo para estos datos el DBSCAN. Para cada algoritmo se ejecuto en dos ocasiones con los valores de EMAS en un proceso y los de MERRA en otro proceso, y así comparar los resultados obtenidos entre ellos.

### 5.1. Visualización de los Datos

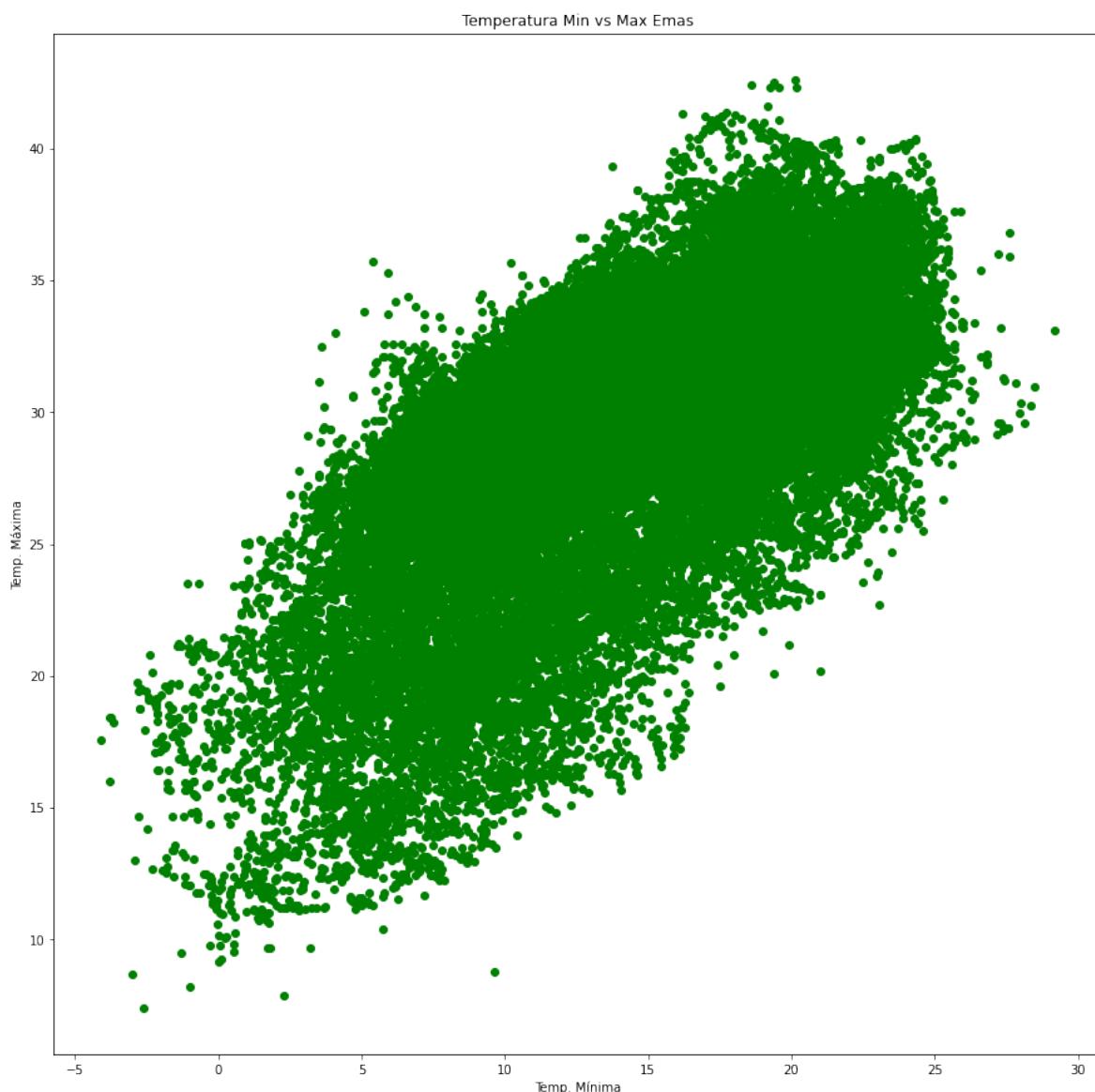
Se muestra mediante una gráfica la representación de las Temperaturas máximas y mínimas de estas dos fuentes.

#### MERRA



**Figura 5:** Temperaturas MERRA

## EMAS



**Figura 6:** Temperaturas EMAS

## 5.2. GMM

Dentro de este modelo, se realizaron dos procesos tanto para EMAS y para MERRA.

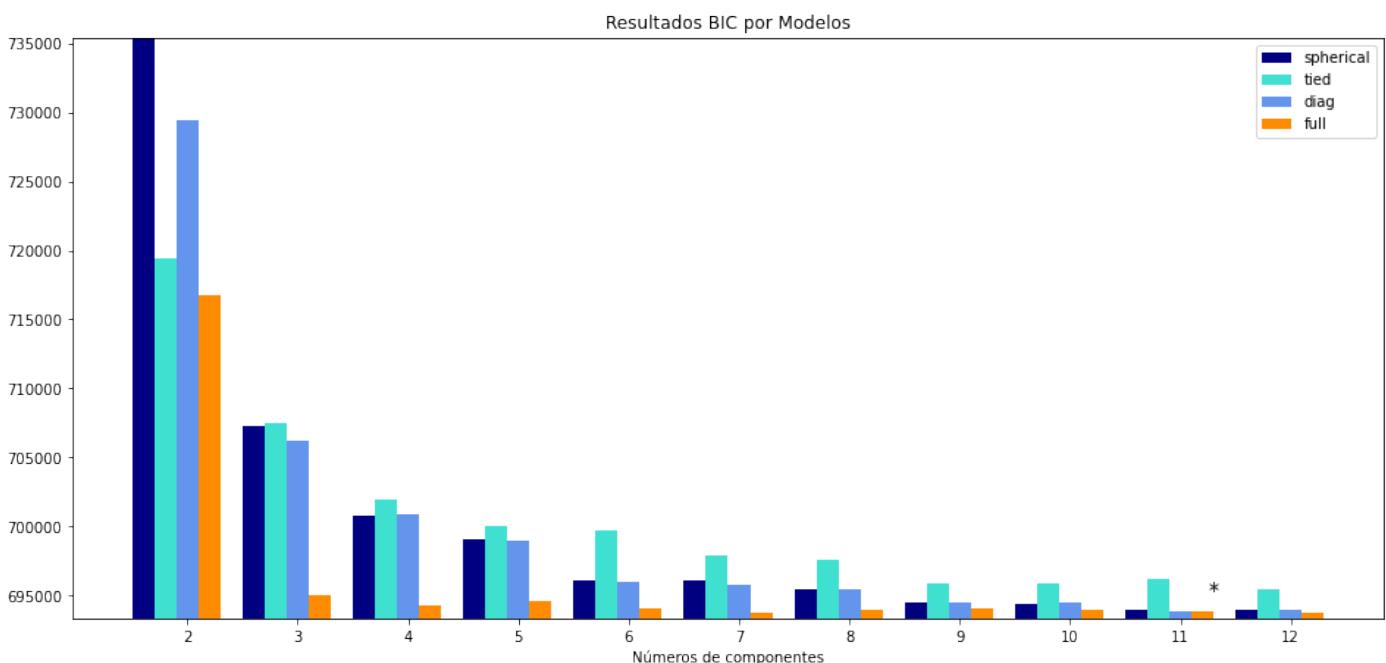
### 5.2.1. MERRA

Para emas se tomaron las columnas de Temperaturas Máximas y Mínimas, como se observa en la Figura 7.

	Temp_max_merra	Temp_min_merra
<b>0</b>	28.54462	8.47449
<b>1</b>	28.94238	8.73407
<b>2</b>	30.52585	9.16574
<b>3</b>	29.86624	9.59552
<b>4</b>	29.81949	8.31488
...	...	...
<b>63429</b>	29.85660	10.20856
<b>63430</b>	26.93439	18.46060
<b>63431</b>	29.45911	10.46365
<b>63432</b>	31.85730	17.40424
<b>63433</b>	27.76379	8.74628
63434 rows × 2 columns		

**Figura 7:** Columnas seleccionadas de MERRA

En la Figura 8 se muestra el numero de k y el modelo que mejor se comporta con este tipo de datos. Para la selección del modelo se refiere tanto al tipo de covarianza como al número de componentes del modelo. En ese caso, AIC también proporciona el resultado correcto (no se muestra para ahorrar tiempo), pero BIC es más adecuado si el problema es identificar el modelo correcto. A diferencia de los procedimientos bayesianos, tales inferencias no tienen antecedentes. De acuerdo a los resultados de BIC para MERRA espera un



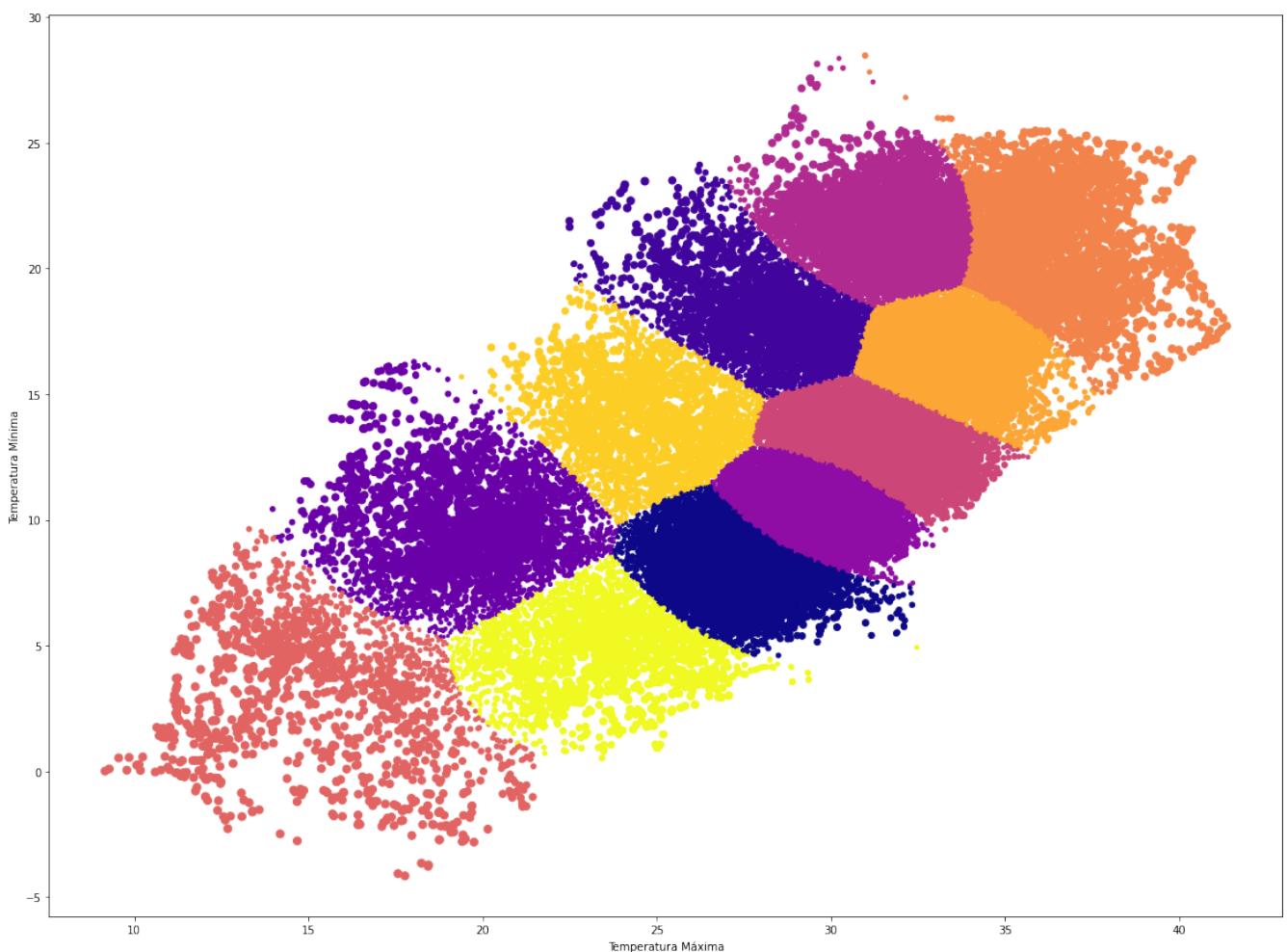
**Figura 8:** Resultados BIC en MERRA

mejor resultado con  $k$  igual a 11 en el modelo de full para realizar un GMM. En la Figura 9 se muestra los elementos a utilizar para generar el clustering de los datos de MERRA.

```
k = 11
modelo_gmm_merra = GaussianMixture(
    n_components      = k,
    covariance_type  = 'full',
    random_state     = 123,
)
modelo_gmm_merra.fit(X=X_merra_clus)
```

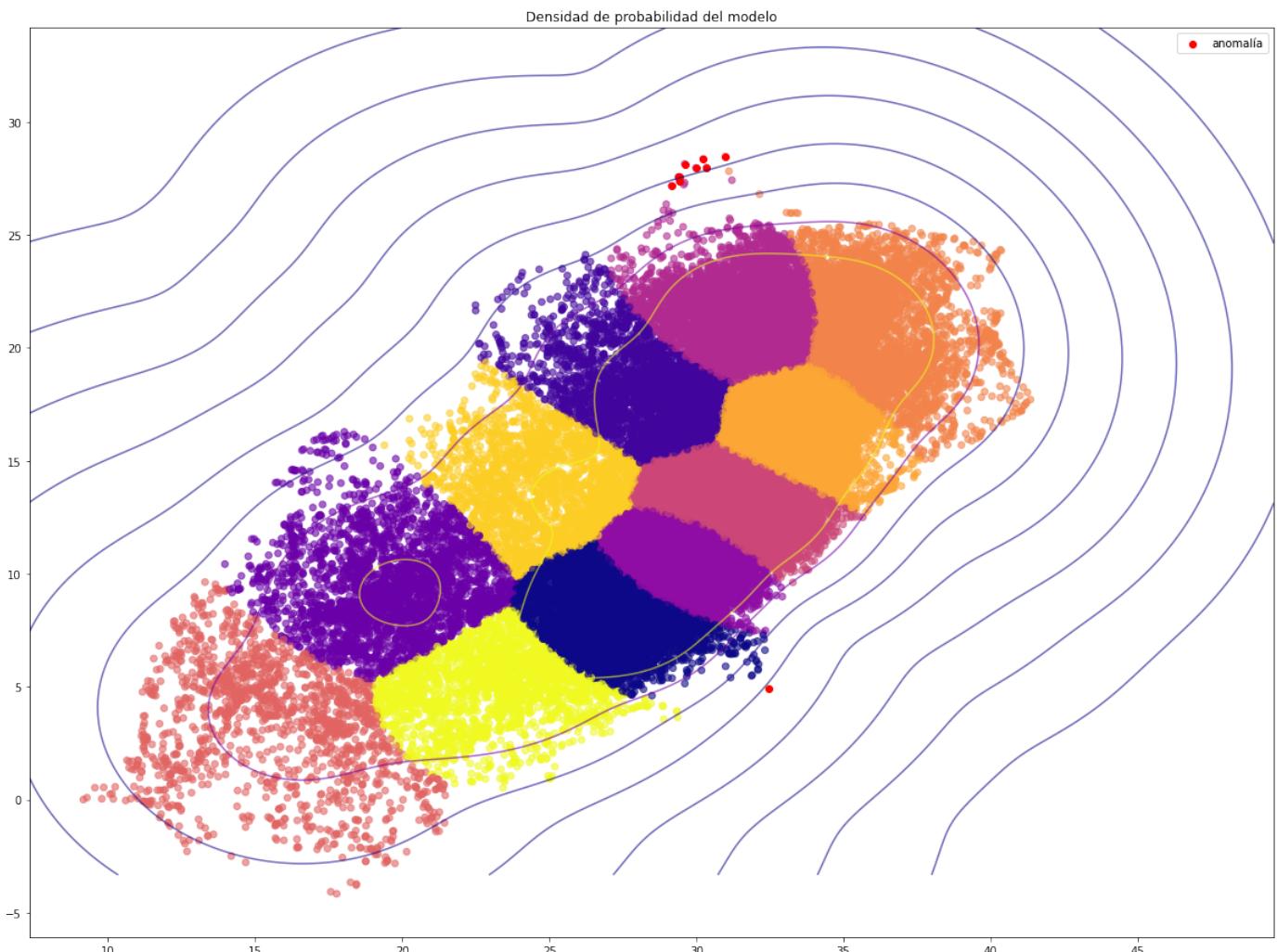
**Figura 9:** GMM Parámetros en MERRA

Se aplicó el algoritmo de GMM a los datos, para ver su resultado en como los acomodo en grupos ver la Figura 10. Como se puede observar las temperaturas que rondan entre los 30 grados como máximo y como mínimo entre 10 y 15 grados son donde se agrupan más grupos y por ende se puede concluir que la densidad de los datos es mayor.



**Figura 10:** Grupos generados por el GMM en MERRA

Para ver la densidad de los datos se puede ver en la Figura 11, estas elipses explican en donde está la mayor cantidad de datos y una detección de anomalías en donde los datos están muy alejados, ahí se puede observar como unos datos en donde fue un día anormal o los sensores fallaron al momento de tomar los registros.



**Figura 11:** Densidad del datos GMM MERRA

### 5.2.2. EMAS

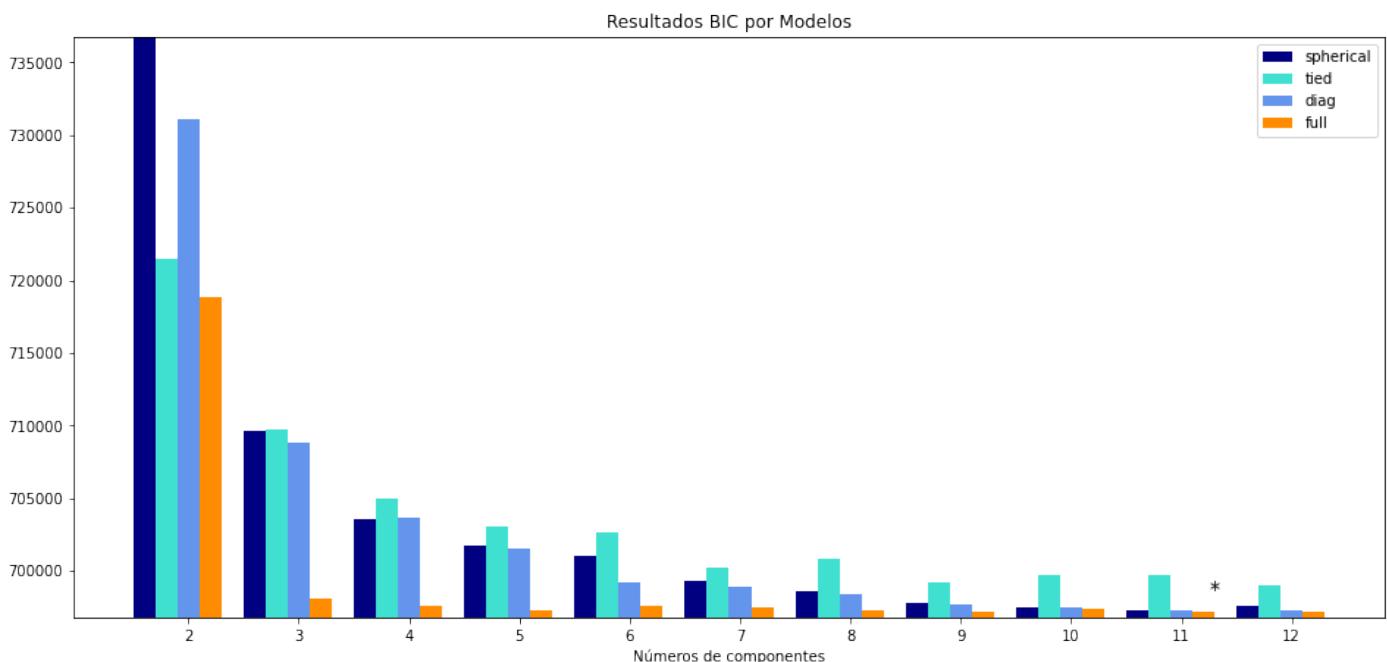
Dentro de la fuente de EMAS se tomaron de igual manera se tomaron las mismas columnas que en MERRa pero correspondientes a esta fuente (ver Figura 12).

	Temp_max_emas	Temp_min_emas
<b>0</b>	28.54462	8.47449
<b>1</b>	28.94238	8.73407
<b>2</b>	30.52585	9.16574
<b>3</b>	29.86624	9.59552
<b>4</b>	29.81949	8.31488
...	...	...
<b>63429</b>	29.85660	10.20856
<b>63430</b>	26.93439	18.46060
<b>63431</b>	29.45911	10.46365
<b>63432</b>	31.85730	17.40424
<b>63433</b>	26.10000	11.40000

63434 rows × 2 columns

**Figura 12:** Columnas seleccionadas de EMAS

En la Figura 13 se muestra el número de k y el modelo que mejor se comporta con este tipo de datos. Los resultados de BIC mostrados indican que el mejor mejor tipo de covarianza a elegir es el tied con un k de 11.



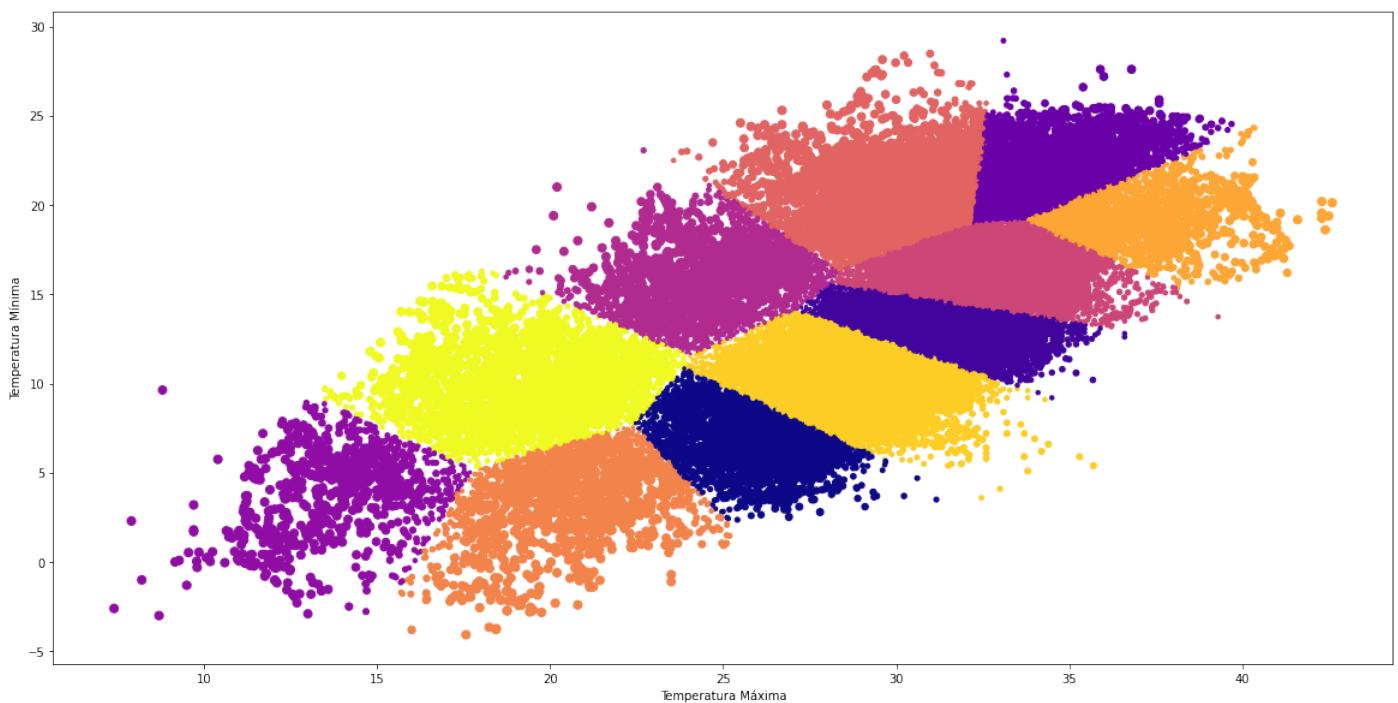
**Figura 13:** Resultados BIC en EMAS

En la Figura 14 se pueden observar los parámetros utilizados para inicializar el modelo de GMM.

```
k = 11
modelo_gmm_emas = GaussianMixture(
    n_components = k,
    covariance_type = 'tied',
    random_state = 100,
)
modelo_gmm_emas.fit(X=X_emas_clus)
```

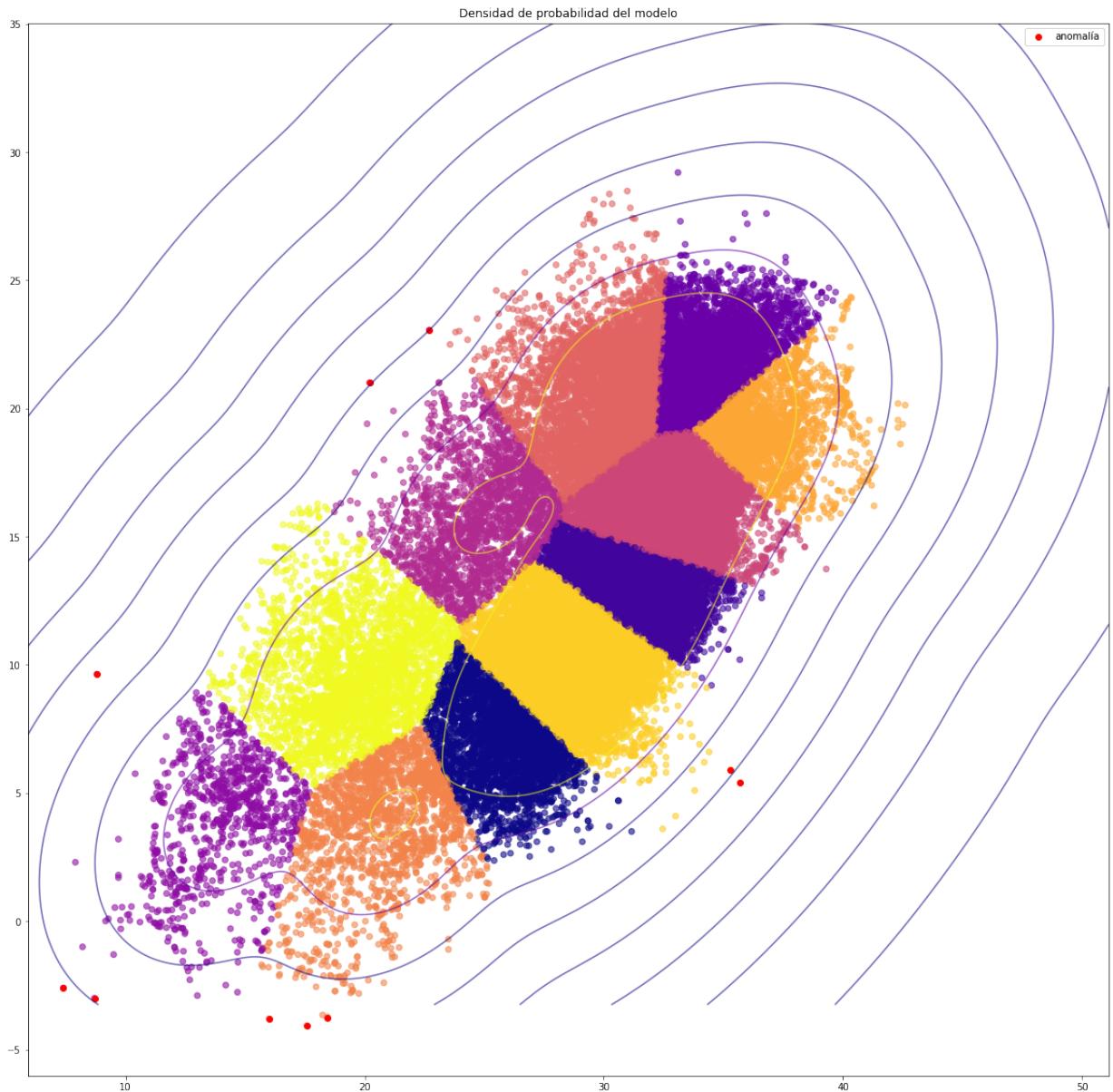
**Figura 14:** GMM Parámetros en EMAS

Se aplicó el algoritmo GMM a los datos de EMAS actualizados con el preprocesamiento, los resultados obtenidos están en la Figura 15 y se puede comparar con los resultados de MERRA al aplicarse el mismo clustering y con su mejor modelo.



**Figura 15:** Grupos generados por el GMM en EMAS

Se está notando que las implementaciones son muy similares, tanto para EMAS como para MERRA, incluso la densidad de los datos de EMAS son en el mismo rango de MERRA, todo esto sucede debido a que los datos fueron actualizados en base a los datos de MERRA (ver Figura 16).



**Figura 16:** Densidad de datos GMM EMAS

Las anomalías presentadas son más visibles en este clustering debido a que la consistencia de los datos de EMAS no es como la de MERRA y hay mas variabilidad de estos.

### 5.3. DBSCAN

Las pruebas realizadas en este algoritmo tuvieron muchas deficiencias y para obtener el mejor resultado se utilizaron más columnas. De igual manera se realizó la función del vecino se calcula y representa las k-distancias para ayudar a identificar el valor óptimo de épsilon.

#### 5.3.1. MERRA

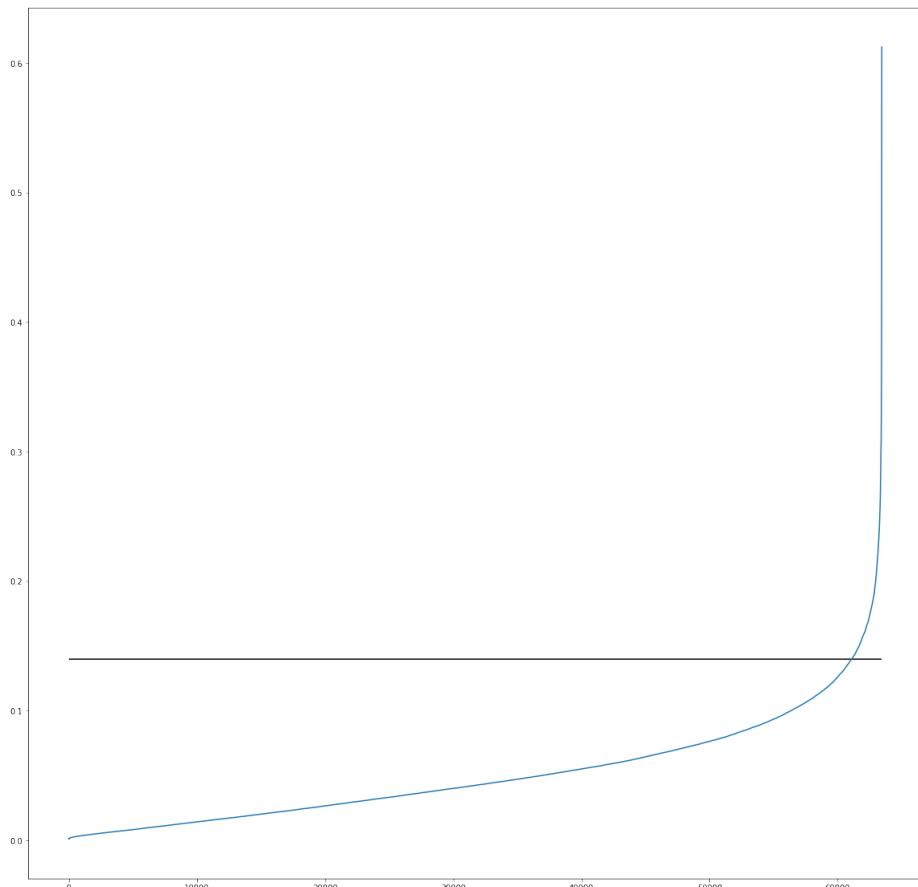
En la Figura 17 se representa las columnas seleccionadas, en este caso se seleccionaron Latitud, Longitud, Temp\_max\_merra, Temp\_min\_merra, Temp\_mean\_merra.

	Latitud	Longitud	Temp_max_merra	Temp_min_merra	Temp_mean_merra
<b>0</b>	21.880000	-102.720000	28.54462	8.47449	17.70834
<b>1</b>	21.880000	-102.720000	28.94238	8.73407	18.08676
<b>2</b>	21.880000	-102.720000	30.52585	9.16574	18.78622
<b>3</b>	21.880000	-102.720000	29.86624	9.59552	19.84659
<b>4</b>	21.880000	-102.720000	29.81949	8.31488	18.53400
...	...	...	...	...	...
<b>63429</b>	20.590000	-100.490000	29.85660	10.20856	19.36948
<b>63430</b>	23.220000	-106.410000	26.93439	18.46060	22.52142
<b>63431</b>	19.720000	-101.180000	29.45911	10.46365	19.04816
<b>63432</b>	24.470556	-99.488611	31.85730	17.40424	24.21817
<b>63433</b>	19.569167	-98.824722	27.76379	8.74628	17.23352

63434 rows × 5 columns

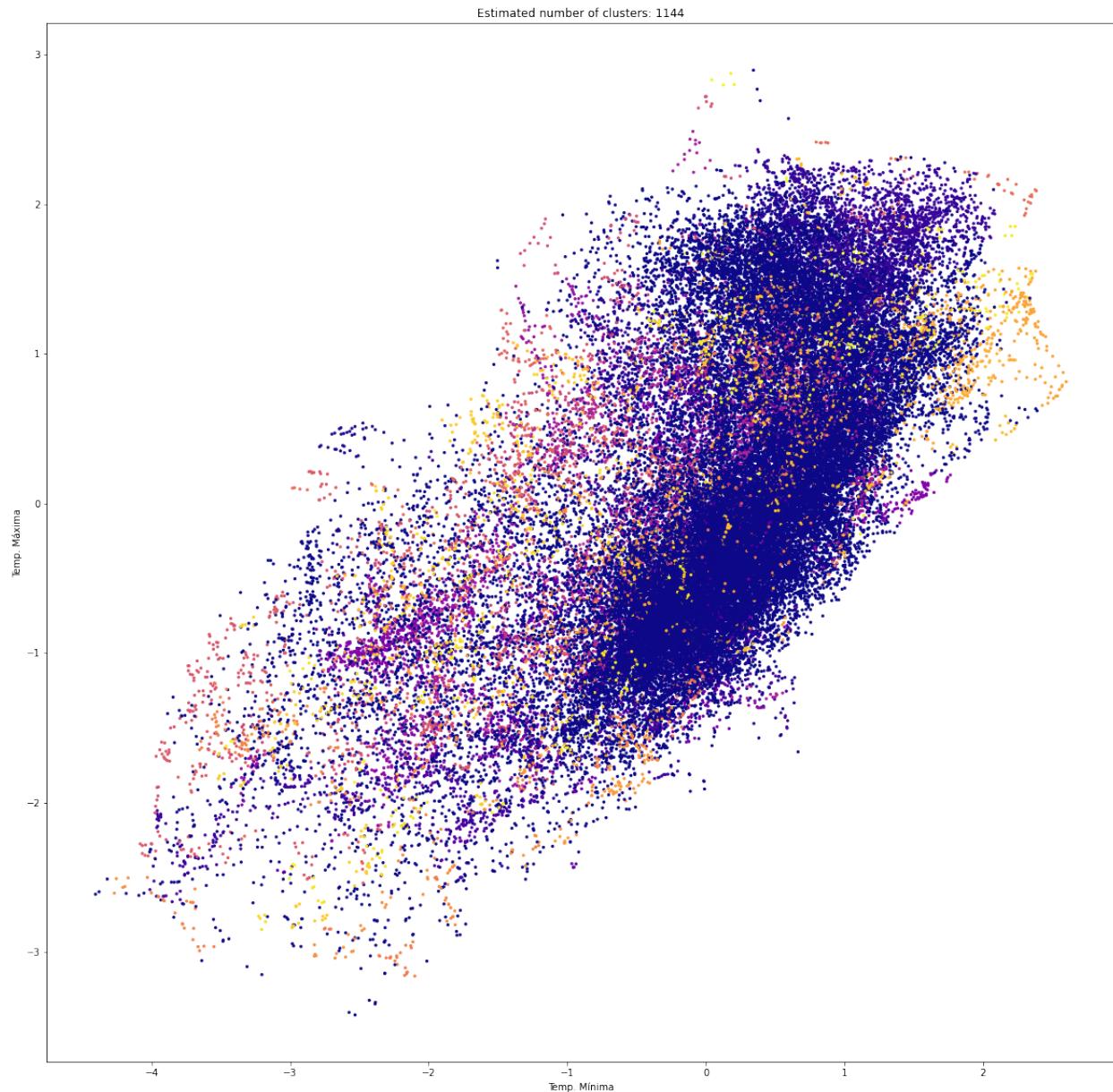
**Figura 17:** Columnas seleccionadas MERRA

Para estos datos se modificaron a un rango de 0 a 1 para normalizar los datos ya que es uno de los requerimientos del algoritmo DBSCAN. Para este algoritmo se válido un numero mínimo de vecinos de 4 y se aplicó la función del vecino. La Figura 18 se nota que el mejor épsilon para estos datos es el 0.14.



**Figura 18:** Mejor epsilon de MERRA

Ya aplicando estos valores dentro del algoritmo se presentaron los siguientes grupos (ver Figura 19). Se nota la gran densidad de un solo grupo, esto debido a una investigación se comenta que una de las desventajas en donde no es muy bueno hacer clustering en este algoritmo es cuando la densidad de los datos es muy alta y demasiada pegada. (4)



**Figura 19:** Agrupamiento de MERRA

## 5.4. EMAS

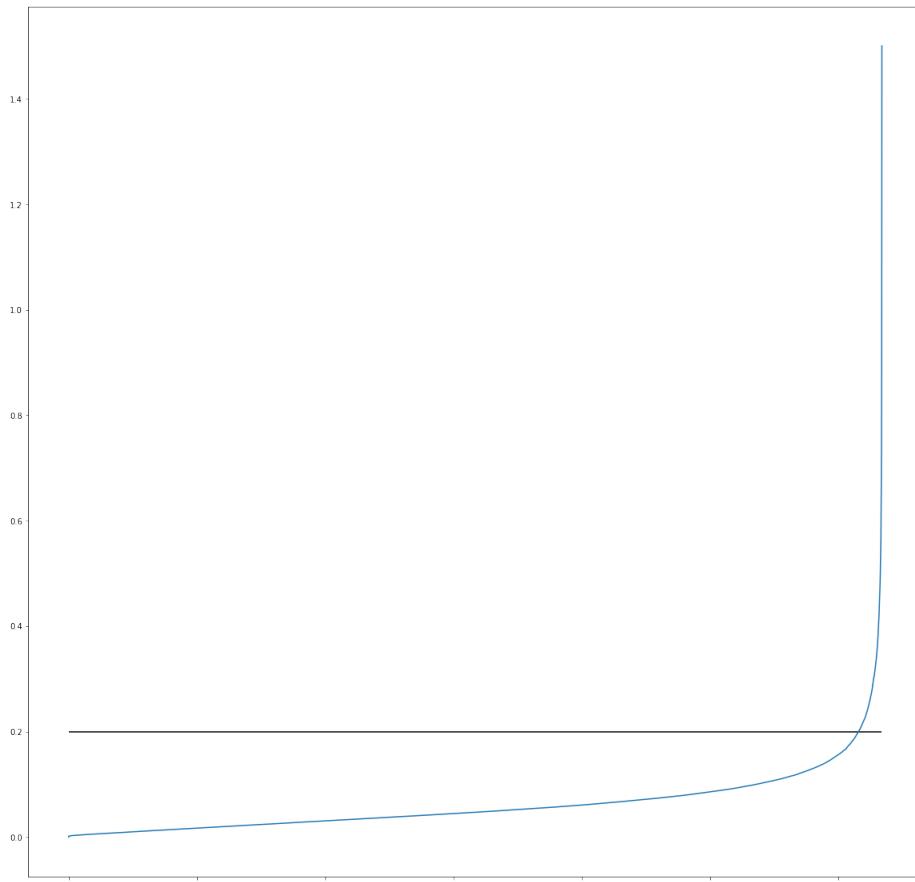
Para EMAS se aplicó la misma técnica que en MERRA, en donde se seleccionaron las mismas columnas correspondientes a su fuente. Se puede notar en la Figura 21.

	Latitud	Longitud	Temp_max_emas	Temp_min_emas	Temp_mean_emas
0	21.880000	-102.720000	28.54462	8.47449	17.70834
1	21.880000	-102.720000	28.94238	8.73407	18.08676
2	21.880000	-102.720000	30.52585	9.16574	18.78622
3	21.880000	-102.720000	29.86624	9.59552	19.84659
4	21.880000	-102.720000	29.81949	8.31488	18.53400
...	...	...	...	...	...
63429	20.590000	-100.490000	29.85660	10.20856	19.36948
63430	23.220000	-106.410000	26.93439	18.46060	22.52142
63431	19.720000	-101.180000	29.45911	10.46365	19.04816
63432	24.470556	-99.488611	31.85730	17.40424	24.21817
63433	19.569167	-98.824722	26.10000	11.40000	17.23352

63434 rows × 5 columns

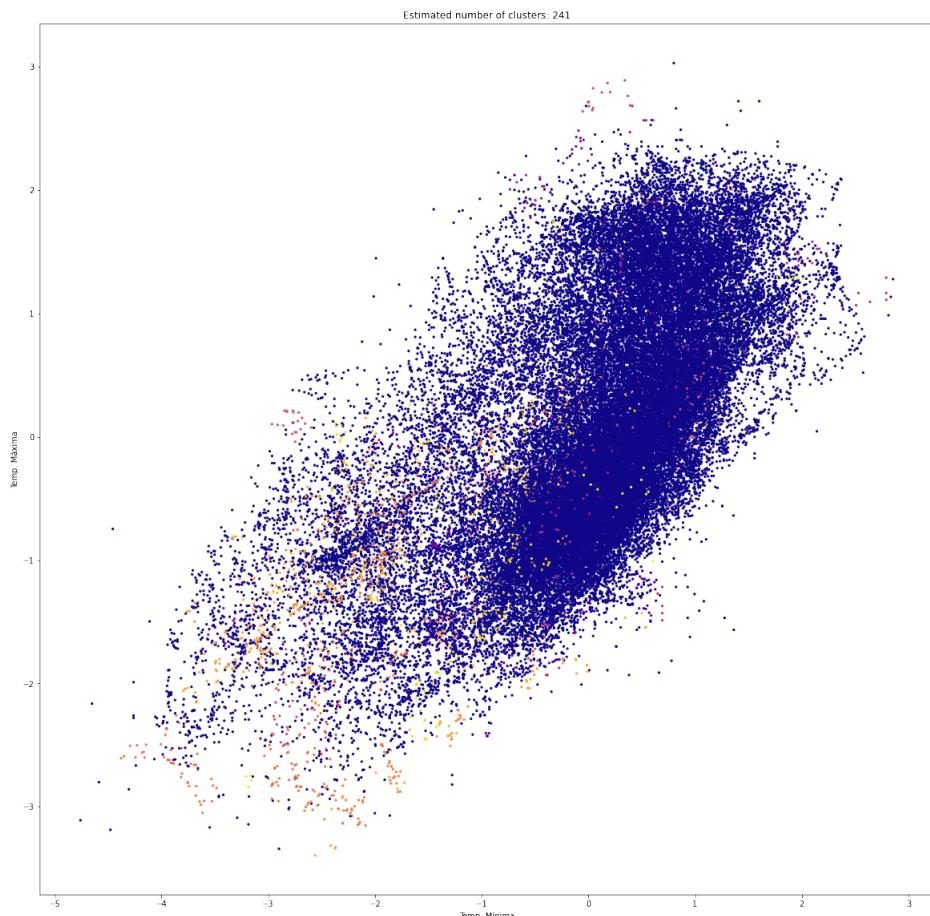
**Figura 20:** Columnas seleccionadas EMAS

Dentro de la aplicación de la función de vecinos, se selecciono la rodilla en la gráfica con el valor de epsilon en 0.2 con de igual manera de 4 vecinos como mínimo.



**Figura 21:** Mejor epsilon de EMAS

En cuestión del clustering generado en estos datos fue más deficiente que el DBSCAN con los datos de MERRA, y era de esperarse y que los EMAS es más inestable que MERRA. En la Figura 22 se puede observar que la mayoría de los datos se fueron en un solo grupo pero la densidad de los anteriores fue mucho menor que en el clustering de MERRA.

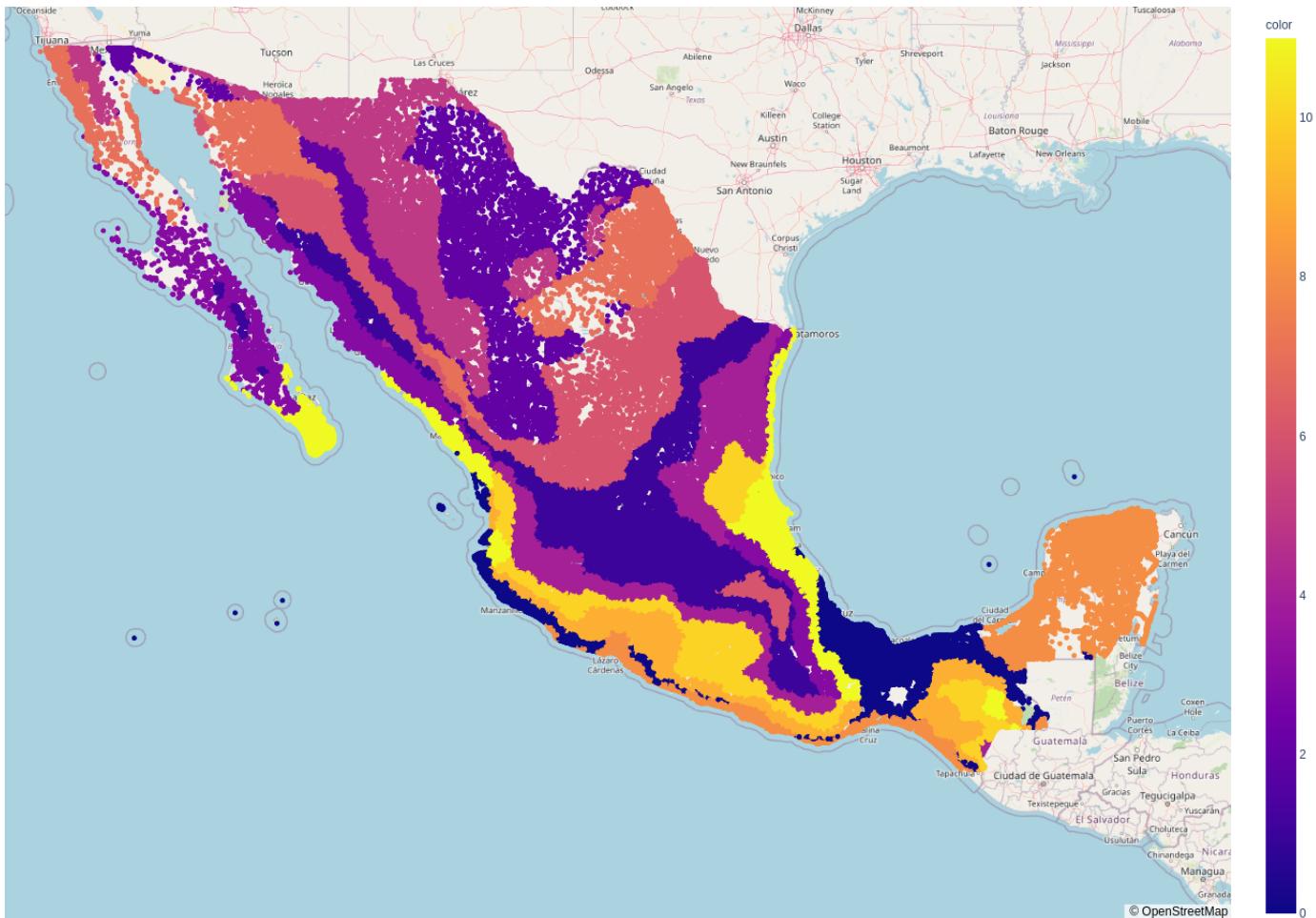


**Figura 22:** Agrupamiento de EMAS

## 6. Resultados

## 6.1. Resultados en Mapa

En primera instancia se espera lograr un agrupamiento lo mas cercano al agrupamiento realizado en estudio mediante el K-means.

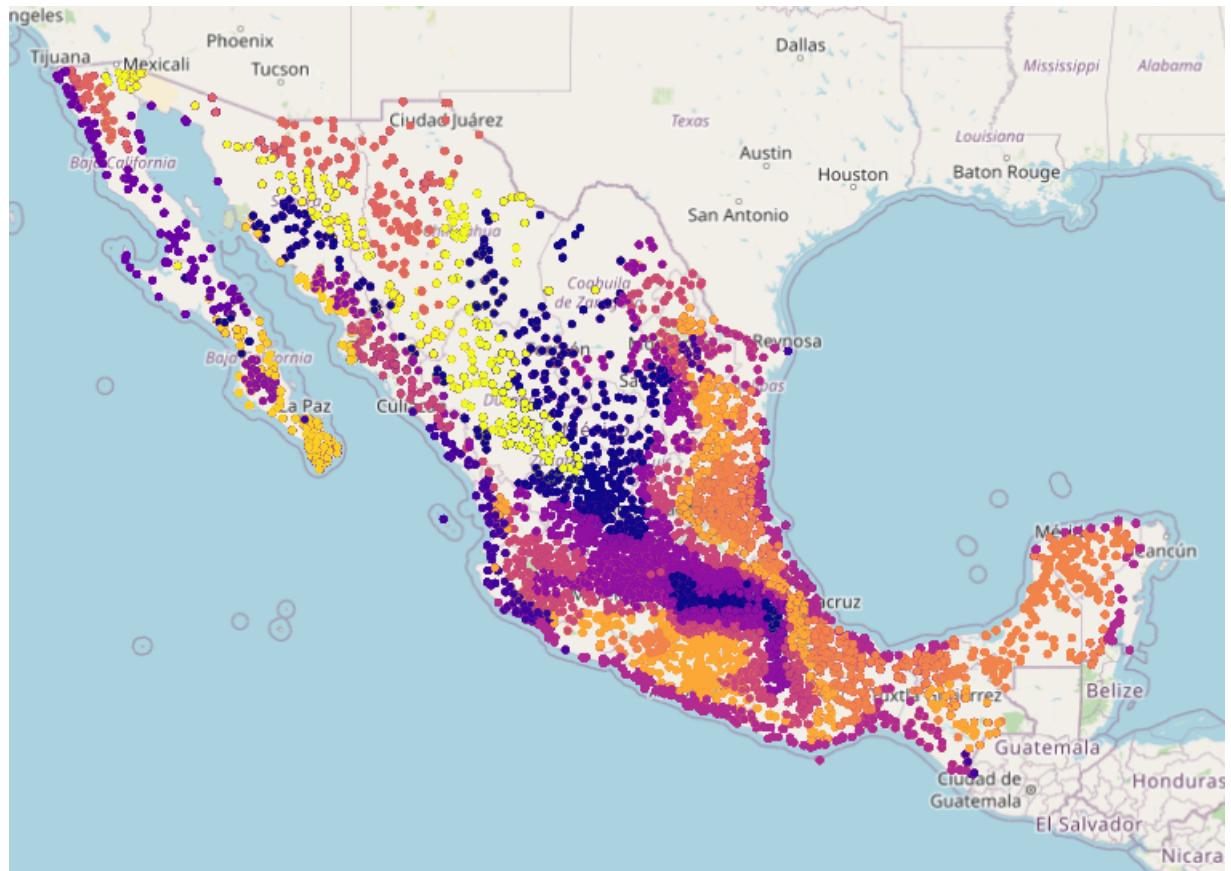


**Figura 23:** Topoformas generadas.

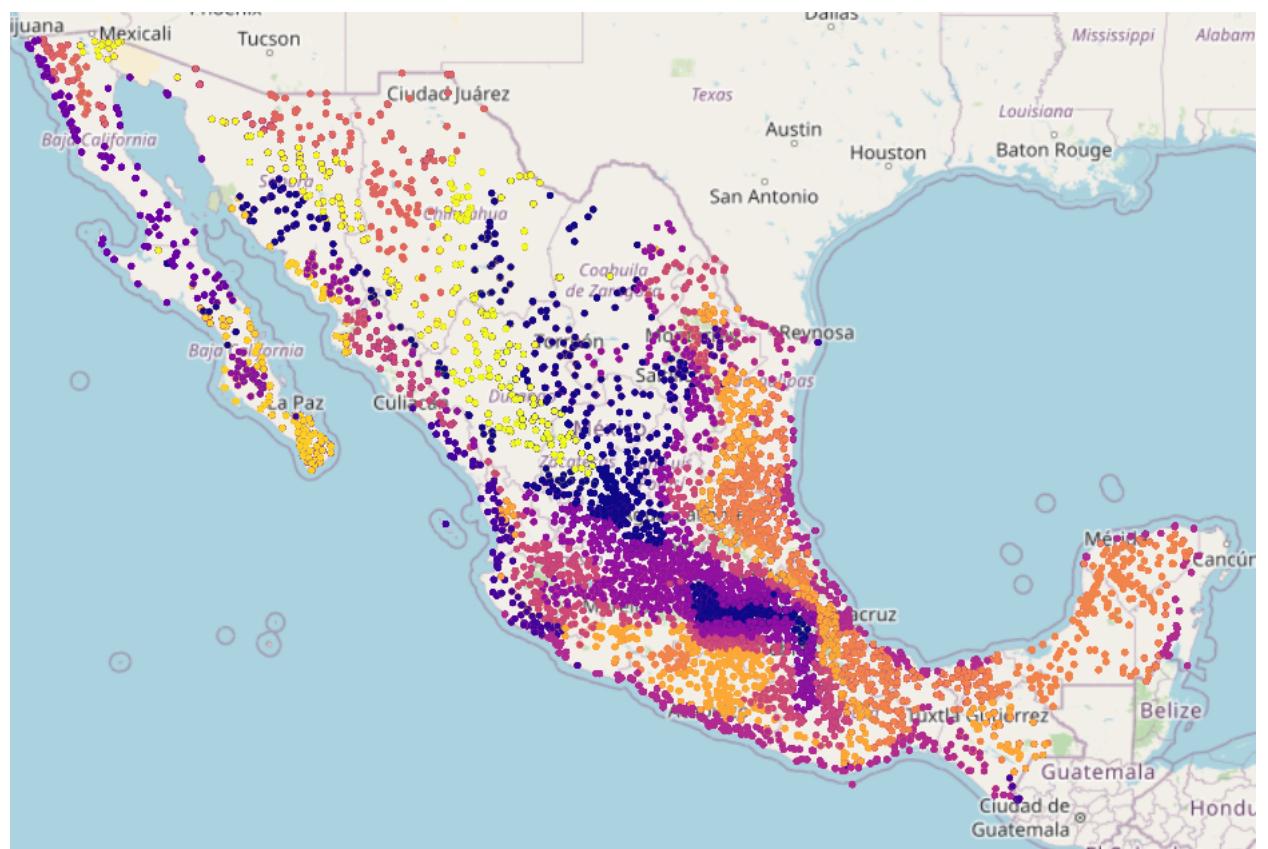
### 6.1.1. GMM

A continuación se muestran los resultados obtenidos dentro del agrupamiento generados en el Gaussian Mixture Model. En esta ocasión se pudo generar un clustering muy semejante al esperado, ya que la mejor optimización que se presentó fue con 11 grupos generados en la Figura 24 y 25.

La cantidad de registros utilizada quedo muy pobre, debido a la densidad de estos datos se pueden ver espacios en donde se espera que haya más grupos, pero a grandes rasgos se puede concluir que estuvo dentro de lo esperado y se puede implementar esta variante dentro del Geoportal. Para tener un resultado más acercado se puede seria seleccionar un temporal más grande y aplicarle este preprocesamiento y agrupamiento.



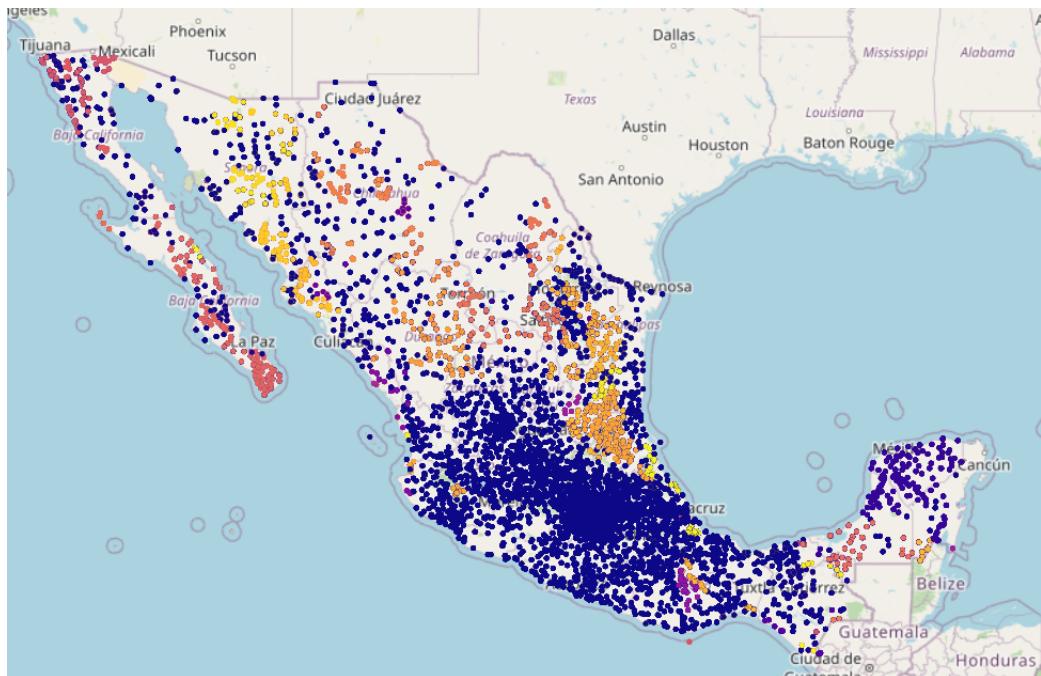
**Figura 24:** GMM mapa de MERRA



**Figura 25:** GMM mapa de EMAS

### 6.1.2. DBSCAN

Se estuvo notando que en este algoritmo no se comporta de una manera satisfactoria debido a la densidad de los datos esta demasiado junta y no se puede encontrar un buen epsilon para generar la mayor cantidad de clusters y así tener un resultado bueno para la selección de topoformas. En la Figura 26 y 27 se nota que la densidad de los datos más semejantes se concentra en la parte central del país porque es donde hay más antenas en ese sector.



**Figura 26:** DBSCAN mapa de MERRA



**Figura 27:** DBSCAN mapa de EMAS

## 6.2. Validación

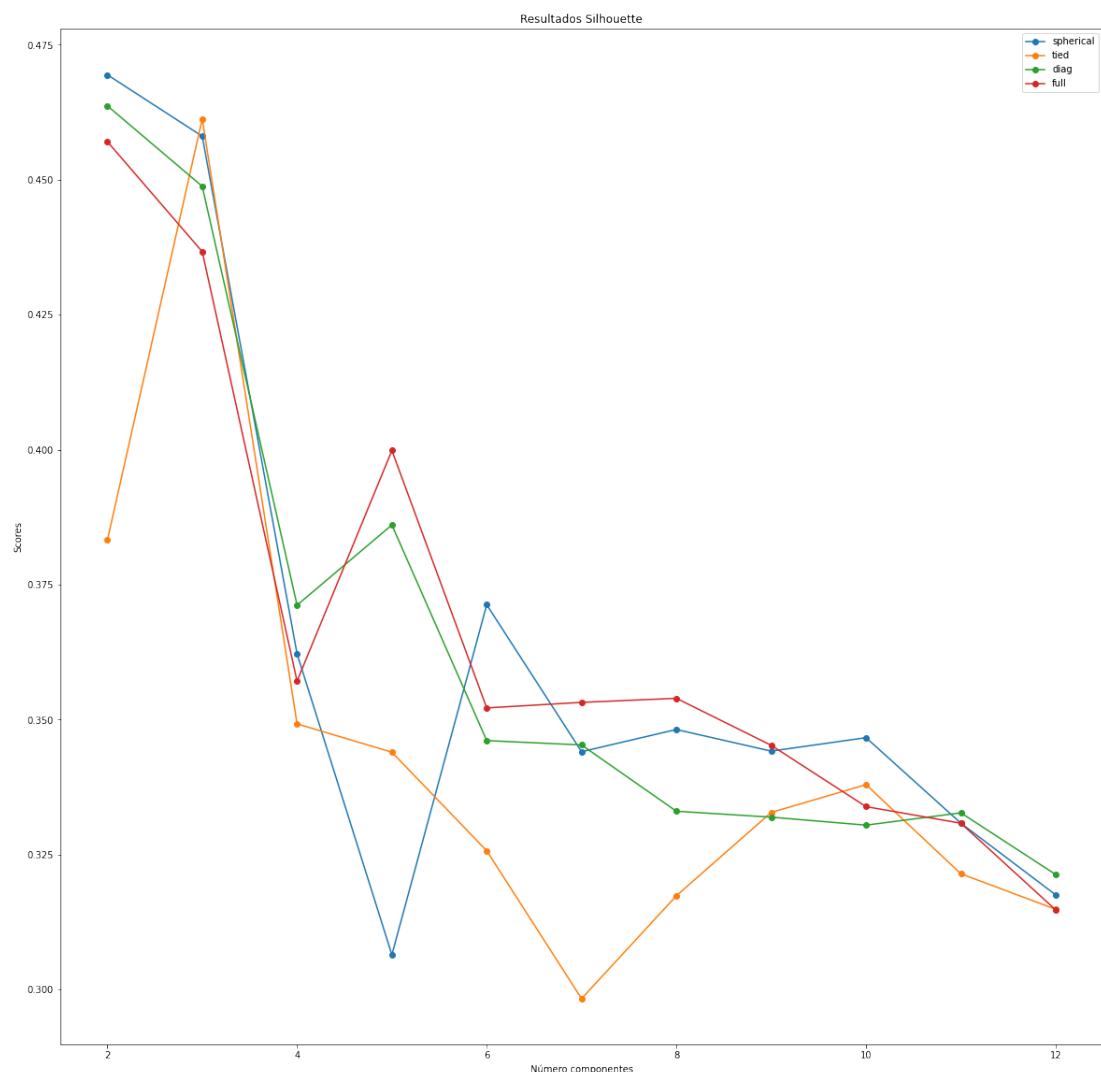
Se aplicaron para cada tipo de clustering un algoritmo de validación para dar efecto de los resultados obtenidos en cada uno de las validaciones previas que se realizaron para el mejor parámetro a utilizar. Los resultados aplicados en cada uno son los siguientes:

- Silhouette
- Calinski–Harabasz
- Davies–Bouldin

### 6.2.1. GMM

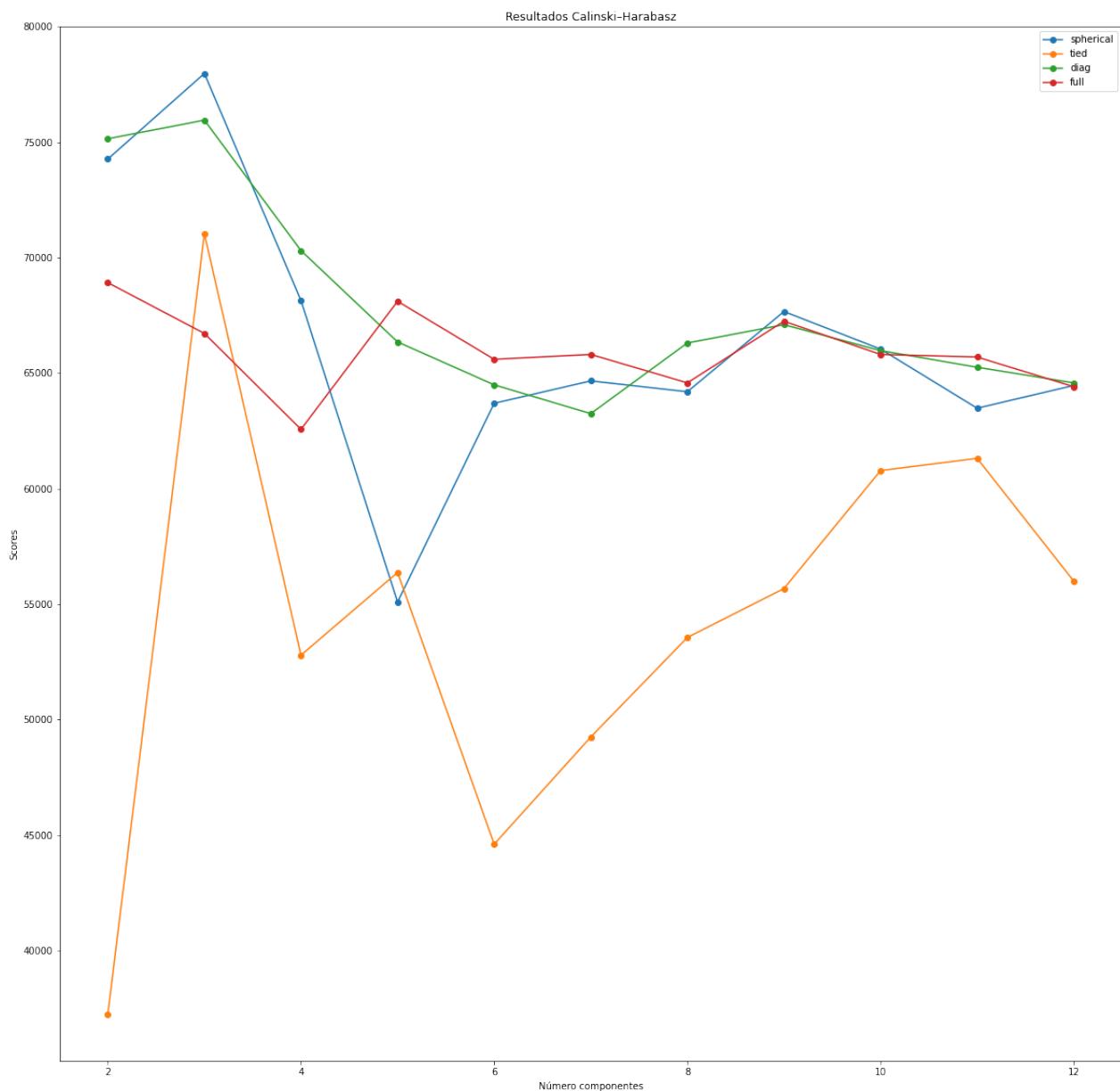
#### MERRA

Para la validación mediante Silhouette los resultados obtenidos fueron esperados para el número de  $k$  contrarresta un poco en los resultados obtenidos mediante BIC, ya que para Silhouette es mejor realizar un algoritmo con  $k = 11$  pero con el tipo de diag. En segunda instancia se corresponde al full.



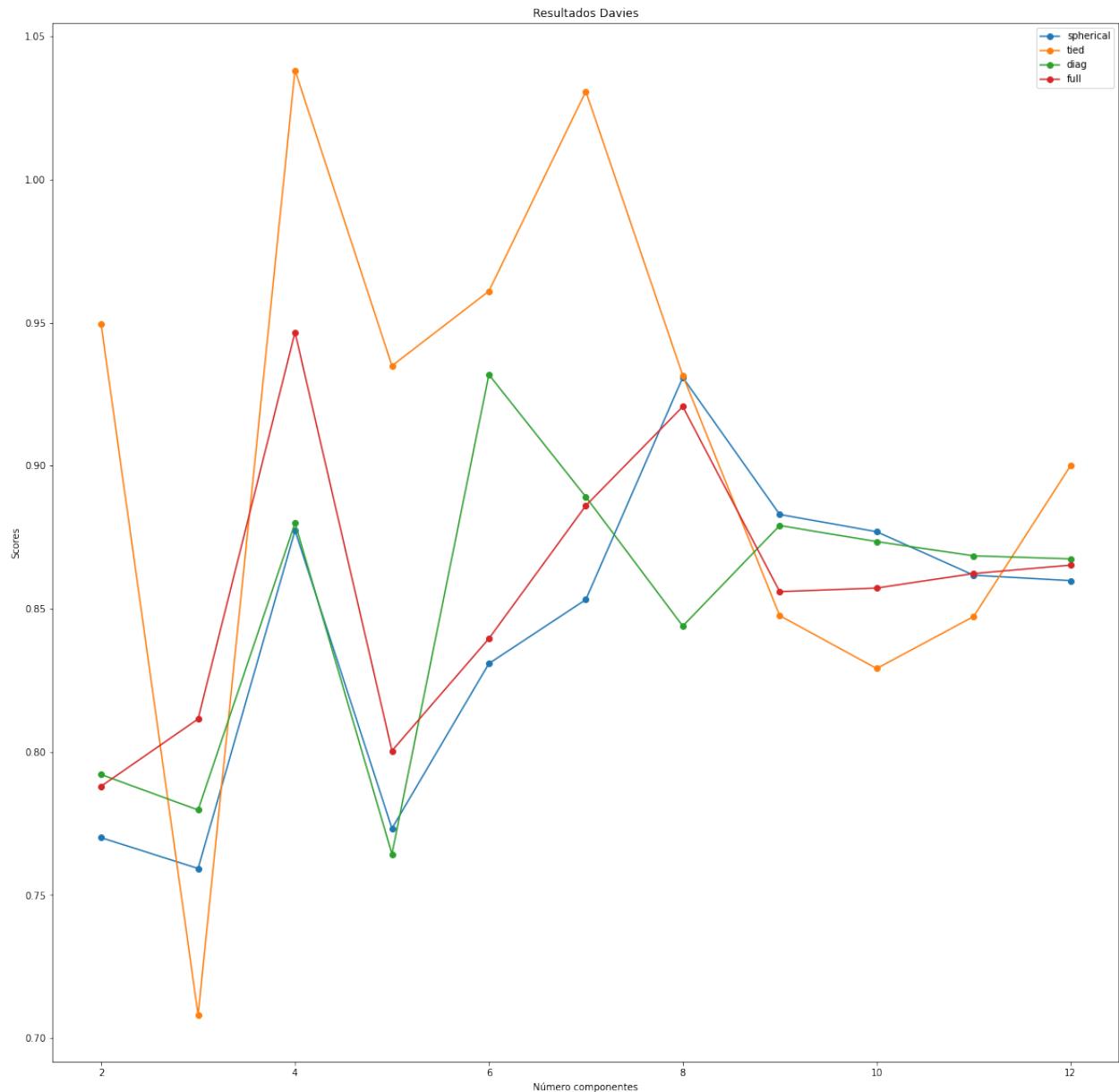
**Figura 28:** Silhouette de MERRA

Para la validación mediante Calinski–Harabasz los resultados obtenidos fueron esperados para el número de  $k$  no contrarresta a los resultados obtenidos mediante BIC, ya que para Calinski–Harabasz es mejor realizar un algoritmo con  $k = 11$  pero con el tipo de full.



**Figura 29:** Calinski–Harabasz de MERRA

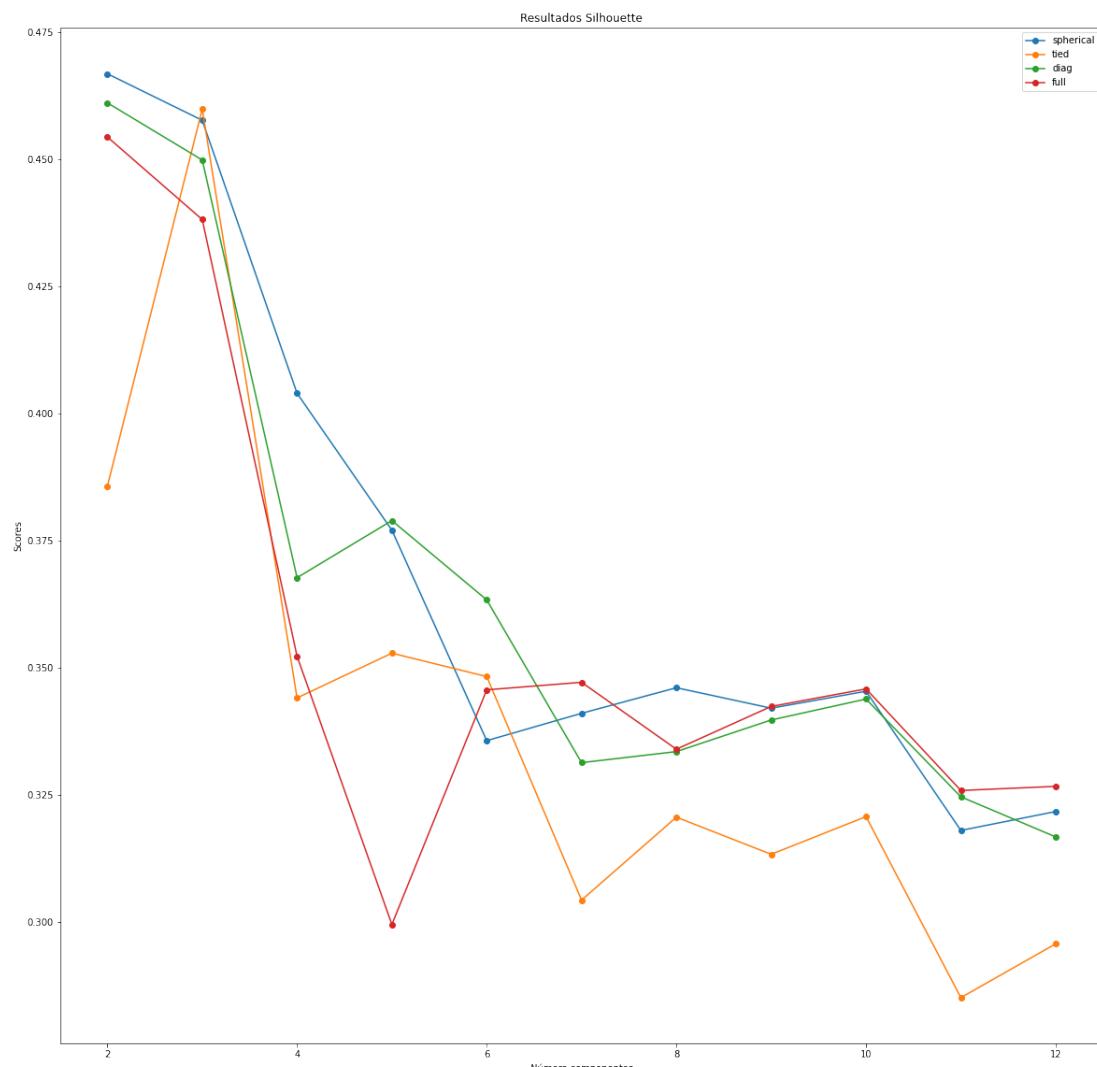
Para la validación mediante Davies–Bouldin los resultados obtenidos fueron esperados para el número de  $k$  contrarresta un poco en los resultados obtenidos mediante BIC, ya que para Davies–Bouldin es mejor realizar un algoritmo con  $k = 11$  pero con el tipo de tied. En segunda instancia se corresponde al full.



**Figura 30:** Davies–Bouldin de MERRA

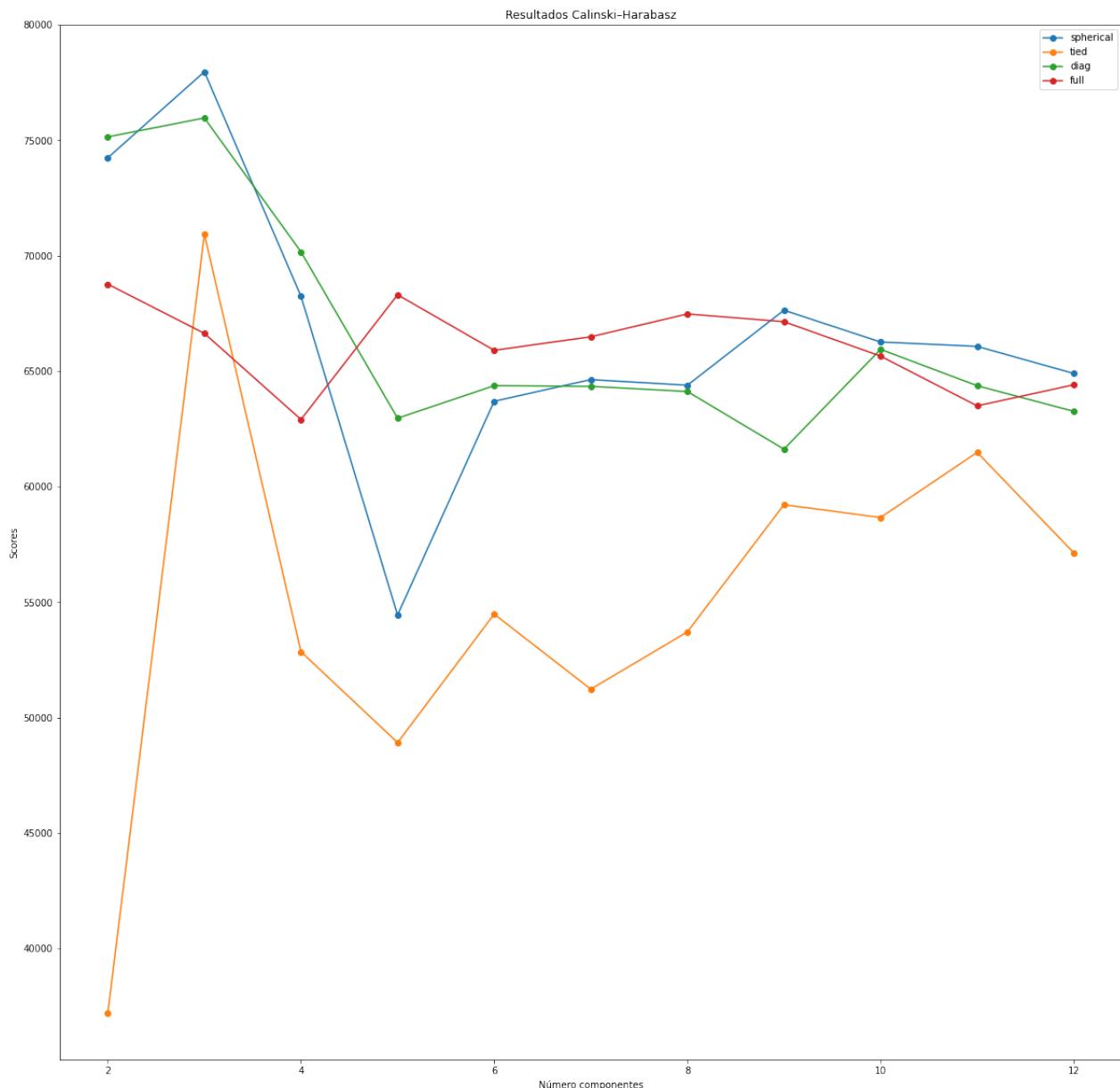
## EMAS

Para la validación mediante Silhouette los resultados obtenidos fueron esperados para el número de  $k$  contrarresta totalmente a los resultados obtenidos mediante BIC, ya que para Silhouette es mejor realizar un algoritmo con  $k = 11$  pero con el tipo de full.



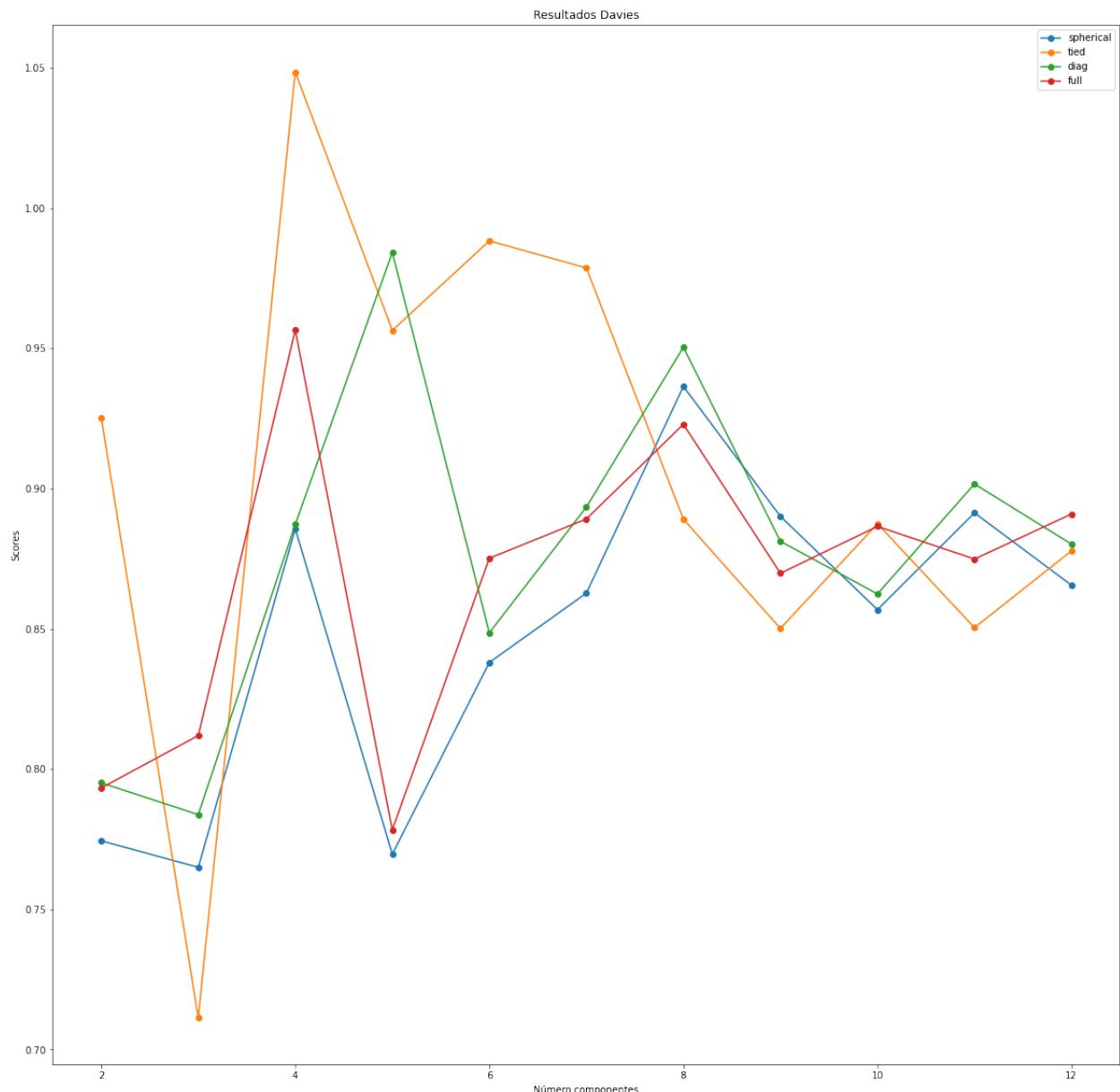
**Figura 31:** Silhouette de EMAS

Para la validación mediante Calinski–Harabasz los resultados obtenidos no fueron esperados para el número de  $k$  obtenidos mediante BIC, ya que para Calinski–Harabasz es mejor realizar un algoritmo con  $k = 11$  pero con el tipo de spherical.



**Figura 32:** Calinski–Harabasz de EMAS

Para la validación mediante Davies–Bouldin los resultados obtenidos fueron esperados para el número de  $k$  contrarresta los resultados obtenidos mediante BIC, ya que para Davies–Bouldin es mejor realizar un algoritmo con  $k = 11$  pero con el tipo de diag.



**Figura 33:** Davies–Bouldin de EMAS

### 6.2.2. DBSCAN

En esta ocasión se presentan los resultados obtenidos para el algoritmo de MERRA y EMAS, pero este clustering resultó muy malo para este tipo de data set ya que para MERRA (ver Figura 34), porque como se ha dicho los valores de Silhouette entre mas cercano a 1 es buen resultado y da un resultado negativo. Y en Davies Buildin da un resultado en donde se pueda tomar como bueno pero esto lo contrarresta en el Calinski–Harabasz ya que este da valores muy por debajo en contra de los obtenidos en el GMM con Calinski–Harabasz.

Para EMAS ocurre lo mismo (ver Figura 35) ya que las dimensiones son muy parecidas dentro del algoritmo DBSCAN con MERRA, pero se contrarrestan con los resultados obtenidos en el GMM.

```
Silhouette = -0.52787296598892
Davies Bouldin = 1.3423473514186597
Calinski Harabasz = 72.70681537300985
```

**Figura 34:** Validaciones de MERRA

```
Silhouette = -0.659923938504237
Davies Bouldin = 1.552694523650105
Calinski Harabasz = 71.84373797153613
```

**Figura 35:** Validaciones de EMAS

## 7. Conclusión

En conclusión para la selección de estos algoritmo me inclinaría con el Gaussian Mixture Model ya que tanto en las pruebas como en la validación obtenida en comparación con los resultados de validación y pruebas de DBSCAN se comporta de mejor manera.

Incluso si colocamos las 12 formas de agrupación se pueden obtener resultados obtenidos de acuerdo a estas investigaciones obtenidas, esto se comprueba en el apartado de validación porque los valores obtenidos en cada uno de los algoritmos utilizados son muy semejantes a los que genera el BIC entre el  $k = 11$  y  $k = 12$ .

Como experiencia personal, como trabajo futuro sería encontrar un método adecuado para el relleno de los valores faltantes mediante un proceso más matemático y dar resultados mas exactos correspondientes para EMAS. Del lado de la elección del tipo de clustering a implementar revisar primeramente los objetivos en los que el clustering es mejor y así sacar deducciones y elegido, ya una vez seleccionado tener algunas funciones para detectar los mejores parámetros y así tener un análisis completo.

Esta área del análisis de datos es muy interesante pero si se necesita tiempo para dar un estudio previo, durante y después para dar resultados consistentes y comprobados.

## Referencias

- [1] E. Morales and H. J. Escalante, “Clustering,” 2017.
- [2] F. S. Caparrini and W. W. Work, “Algoritmos de clustering.”
- [3] “Detección de anomalías con gaussian mixture model (gmm) y python.”
- [4] “clustering y heatmaps: aprendizaje no supervisado.”