Assignment: Clasificación Climática con datos del CMIP6

CLARA ALEJOS

YAGO GARCÍA-ARGÜELLES

SERGIO ORDÁS

ÍNDICE

1. INTRODUCCIÓN

1.1 Objetivo

La energía solar es una de las fuentes renovables más prometedoras actualmente. Sin embargo, la viabilidad de su implementación depende de la correcta selección de la zona geográfica, ya que hay que tener en cuenta muchos factores climáticos que afectan a su adecuado funcionamiento.

Nuestro proyecto tiene como objetivo identificar y clasificar las regiones del mundo que tengan las condiciones climáticas idóneas para instalar paneles solares. Para ello, hemos seleccionado siete modelos climáticos globales distintos, de cinco variables distintas, que recogen datos mensuales desde 1850 a 2014. Tras un preprocesamiento, hemos construido un ensemble multi-modelo y hemos aplicado PCA, con el fin de concentrar la mayor parte de variabilidad climática en un conjunto reducido de componentes. Finalmente haríamos una clasificación en clusters con K-means, agrupando de esta manera regiones con patrones climáticos similares, permitiéndonos así identificar las zonas del planeta con condiciones favorables para el aprovechamiento de la energía solar. Todo este procedimiento se desarrollará con detalle más adelante.

1.2 Las variables

- Radiación solar (rsds): es la variable más directamente relacionada con la energía solar. Representa la cantidad de energía solar que llega a la superficie terrestre.
- Nubosidad (clt): influye de manera directa sobre la radiación solar recibida. Una mayor cobertura de nubosidad reduce la cantidad de energía solar aprovechable.
- Temperatura del aire (tas): afecta en lo que respecta a la eficiencia de los paneles solares. Las altas temperaturas pueden disminuir el rendimiento de las celdas fotovoltaicas.
- Velocidad del viento (sfcWind): el viento puede contribuir a la refrigeración natural de los paneles, ayudando a mantener su eficiencia; pero también valores altos de viento pueden afectar a la estabilidad de las estructuras.
- Precipitación (pr): influye tanto en la radiación incidente como en el mantenimiento de los paneles.

En conjunto, estas cinco variables nos permiten realizar un análisis global de las zonas óptimas para el aprovechamiento de energía solar.

1.3 Los modelos

La idea original era la selección de diez modelos distintos, pero nos enfrentamos a dos problemas:

- Una vez estuvieron escogidos los diez, fuimos variable por variable, pero al ir a la última variable (sfcWind), nos dimos cuenta de que esa variable no estaba estudiada por dos de los modelos elegidos, así que ya recortamos a ocho.
- El segundo problema lo tuvimos al empezar el preprocesado, ya que nos vimos ante la situación de que cuatro modelos tenían un tamaño de rejilla distinto al de los otros. Para tres modelos pudimos arreglar el problema (se verá más adelante), pero para el modelo que quedaba nos daba un problema de diferencia de calendarios y además contaba con una resolución muy baja, así que decidimos descartarlo.

Los siete modelos finales fueron: ACCCESS-CM2, ACCESS-ESM1-5, CESM2-WACCM, CMCC-CM2-SR5, FIO-ESM-2-0, MRI-ESM2-0, TaiESM1

2. METODOLOGÍA

2.1 PREPROCESADO

El objetivo de esta parte es unificar y preparar la información de los distintos modelos para las cinco variables seleccionadas antes de aplicar PCA y Clustering.

2.1.1 Introducción de Datos

Cargamos los siete modelos climáticos seleccionados para nuestras cinco variables de interés.

Para el análisis temporal, quisimos tomar de referencia el periodo comprendido desde el 1 de enero de 1950 hasta el 31 de diciembre de 1980. Para cada modelo y variable, calculamos la media mensual.

2.1.2 Regridding

Para solucionar el problema comentario anteriormente de los distintos tamaños de rejilla, proyectamos todos los modelos a un único tamaño de malla de referencia.

Primero, identificamos que el tamaño mayoritario era de 192 x 288 (latitud x longitud). Después, el regridding lo realizamos utilizando la librería xesmf, y utilizando interpolación bilineal.

2.1.3 Construcción Ensemble

Teniendo ya todos los modelos con el mismo tamaño y mismo formato temporal, hicimos el ensemble multi-modelo. El valor final para cada variable en cada celda de la rejilla y para cada mes lo calculamos como la media simple de los siete modelos. Esto dio lugar a cinco DataArrays, cada uno con 12 meses, 192 latitudes y 288 longitudes.

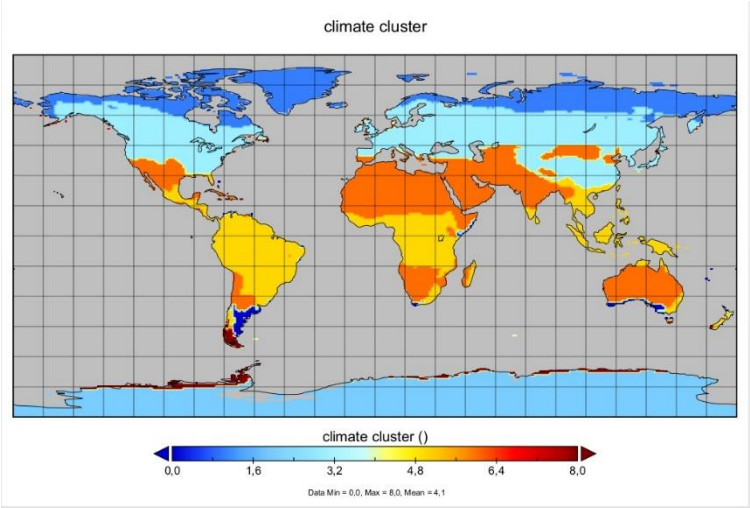
2.1.4 Exportación Final

Generamos dos formatos de salida distintos:

- CSV: Creamos un DataFrame, aplanando los datos para tener una fila por cada combinación de mes, latitud y longitud. Las columnas representan las cinco variables.
- NetCDF: Lo creamos conservando la estructura (mes, lat, lon)

Finalmente, aplicamos un filtro espacial para centrar el análisis en las regiones relevantes:

- Excluyendo regiones oceánicas: Utilizando "cfeature", y una máscara binaria, quitamos las celdas de la rejilla que caían sobre el mar, asegurando así un análisis sobre la superficie terrestre únicamente. Esto es debido a que nos parecía irrelevante tener en cuenta estas regiones, ya que siendo nuestro objetivo entender que zonas son más idóneas para la instalación de paneles solares, no tenía sentido.
- Recortando latitudes extremas: Para evitar el sesgo de datos debido a la baja resolución de la rejilla en los polos, y porque pensamos que dichas zonas no suelen ser características del aprovechamiento solar, excluimos las regiones al sur de -60º de latitud y al norte de 45º de latitud.



En este primer análisis de clustering que hicimos al principio, podemos observar como toda la parte del norte y sur, no nos dio información relevante, haciendo además que el resto del mapa sea más general.

2.2 PCA Y CLUSTERING

Teniendo ya el ensemble multi-modelo y el dataset final listo, procedimos a hacer la reducción de dimensionalidad y la clasificación en regiones climáticas.

2.2.1 Estructuración y Estandarización

Antes de aplicar PCA, que es sensible a las diferentes escalas de las variables, la matriz de datos fue normalizada utilizando MinMaxScaler. De esta manera, tenemos los datos de cada variable siguiendo un rango entre 0 y 1, asegurándonos así de que ninguna variable domine el cálculo de la varianza con su magnitud.

2.2.2 Aplicación PCA

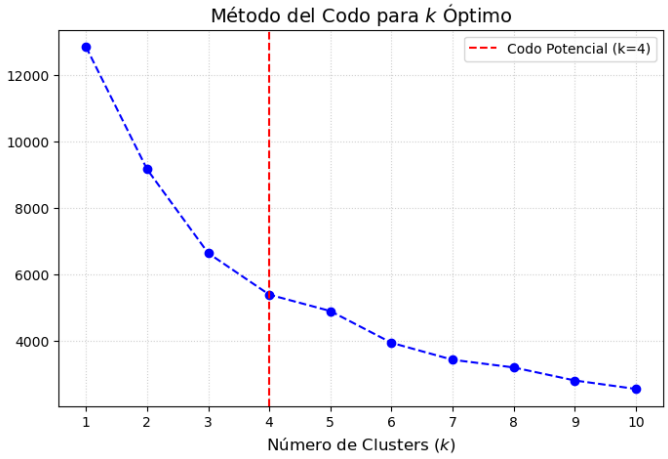
Objetivo: Concentrar la mayor parte de la variabilidad en un número reducido de componentes no correlacionadas.

Para determinar el número de componentes, utilizamos el criterio de varianza acumulada, seleccionando la cantidad mínima de componentes principales necesarias para explicar al menos el 90% de la varianza total.

Dieron lugar a cinco componentes principales, por lo que la matriz de entrada se redujo de 60 dimensiones (12 x 5) a 5 dimensiones.

2.2.3 Clasificación con K-means

Aplicamos K-means sobre el conjunto de datos. Para el número de clusters utilizamos el método del codo.



Observando, el descenso significativo se hace mucho menos pronunciado a partir de k=4, lo que hizo que escogiéramos 4 clusters en un principio para nuestro análisis.

2.2.4 Construcción Mapa

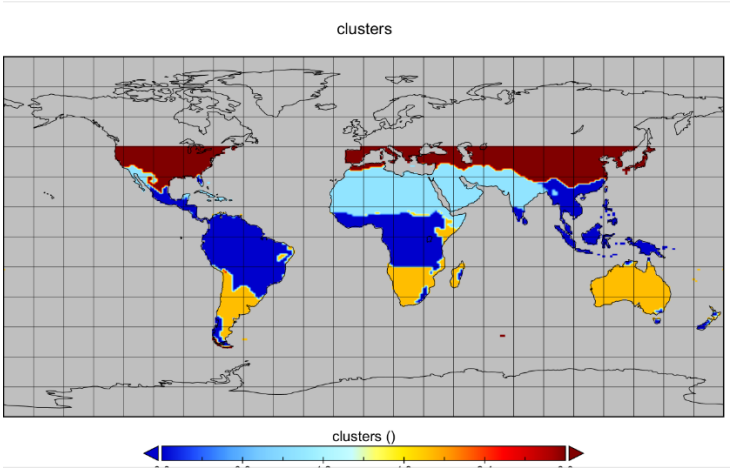
Las etiquetas de los clusters de mapearon de nuevo en la rejilla original.

Creamos un DataArray que reproduce el grid espacial, asignando la etiqueta del cluster correspondiente a cada celda de tierra.

3. RESULTADOS

3.1 JUSTIFICACIÓN NÚMERO CLUSTERS

Inicialmente hicimos caso al Método del Codo para determinar el número de clusters, y este fue el resultado.

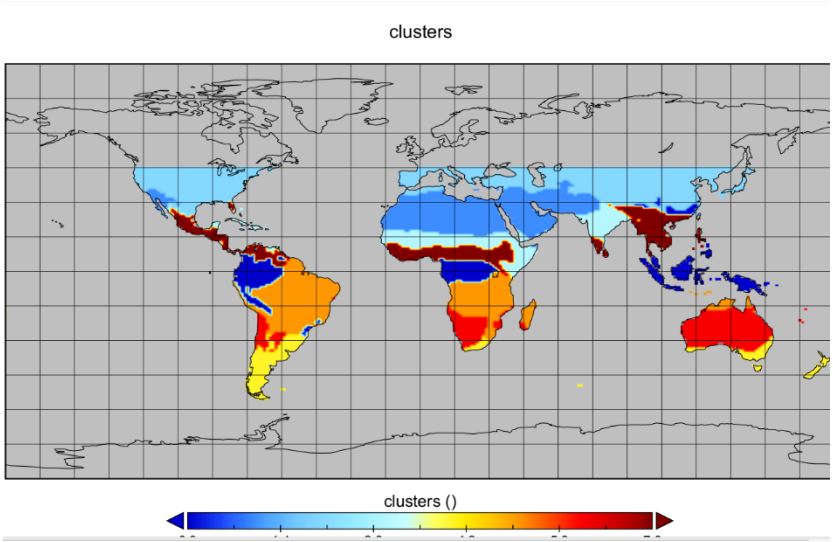


Al analizarlo, aunque se puede observar cual sería la zona más idónea para la instalación de las placas, pensamos que en algunas regiones estaba generalizando. Por ejemplo, la zona de oriente medio se encuentra en el mismo grupo que el sur de Asia, cuando en el sur de Asia predominan fuertes lluvias de un clima monzónico; y en el medio oriente no.

Ante esto, decidimos probar con diferentes valores de k por nuestra cuenta, y finalmente seleccionamos 8 clusters.

3.2 CONCLUSIONES

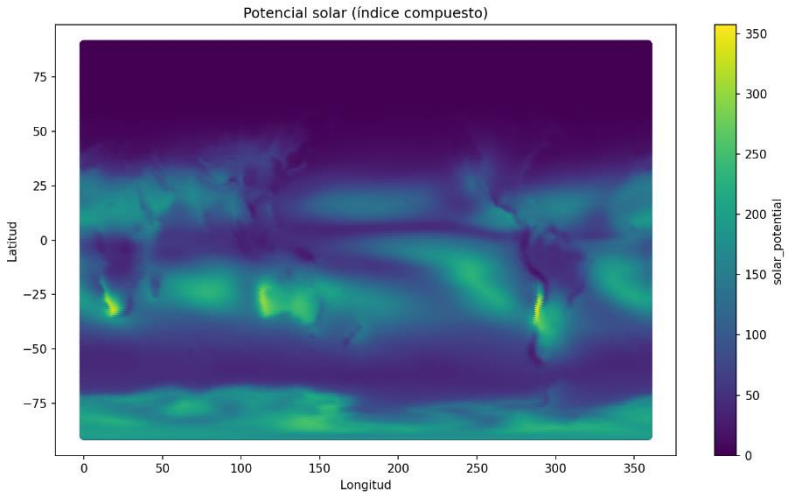
3.2.1 Conclusión general



Con 8 clusters, podemos observar como el “Cluster Óptimo” podríamos ubicarlo perfectamente, por el Sahara, la Península Arábiga, seguido también del cluster que comprende el Atacama y desierto de Australia central. Esto explica que las zonas más idóneas para colocar paneles solares son regiones caracterizadas por recibir gran cantidad de energía solar, tener poca nubosidad, y pocas precipitaciones. Sí que es verdad que en estas regiones las temperaturas suelen ser muy elevadas, haciendo que los paneles se calienten y puedan rendir menos, pero aún así el beneficio del gran aprovechamiento solar compensa esa pequeña pérdida de rendimiento.

3.2.2 Refuerzo conclusión

Para fortalecer nuestra decisión de interpretación de clusters, hicimos un PCA y K-means extra, pero solo teniendo en cuenta la nubosidad y radiación solar, para valorar con un modelo aparte que zonas son mejores para paneles solares.



En la imagen podemos observar como las zonas más claras representarían las zonas más idóneas, siendo estas el Atacama, y Australia, principalmente. Lo que refuerza nuestra conclusión anterior.