Hotspot Mapping with Clustering: Identifying Climate Risk Zones in Europe

Andrea Baschiera
Ca' Foscari University of Venice
Venice, Italy
878612@stud.unive.it

Contents

NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research.

1 Introduction

Climate change is looming on the 21st century, impacting ecosystems, economies, and societies worldwide. Among the most critical climate-related hazards, heatwaves, extreme precipitation events, droughts, and strong winds have significant socio-economic consequences, particularly in Europe, where the frequency and intensity of these phenomena are increasing [3].

To mitigate and adapt to these risks, it is crucial to identify the most exposed regions, or "hotspots," that are susceptible to multiple climate hazards. In this study, I employ an unsupervised machine learning technique (**K-means clustering**) to classify European regions based on their exposure to five key climate risk drivers:

- **Heatwaves**
- **Extreme precipitations**
- **Dryness**
- **Droughts**
- **Extreme winds**

retrieving the data from *Copernicus Climate Data Store* (CDS), specifically from the database "Climate indicators for Europe from 1940 to 2100 derived from reanalysis and climate projections" (link: `https://cds.climate.copernicus.eu/datasets/sis-ecde-climate-indicators?tab=overview`)[1].

The goal of this study is to create a **spatial classification of future climate hazard hotspots**, which could support policymakers and researchers in long term risk assessment and adaptation strategies (the long term is referring to the period 2071 - 2100).

2 Related Work

Several studies have applied clustering techniques to climate-related datasets to classify regions based on environmental conditions. Unsupervised learning methods, such as K-Means and hierarchical clustering, have been used for climate zone classification [4], extreme event pattern recognition [5], and drought risk assessment [7].

For instance, [6] analyzed precipitation patterns across Europe using K-Means clustering, identifying distinct rainfall regimes. Similarly, [2] applied unsupervised techniques to heatwave duration and frequency data, revealing increasing clustering trends in Southern Europe due to climate change.

In this work, I build on previous methodologies by integrating multiple climate hazards into a **multivariate clustering framework**. Unlike studies that focus on individual risk drivers, my approach provides a more holistic perspective by identifying regions exposed to compound climate hazards.

3 Data Collection & Preprocessing

3.1 Climate Data Sources

The period from 1950 to 2100 is covered using historical data from reanalyses and projections using the *MPI-ESM-LR* (MPI, Germany) climate model. While extracting the dataset, five climate hazard indicators were selected (hereby presented):

- **Heatwave days**: n° of days per year exceeding the 99th percentile of daily maximum temperatures (May–September);
- **Extreme precipitation frequency**: n° of days per year with precipitation above the 95th percentile (1981-2010);
- **Consecutive dry days**: Longest consecutive period (in days per year) with daily precipitation below 1 mm;
- **Meteorological drought duration**: n° of months per year with an SPI-3 index indicating severe drought;

- **Extreme wind speed days**: n° of days per year with 10m wind speed above the 98th percentile.

Each dataset is stored in *NetCDF* format and provides values over a spatial grid covering the whole European region. However, differences in temporal resolution (daily vs. monthly vs. nanoseconds) and data encoding required specific preprocessing steps.

3.2 Data Preprocessing

3.2.1 Loading and Inspecting NetCDF Files

Each dataset was loaded using the `xarray` library in Python, which allows efficient handling of multi-dimensional climate data. Below is an example of how the heatwave dataset was loaded:

```
import xarray as xr
from google.colab import drive

# Mount Google Drive (if not already mounted)
drive.mount('/content/drive', force_remount=True)

# Load heatwave dataset
path_hw = "/content/drive/MyDrive/.../09_heat_waves_climatological.nc"
hw = xr.open_dataset(path_hw, engine="h5netcdf")
hw  # Display dataset information
```

This procedure was repeated for the other climate variables aforementioned.

3.2.2 Handling Temporal Inconsistencies

I standardized all data to an annual timescale by computing the mean values over the period 2071-2100. Since some data were computed in days (per year) while others were in nano-seconds, I standardized all data to days per year. Additionally, missing values were replaced with zeros, ensuring that gaps in the dataset did not affect the clustering results. Below is the preprocessing applied to the Extreme precipitation frequency data:

```
precipdata = precip["data"]
a = precipdata.values.astype('timedelta64[D]')  # Convert to days
a = np.where(np.isnan(a), 0, a) #Setting missing data to 0.
precip_int = a.astype(int)#Changing the format from timedelta64[D] to int[64]

# Create an xarray DataArray with explicit dimension names and coordinates
new_precip_int = xr.DataArray(
    precip_int,
    dims=['time', 'lat', 'lon'],
    coords={'time': precip['time'], 'lat': precip['lat'], 'lon': precip['lon']},
    name="data_int")

# Add the new variable to the dataset
precip["data_int"] = new_precip_int
```

A similar process was applied to the Extreme wind speed days data, which was also stored in nano-seconds. The other variables, fortunately, were already in days and in the *int64* format.

3.2.3 Standardization of Variables

To ensure fair weighting among climate variables, I applied Z-score normalization, which transforms each variable to have zero mean and unit variance:

```
# Convert to NumPy array
stacked=mergeddata.to_array(dim="variables").stack(spatial=("lat", "lon"))
```

```
X=stacked.values.T
# Normalize data
X = (X - np.mean(X, axis=0)) / np.std(X, axis=0)
```

3.3 Visualization of Preprocessed Data

To verify the correctness of preprocessing, I generated visualizations for each climate hazard variable. Here below is shown an example: the plot for the average heatwave days per year averaged over 2070-2100.

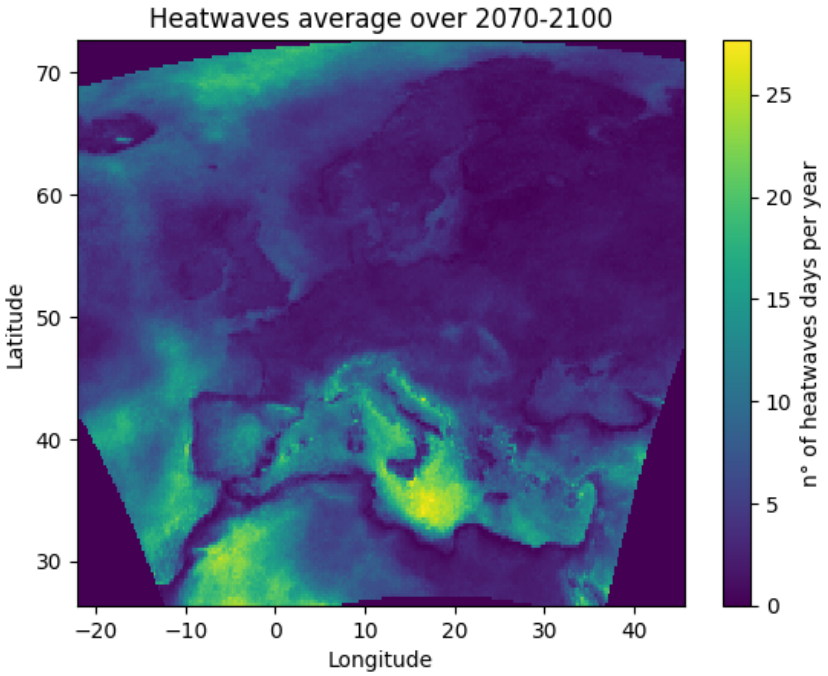

Figure 1: Average heatwave days per year (2070-2100)

As can be seen from the plot, the dark blue area roughly located at the four corners is comprised by gridded points all set at zero. Those were the missing data I retrieved in the original datasets. The extent to which all those points weighted on normalisation was not deemed enough to skew the conclusions later reached with the clustering excercise.

4 Methodology

4.1 Clustering with K-Means

The **K-Means clustering algorithm** partitions data into k distinct groups by minimizing the sum of squared distances between data points and their respective cluster centroids. The objective function for K-Means is defined as:

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - \mu_i||^2 \tag{1}$$

To determine the optimal number of clusters, I based my choice on the **Elbow Method**, which examines the reduction in inertia (sum of squared errors, SSE) as k increases. Initially, the method suggested an optimal number of $k = 3$, as the inertia (SSE) exhibited a significant drop before stabilizing (shown in the Figure below).
However, the classification with **3 clusters** was too coarse and not representative of the diverse climate risk factors across Europe. The regions were grouped into broad areas that did not effectively differentiate their exposure to various climate hazards. Therefore, after careful consideration it was opted for *k=6*. This choice was not only justified by the need for a more granular inspection of the hotspots, but also by the fact that with k=3 there was still much inertia left unexplained.

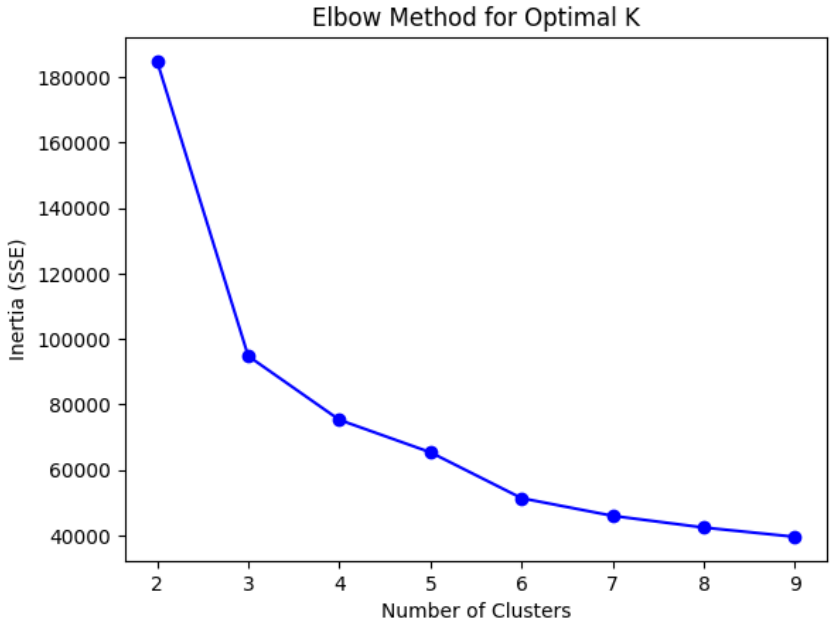

Figure 2: Elbow method applied to determine optimal k

To assess the quality of the clustering, I also evaluated the **Silhouette Score** for different values of k. The highest score was found for $k = 3$ (0.582). Although the Silhouette Score slightly decreased for $k = 6$ (0.397), the clustering outcome was still more informative and interpretable.

Furthermore, a classification of the clusters was needed. Naming the clusters required the analysis of the **normalized mean values** of each variable for each clusters. The following table shows these values:

After cross-checking the values of the table with the visual results of the clustering excercise, I proceeded to assign names based on the extent to which the cluster in question manifested elevated values for one or more of the hazards. The results of this classification are presented and discussed in the next section.

labels	heatwaves	precipitation	wind	dry	drought
0.0	2.236554	-1.047246	0.207850	1.269914	1.077537
1.0	-0.610848	0.298248	0.475789	-0.429221	-0.399490
2.0	-1.043088	-1.553179	-2.949304	-0.730567	-1.373968
3.0	-0.024930	-1.404882	0.217352	2.664907	1.827610
4.0	0.264184	1.315289	-0.084715	-0.506595	-0.494304
5.0	0.602273	-0.386398	0.240735	0.149321	0.873206

Figure 3: Averages per clusters

4.2 Limitations of the approach

The clustering results were computed used only the K-means algorithm. From a machine learning perspective, the analysis should have included comparative results using other clustering techniques as well, like, for instance, hierarchical clustering. Nevertheless, the results obtained with K-means were judged satisfactorily.

5 Experimental Results and Discussion

The clustering analysis successfully identified **five macro-climatic risk zones** across Europe, each characterized by distinct exposure to different climate hazards. Moreover, qualitatively interpreting the mean values of the hazards enabled me to assess the relative magnitude of the danger of each cluster. Hereby are presented the results (the cluster purposely set in black representing the missing values):

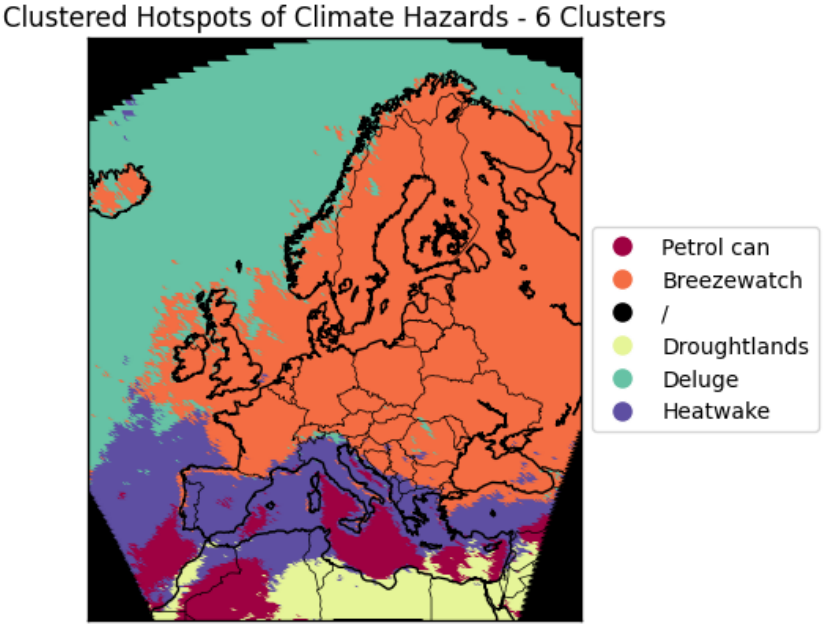

Figure 4: Clustering results of macro-climatic risk zones

The distinctive feature of each of the five macro-climatic zones is encapsulated by their assigned names:

- **Petrol can**: spanning most of the Mediterranean sea and parts of Algeria and Morocco, this area will be characterized by several heatwaves in its climatic future. A warming Mediterranean has the potential to trigger countless climatic consequences, as the recent events in Valencia testified (the name *petrol can* derives precisely from that catastrophe). Therefore, we should regard this area as high-danger, especially for what it could cause to Southern Mediterranean countries;
- **Breezewatch**: covering most of Scandinavia, the UK and most of continental Europe, this region will experience relatively moderate temperatures but will be characterized by slightly increased extreme wind events. The name *Breezewatch* suggests a climate dominated by these strong winds, but overall less endangered from hazards in comparison of the others;
- **Droughtlands**: found mainly across the northern African countries (i.e. Egypts, Lybia, part of Tunisia and Algeria), this zone will be primarily exposed to dry weather, prolonged drought periods and low precipitation levels. The name *Droughtlands* implies regions that are highly susceptible to water scarcity and desertification, thus warranting their inclusion in the high-danger zones.
- **Deluge**: this oceanic area, looming over parts of France, the UK, and Ireland, will experience high precipitation levels and extreme rainfall events. *Deluge* suggests the frequent heavy rainfall and risk of flooding, not to mention coastal erosion risks. A medium-level danger classification is therefore apt;
- **Heatwake**: covering high density geographical areas (i.e. Spain, South of France, Italy, Greece, Turkey), the risks associated with this zone should ring a bell to policymakers.

Notwithstanding the somewhat mild values that can be seen from the table above, these areas are at risk of a mix of heatwaves and high temperature anomalies, often accompanied by dry conditions in countries where populations, economics and biodiversities are the most vulnerable. The term *Heatwake* reflects the fact that countries should realise as soon as possible the perils of the combination of risks just mentioned. Based on the quantitative results, these areas should warrant a medium-level danger, but my interpretation given the results examined of the other clusters hints to a rather medium-to-high danger risk.

6 Conclusions

This study successfully applied K-Means clustering to identify five macro-climatic risk zones across Europe, each characterized by distinct exposure levels to various climate hazards. By analyzing the mean values of key climate indicators for each zone, I provided an assessment of the relative magnitude of risk for each region. The final classification highlights critical geographical vulnerabilities, emphasizing regions that may require urgent adaptation strategies.

The clustering process proved effective in distinguishing high-risk areas such as the *Petrol Can* and *Droughtlands*, which will likely exhibit extreme heat and drought conditions, respectively. Meanwhile, *Breezewatch* emerged as a relatively less hazardous zone, experiencing only moderate wind variations. The *Deluge* and *Heatwake* clusters underline the importance of considering the interplay between precipitation extremes and temperature anomalies in densely populated regions, especially as far as the latter is concerned.

While the clustering approach provided meaningful insights, several limitations should be acknowledged:

- **Single Climate Model Dependency:** The analysis was based on projections from a single climate model (Germany). Incorporating multiple climate models could enhance the robustness of the results by reducing potential model biases.
- **Lack of Historical Data Comparison:** The study focused solely on future climate projections. A comparison with historical data (e.g., 1970-2000 averages) could help assess whether climate change is altering and will significantly reshape regional risk distributions.
- **Resolution and Local Factors:** The clustering methodology captures large-scale climate trends but does not account for localized variations (e.g., urban heat islands, microclimates, and socio-economic factors), which could refine risk assessments.

Future research should aim to address these limitations by integrating multi-model ensembles, incorporating historical data comparisons, and refining clustering techniques with additional environmental and socio-economic variables. Such advancements would improve the accuracy of climate risk assessments and provide more comprehensive support for policy decisions.

Ultimately, the study demonstrates the utility of machine learning techniques in climate risk assessment, providing a data-driven foundation for proactive adaptation measures. However, further refinements and interdisciplinary approaches are needed to fully capture the complexity of climate hazards in Europe.

References

[1] Copernicus Climate Change Service (C3S). *Climate indicators for Europe from 1940 to 2100 derived from reanalysis and climate projections*. URL: https://cds.climate.copernicus.eu/datasets/sis-ecde-climate-indicators?tab=download. (accessed: 02.02.2025).

[2] Ryan E Adams et al. "The relationship between atmospheric circulation patterns and extreme temperature events in North America". In: *International Journal of Climatology* 1 (2021), pp. 92–103.

[3] Intergovernmental Panel on Climate Change (IPCC). "Summary for Policymakers. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the IPCC". In: *Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA* (2021). URL: https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_SPM.pdf.

[4] Geo McLachlan and David Peel. "Mixture models and neural networks for clustering". In: *URL http://en. scienti ccommons. org/43159010* (2007).

[5] Nikola Milojevic-Dupont and Felix Creutzig. "Machine learning for geographically differentiated climate change mitigation in urban areas". In: *Sustainable Cities and Society* 64 (2021), p. 102526. ISSN: 2210-6707. DOI: https://doi.org/10.1016/j.scs.2020.102526. URL: https://www.sciencedirect.com/science/article/pii/S2210670720307423.

[6] Verónica Torralba et al. "Seasonal Climate Prediction: A New Source of Information for the Management of Wind Energy Resources". In: *Journal of Applied Meteorology and Climatology* 56 (Feb. 2017). DOI: 10.1175/JAMC-D-16-0204.1.

[7] Huiqian Yu et al. "Hotspots, co-occurrence, and shifts of compound and cascading extreme climate events in Eurasian drylands". In: *Environment International* 169 (2022), p. 107509.