Statistical analysis of weather in London

Adama Koita
J12526099014
Télécom Paris, IP Paris
Paris, France
adama.koita@telecom-paris.fr

Arnaud Brisset
J12526099007
ENSTA, IP Paris
Paris, France
arnaud.brisset@ensta.fr

Théo Couerbe
J12526099011
Ecole des Mines
Saint-Etienne, France
theo.couerbe@etu.emse.fr

February 9, 2026

Contents

1 Introduction

1.1 Background and description of the problem

We used a dataset provided by the European Climate Assessment (ECA) (source), which consisted in measurements recorded by a weather station near Heathrow airport in London, UK. For each of the 10 weather features, there is one value per day from 01/01/1979 to 31/12/2020, totaling 15341 samples.

To be able to use the statistical tools we learned in class, we considered each weather feature as an independent random variable and that all the samples are the independent realizations of the random variables, ie. each daily value is independent of any other daily value.

The weather features are detailed in the following table:

Name	**Description**	**Unit**
cloud_cover	cloud cover measurement	oktas
sunshine	sunshine measurement	hours
global_radiation	irradiance measurement	W/m^2
max_temp	maximum temperature recorded	°C
mean_temp	mean temperature	°C
min_temp	minimum temperature recorded	°C
precipitation	precipitation measurement	mm
pressure	pressure measurement	Pa
snow_depth	snow depth measurement	cm

Table 1: Weather features

Note: cloud cover is measured in oktas, which ranges from 0 (no clouds) to 8 (completely overcast), where 1 okta corresponds to an eighth of the sky being obscured by clouds.

1.2 Motivation

This statistical analysis may help to answer the following questions:

- How precisely can we estimate different statistics of the daily mean temperature from 1979 to 2020?
- Does the London daily meteorological data present particular temporal structures and variance due to seasonality or long-term phenomena?
- Can we find links between temperature and pressure or between seasons and day weather types?

2 Methodology

The statistical analysis follows a structured workflow beginning with data preprocessing, where missing values are removed and the ‘date‘ variable is converted to extract temporal features (year, month, day). A categorical ‘season‘ variable is defined (Winter: Dec-Feb, Spring: Mar-May, Summer: Jun-Aug, Autumn: Sep-Nov), and for long-term analysis, the dataset is split into three equal chronological periods.

Given the heterogeneous units of the weather variables (e.g., °C, mm, hours), all quantitative features are standardized (centered to mean 0 and scaled to variance 1) to ensure comparable contributions to the variance. The analysis then proceeds with Principal Component Analysis (PCA) to explore global data structure, Correspondence Analysis (CA) for categorical relationships, and Discriminant Factorial Analysis (DFA) to investigate the separability of seasons and time periods.

3 Descriptive Analysis

We conducted a brief descriptive analysis of the daily mean temperature.

Here are visualizations of the distribution of this variable:

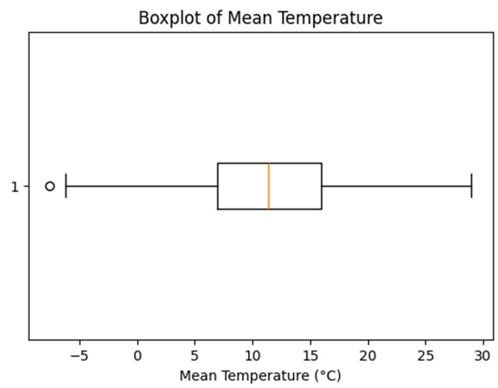

Figure 1: Boxplot of the daily mean temperature.

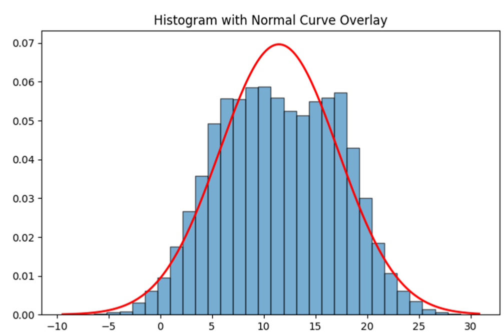

Figure 2: Histogram of the daily mean temperature with normal curve overlay.

Point and interval estimation were conducted for the mean, median and variance of the daily mean temperature.

Here are the results for point estimation:

Statistic	**Value estimated (in °C)**
Mean	11.47
Median	11.4
Variance	32.8
Standard deviation	5.7

Table 2: Point estimations

The confidence interval for the estimated mean was obtained assuming the estimator of the mean is asymptotically a normal distribution or a T distribution. Both methods resulted in the same confidence interval equal to [11.38,11.57] (95% confidence level), which is expected since the T distribution converges to a normal distribution (here, $n = 15341 \geq 30$). Using the bootstrap method, the confidence interval was equal to [11.39,11.57].

The confidence interval for the estimated median was obtained assuming the estimator of the median is asymptotically normal and is equal to [11.1,11.7] (95% confidence level). Using the bootstrap method, the confidence interval differed slightly and is equal to [11.2,11.6].

The confidence interval for the estimated variance was obtained using the bootstrap method and is equal to [32.2,33.4]. Below is the bootstrap distribution of the sample variance:

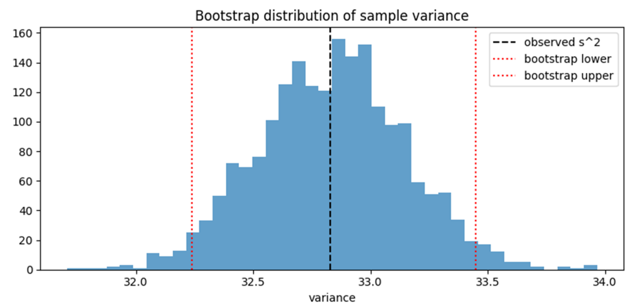

Figure 3: Bootstrap distribution of the sample variance

Here are some additional statistics to better describe the daily mean temperature distribution:

IQR	9
Skewness	-0.013
Kurtosis	-0.656

Table 3: Other statistics of the distribution

4 Principal Component Analysis

4.1 Description

Principal Component Analysis (PCA) is an unsupervised multivariate statistical method used to reduce the dimensionality of a dataset while preserving as much variance as possible. It transforms the original correlated variables into a new set of orthogonal variables, called principal components, which are linear combinations of the original features.

In this study, PCA was applied to the standardized meteorological variables in order to:

- identify the dominant sources of variability in the data,
- explore potential temporal structures such as seasonality,
- reduce redundancy between correlated variables,
- facilitate visualization and interpretation of high-dimensional data.

Each principal component is associated with a decreasing amount of explained variance. The first component captures the largest possible variance, while each subsequent component captures the maximum remaining variance under orthogonality constraints.

4.2 Results and Analysis

Cumulative contribution and dimensionality reduction To quantify how many principal components are needed to represent the dataset, we report the cumulative explained variance (cumulative contribution) in Table 4. The rapid increase of the cumulative contribution for the first components indicates strong correlations between variables (temperature, radiation, sunshine, etc.), meaning that the dataset effectively lies in a lower-dimensional manifold.

In practice, this cumulative table justifies focusing on the first two to three components for visualization and interpretation: they already capture the dominant structures, while higher-order components mainly represent finer-scale variability and rare events.

```
array([0.42130005, 0.63873372, 0.75780648, 0.86301276, 0.9356567 ,
       0.97409381, 0.99087009, 0.9982293 , 1.        ])
```

Figure 4: Cumulative contribution.

Interpretation of the first principal component The projection of the observations onto the first principal component reveals a clear ordering of the data consistent with the annual seasonal cycle. When the points are colored by season, winter and summer appear at opposite ends of the axis, while spring and autumn occupy intermediate positions.

This structure naturally emerges from the covariance between temperature-related variables (mean, minimum, and maximum temperature), solar radiation, and sunshine duration. Since these variables increase and decrease together throughout the year, PCA aggregates them into a single dominant direction. As a result, the first principal component can be interpreted as an energy-related axis reflecting the global seasonal variability of the climate in London.

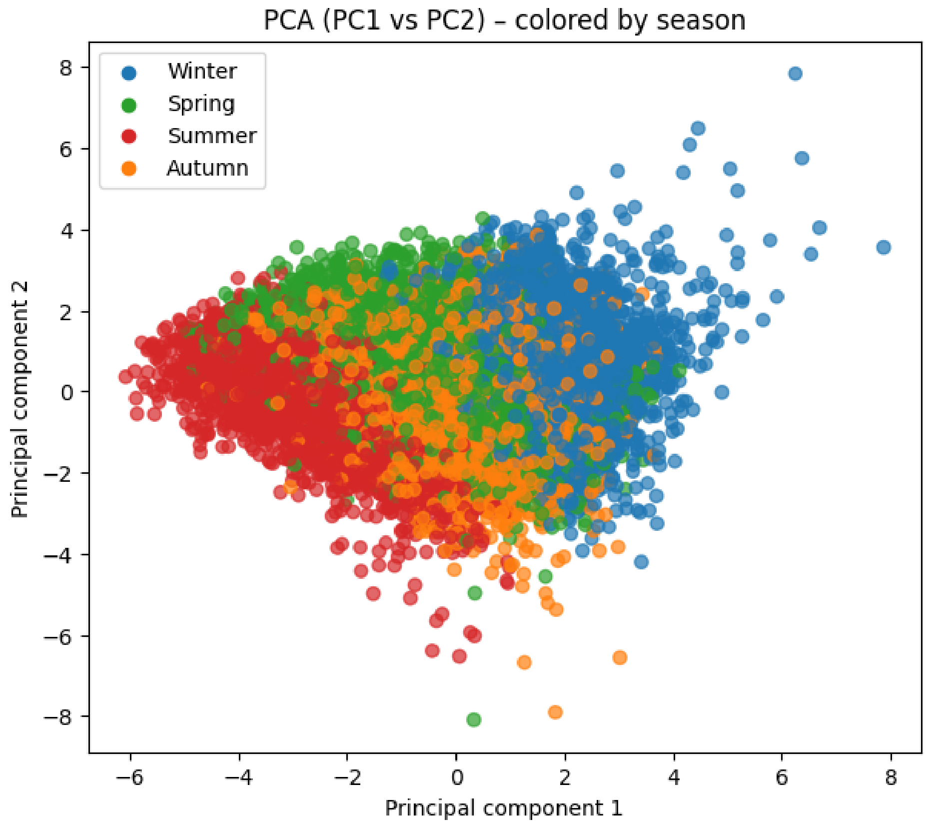

Figure 5: PC1–PC2 projection colored by season.

Role of the second principal component While the first component captures the overall seasonal level, the second principal component highlights differences that remain after removing this dominant effect. In the PC1–PC2 projection, days belonging to spring and autumn do not fully overlap, even when their PC1 values are similar.

This separation indicates that meteorological conditions with comparable temperature levels may still differ in other aspects, such as radiation balance, cloud cover, or atmospheric pressure. In particular, PC2 separates the ascending and descending phases of the annual cycle, reflecting asymmetries between warming and cooling periods. This behavior is consistent with physical phenomena such as thermal inertia and seasonal atmospheric circulation patterns.

Temporal continuity and gradients When the PCA projection is colored using a continuous temporal gradient starting from the spring equinox, the data exhibit a smooth and continuous trajectory through the principal component space.

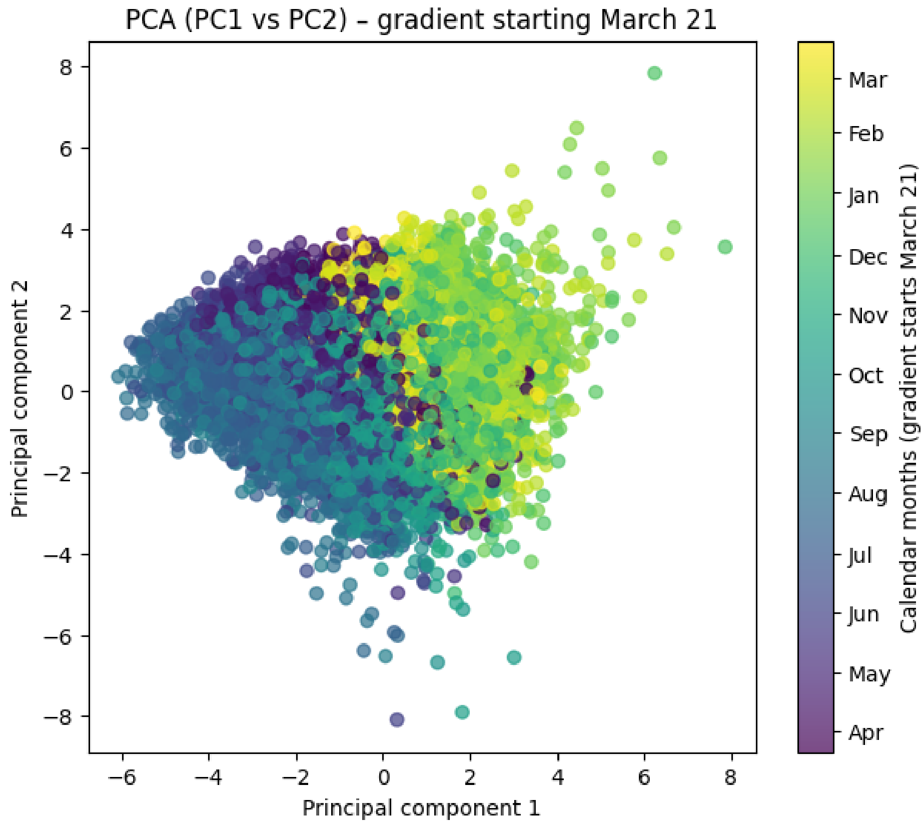

Figure 6: PC1–PC2 projection colored by a temporal gradient starting March 21.

The absence of sharp transitions or isolated clusters confirms that daily meteorological variability evolves progressively over time. This observation supports the interpretation of weather as a continuous dynamical system rather than a collection of discrete states.

Contribution of higher-order components Extending the analysis to the third principal component provides additional insight into less dominant sources of variability.

While PC3 explains a smaller fraction of the variance, it highlights days that deviate from typical seasonal patterns. These points often correspond to atypical or extreme meteorological situations, such as unusual precipitation or pressure conditions. Therefore, higher-order components capture localized or episodic phenomena rather than large-scale temporal structures.

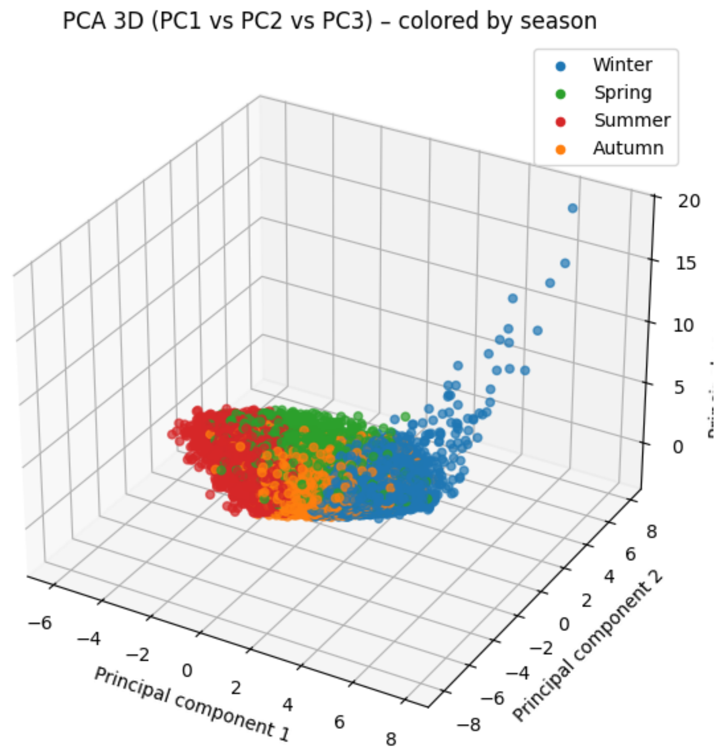

Figure 7: 3D projection onto the first three principal components.

Synthesis Taken together, these observations show that PCA decomposes the meteorological variability into a hierarchy of effects. The dominant seasonal cycle explains most of the variance, secondary components capture asymmetries and intra-seasonal dynamics, and higher-order components represent rare or extreme events. This hierarchical organization provides a coherent framework for understanding the structure of the dataset and motivates the use of complementary methods, such as Correspondence Analysis and Discriminant Factor Analysis, to further explore categorical relationships.

5 Correspondence Analysis

5.1 Description

Correspondence Analysis (CA) is a multivariate statistical analysis technique similar to PCA, but handles a population of individuals described by two qualitative variables. Each variable can take a finite number of values and the data can be represented in a table, where the columns correspond to the categories of the first variable and the rows correspond to the categories of the second variable. In other words, it is a technique that can summarize a two-dimensional table and display it in a biplot.

Here, two CA approaches were conducted.

1. **pressure vs. temperature**: each feature was split into 3 equal bins (low, medium, high)

pressure_cat	Low	Medium	High
mean_temp_cat			
Low	35	916	1014
Medium	175	6217	3857
High	0	1868	1219

Figure 8: Contingency table (pressure vs. temperature CA)

2. **season vs. day type**: day types are defined as followed:

Precipitation	Cloud cover	Day type
Low	Low	Sunny
	High	Cloudy
High	Low or high	Rainy

Figure 9: Definitions of day types

day_type	cloudy	rainy	sunny
season			
Autumn	1845	808	1169
Spring	1967	713	1184
Summer	1964	649	1251
Winter	2060	829	902

Figure 10: Contingency table (season vs. day type CA)

5.2 Results and analysis

5.2.1 Pressure vs. temperature

Here is the biplot obtained:

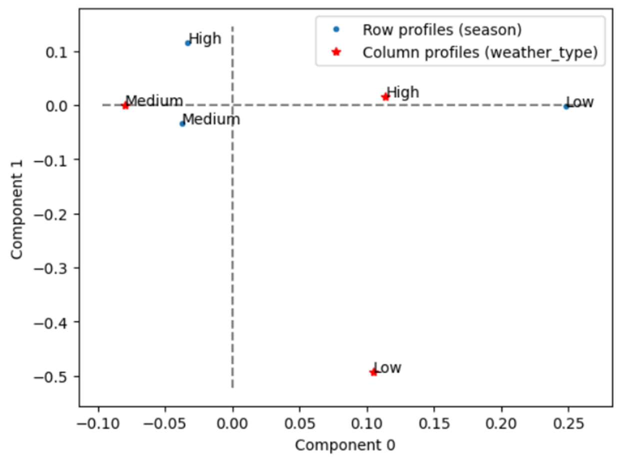

Figure 11: Biplot for pressure vs. temperature CA

Interpretation

On the "component 0" axis, values are very positively correlated with the low temperature category and, less so, positively correlated with high and low pressure.

On the "component 1" axis, values are not so correlated with temperature categories, because they are close to the horizontal axis, except slightly correlated with high temperatures. On the other hand, it is highly negatively correlated with low pressure, but not so much with other pressure categories.

To summarize, we can say that one axis is more related to mean temperature (component 0) and the other more to pressure (component 1). This shows that pressure and temperature are rather independent, especially when temperatures are low.

To check if the two variables are decorrelated, we plotted the pressure as a function of the mean temperature:

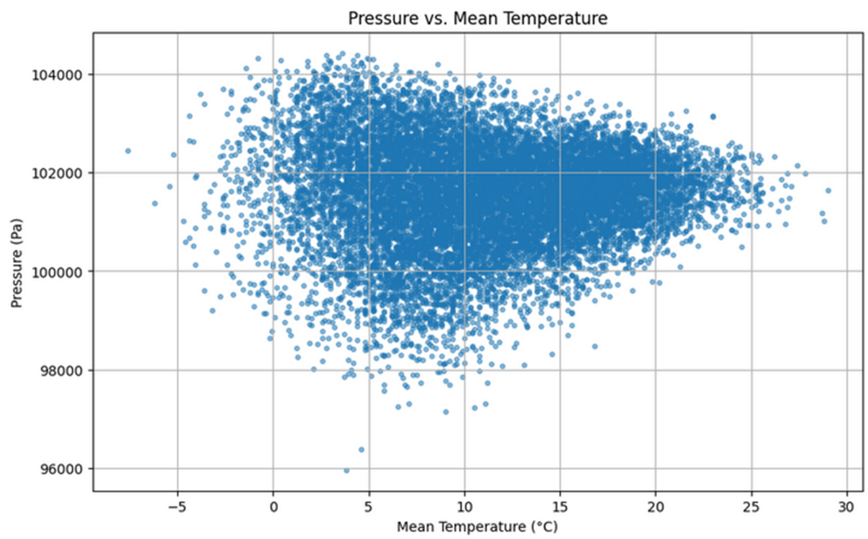

Figure 12: Pressure as a function of mean temperature

There seems to be no relation between both variables, since we obtain a cloud of points. This can be confirmed by the Spearman correlation coefficient, which ranges from -1 (perfectly negatively correlated) to 1 (perfectly positively correlated). In our case the coefficient is close to 0, which means both variables are decorrelated.

5.2.2 Season vs. day type

Here is the biplot obtained:

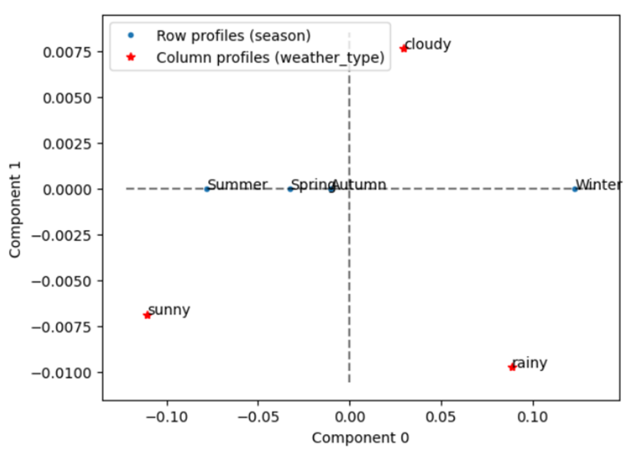

Figure 13: Biplot for pressure vs. temperature CA

Interpretation

Regarding the component 0 (horizontal) axis, lower values on the axis correspond more to Summer and sunny days, while higher values correspond to Winter and rainy or cloudy days. Therefore, we can say that rainy and cloudy days are more typical of Winter than Summer in London and that sunny days are more typical of Summer than in Winter.

Regarding the component 1 (vertical) axis, we could say that this axis measures the variance of the parameters used to define a Day Type. (cf. Figure 9).

Lower points on the axis (sunny and rainy) correspond to equal levels of precipitation and cloud cover, since a sunny day is defined as having "low" precipitation and "low" cloud cover and since a rainy day has "high" precipitation and "high" cloud cover. Let us note that we defined a rainy day as just having "high" precipitation disregarding the cloud cover level, but a rainy day is in fact a day with both "high" precipitation and "high" cloud cover, which is logical since it cannot rain without some cloud cover. This can be confirmed in the following contingency table where we separated "rainy" into two subcategories "rainy and cloudy" and "rainy and not cloudy".

day_type	cloudy	rainy_cloudy	rainy_not_cloudy	sunny
season				
Autumn	1845	713	95	1169
Spring	1967	632	81	1184
Summer	1964	570	79	1251
Winter	2060	719	110	902

Figure 14: Contingency table separating rainy days into cloudy and not cloudy

On the other hand, higher points on the axis correspond to distinct levels of precipitation and cloud cover, ie. a "cloudy" day is defined as "low" precipitation and "high" cloud cover.

6 Discriminant Factorial Analysis

6.1 Description

Discriminant Factorial Analysis (DFA), or Linear Discriminant Analysis (LDA), is a supervised method that seeks to identify linear combinations of features that best separate predefined classes. Unlike PCA, which maximizes total variance, DFA maximizes the ratio of inter-class variance to total variance.

Let G be the global center of gravity and G_k the center of class k. The analysis solves the generalized eigenvalue problem $A^{-1}Bu = \lambda u$, where B is the inter-class covariance matrix and A is the total covariance matrix. The resulting axes optimally discriminate the classes.

6.2 Results and analysis

6.2.1 Seasonal Discrimination

DFA was applied to test the separability of the four seasons. The projection onto the first two discriminant axes is shown in Figure 15. The first discriminant factor (DF1), which explains approximately 66% of the discriminating power, clearly separates Winter (in blue) from Summer (in red). This axis likely corresponds to the dominant temperature and radiation gradient. Spring (green) and Autumn (orange) occupy intermediate positions and exhibit a strong overlap. This indicates that while the "extreme" seasons are statistically distinct, the transition periods share very similar statistical profiles in terms of daily weather metrics.

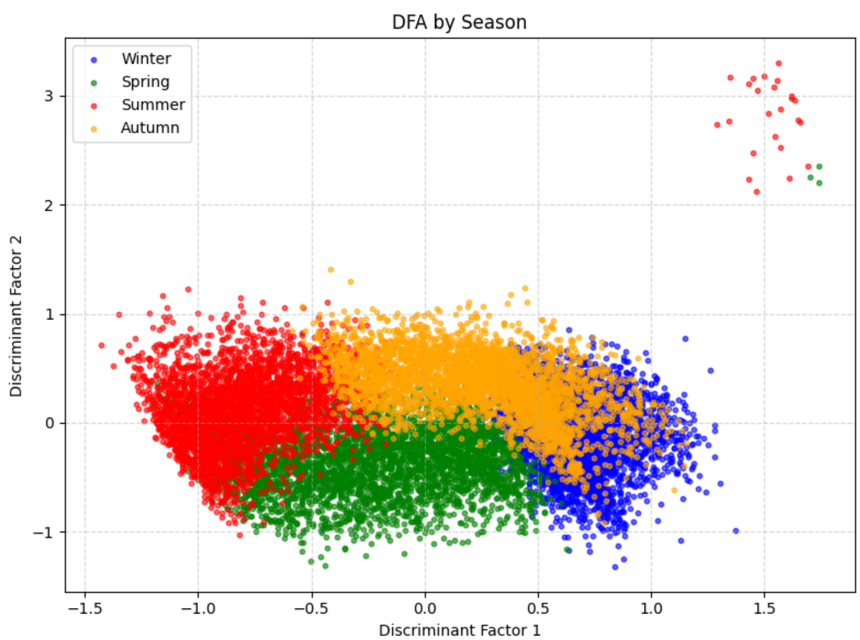

Figure 15: Projection of daily weather data on the first two discriminant axes, colored by season. DF1 separates Winter and Summer, while Spring and Autumn overlap.

6.2.2 Temporal Stability

A second DFA was performed to discriminate between three chronological periods: 1979-1992, 1993-2006, and 2007-2020. As shown in Figure 16, the distributions of observations for the three periods overlap almost completely. Unlike the seasonal analysis, the discriminant analysis extracts axes that fail to significantly separate the time periods. This suggests that the internal structure of the relationships between weather variables (covariance structure) and their mean values have not undergone drastic shifts large enough to be detected by this linear classification method. This supports the hypothesis of stationarity for the purposes of this study, although subtle climate change trends might exist that require more sensitive time-series analysis to detect.

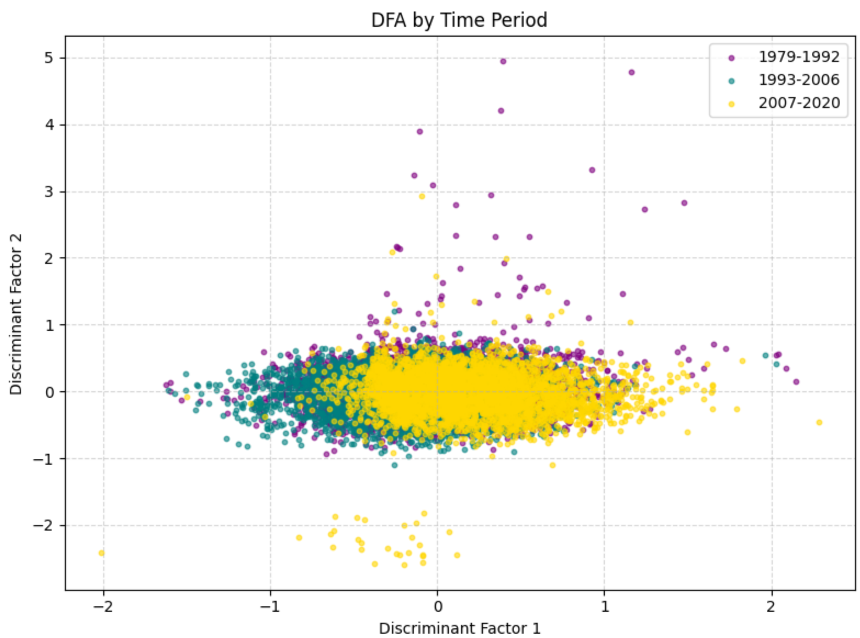

Figure 16: Projection of daily weather data on the first two discriminant axes, colored by time period. The significant overlap indicates stability in the weather statistical structure over the decades.

7 Conclusion and Future Work

This work analyzed daily meteorological data from London between 1979 and 2020 using descriptive statistics, PCA, CA, and DFA. The large sample size enables precise estimation of the daily mean temperature, with narrow and consistent confidence intervals for the mean, median, and variance, confirming the robustness of the estimators.

PCA reveals that the data is strongly structured by seasonality. The first principal component captures a dominant thermal and energetic contrast driven by temperature and solar-related variables, while the second component reflects secondary, transitional variability. The cumulative contribution indicates that most of the variance is explained by a small number of components, highlighting strong correlations among variables.

Correspondence Analysis shows weak associations between temperature and pressure categories, suggesting near-independence between these variables. Seasonal tendencies in day weather types are observed, but the proximity of categories to the origin indicates moderate relationships rather than strong dependencies.

DFA confirms that seasons can be statistically discriminated, with clear separation between winter and summer, while spring and autumn strongly overlap. No significant discrimination between time periods is observed, indicating stability of the overall statistical structure over the studied decades.

Future research could incorporate temporal dependence through time-series modeling to study long-term trends and persistence effects. Finer seasonal or monthly analyses may help characterize intra-seasonal variability, and extending the study to other locations would allow for comparative climatic analysis.

Team Contributions

Each member took part in writing sections within the report and preparing the final presentation.

Member	**Key Contributions**
Arnaud Brisset	Team management, writing of the introduction and descriptive analysis, writing and implementation of correspondence analysis.
Théo Couerbe	Writing of the methodology, implementation and writing of the DFA.
Adama Koïta	Help to find a data set, implementation of PCA, writing of PCA and conclusion.