TRIED TPA-05 Report

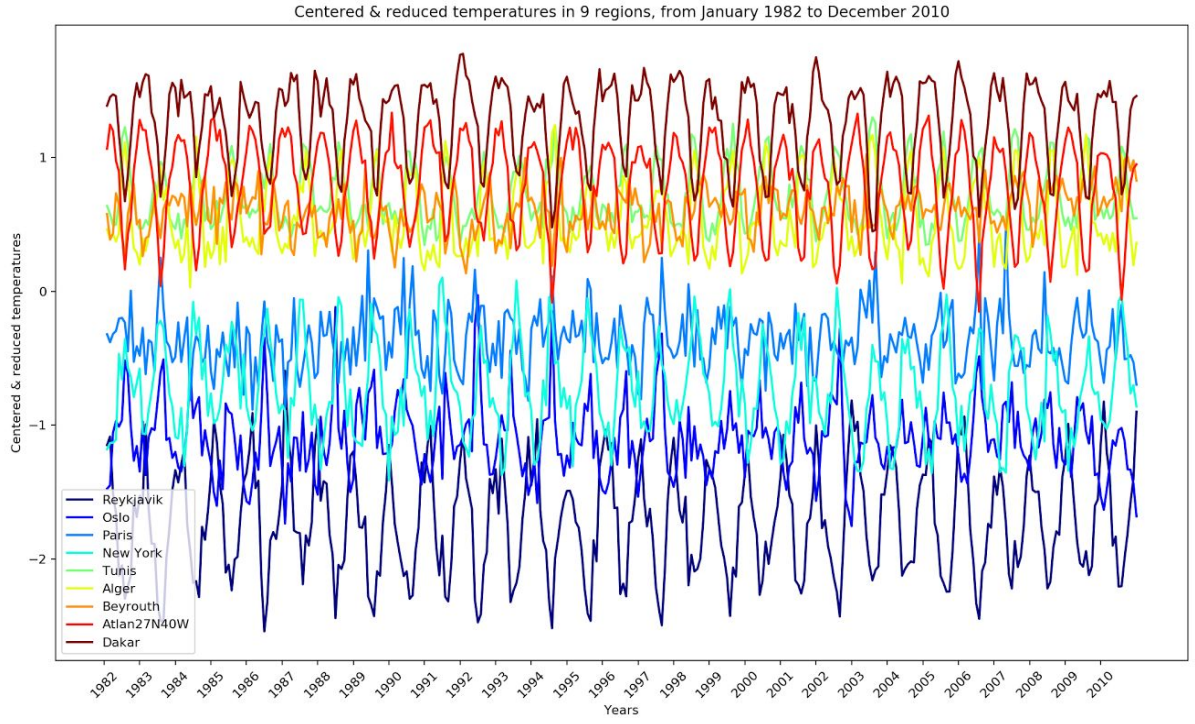
Carl Robinson - 26th Oct 2017

V – 2nd Part : PCA of temperatures

0) Objective and Data

To use principal component analysis and linear regression to analyse and interpret temperature climate data. The data file consists of: 2 date columns for year and month; 9 temperature data columns, one for each of nine geographical locations in the northern hemisphere, arranged in order from north to south; 348 rows, one temperature measurement per city each month, from Jan 1982 to Dec 2010 (29 years). The instances are the cities, and the variables are the months, so each column represents a chronological series of temperature data for a city.

1) Data representation



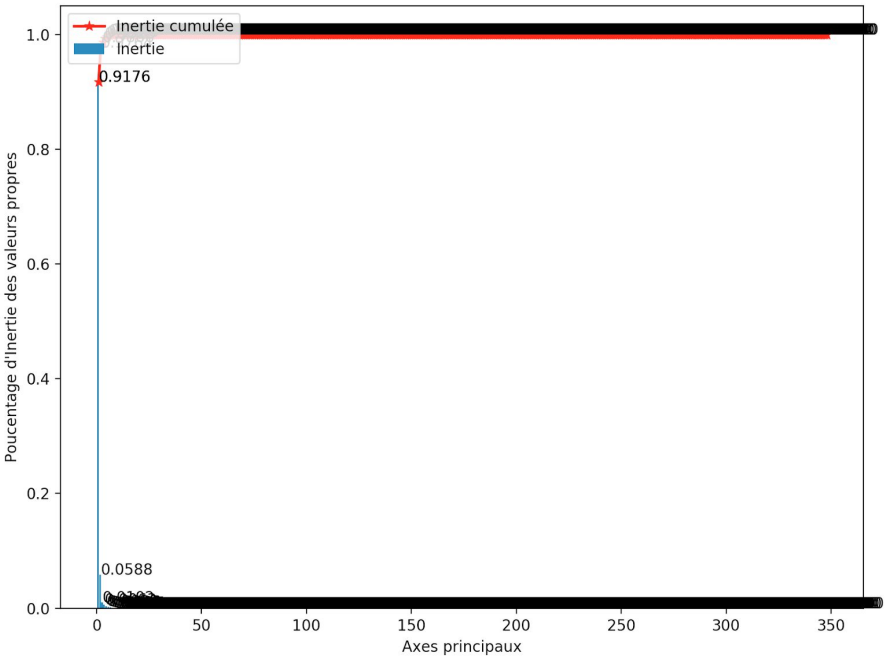
- Linear regression gradients for raw data
  - |Max| = 0.0869202859654, |Min| = 0.00837845575677, Difference = 0.07854183021
- Linear regression gradients for centered and reduced data
  - |Max| = 0.00538608709344, |Min| = 0.00532582289158, Difference = 0.00006026420186
- The difference between gradients is much smaller for the centered and reduced data. This shows that centering and reducing the data effectively cancels the global inter-year tendency for temperatures to rise. This is apparent in the overall flat horizontal progression of all 9 plots.
- The temperature plots have a clear vertical arrangement, with the colder, more northern cities in the lower half of the figure, and the hotter, more southern cities in the upper half. The colour scheme chosen supports this visualisation, with warmer colours chosen for the hotten cities.
- The 9 plots all clearly follow an oscillating pattern of rising and falling over the course of each year. This shows that the seasonal effect has been preserved, allowing for further analysis.
- The oscillation of the colder cities has a greater magnitude than the hotter cities.
- There can also been seen a clear gap between the colder and warmer cities - a north-south divide across the northern hemisphere.

2) PCA for centered and reduced data

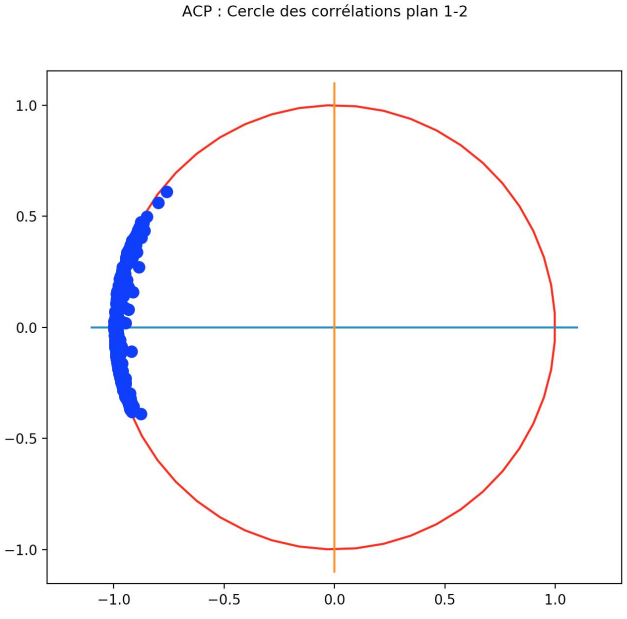
2.1) Percentages of inertia for the eigenvalues associated with the principal components

Individual and cumulative inertia (instances = 9 locations, variables = 348 months):

- The sum of the inertia from all 348 principal axes equals 1
- The sum of the inertia values of principal axes 1 and 2 equals 0.9784
- As 97.8% of the inertia is captured in the first two principal axes for either projection, it can be said that almost all of the available information is captured and represented in these two alone. Therefore a projection of the instances onto these two axes can be treated as a reliable representation of the measured phenomenon.

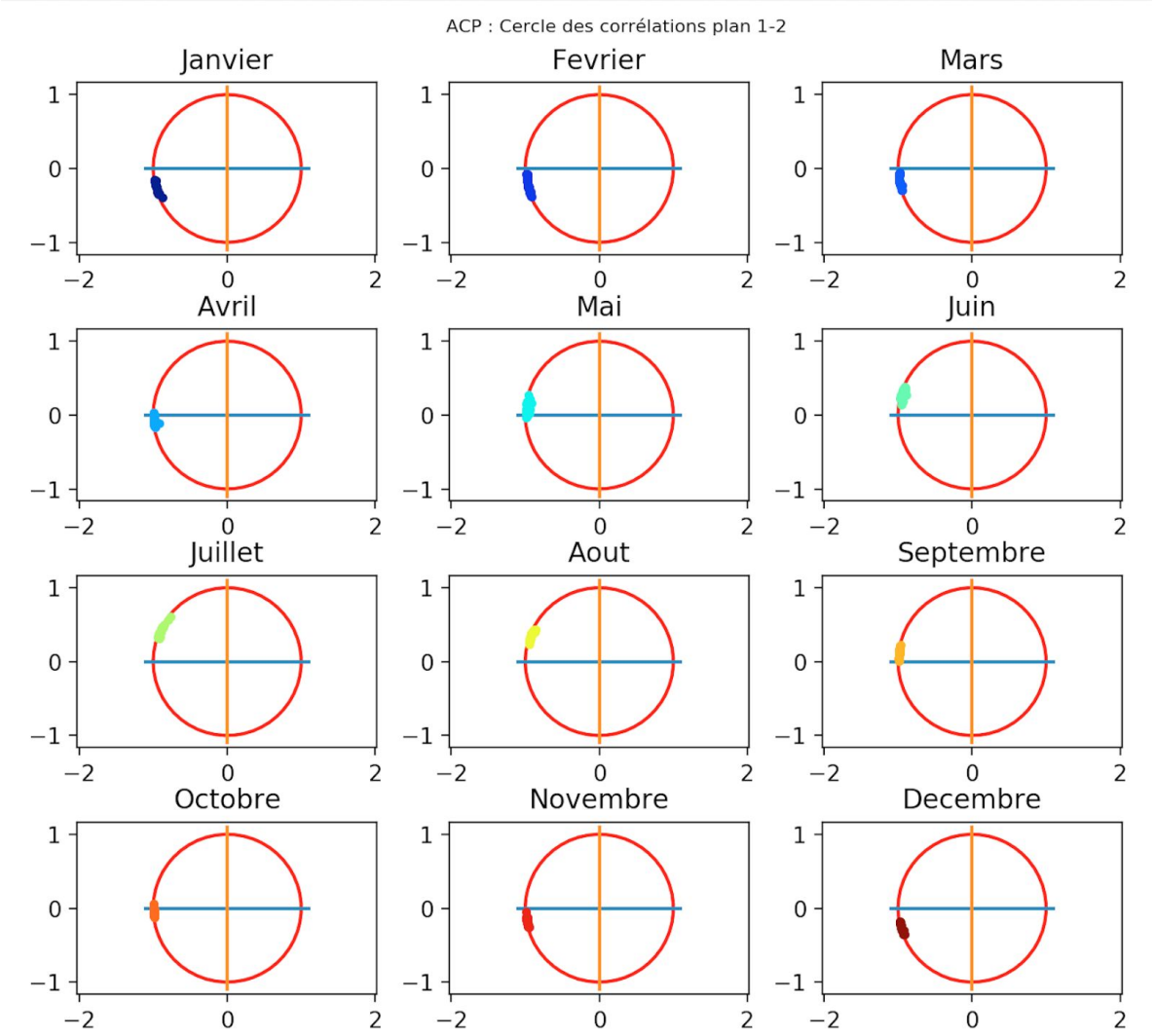


2.2) Correlation circles

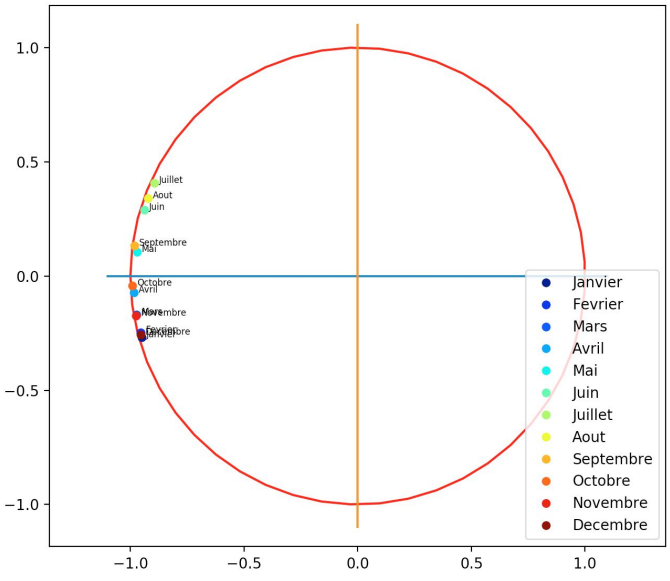


- All month variables are aligned in the same direction along the first principal component, indicating that they are strongly correlated with each other i.e. their mean monthly temperatures increase and decrease together.
- The variables all lie on or very near the unit circle's edge, which shows that all or almost all of their variance (information) is expressed as a combination of the first two principal components. We can therefore rely on this figure as an accurate representation of the monthly temperature data.
- The vertical spread is mostly caused by variation along the second principal component. There are too many points on the figure to easily identify which variables have positive and negative correlations, so we must choose an alternative representation in order to detect a pattern.

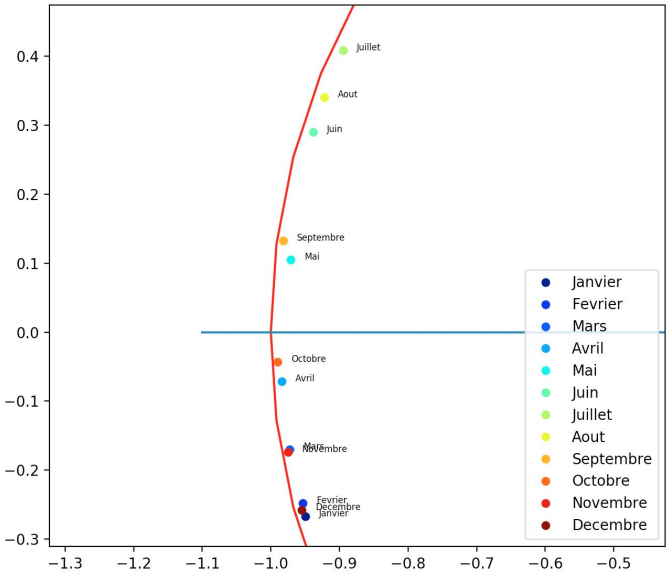
- By plotting the correlation circle variable points for each month separately (see figure below), we expose an oscillating pattern of movement that has a period of 12 months.
- January's points are grouped low down, with a negative value for the second principal component. As the months pass, these points rise steadily up until they reach a peak in July. The points then migrate back down until December, after which the pattern surely repeats itself.
- It must be noted that as the points stick to the edge of the circle during this 12 month oscillation, they also move back and forth along the first principal component to a lesser degree.
- This pattern exposes the seasonal effect on temperature that is experienced globally, and reflects the oscillating pattern seen in the plot for question 1).



Average vector values for each month:



Average vector values for each month (zoomed in):



- By plotting only the mean temperature values for each month, it is possible to spot clusters of months which are closely correlated to one another.
- The summer months, June, July and August have the highest mean temperatures and are clustered together at the top, while the slightly cooler months that lead into and out of summer, September and May, are grouped together below these.

- While October and April are opposite sides of the calendar, 6 months apart, they have similar mean temperatures, so are closely correlated here.
- Finally, the cold winter months, January, February and March have very similar low average temperatures, and so are represented very close to one another, low down on the figure. March and November are the months that lead into and out of winter, and these are tightly grouped just above.
- From this pattern we can discern that the second principal component's values are highly related to mean temperature; a large positive value indicates a high mean temperature, whereas a large negative value indicates a low mean temperature.

2.3) Contributions of instances to the two retained axes, PC1 and PC2

Contributions using Centered and Reduced data:

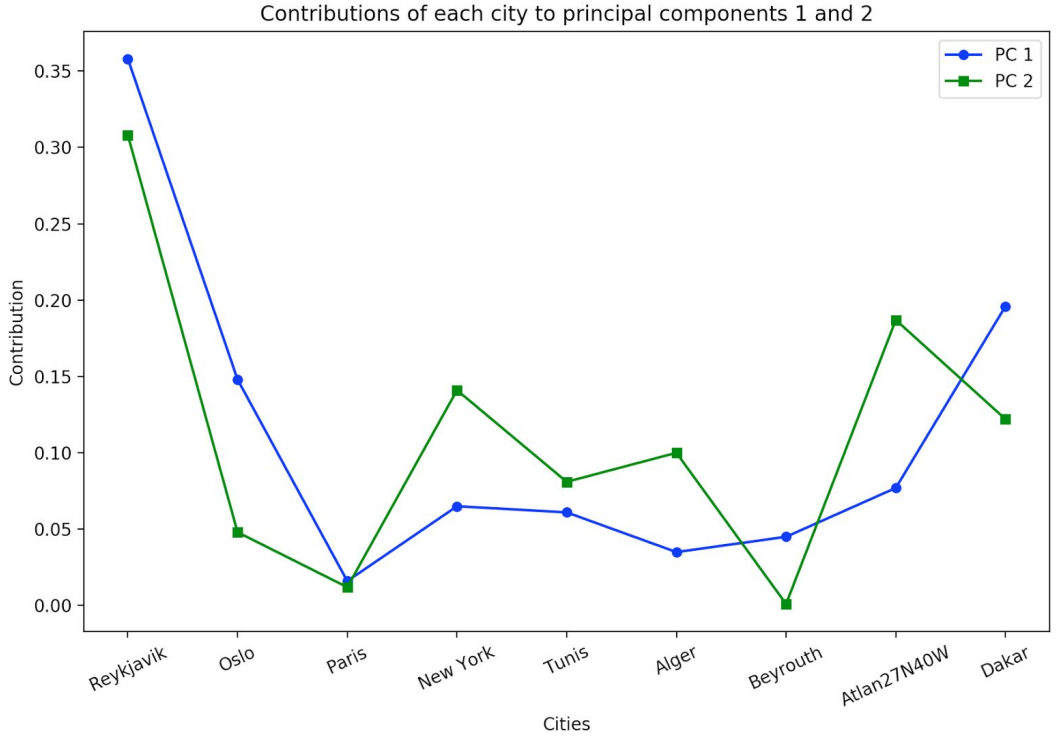
Location	PC1	PC2
Reykjavik	0.358	0.308
Oslo	0.148	0.048
Paris	0.016	0.012
New York	0.065	0.141
Tunis	0.061	0.081
Alger	0.035	0.1
Beyrouth	0.045	0.001
Atlan27N40W	0.077	0.187
Dakar	0.196	0.122
Sum Total	1.000	1.000

Contributions using original data (for reference):

Location	PC1	PC2
Reykjavik	0.008	0.124
Oslo	0.028	0.29
Paris	0.069	0.071
New York	0.047	0.24
Tunis	0.162	0.001
Alger	0.144	0.007
Beyrouth	0.151	0.002
Atlan27N40W	0.168	0.099
Dakar	0.222	0.165
Sum Total	1.000	1.000

- In both tables, summing by column shows the contributions for each principal component (PC) sum to 1.
- However, unlike the original data, the centered and reduced data is useful for comparing the contributions of the cities. This is because it eliminates both the dominant effect cities with higher mean temperatures have over cities with lower temperatures, and the dominant effect of cities a greater variance in temperature.

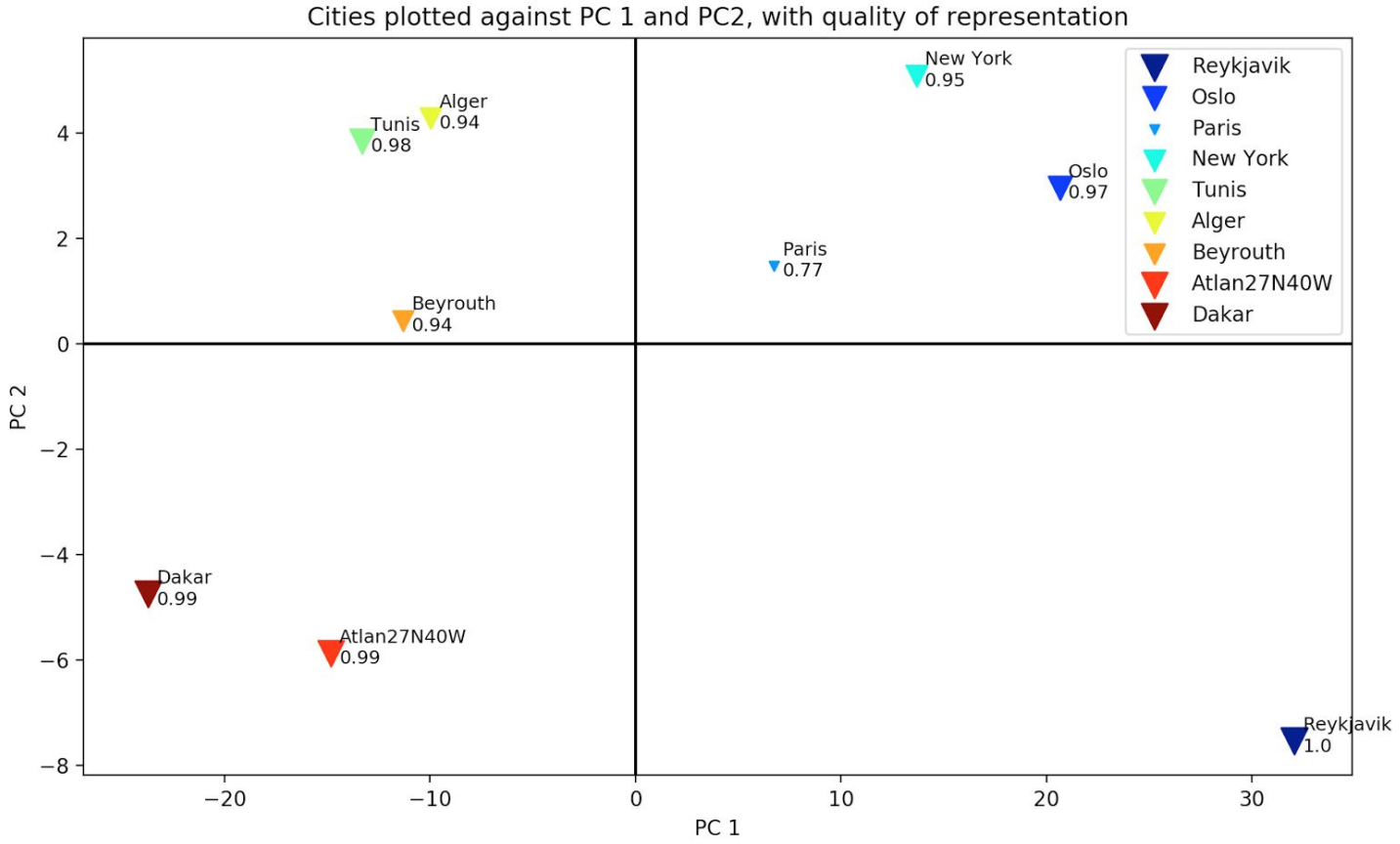
Plot of contributions using centered and reduced data:



- From the graph produced with the 'centered and reduced' data, we can see that the cities at the far left and right of the graph (i.e. the most northern cities with extreme high temperatures, and the most southern cities with extreme low temperatures), contribute the most towards the creation of principal components 1 and 2. Cities in the middle contribute much less.
- This is because the principal axes are defined by maximising the variance of the points along them. When calculating this variance, the extreme temperature values are taken into consideration more than those in the middle of the range, as they lie at the ends of the axes. Any change to the orientation of a principal component axis affects the distances between the extreme values and the axis much more than for other values, so temperature values from these cities play an outsized role in determining the optimal axes.

2.4) PCA cloud of instances on principal axes 1 and 2

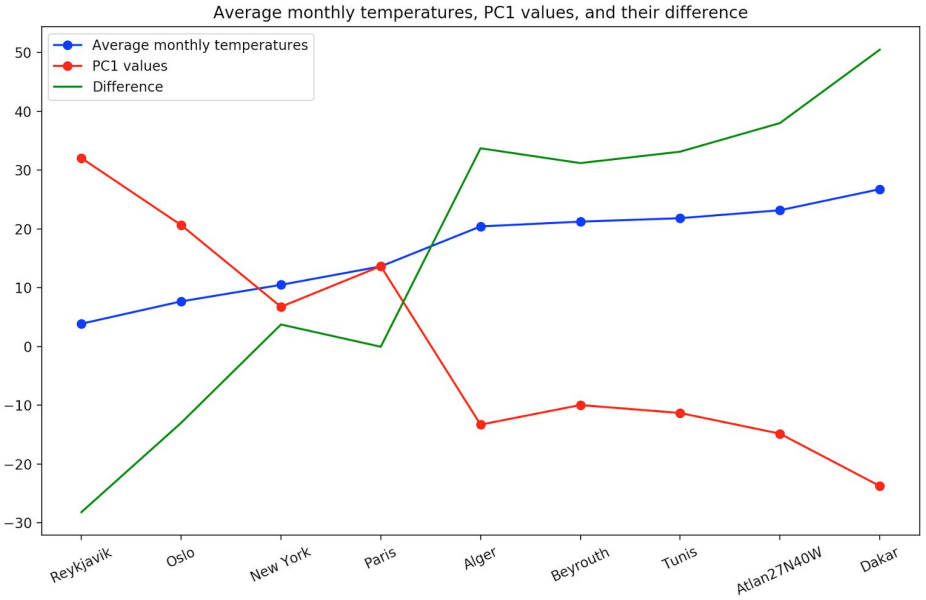
a) Cloud of points



- Here we see the colder northern cities (e.g. Reykjavik, Oslo) on the far right of the plot, with large values for PC1. The warmest cities (e.g. Dakar, Atlan27N40W) are on the far left of the plot, with large negative values for PC1. The temperate cities such as Paris are positioned more centrally in the PC1 axis. From this pattern we discern that the first principal component's values have a fairly constant negative correlation to mean temperature; a large positive PC1 value indicates a low mean temperature, whereas a large negative value indicates a high mean temperature. As PC1 represents around 91.8% of the total inertia, movement along this axis is much more discriminatory than movement along the vertical PC2 axis.
- We recall that the second principal component only represents around 5.8% of the total inertia, so any interpretation of a trend should take this into account. Reykjavik, Dakar and Atlan27N40W have very negative PC2 values, whereas Oslo, New York, Paris, Alger, Tunis, Beyrouth, all have positive PC2 values. This suggests that negative values of PC2 are correlated with cities with extreme hot or extreme cold climates, whereas positive PC2 values indicate more temperate climates, but more investigation is required to confirm/refute this.

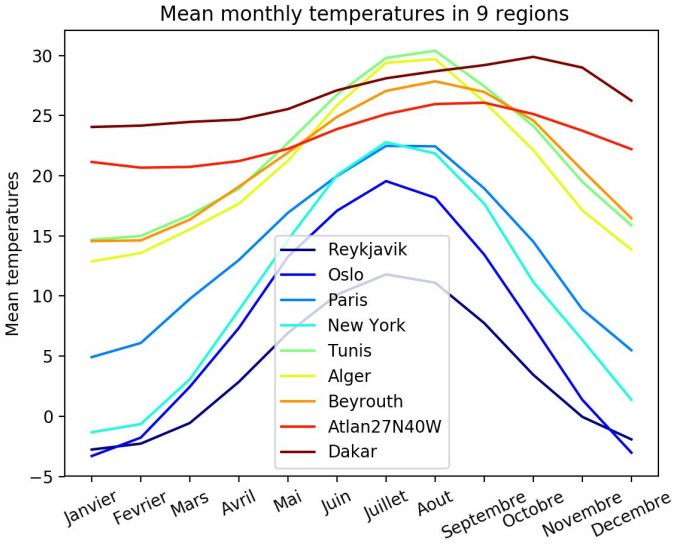
- Clustering is clearly apparent in this plot, as just described. Cities with similar temperatures are positioned close together on the plot, and far away from those with dissimilar climates. Used in conjunction with the correlation circle, this is useful in identifying previously unknown relationships between entities.
- Each city is accompanied by a quality value, calculated as the sum of the quality representations by PC1 and PC2, as these are the two axes used in the plot. The size of the triangle marker of each city indicates this quality of representation - the larger its marker, the better represented a city is by the combination of PC1 and PC2. The plot shows that the sizes follow a similar pattern to the contributions of the cities to each axis, in that the cities with the most extreme hot/cold climates have the largest values for quality representation. Paris is positioned in the middle, and is significantly less well represented by PC1 and PC2, indicating that much of its information is captured by other principal components not displayed on this plot.

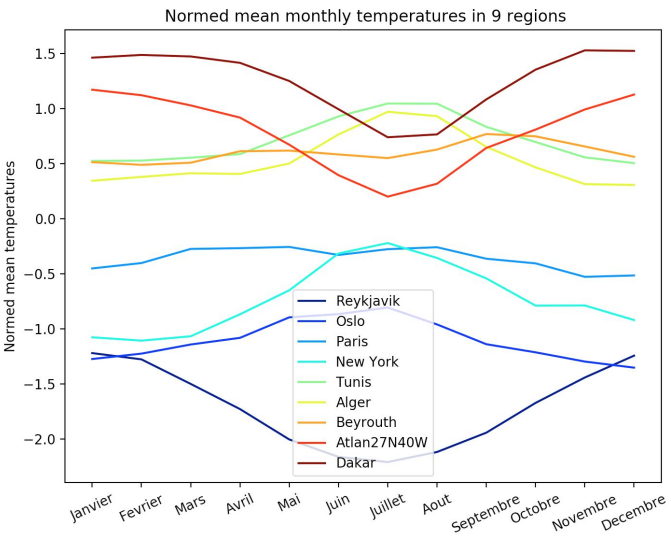
b) Comparison of plots of average monthly temperatures and first principal component



- Plotting the cities in ascending mean monthly temperature order clearly shows that a high PC value indicates a low mean temperature, and vice-versa, which is consistent with our interpretations from the cloud of points. By plotting the difference we can easily spot the linear relationship between the two.
- The correlation coefficient between average monthly temperatures and PC1 values = 0.99. This is very close to a perfect correlation of 1.0, so we can say there that the PC1 values are simply a result of a linear combination of the mean monthly temperatures.

c) Monthly climate graphs





- The first graph shows that the mean monthly temperature of the colder cities rises and falls dramatically throughout the course of the year, whereas it is far more constant in the hotter cities. This, however, does not explain what we see in the cloud of PCA points.
- The second graph has the mean monthly temperatures normalised by subtracting the global monthly averages from the mean monthly averages for each city, then dividing by the monthly standard deviation (across all cities).
- Here, we can see that the normalised mean temperatures drop considerably in the middle of the year for Reykjavik, Dakar and Atlan27W40W, the cities with the most extreme hot/cold temperatures. In the PCA plot, these three cities have the largest negative PC2 values in the cloud of points.
- On the other hand, the normalised mean temperatures rise somewhat in the middle of the year for Oslo, New York, Tunis and Alger, and these are the cities with the largest positive values for PC2.

We can conclude that the second principal component represents the direction and the degree to which a city's monthly average temperatures vary relative to the other cities.

- Cities with negative PC2 values all simultaneously become warmer or cooler relative to other cities at the same time of the year. E.g. Reykjavik, Dakar and Atlan27W40W all become cooler relative to other cities during summer, and all become warmer relative to other cities during winter.
- The same is true for cities with positive PC2 values, only their change in relative temperature goes in the opposite direction. E.g. Oslo, New York, Tunis and Alger all become warmer relative to other cities during summer, and all become cooler relative to other cities during winter. New York has the highest PC2 value as it varies the most with respect to the other cities, as can be seen by the steepness of its curve on the monthly mean temperatures plot.

This is reflected on the normed mean monthly temperature graph by the distance of the plots from the y=0 line; the closer a plot is to this line, the less different that city's mean monthly temperatures are from the mean monthly temperatures of the other cities.

- Dakar and Atlan27W40W have winter (Nov to Feb) temperatures that are significantly higher than the other cities, whereas their summer temperatures are much more similar. We know that their mean temperatures are fairly constant all year round, so this is more of a reflection of the fact that their summers temperatures are less different due to the other cities temperatures rising to meet them. PC2 values for Dakar and Atlan27W40W are strongly negative, indicating that change in relative temperature is correlated.
- Conversely, Reykjavik's summer (Jun, July, Aug) temperatures vary the most significantly from the others, being much lower than the rest, suggesting it has relatively cold summers compared to other cities. The PC2 value for Reykjavik is also strongly negative, indicating that its change in relative temperature is correlated with Dakar and Atlan27W40W.
- Oslo and New York become more different to the others during winter, which suggests they have relatively cold winters. Tunis and Alger plots show the greatest difference in the summer, suggesting they have relatively warm summers. These cities all have strongly positive PC2 values, indicating that their change in relative temperature is correlated, and it moves in the opposite direction to Reykjavik, Dakar and Atlan27W40W at the same time of the year.