Cycle Plot Revisited: Multivariate Outlier Detection Using a Distance-Based Abstraction

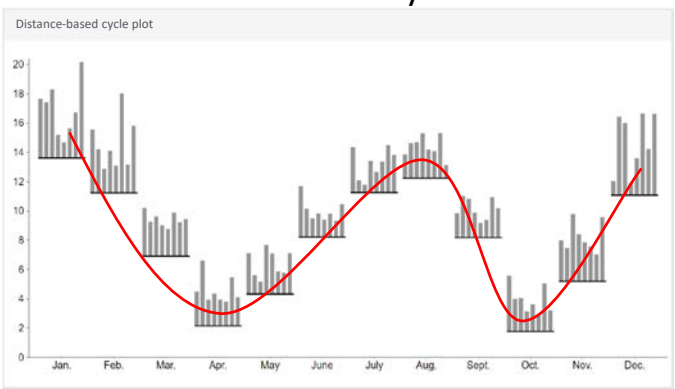
M. Bögl[1], P. Filzmoser[1], T. Gschwandtner[1], T. Lammarsch[1], R. A. Leite[1], S. Miksch[1], & A. Rind[2]

[1]Vienna University of Technology, Austria [2]St. Pölten University of Applied Sciences, Austria

Supplementary material to support the reader in the usage scenario (Section 6) of our paper.

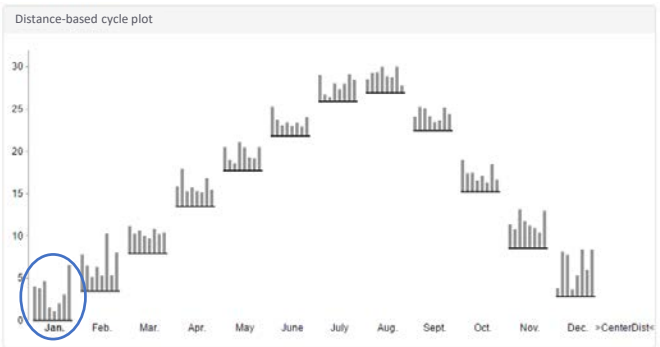
Distance-Based Cycle Plot

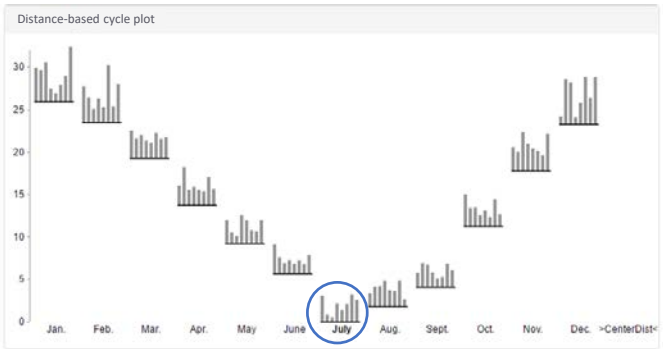
Considering the transitions between high and low peaks of the season in the original cycle plot representation of each variable, the seasonal pattern of the Mahalanobis-distance-based abstraction from multivariate space follows a similar smooth behavior compared to the underlying univariate cycle plots.



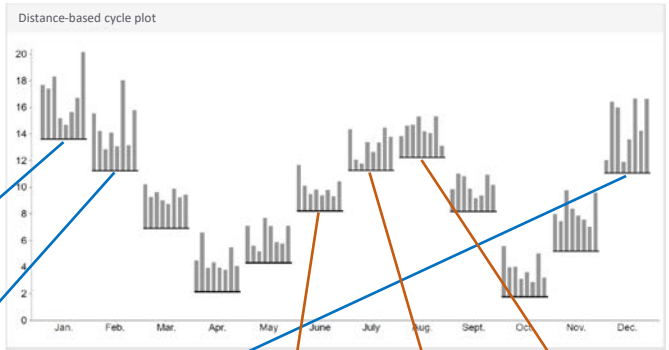
Our prototype allows to change the global reference point either to the **global** center (default) or to any of the **group** centers.

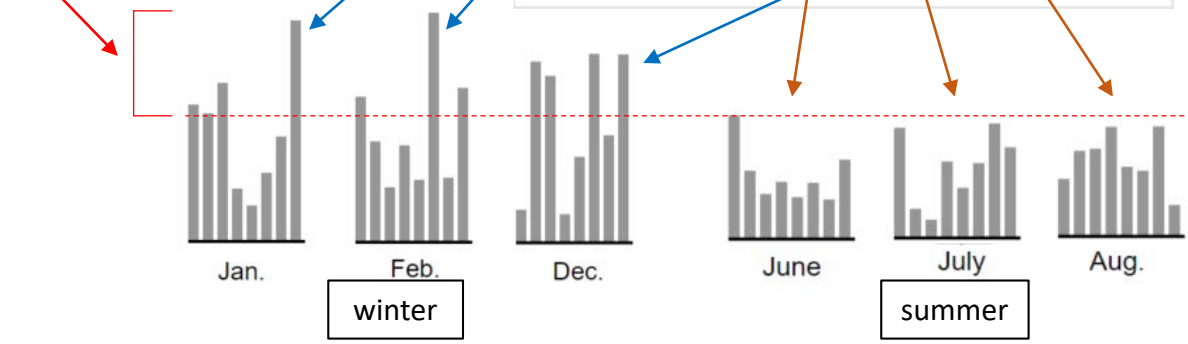
Selecting for example January as global reference point shows that the other winter months are closer to January than the summer months in the multivariate space. Likewise selecting July, summer months are closer.



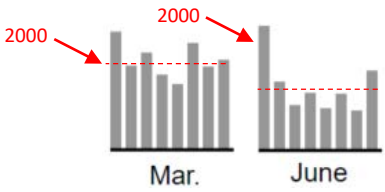


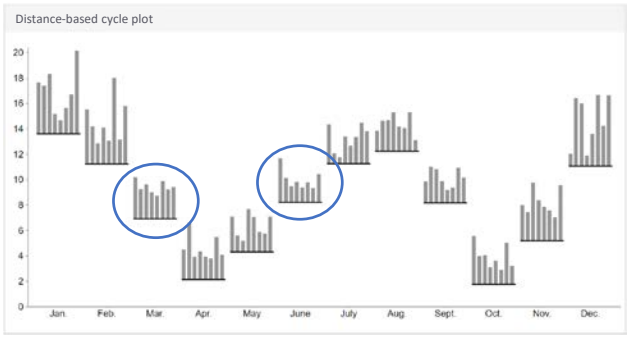
The user then considers the behavior within the groups according to their position in the seasonal cycle (cf. tasks T2 & T3). He/She identifies a tendency in some of the peak months (Dec., Jan., & Feb.), that the data values within the group vary more than in others. Especially, when comparing to the other peak in summer, the user detects this additional variation with larger distances to the group center (T4).



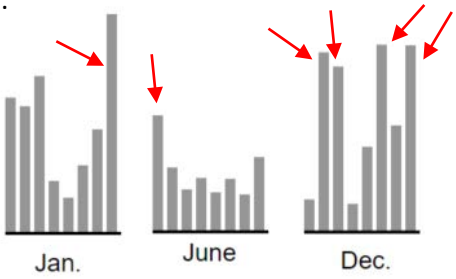


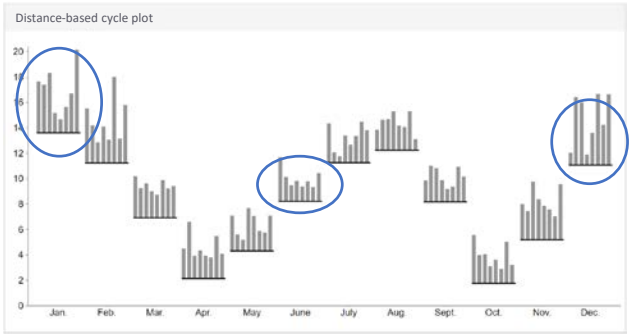
Next, the user compares the variations within the groups and across groups in more detail (T2 & T3). When looking at the months June and March, he/she spots distances with roughly the same length, except for the first year. According to this pattern, these months seem to be quite stable months across all dimensions.



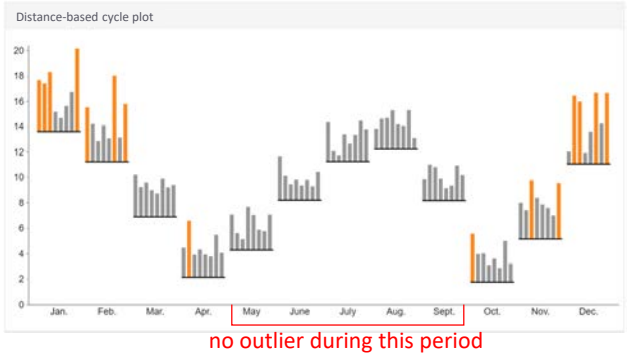


Even without highlighting the user can easily identify extreme values by large bars, that may be possible outliers (T4). Amongst others, the user considers the last year in January, first in June, and several in December, as possible outliers.





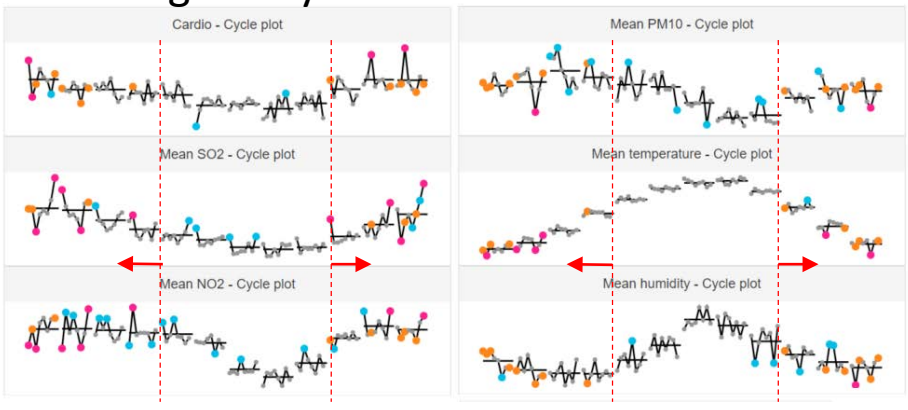
For example, the user finds interesting that there are multivariate outliers only in months Oct.–Apr., and an exceptionally large number in Dec.–Feb.



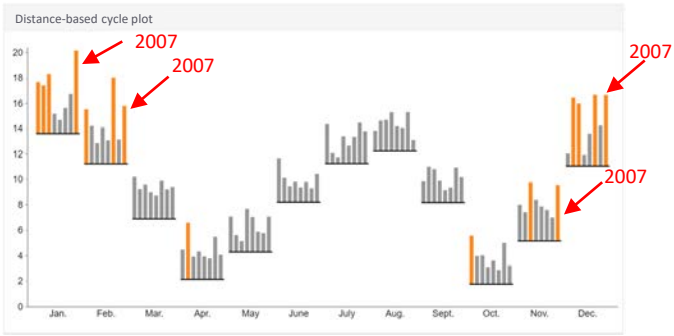
no outlier during this period

Knowing that, the user detects the same pattern in the original cycle plots and recognizes that there are more data points in these winter months highlighted in magenta (T7), indicating outliers in both, uni- and multivariate space.

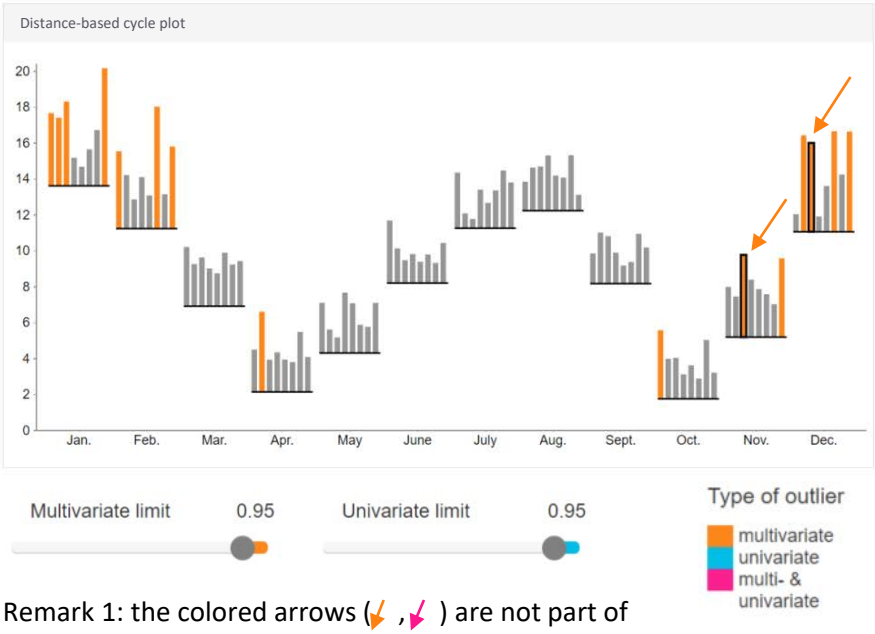
Original Cycle Plot for each Variable

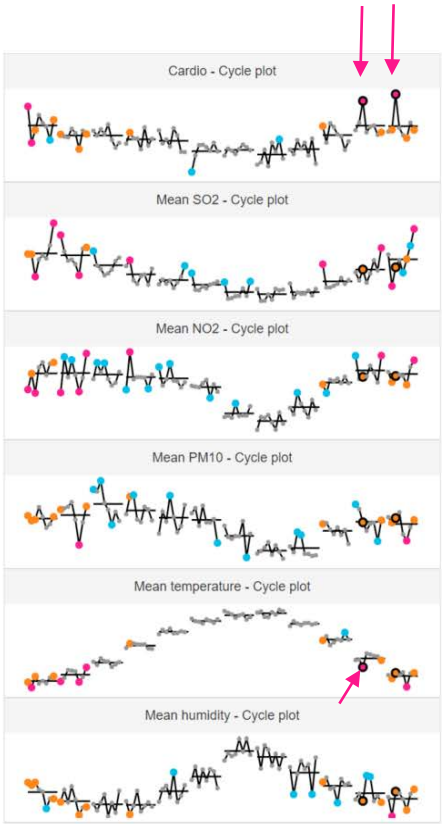


The user immediately recognizes that the last year (2007) in Nov.–Feb. are all multivariate outliers.



Looking at the original cycle plot for the variable **cardio**, the user detects two extreme data points in Nov. and Dec., highlighted in magenta. Selecting them shows that in the distance-based cycle plot, they can also be recognized as data points with large distance to the center (T9). The user also recognizes that besides being multivariate outliers, the variable cardio is also an univariate outlier in Nov. and Dec., but the variable temperature is an univariate outlier only in Nov. not in Dec.





Remark 1: the colored arrows (,) are not part of the prototype.
Remark 2: selected items have their border increased (Nov. 2002 and Dec. 2002).

By changing the outlier boundary with the slider, the user can track the data points that are borderline and are indicated as outliers, when the boundary is decreased. For example, the first bar in month Mar. and Jun. in the distance-based cycle plot are only highlighted as outliers, when changing the threshold from the 0.95 to the 0.9 quantile (T10). This allows to interactively get an impression about how extreme the outliers are.