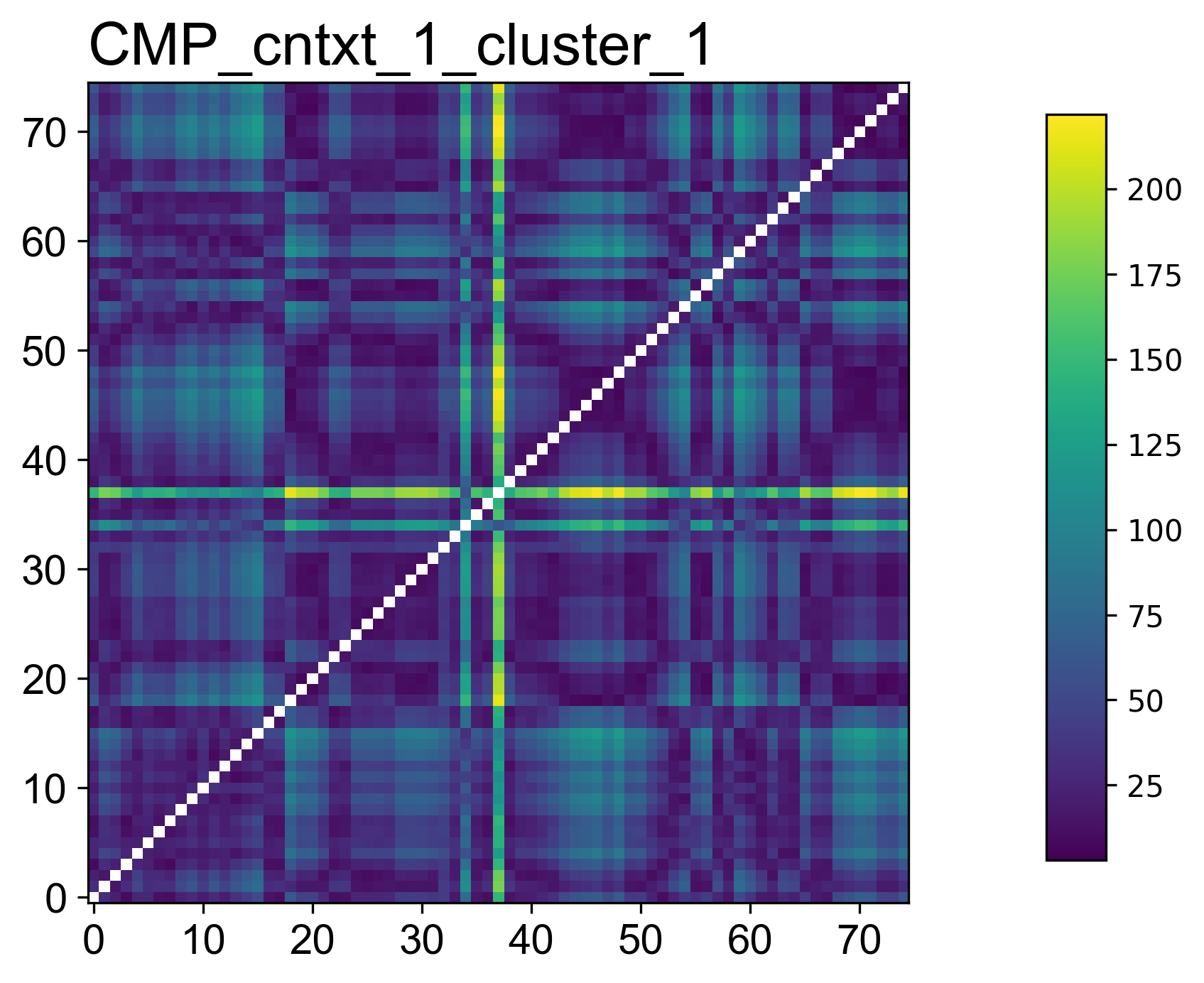
**OUTLIERS DETECTION (METHODS DESCRIPTION)**



Once the CMPs for contexts and clusters were obtained, the next step concerns finding days, belonging to the same cluster, that are far away distant from the rest ones. This process is well known as Outliers Detection.

Outlier definition of Hawkins [Hawkins 1980]:

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”.

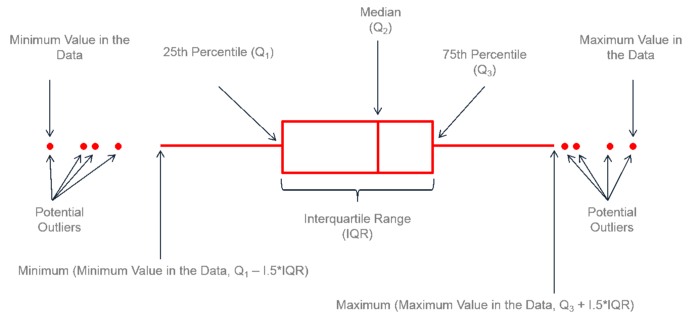
In this case, the outliers detection was performed taking into account the median of each column (representative of a distance of a certain day from the other ones) and then applying some techniques to discover which median (distance and the day to belong to) is an outlier. Higher is the distance higher is the probability of a day(context) to be an outlier. The methods employed to detect outliers are four: boxplot, z-score transformation, elbow-method and the last, Generalized Extreme Studentized Deviate. Gli outliers presi in considerazione sono di tipo puntuale e I metodi sono di tipo statistico.

Below, a brief introduction of the main features of these statistical methods:

BOXPLOT (BOX AND WHISKERS PLOT)

A box and whisker plot displays the five-number summary of a data set. The five-number summary is the minimum, first quartile, median, third quartile, and maximum.

It also shows the [spread](https://www.statisticshowto.com/measures-of-spread/)and [centers](https://www.statisticshowto.com/center-of-a-distribution/)of a data set and is useful for indicating whether a distribution is skewed and whether there are potential unusual observations (low probability to happen) in the data set!

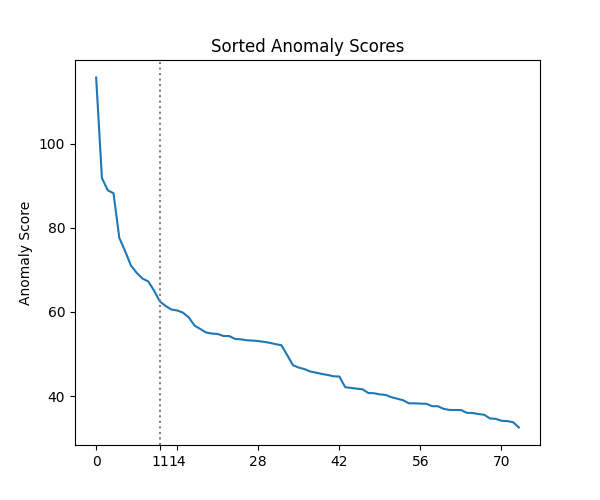


Z-SCORE TRASFORMATION

The process of changing data set values into z-scores involves creating signed numbers, called a z-score, such that, the sign of the z-scores (+ or –) identifies whether the values are located above the mean (positive) or below the mean (negative). The numerical value of the z-score corresponds to the number of standard deviations between a value and the mean of the distribution. z = 0 is in the center (at the mean), and the extreme tails correspond to z-scores of approximately –2.00 on the left and +2.00 on the right. If an entire distribution of values is transformed into z-scores, the resulting distribution of z-scores will always have a mean of zero and a standard deviation of one. The shape of the original distribution doesn’t change if transformed in z-score one and the location of any individual score relative to others in both distributions is the same. In conclusion advantage of standardizing distributions is that the values of different distributions can be compared. Those value that falls out]-2;2[ are possible outliers.

Immagine che contiene testo, luce, scuro

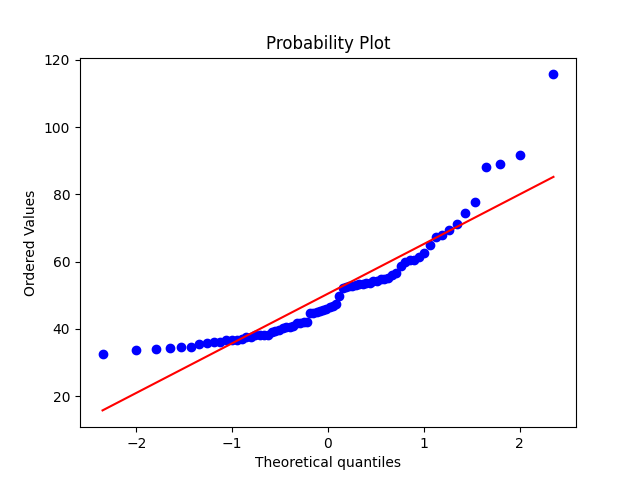
Descrizione generata automaticamente

ELBOW-METHOD

To detect those distances that are extremally large, another possible method is the elbow method. It sorts the medians of columns and gives them an Anomaly Score. When the rate of change of Anomaly score doesn’t vary significantly, the elbow is found. All distances (context or days) below the threshold are “outliers” or possible anomalies.

GENERALIZED EXTREME STUDENTIZED DEVIATE (GESD)

GESD is a simple statistical approach used to detect one or more outliers in a univariate data set that follows an approximately normal distribution. Statistical approaches assume that regular data follow some statistical model and the data not following the model are outliers. The GESD test only requires that an upper bound for the suspected number of outliers be specified. Given the upper bound, r, the generalized ESD test essentially performs r separate tests: a test for one outlier, a test for two outliers, and so on up to r outliers.



The generalized ESD test is defined for the hypothesis:

H0: There are no outliers in the data set

Ha: There are up to r outliers in the data set

Our test statistic is given by the formula below:

Here, and σ denote sample mean and sample standard deviation, respectively. In GESD we remove the observation that maximizes |xi — | and then recompute the above statistic with n-1 observations. We repeat this process until r observations have been removed. This results in the r statistics R1, R2 ………., Rr. Corresponding to the r test statistics, compute the following r critical values:

where is the 100p percentage point from the [t distribution](https://www.itl.nist.gov/div898/handbook/eda/section3/eda3664.htm) with ν degrees of freedom and

Our Significance level will be denoted by α.

The number of outliers is determined by finding the largest I such that Ri > λi.