1 Introduction

Transitions between energy sources sometimes create potential uncertainties and risks, such as risks affected by solar panels and wind turbines. In energy systems, outliers can often reflect such potential hazards and risks. In Erik Duijm's master thesis Outlier Detection in Energy Climate Data[2], he proposes the use of the Maximally Divergent Intervals(MDI) algorithm[1] to detect temporal outliers and finds that these outliers are potentially risky for the energy grid.

In the MDI algorithm, both Full covariance (Full method) and Identity covariance distribution model (ID method) can be used to detect outliers. In their work Duijm used the Full covariance method to detect outliers, as this is expected to take the specific properties of the outlier into account. However, Since the ID method will result in a simple comparison of the mean vectors of a given interval and the remainder of the time-series, it will have a faster of operation. Figure.1 shows the total running time for the period 1950-2019, a significant speedup is seen for the ID method in relation to the Full method (Here CE, KL denote Cross-entropy and Kull-back leibler criterion respectively). Therefore, we would focus on: To what extent the ID method can have approximately the same performance as the Full method.

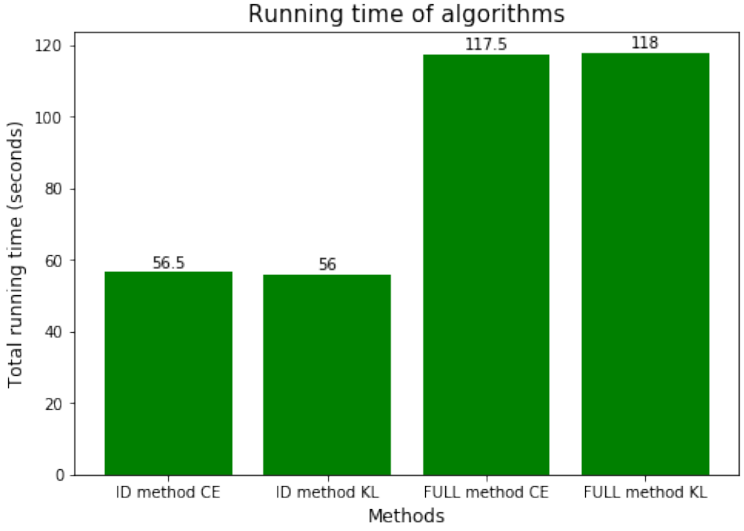

Figure 1: Running time of ID and Full method under Cross-entropy and Kull-back leibler criteria

To investigate our research problem, we conducted various experiments to compare the results of the two algorithms under Cross Entropy and Kullback-Leibler criteria. These two criteria are the metric used to quantify the difference between to intervals. We will first start with a Intersection over Union based matching analysis to investigate the overlap between the intervals of outliers generated by the ID and Full methods over multiple time spans, and based on this we will perform a correlation visualization of the order of these outliers in section 2. This allows us to visualize the overlap between the outliers generated by the two methods, and to some extent to answer the question: how many outlier intervals detected by the Full method can be also detected by the ID method based on Intersection over Union (IoU).

Afterwards, we combined the outlier data from different time spans, and after removing the overlapping outlier intervals using IoU in section 3, we visualized the Rank correlation and distribution of overlaps to investigate how close the outliers detected by the ID method are to those detected by the Full method for all time spans.

In addition, in section 4 we implemented the MDI program to re-calculate the divergence scores of different methods using the outlier intervals already obtained. For example, using the outlier intervals we detected by ID method to re-calculate the divergence scores of Full method. Moreover, we performed statistical correlation analysis on the re-computed divergence scores to investigate their correlation relationship. Finally,

we look into how many outliers at the initial set-up to obtain a desired set size after all the filtering done in section 5. For these related experiments are conducted to investigate how many outliers can be generated and the correlation between the length of the outlier intervals and their divergence scores. The code and experiment details can be found here: `https://git.science.uu.nl/s.tan3/outlier_detection-project`

2 Matching analysis and Rank correlation over multiple time spans

2.1 Matching analysis over multiple time spans

The ERA5 dataset contains historical climate data, spanning the period from 1950 to 2019. The Meteorological data from ERA5 was transformed to Energy variable as described in chapter 3 of Duijm's work[2]. We ran the ID and Full method MDI algorithms on Cross Entropy and Kullback-leibler criteria with the ERA5 dataset, at which point we have multiple time spans of outlier interval data. For either of the ID and Full methods, we have outlier results for 7 decades from 50s to 10s, while for each decade, we have 16 files corresponding to the results of running different criteria for each of the 8 different time spans. Each file contains 50 outlier points.

Specifically, each outlier interval contains a tuple of three elements: (a, b, Outlier score), where *a* is the first left point within a detected interval, *b* is the first right point after the interval and 'score' is the outlier divergence score. Table 1 shows the forms of outlier intervals, *Criterion*, *Method* and *Ranking* were added to identify the corresponding characters.

Table 1: Examples of outlier interval output

a	b	Outlier Score	Criterion	Method	Ranking
51831	51880	8.844	Cross-Entropy	Full	1
54649	54697	12.637	Cross-Entropy	ID	1

To investigate the overlap between ID and Full methods over multiple time spans, we conducted Matching analysis, the pseudocode of the algorithm is shown in Algorithm 1. The IOU criterion is implemented in matching analysis to compare the result lists (ID and Full methods results). When the outlier interval results produced by the two algorithms overlap, we call this situation a matching event. To illustrate this, since the threshold of IoU is set to 0, that is, when the iou scores of sublists i and j exceed 0, we will say that a corresponding matching event is generated. When running the matching algorithm, we find that there are multiple matching events detected by an single outlier interval, for example: one outlier interval of ID method matches multiple intervals of Full method.

To avoid this overlapping situation, we first sort the matching events according to their IoU values, and then keep only the one with the largest IoU score and remove the other overlapping events. Furthermore, the matching algorithm has two directions (loop from different lists): Matching ID method result list to Full method result list, and vice versa. Figure.2 illustrates the proportion of matching events to the number of all outliers, where for the Kull-back Leibler criterion, 60% of the ID outlier intervals can find a unique overlap with the Full method outliers, and vice versa. Meanwhile, for the Cross Entropy criterion, the proportion is less than 50%. More detailed visualization results can be found in appendix (Figure.A.1 and Figure.A.2).

Algorithm 1 Matching Algorithm

```
IoU_threshold = 0
Initialize matching_event as Dictionary
for i in list1 do
    for j in list2 do
        if caclulate_iou(i,j) > 0 then
            matching_event[i].append(j)
        end if
    end for
end for
Sort the elements for all keys in the matching_event dictionary
Keep the element in matching_event with the highest IoU score, remove others
Return matching_event
```

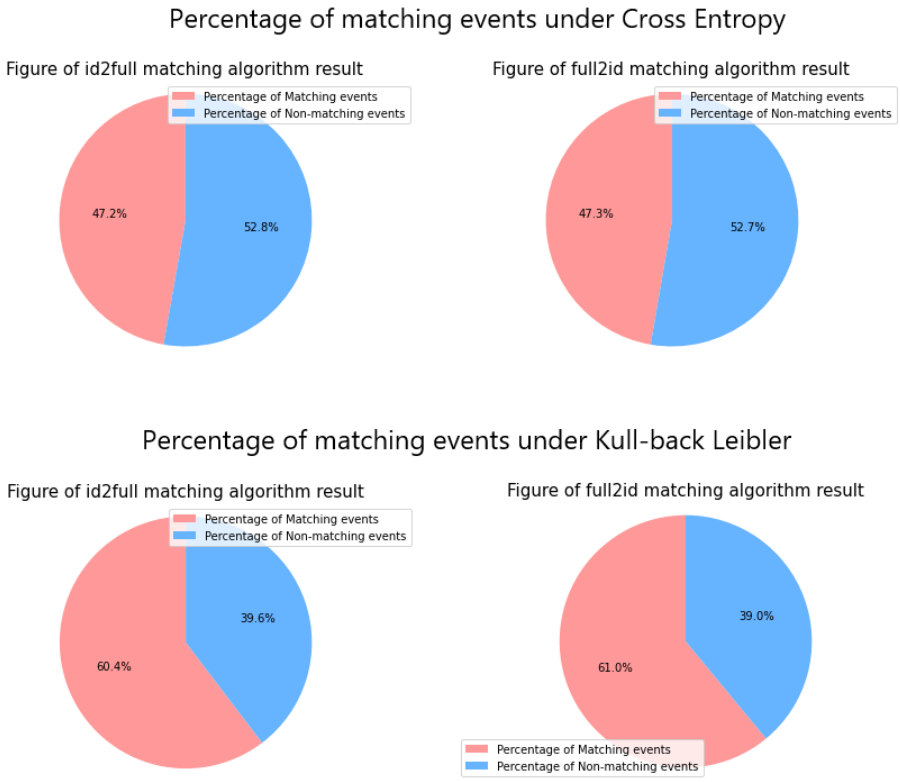

Figure 2: Percentage of matching events

2.2 Rank correlation of matching algorithm

Analyzing the ranking of the outlier intervals matched by the matching algorithm allows us to observe the distribution of the matching events. For example, after running the matching algorithm, we can observe whether the matched outliers conform to a positive correlation distribution, and if so, the ranking of the outliers generated by the ID method corresponds to a similar distribution in the results of the Full method. The visualization of the results for all decades (Figure.B.2, Figure.B.1) can be found in appendix. Figure.3 demonstrates one example of Rank correlation during 2000 decade. As can be seen, we can observe a certain tendency of positive correlation of matching events in the scatter plot, however, it is difficult to visually analyze their correlation because we have a total of 7 decades of data and 8 different time spans in each decade. Therefore, we will combine multiple time spans of data from each decade for further analysis.

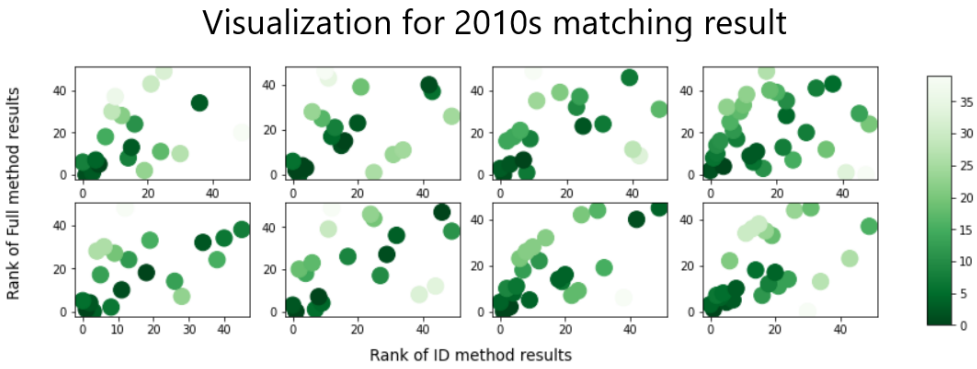

Figure 3: An example of Rank correlation during 2000 decade

3 Matching analysis and Rank correlation after merging time spans

In order to merge outlier data with different time spans, we use the steps in algorithm 2. Specifically, we first merge outliers over multiple time spans, and this operation should be operated four times because we have data from both the cross entropy and kull-back leibler criteria for the ID and Full methods. After that, for

each merged list, we perform step 2-19 to keep only the outlier with the largest outlier score in the list, and re-rank them by their divergence scores. This procedure allows us to ensure that there will be no overlap in the merged outlier data. Table.2 shows the average number of outlier intervals before and after combining the outlier data, here the average means the average number among 7 decades data.

Afterwards, Matching analysis and Rank correlation analysis were applied in the same steps as in the previous experiment. We visualized the results and colored the outliers by their IoU scores to analyze the correlations for better understanding. The rank correlation visualization graphs from ID to Full method under both criteria are shown in the Figure.4 and Figure.5, the reversed graphs(from Full method to Id method) can be found in appendix(Figure.C.1, Figure.C.2). In addition, to analyze the IoU scores of matching events, we also visualized the distribution of IoU scores of matching events in each decade, as shown in Figure.6 and 7.

To demonstrate the results, the rank of matching events under the cross entropy criterion exhibits stronger positive correlations, with average correlation coefficients exceeding 50%. Specifically, for the 1970s and 2000s, the coefficients of spearman's and pearson's correlations exceed 70%, and the p-values are less than 0.05. However, for the 1990s, the correlation coefficients of matching events was only 11% and did not pass the significance test. In contrast, for the matching events under the kull-back Leibler criterion, even though their mean and median IoU scores reach 0.83 and 0.93 respectively, yet their correlations are not as significant as those of the Cross entropy criterion, reaching only 45.48% and 41.14%. It indicates that for this criterion, although the outlier intervals generated by the ID method are very close to those generated by the Full method, the ranking of the outliers is indeed not close to the distribution of Full method's output.

Algorithm 2 Merging approach: Merging outliers for ID and Full method results over different time spans

```
Merge outlier intervals over different time spans as list1
Filter overlapping intervals using on IoU(step3 - 17)
Initialize Duplicates as Dictionary
for i in list1 do
    for j in list1 (i != j) do
        if caclulate_iou(i,j) > 0 then
            Duplicates[i].append(j)
        end if
    end for
end for
Sort the elements for all keys in the Duplicates dictionary by their Outlier scores
for keys in Duplicates do
    for k = 1 to #(elements in keys) (iteration from the second element) do
        for list1.remove(Duplicates[keys][j]) do
        end for
    end for
end for
Re-rank the outliers in list1 by their outlier scores
Return list1
```

Table 2: Average number of outliers before and after Merging approach

Case	avg.#(outliers) before Merging	avg.#(outliers) after Merging
ID method (Cross Entropy)	400	77.28
ID method (Kull-back Leiber)	400	74.0
Full method (Cross Entropy)	400	69.71
Full method (Kull-back Leiber)	400	81.71

4 Re-calculating the divergence scores using different method

we implemented the MDI program to re-calculate the divergence scores of different method using the outlier intervals already obtained. For example, using the outlier intervals detected by ID method to re-calculate the divergence scores of Full method. The main intuition of this experiment is that if we use the obtained ID outlier intervals to re-compute the new outlier scores using the Full method, how would the results correlate with the outliers obtained directly using the Full method. Figure.8 shows the correlation scatter of the outlier

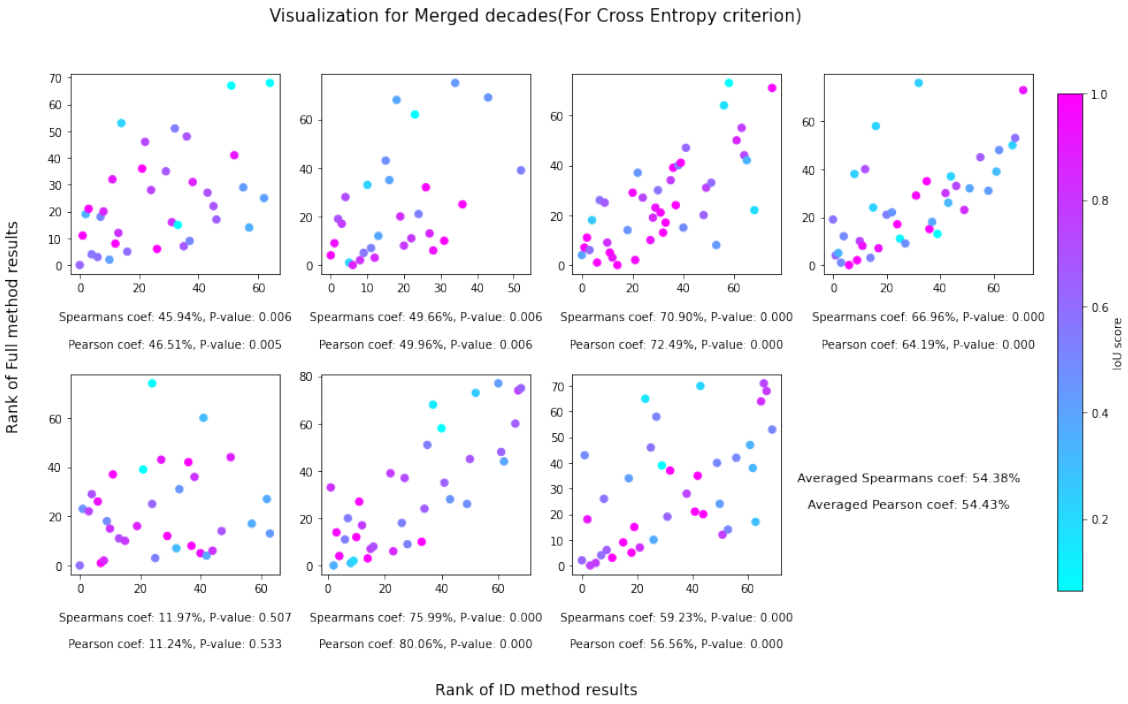

Figure 4: Visualization of rank correlation after megering outliers over different time span

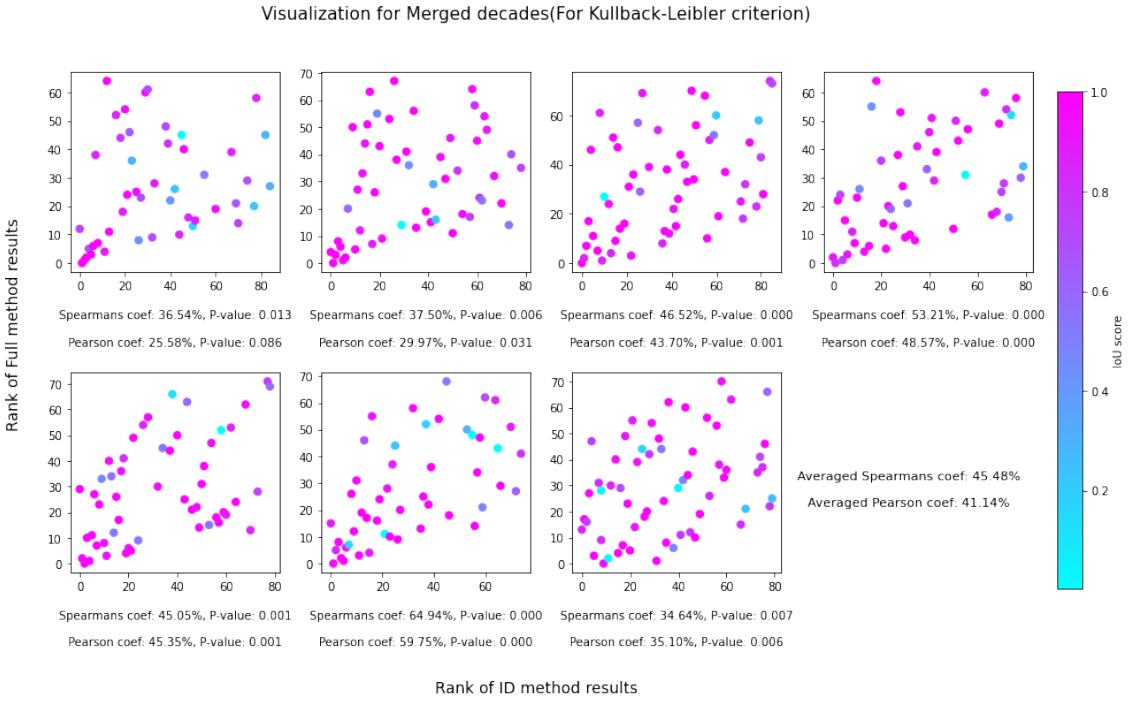

Figure 5: Visualization of rank correlation after megering outliers over different time span

scores obtained by using the outlier intervals which detected by the ID method(under cross entropy criterion). The color represents the ”original ranking” of the outliers, where it denotes the ranking before re-calculation. It can be observed that there is a strong positive correlation(both coefficients are over 80%) between the re-computed outliers using the outlier intervals from ID results and the outliers detected by the Full method.

As for the kull-back leibler criterion, it can be observed from Figure.9 that more outliers are stacked together with lower divergence scores, and this stacking property leads to a lower correlation coefficient. This means when outliers detected by the ID method are re-calculated using the Full method, the smaller their divergence scores are, the lower the positive correlation is. As a result of the rank information, outliers with lower rank (larger rank value) have higher full method outlier scores after recalculating the full method scores,

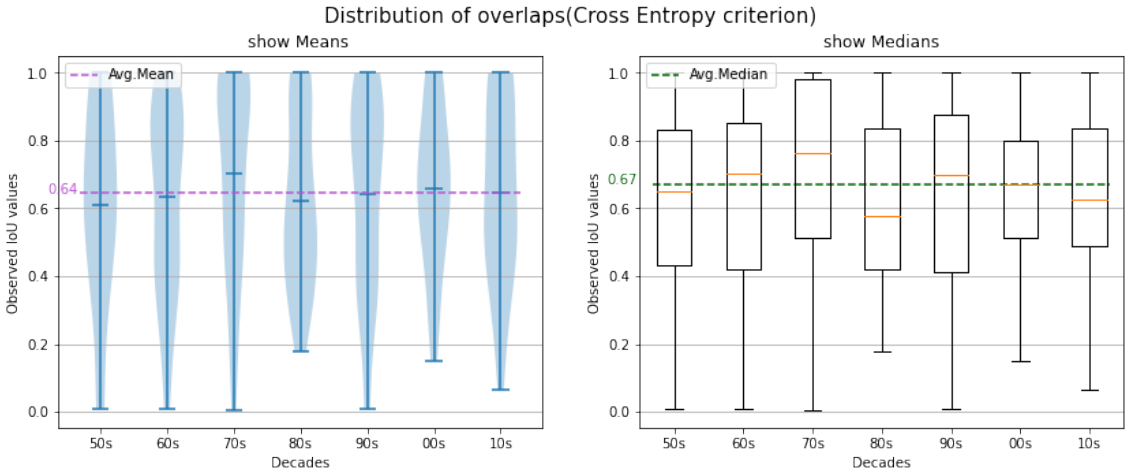

Figure 6: Distribution of Merged matching events (Cross Entropy)

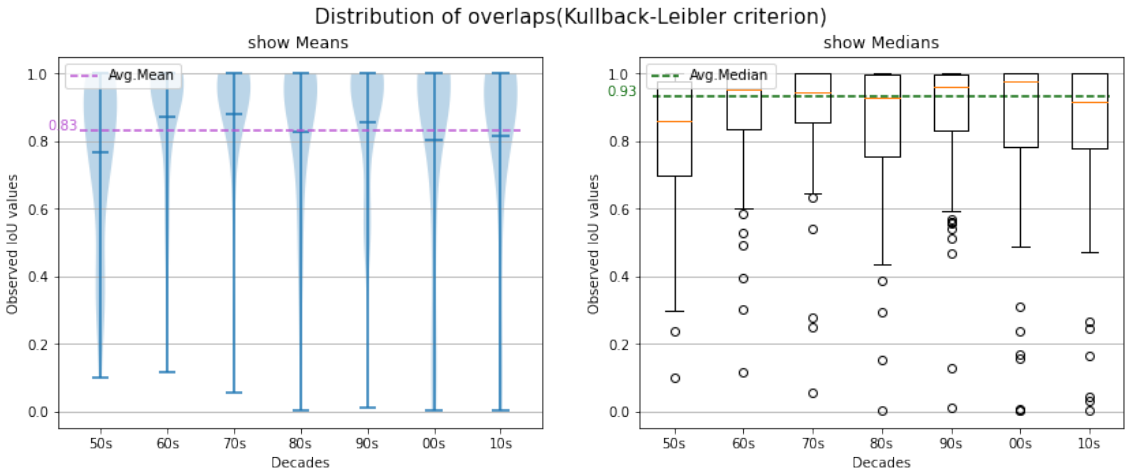

Figure 7: Distribution of Merged matching events (Kull-back Leibler)

while outliers with higher rank have higher full method outlier scores after recalculating the full method scores. This suggests, to some extent, that the ID method can detect outliers that are more deviant from the normal value, but inaccurate for outliers that are less deviant.

5 How many outliers do we need

To investigate how many outlier points are sensible for the MDI algorithm to output, the following experiments were conducted. The pseudocode for this experiment can be found in algorithm 3. First, the MDI algorithm was set to output 1000 outliers for both ID and Full methods in the experiments, which ensures that we will have enough outliers for analysis. After this, we count the number of outliers we actually get, because in reality we will not have 1000 outliers. One possible reason is: in the MDI algorithm, when all the outliers are obtained, the program runs the IoU algorithm to remove the overlapping intervals. Hence, even if the program is set to generate 1000 outliers, those overlapping intervals will be movoed. To illustrate the results, Figure.10 and Figure.11 shows only 150 to 500 outliers for 7 decades are actually generated. The distribution is due to the fact that we do not combine data from multiple time spans. All methods have a mean number around 270 for different criteria, while their median is around 240, which indicates in general, we can only detect less than 300 outliers using the ERA5 dataset.

Furthermore, As shown in work of Barz et al.[1], a statistical test can be used to detect unbiased Kull-back Leibler divergence whose distribution is consistent with chi-squared with $d + d(d+1)/2$ degrees of freedom (where d is the number of variables). In our experiments, we checked how many outlier intervals can pass the statistical significance test, and we first tested the default program that generates 50 outliers, and then implemented the version that generates 1000 outliers. For the former experiment, all outliers passed the significance test for both ID and Full methods. Nevertheless, for the latter experiment, only about 30%-40% of

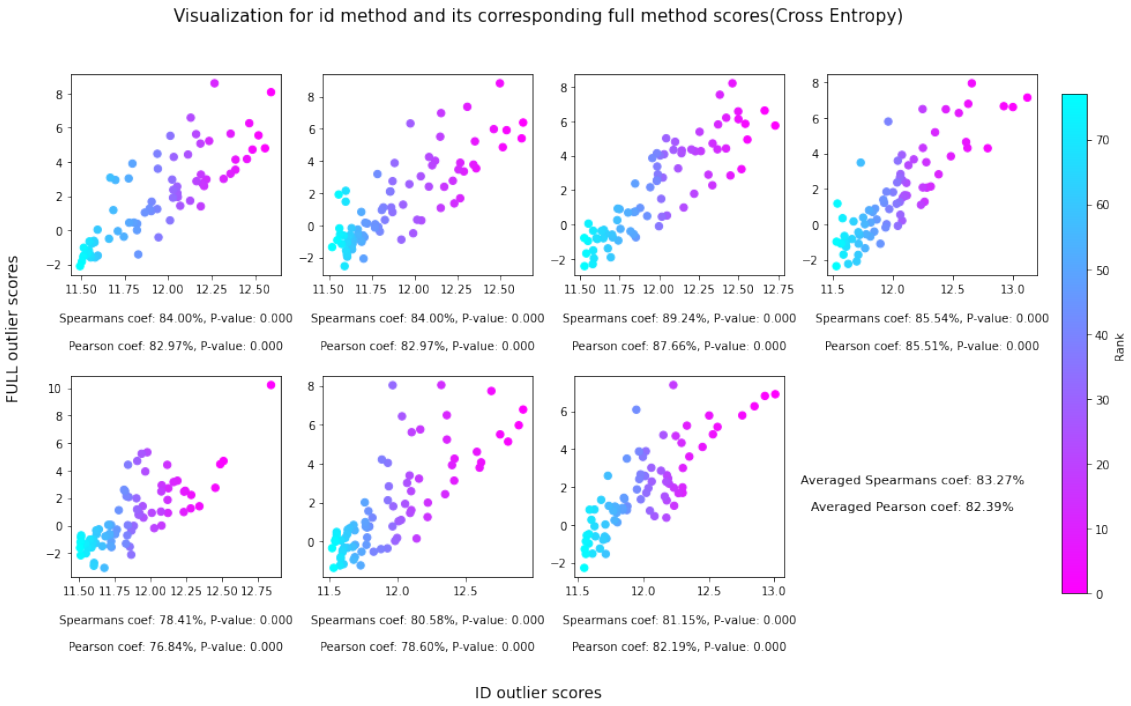

Figure 8: Visualization of rank correlation after re-calculating outlier scores (Cross Entropy)

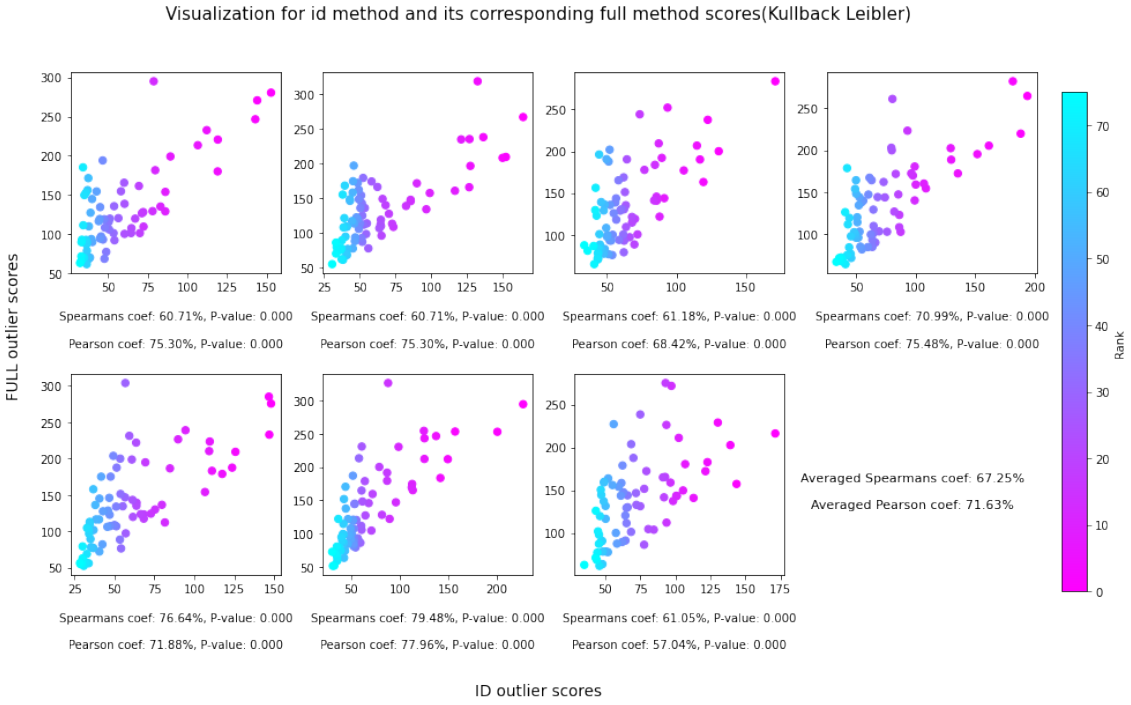

Figure 9: Visualization of rank correlation after re-calculating outlier scores (Kull-back Leibler)

the outlier intervals generated by the ID method passed the significance test, while almost all of the results of the full method passed the test. Specifically, when alpha is 0.05, only 43.77% oputlier intervals passed the significance test, while alpha is 0.01, the percentage becomes to 34.83%. Both percentages for Full method are above 98%. The details can be seen in Table.3. Overall, all outliers generated by Full method almost passed the significance test, while only less than half of the outliers generated by the Full method passed it.

Algorithm 3 How many outliers do we need: Counting the actual number of outliers output by MDI

1: Run MDI program to output 1000 outlier intervals for ID and Full methods (Under Cross entropy and Kull-back Leibler criteria)
2: Count the real number of outliers in the outputs
3: Calculat the mean and median values of real number of outliers
4: Visualize the results

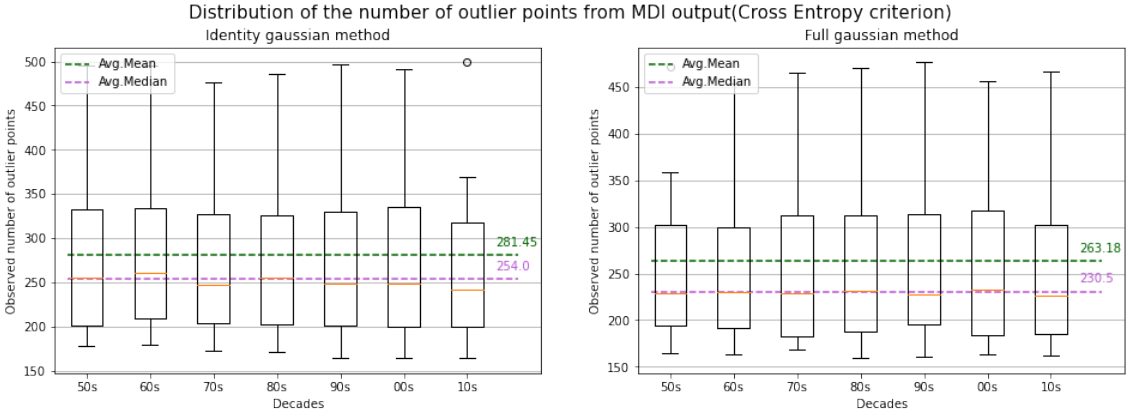

Figure 10: Distribution of the number of outliers from MDI outputs (Cross entropy)

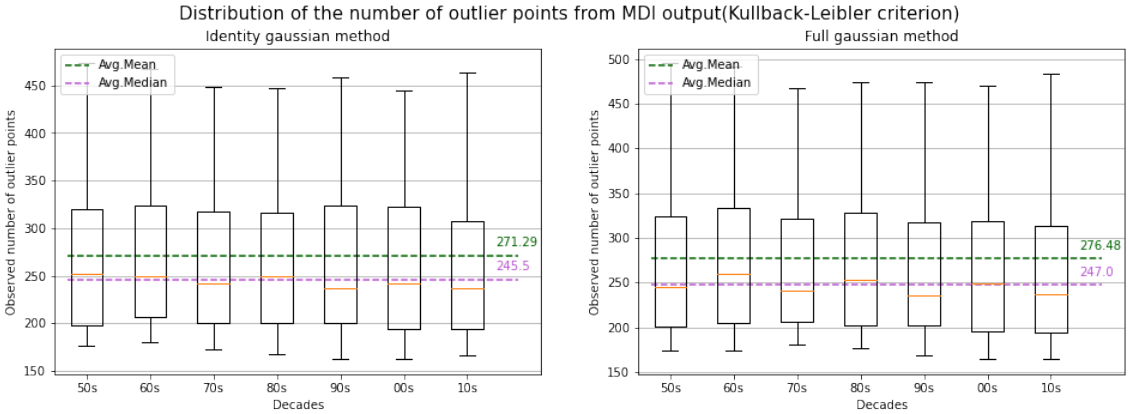

Figure 11: Distribution of the number of outliers from MDI outputs (Kull-back Leibler)

Table 3: Significance test for results under Kull-back Leibler

	ID method	Full method
Number of outliers	15192	15483
Number of outliers passed the test(alpha=0.05)	6649	15297
Percentage of outliers passed the test(alpha=0.05)	43.77%	98.80%
Number of outliers passed the test(alpha=0.01)	5291	15070
Percentage of outliers passed the test(alpha=0.01)	34.83%	97.33%

6 Correlation between outlier lengths and their scores

The length of an outlier interval is the distance between the left and right points of the outlier interval, for instance, the length of the outlier interval (a, b, Outlier score) is b - a. As the last experiment in this paper, we investigate whether there is a correlation between the outlier interval length and the outlier divergence score. Figure.12 and Figure.13 show the visualization results of the ID method experiments, the results of the Full method can be found in the appendix(Figure.D.1, Figure.D.2). A noteworthy point is when running the full method under the kull-back leibler criterion, it produces outliers with negative divergence values.

To illustrate the results, firstly, for the outliers using cross entropy, there is a negative correlation between their lengths and scores. Coefficients of both ID and Full methods are from 15% to 18%, which indicates divergence

scores will probably slightly become lower when outlier lengths get larger. Meanwhile, for the kull-back leibler criterion, opposite correlation observed (from 24% to 27% for ID method and from 37% to 38% for the Full method). Overall, correlations under all methods and criteria passed the spearman and pearson statistical test(p-valuei0.05), however, all correlation coefficients are below 40%.

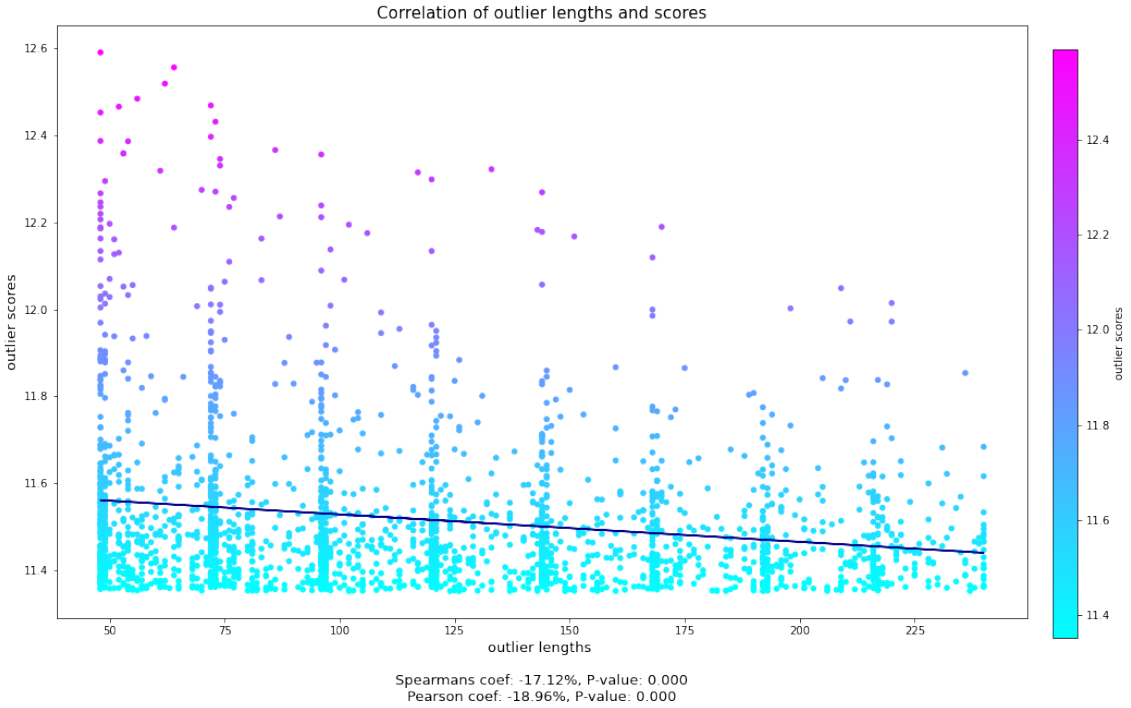

Figure 12: Correlation between outlier lengths and their scores (ID method under Cross Entropy)

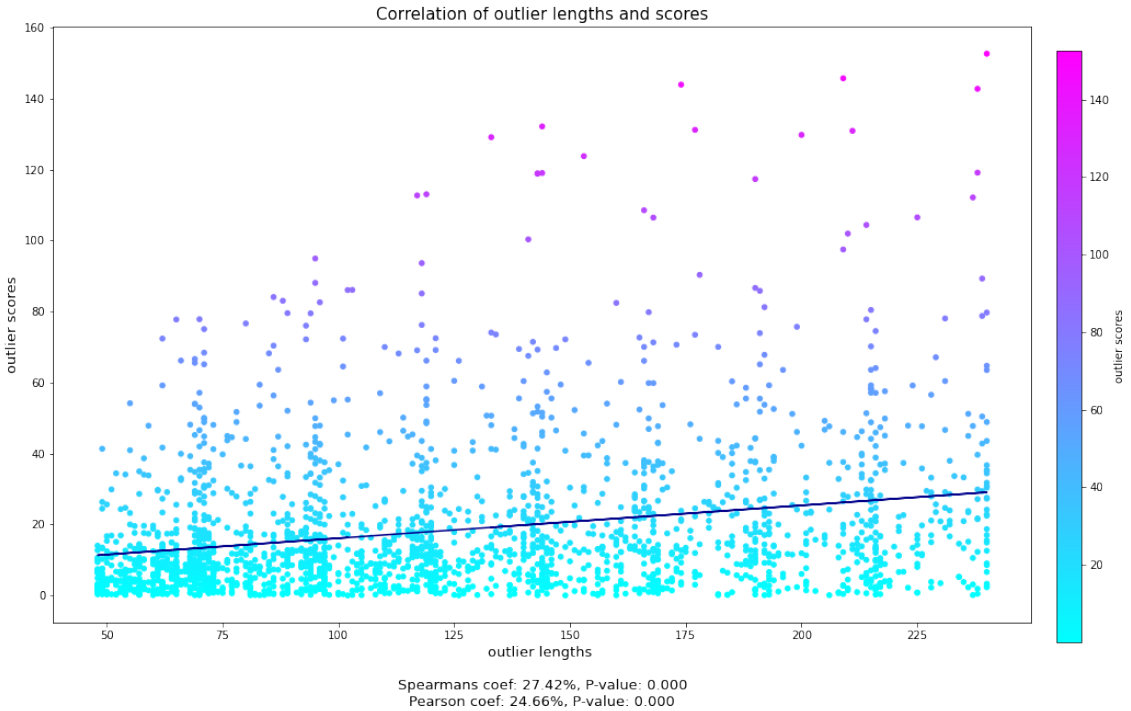

Figure 13: Correlation between outlier lengths and their scores (ID method under kull-back Leibler)

Conclusion

In this report, we conducted an in-depth study on the evaluation for outlier detection in energy climate data. First, we designed and implemented Matching analysis as well as rank correlation in an attempt to explore the correlation between ID and Full method results in a convenient way. The matching algorithm compares the outlier intervals output by two methods in a simple and intuitive way, and determines whether the ID method has the ability to generate the outliers detected by the Full method by calculating their overlap. In terms of results, the ID method can detect about 40% to 60% of the outliers generated by the Full method.

The rank correlation allows us to consider the correlation of the rankings of matched events by sorting the outliers by their detected scores. This experiment reveals whether the ID method generates similar outliers as the Full method and whether their rankings are also similar. It is important in real-world applications because the higher ranked outliers represent more significant characteristics, which means more attention should be paid to them. Furthermore, We examined the IoU scores for these matching events to analyze to what extent they are similar. Overall, similarities are above 60% in the matching events (83% for kull-back leibler and 64% for cross entropy).

However, this correlation is difficult to observe visually from the results, so we performed a strategy of combining outliers over different time spans to observe the similarity of outliers generated by different methods at ”the decadal level”. The results show outliers generated by the two methods are similar in ranking to a certain extent (45%-50%), which indicates that the outliers generated by both methods have similar rankings to some extent. Moreover, we implemented re-calculation process to re-compute outlier scores of other methods using the outlier intervals obtained by one method, and then ran the rank correlation experiment to observe their correlation.The results show that under the cross entropy criterion, there’s a noticeable positive correlation relationship(over 80%) for ID method and its corresponding Full method scores, while this tendency is weakened under the kull-back leibler criterion(67%).

Another question we investigated is how many outliers are appropriate to generate for the ERA5 dataset? To solve this problem, we first set the output of 1000 outliers in the MDI program, and then counted the actual number of output outliers. It turns out that in practice we only get outliers with a mean value of about 270. Finally, we analyzed whether there is a correlation between the length of the outlier intervals and their divergence scores. In general, no significant correlation was found, even though outliers with higher scores tended to exhibit stronger correlations. Specifically, for outliers with higher scores, the greater the length of the outlier interval, the lower the outlier score under the cross entropy criterion, while the greater the length of the outlier interval, the higher the outlier score under the kull-back leibler.

To conclude, we have conducted evaluation for outlier detection in energy climate data in this study. In general, The ID method is able to generate outliers that overlap with the Full method, and their distribution is close to that of the full method, to some extent. In addition, Using the ID method is more advantageous to predict outliers with larger outlier scores, while it is less obvious for outliers with smaller outlier scores.

References

[1] Björn Barz et al. “Detecting regions of maximal divergence for spatio-temporal anomaly detection”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.5 (2018), pp. 1088–1101.

[2] Erik Duijm. “Outlier Detection in Energy Climate Data”. MA thesis. 2021.

Appendices

A Visualization result of matching algorithms

Figure.A.1 and Figure.A.2 show two visualizations of the matching analysis, including the ID to Full method under the Cross entropy and kull-back Leibler criterion and the opposite experiment, respectively.

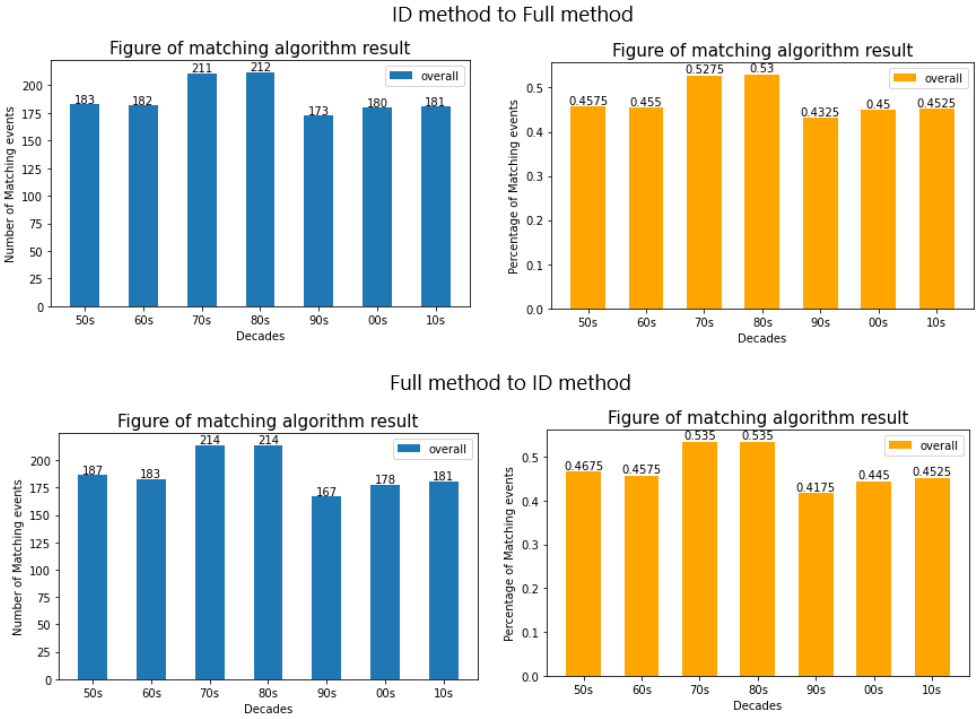

Figure A.1: Visualization result of matching algorithm under Cross Entropy

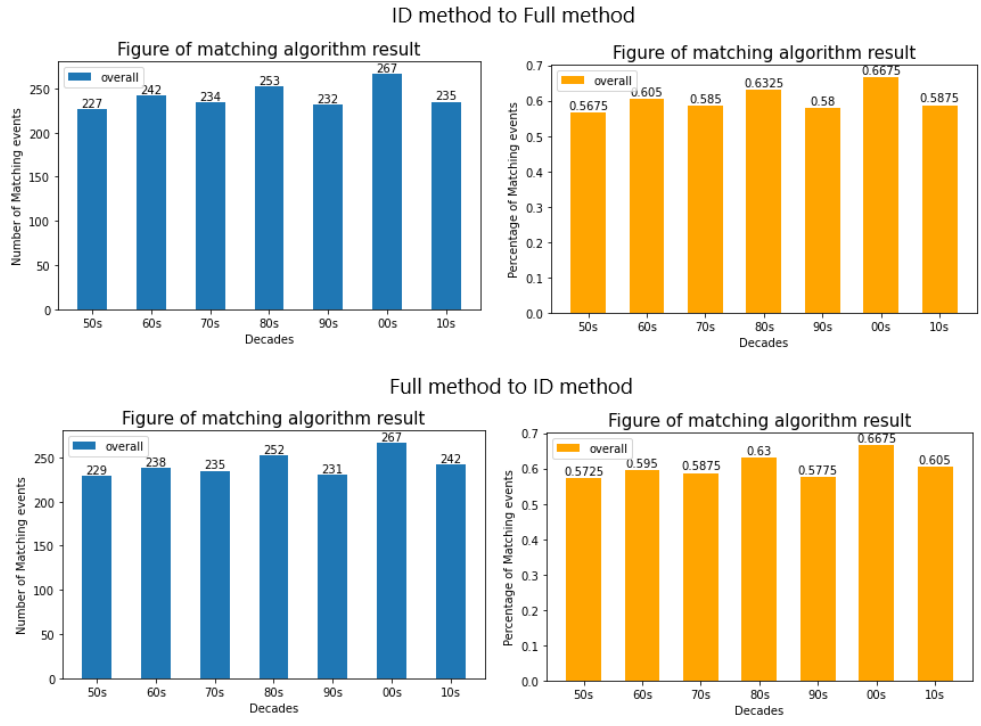

Figure A.2: Visualization result of matching algorithm under Kull-back Leibler

B Visualization result of Rank correlation

The following two figures(Figure.B.1 and Figure.B.2) show the results of rank correlation visualization after implementing matching analysis for both criteria, without merging.

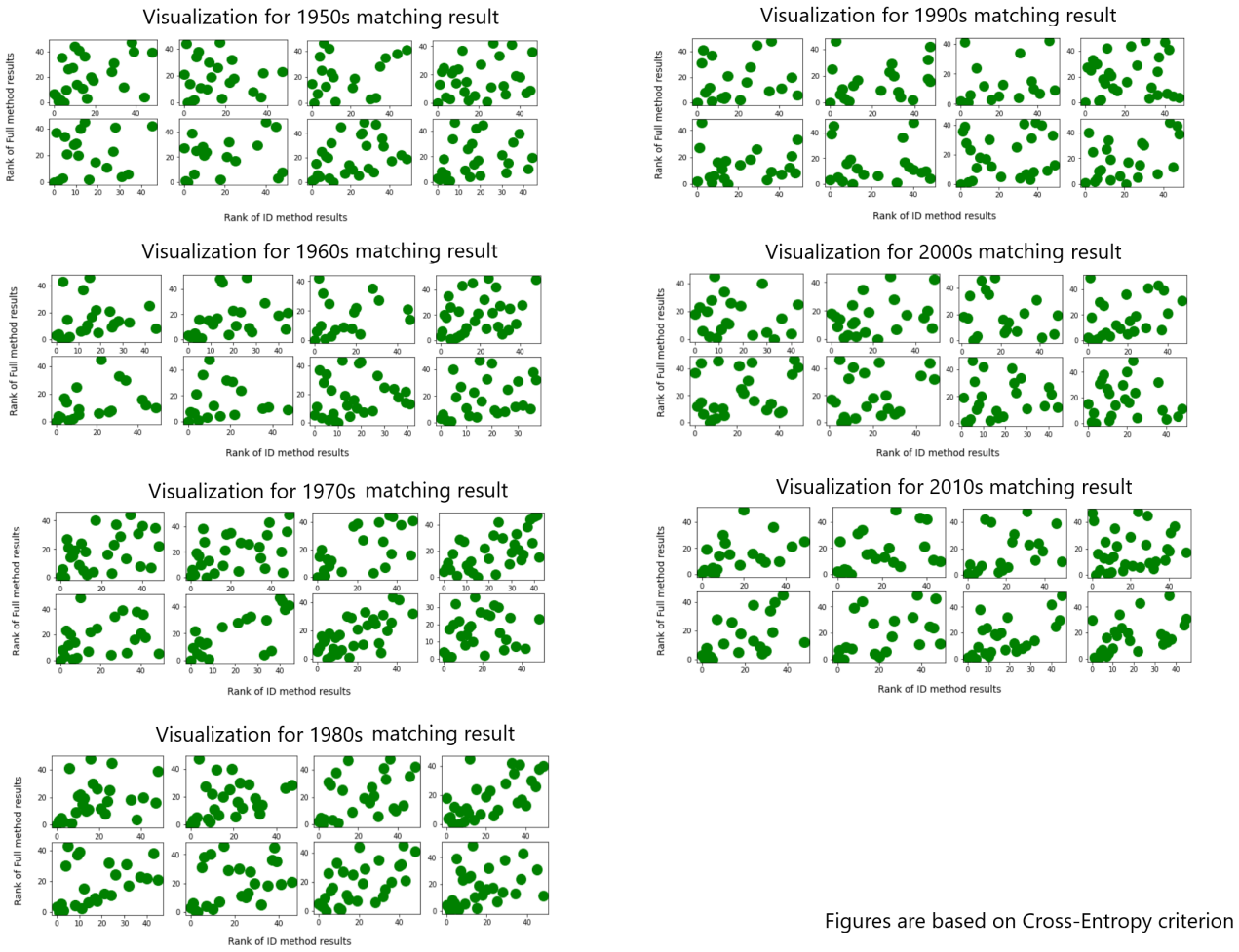

Figure B.1: Rank correlation of Matching analysis under Cross Entropy

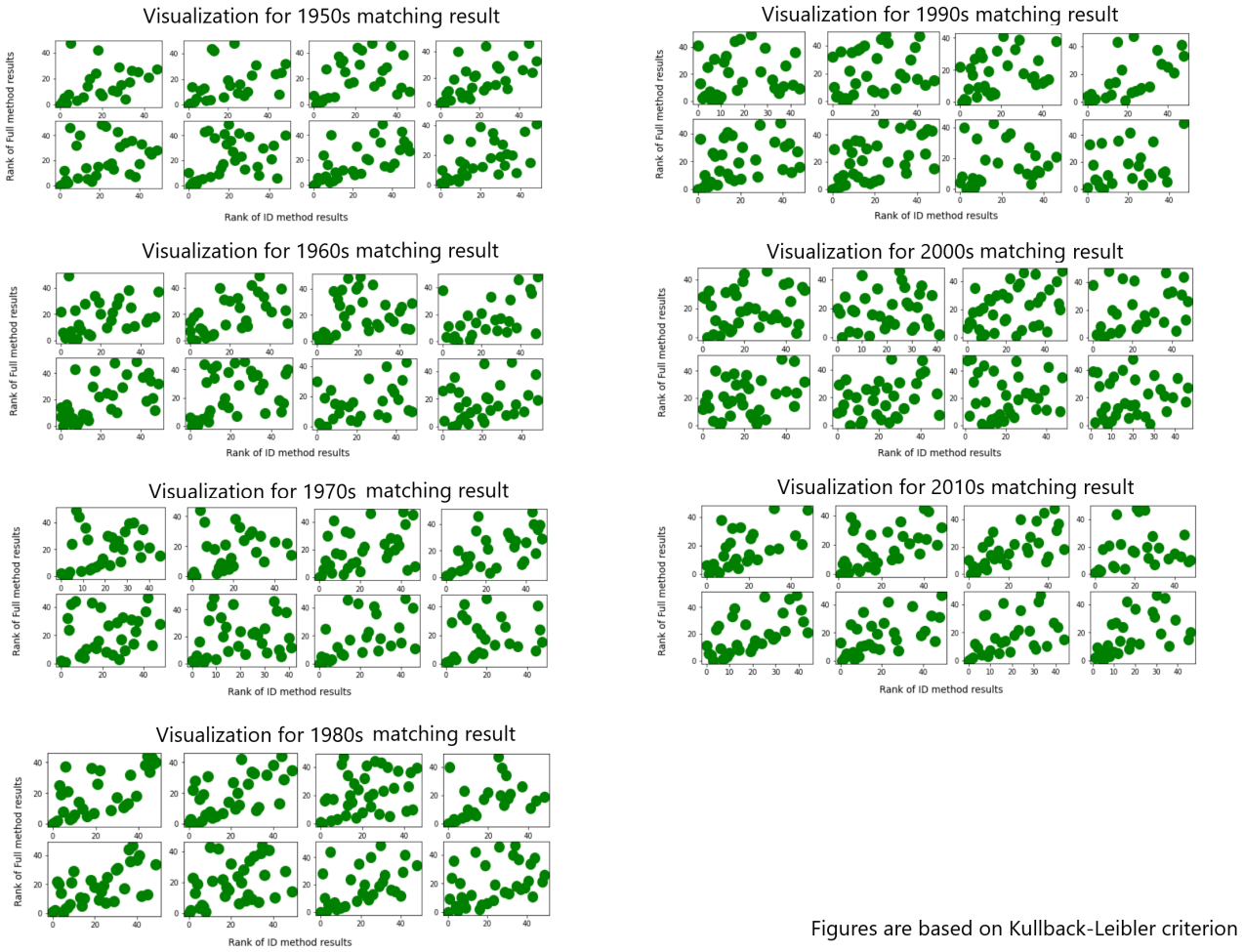

Figure B.2: Rank correlation of Matching analysis under Kull-back Leibler

C Rank correlation after megering outliers over different time span

The following two figures(Figure.C.1 and Figure.C.2) show the correlation from Full method and ID method.

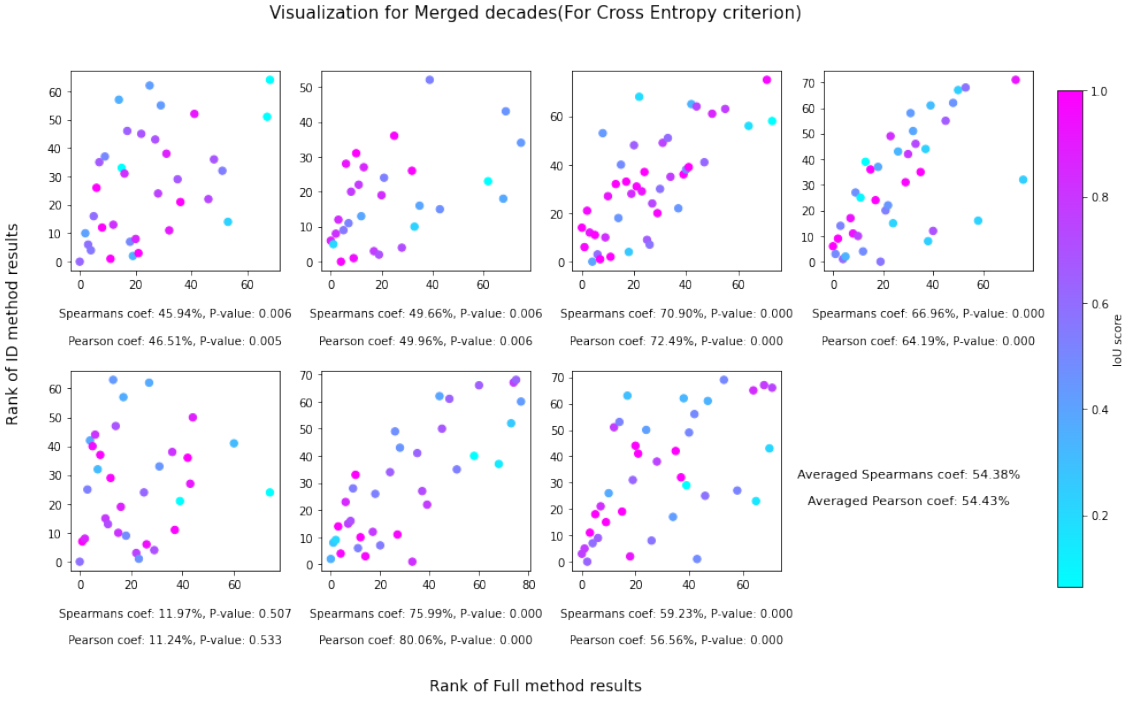

Figure C.1: Visualization of rank correlation after megering outliers over different time span

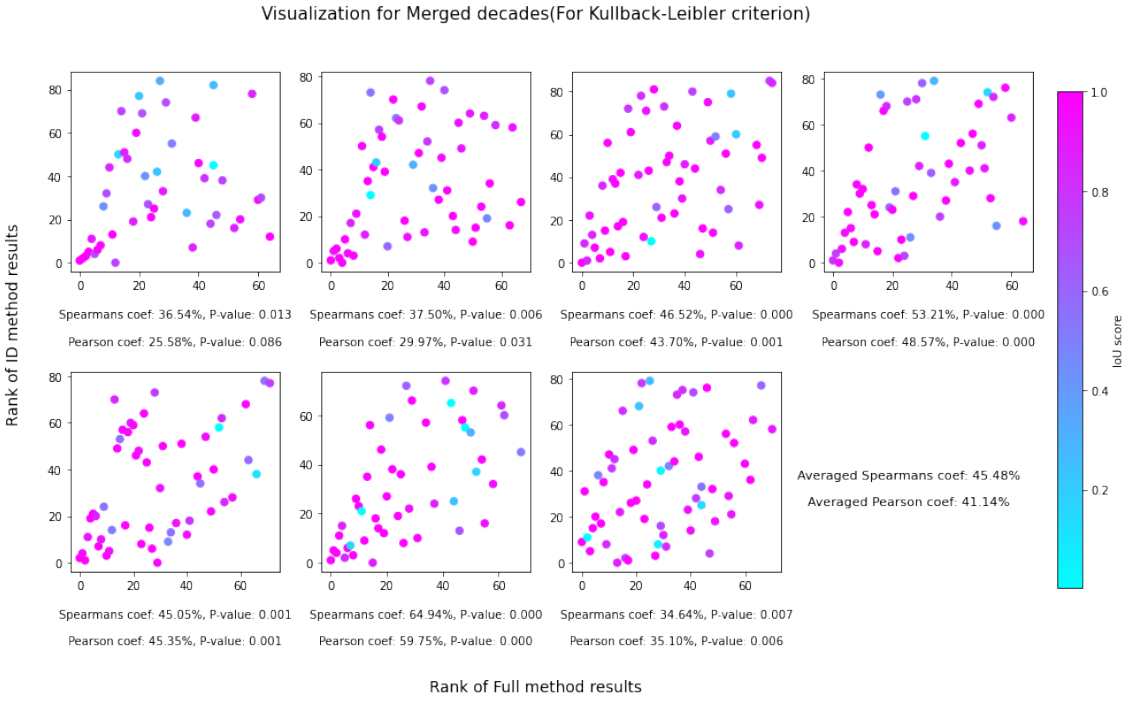

Figure C.2: Visualization of rank correlation after megering outliers over different time span

D Correlation between outlier lengths and their scores

Figure.D.1 and Figure.D.2 shows the Correlation between outlier lengths and their outlier scores Visualization results. It also includes the results of fit lines and Pearson and Spearman correlation analysis.

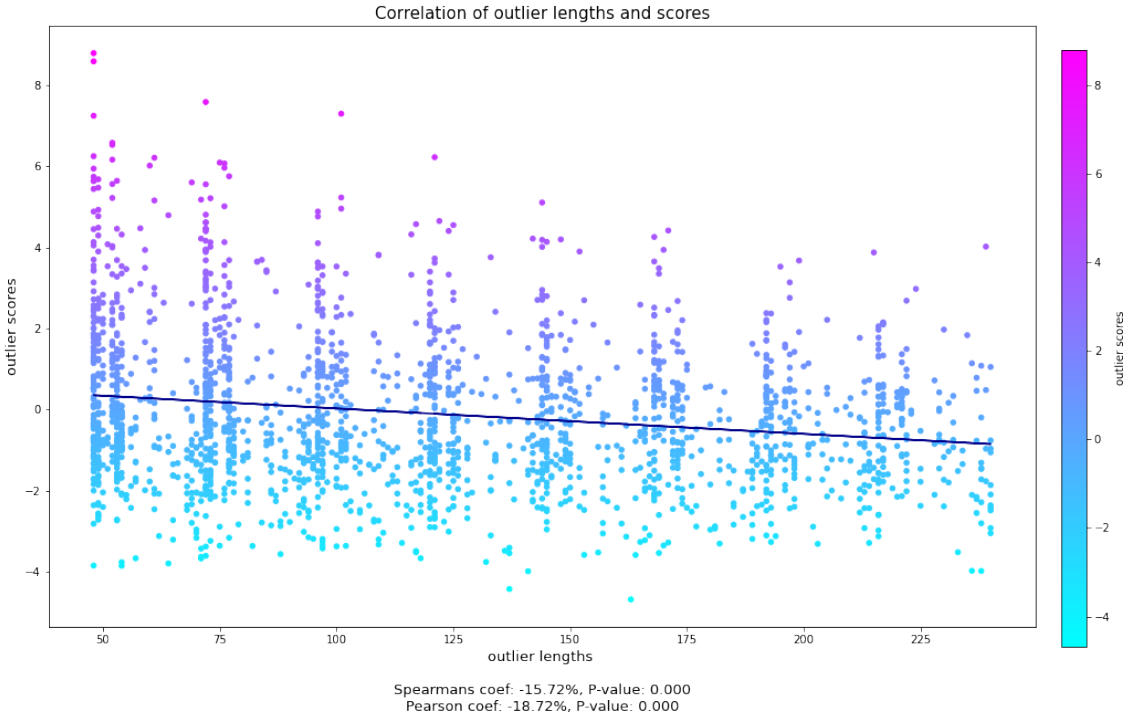

Figure D.1: Correlation between outlier lengths and their scores (Full method under Cross Entropy)

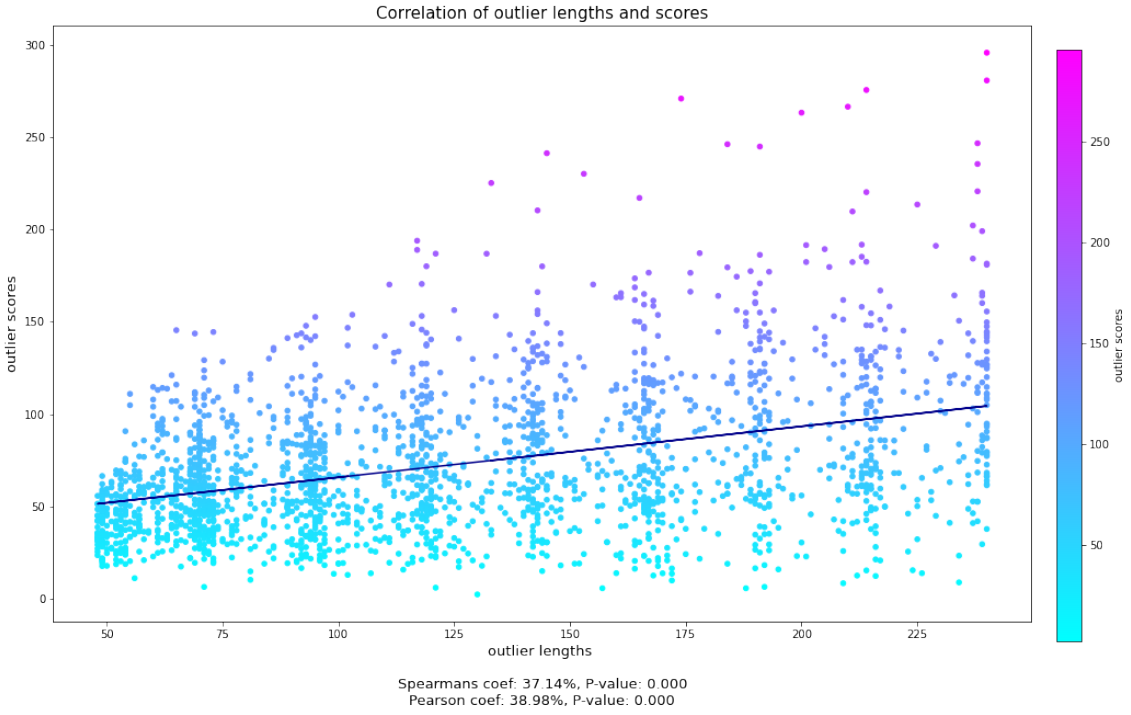

Figure D.2: Correlation between outlier lengths and their scores (Full method under kull-back Leibler)