IASD - NoSQL Project - Report
Debunking Time Series Distance Myths

Alexandre Olech

March 2025

1 Method and Main Results

The goal of this work is to challenge widely adopted beliefs about Distance Measures in the Time-Series community (which mainly come from the works of Ding et al. [2] and [3]), by confronting them to a comprehensive experimental and statistical evaluation.

In particular, authors challenge the four beliefs, which from their work appear as misconceptions :

- z-normalization is the go-to normalization.
- ED is the go-to lock-step measure.
- Elastic measures are better than sliding measures.
- DTW is the best elastic measure.

To challenge these claims, they consider a set of 71 distance measures (lock-step, sliding, elastic, kernel, embedding) and 8 normalization methods. Then, they evaluate as follows :

Algorithm 1 Experimental Evaluation Setup

```
for all Distance Measure do
    for all Normalization do
        Compute the average accuracy of 1NN over 128 datasets of the UCR archive
    end for
end for
Assess the significance of performance differences with statistical tests :
  • Wilcoxon test (95% confidence level) [7](better than the t-test according to Rice
    2006 [6])
  • Friedman test [4]+ Nemenyi test (90% confidence level) [5]
```

Now let's list the main results :

- Normalization significantly affects performance. For some distances, z-score is not the best.
- ED is not the best lock-step distance.
- Elastic measures are not ALWAYS better than sliding measures.
- DTW is not the best elastic measure.
- Best results : elastic and kernel configurations
- Embedding methods are almost as good and provide a very promising runtime-to-accuracy trade-off.

2 Strengths and limitations:

2.1 Weak points

I am convinced that the chosen method is the correct method for the given goal. It uses the appropriate tools for evaluation, based on other works in the community (1NN, Statistical tests...). For weak points, I only have a few remarks to improve the reproducibility :

- Not very clear on the fact that the data is pre-normalized with z-score in most of the UCR Archive. (Also, it might not be the case in the current version of the UCR, from my limited observations). They do not assess to what extent this limits the scope of the results. It would be beneficial to do so.
- It would have been nice to upload a checkpoint of the UCR archive to further ensure reproducibility.
- The website for exploration of the full results is not accessible anymore. It simply redirects to the GitHub code. It would be beneficial for researchers and reviewers to be able to access the website with full results.
- 1NN works decently on 95% of datasets in the UCR archive, but for 4 of them, the accuracy is 0.00%, which could indicate they are not relevant datasets. Clarifying whether and why they included these or not in the comparison would help in reproducing their results.

2.2 Strong points

- Comprehensive experimental setup (128 datasets, 71 measures, 8 normalizations).
- A very valuable effort to evaluate, not just the accuracies, but also the statistical significance of observed accuracies, using works on the subject by [1], [4], [5], [6] and [7]. This behavior should be further encouraged in the community.
- The courage to drop a *"Pavé dans la marre"* by showing that a great part of the community has widespread misconceptions.

3 Questions

3.1 Is the paper solving an ongoing research problem?

The paper is not exactly solving the question of finding the best distance measure and normalization for Time Series, but rather challenging widely spread beliefs and widely-unquestioned results of previous works.

By doing so, it contributes importantly to the research problem of understanding and comparing different distance measures and normalizations. And it does so by providing a more comprehensive experimental evaluation, backed by rigorous statistical tests.

3.2 Is the paper improving existing state-of-the-art performances of a task?

Yes, the paper successfully identifies normalizations and distances measures which are statistically better than the state-of-the-art.

3.3 Is the paper opening new research directions and problems?

Yes, the paper opens many research directions which were regarded as not very important previously, or seen as "solved" despite lack of rigorous evaluations, like identifying the best distance measure and normalization depending on the type of data, and evaluating the effect of adopting a greater variety of distance measures into Time Series projects. It also showcases the very interesting runtime-to-accuracy of embedding methods, which opens up promising research areas based on improving these methods.

3.4 Are the experimental results covering the claim of the paper?

Authors provided a very comprehensive evaluation of 71 distance measures, 8 normalization methods, over 128 time-series datasets, using the mainstream evaluation method (1NN accuracy) and well-sourced statistical tests (based on [1], [4], [5], [6] and [7]) to assess the significance of the results. For these reasons, my answer is yes, the experimental results cover the claim of the paper. I have just noted one simplifying assumption, which is that, according to authors, a great part of the datasets of the UCR archive are pre-normalized with z-score, which might limit the scope of the results concerning normalization methods.

3.5 Are there missing related work?

Based on my knowledge and research, the authors correctly identified the main references on the question, notably the decade-old very influential study of Ding et al. [2] and their more recent work [3] which takes as granted the previous inconclusive findings. As they point out, works on the subject are very limited, and the references are mostly outdated, and with very limited statistical analysis. For these reasons, the community has widely endorsed beliefs about time series distance measures, which were backed up by very limited evidence, which has reduced the identified need for research on this question.

4 Implement the proposed methods (if there is any new proposed method).

4.1 Find (if possible) the implementation used by the author.

All the scripts of the evaluation method are easily accessible on the GitHub. However, it is implemented in Matlab, with imported C and Java. Since I have no experience with Matlab, and since the evaluation method is well explained and not very complex, I decided instead to re-implement the evaluation method from scratch with a language I'm more comfortable with, rather than running a code in a language I'm not comfortable with. It's also good for reproducibility. But I used their code as a reference for my implementation of distance measures.

Actually, there is one small issue in terms of accessibility and reproducibility of the evaluation method, which doesn't come from the authors code, but rather from the UCR archive. It seems it has been updated since the article was written. It doesn't contain exactly the same number of datasets. Also, the datasets do not exactly match the data summary. It seems the data summary is a little bit outdated.

4.2 Reproduce the experimental evaluation.

My goal was to reproduce the experiment on lock-step distances of section 5, with results presented in table 2. I implemented a grid of experiments, for five normalizations and six lock-step distance measures.

Normalizations :

- z-score
- min-max
- Unit Length
- Mean Normalization
- Tanh Normalization

Distances :

- ED
- Lorentzian
- Manhattan
- Average L_1, L_{inf}
- Jaccard
- Minkowski

I then implemented algorithm 1 of the article, at the batch level, to compute the 1NN accuracy for each distance, normalization for each dataset.

4.2.1 Did you succeed?

Yes, all signs indicate my implementation successfully runs the experiment. However I had to make some choices which may differ with the setup of the article

- I only consider datasets which are both in the folder and in the data summary (there seems to be inconsistencies between the two). Thus the set of datasets might not be exactly the same as in the article.
- I only consider datasets in the UCR archive which have a rough cost estimate of $c = n_{test} * n_{train} * d < 1e8$. It resulted in a runtime of 4h30.
- I only consider datasets with fixed length according to the data summary.
- It resulted in 80 datasets, vs. 128 in the original article.
- I only computed the accuracies, I did not reproduce the statistical tests.

4.2.2 Are the results consistent with those in the paper?

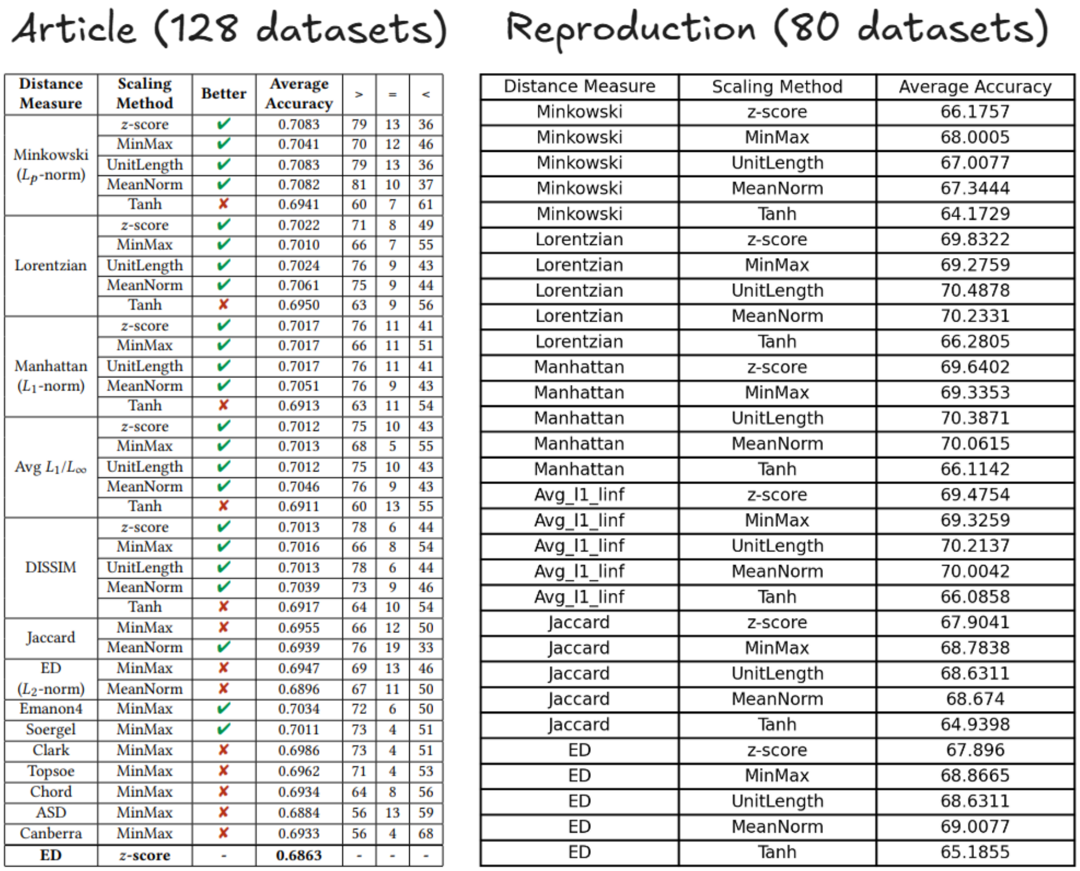
Figure 1 shows the comparison between table 2 of the article and the table obtained from my experiment.

Consistent with the paper (excluding Minkowski : requires tuning):

- ED and Jaccard outperformed by Lorentzian, Manhattan and Average L_1, L_{inf}.
- z-score not always the best.
- z-score is systematically outperformed by MeanNorm and UnitLength.
- Tanh normalization offers the poorest results.
- Excluding Tanh, the global range of accuracy values is very close.

Slightly different from the paper :

- Minkowski is not as competitive (doesn't really count : requires tuning and was not tuned).
- Tanh leads to lower accuracies (68% vs. 65-66% in the article).



Article (128 datasets)

Distance Measure	Scaling Method	Better	Average Accuracy	>	=	<
Minkowski (L_p-norm)	*z*-score	✓	0.7083	79	13	36
	MinMax	✓	0.7041	70	12	46
	UnitLength	✓	0.7083	79	13	36
	MeanNorm	✓	0.7082	81	10	37
	Tanh	✗	0.6941	60	7	61
Lorentzian	*z*-score	✓	0.7022	71	8	49
	MinMax	✓	0.7010	66	7	55
	UnitLength	✓	0.7024	76	9	43
	MeanNorm	✓	0.7061	75	9	44
	Tanh	✗	0.6950	63	9	56
Manhattan (L_1-norm)	*z*-score	✓	0.7017	76	11	41
	MinMax	✓	0.7017	66	11	51
	UnitLength	✓	0.7017	76	11	41
	MeanNorm	✓	0.7051	76	9	43
	Tanh	✗	0.6913	63	11	54
Avg L_1/L_∞	*z*-score	✓	0.7012	75	10	43
	MinMax	✓	0.7013	68	5	55
	UnitLength	✓	0.7012	75	10	43
	MeanNorm	✓	0.7046	76	9	43
	Tanh	✗	0.6911	60	13	55
DISSIM	*z*-score	✓	0.7013	78	6	44
	MinMax	✓	0.7016	66	8	54
	UnitLength	✓	0.7013	78	6	44
	MeanNorm	✓	0.7039	73	9	46
	Tanh	✗	0.6917	64	10	54
Jaccard	MinMax	✗	0.6955	66	12	50
	MeanNorm	✓	0.6939	76	19	33
ED (L_2-norm)	MinMax	✗	0.6947	69	13	46
	MeanNorm	✗	0.6896	67	11	50
Emanon4	MinMax	✓	0.7034	72	6	50
Soergel	MinMax	✓	0.7011	73	4	51
Clark	MinMax	✗	0.6986	73	4	51
Topsoe	MinMax	✗	0.6962	71	4	53
Chord	MinMax	✗	0.6934	64	8	56
ASD	MinMax	✗	0.6884	56	13	59
Canberra	MinMax	✗	0.6933	56	4	68
ED	***z*-score**	-	**0.6863**	-	-	-

Reproduction (80 datasets)

Distance Measure	Scaling Method	Average Accuracy
Minkowski	z-score	66.1757
Minkowski	MinMax	68.0005
Minkowski	UnitLength	67.0077
Minkowski	MeanNorm	67.3444
Minkowski	Tanh	64.1729
Lorentzian	z-score	69.8322
Lorentzian	MinMax	69.2759
Lorentzian	UnitLength	70.4878
Lorentzian	MeanNorm	70.2331
Lorentzian	Tanh	66.2805
Manhattan	z-score	69.6402
Manhattan	MinMax	69.3353
Manhattan	UnitLength	70.3871
Manhattan	MeanNorm	70.0615
Manhattan	Tanh	66.1142
Avg_l1_linf	z-score	69.4754
Avg_l1_linf	MinMax	69.3259
Avg_l1_linf	UnitLength	70.2137
Avg_l1_linf	MeanNorm	70.0042
Avg_l1_linf	Tanh	66.0858
Jaccard	z-score	67.9041
Jaccard	MinMax	68.7838
Jaccard	UnitLength	68.6311
Jaccard	MeanNorm	68.674
Jaccard	Tanh	64.9398
ED	z-score	67.896
ED	MinMax	68.8665
ED	UnitLength	68.6311
ED	MeanNorm	69.0077
ED	Tanh	65.1855

Figure 1: Comparison between table 2 of the article and the table obtained from my experiment.

4.2.3 Are the results different, but the trends and the claims still hold?

Even with the slight differences in the results, the claims about beliefs 1 and 2 being false completely hold since they show that :

- z-score is not always the best normalization.
- ED can be outperformed by other lock-step measure.

5 Final Verdict

Authors made an effort at developing a comprehensive experimental setup, backed by rigorous statistical tests. I am convinced their method is correct, and I was able to reproduce a part of the experimental results, in a way which is fully consistent with their claims. Thanks to this work, they have debunked myths based on inconclusive findings, about a very crucial aspect of Time Series processing. Based on the great quality of their work and on their important contribution to the time series research community, I fully accept the article for publication.

References

[1] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[2] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

[3] Rui Ding, Qiang Wang, Yingnong Dang, Qiang Fu, Haidong Zhang, and Dongmei Zhang. Yading: Fast clustering of large-scale time series data. *Proceedings of the VLDB Endowment*, 8(5):473–484, 2015.

[4] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32:675–701, 1937.

[5] Peter Nemenyi. *Distribution-free Multiple Comparisons.* PhD thesis, Princeton University, 1963.

[6] John Rice. *Mathematical Statistics and Data Analysis.* Cengage Learning, 2006.

[7] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, pages 80–83, 1945.