The Title

by

Magnus E. Bugge

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Undergraduate Research Project
in Data Science.

Victoria University of Wellington
2024

Abstract

Investigation of tie breaking and omitting Na values on differenet libraries performance. New tie breaking idea is implemented and compared with other tie breaking methods. Data used for this is pulled from previously articles in the field.

Contents

Chapter 1

Introduction

Time series is an important category of data, since climate, finance, and health data all can be structured as time series. Classical regression, clustering and classification of time series data often works by analysing the observed values directly. Ordinal Patterns(OPs) take a different approach to time series analysis. Instead of using the observed values, OPs transform those values into symbols. These symbols retain information about the relative values, and aspire to capturing relevant information about the time series and the process that produced it.

Ordinal Patterns have been around for twenty years. They were proposed by Bandt and Pompe in 2002 [1]. This seminal article has received, to date, over 3000 citations on the Web of Science. Their main properties are invariance, robustness, and being model-agnostic.

OPs have been used in applications in several domains. Many of these applications involve comparing time series, signals, and images through features extracted from their Ordinal Patterns. Since results about the distribution of such features are recent, our initial interest was comparing existing conclusions with those supported by the new statistical tests. While approaching this research avenue, we found indications of reproducibility crisis [2] in this field, mostly related to data availability. We then moved into the application of these new tests to already published results, and

found evidence of the other type of reproducibility crisis: diverging results depending on the type of library employed. MAGNUS; then, we decided to study the effect of preprocessing of the data by a combination of omitting Na values and adding noise to break ties. A new tie breaking method is proposed and its performance on computing entropy and p values is investigated.

We are interested in verifying the reproducibility of results of Ordinal Patterns in the scientific literature. Initially, we wanted to check how many of the 3000 articles, were reproducible. This was a big task, so at first it was narrowed down to just papers focusing on climate. In the end, around 31 papers were investigated deeply.

Chapter 2

Ordinal Patterns

2.1 History

Bandt and Pompe's first paper about ordinal patterns is from 2002. They introduced the concept of turning a time series or signal into a sequence of symbols. They furthermore used the Shannon entropy on the symbol distribution [1]. The Shannon Entropy was developed in 1948 by C. E. Shannon [3] as a way to quantify uncertainty and unpredictability.

The 1995 paper "A statistical measure of complexity" by López-Ruiz et al. [4] introduced the concept of a statistical measure of complexity. This idea was applied to OPs in 2003 [5], and in 2004 the version of it using Jensen-Shannon divergence was published [6]. We use the latter in this work.

2.2 Definition

Every analysis that employs OPs starts by defining the (integer-valued) word size $D \in \{2, 3, \dots\}$ and time lag τ. We will only use $\tau = 1$, and thus will drop, it from the following discussion.

Consider the finite time series of real values $\boldsymbol{x} = (x_1, x_2, \dots, x_{n+D-1})$.

The word size D is chosen, as seen in the bibliometric analysis below, a word size between 3-6 is most common. The Bandt & Pompe transformation (also called "BP Symbolisation") converts each subset of size D of contiguous and different values into a symbol, i.e., the first tuple of D values will be $(x_1, x_2, \ldots, x_D)$, the second tuple will be $(x_2, x_3, \ldots, x_{D+1})$, and so on.

There are at least two ways of computing OPs. We will describe one of them. Each tuple is transformed into a pattern by ranking them by numerical order. The lowest observation gets assigned the number 0 and the highest observation gets the number $D - 1$. The pattern can then be written as a string of these numbers. The tuple $(0.51, 0.79, 0.14)$ will have pattern $1, 2, 0$. How we write the patterns is transparent, provided there is only one pattern for each possible sorting of the tuple. For example, we could have defined that $(0.51, 0.79, 0.14)$ becomes π^3 or c or $\mathfrak{c}$. Assuming there are no ties in each word, there are $D!$ possible patterns that we will denote $\mathbf{\Pi} = \{\pi^1, \pi^2, \ldots, \pi^{D!}\}$.

By applying the BP symbolisation, we transform the time series $\boldsymbol{x} = (x_1, x_2, \ldots, x_{n+D-1})$ into the sequence of symbols $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)$. The observed frequency pattern $\widehat{p}_i$ is defined as

$$\widehat{p}_i = \frac{\#\{t : \pi_t = \pi^i\}}{n},$$

with which we define the vector of observed proportions:

$$\widehat{\boldsymbol{p}} = (\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_{D!}).$$

A tie is when a tuple contains identical values. In practice, there may be ties, for instance when observing times series registered with finite precision. There are several ways to handle ties. They can be ignored, i.e., no pattern is computed for tuples with ties, or they can be broken by adding small random perturbations. Techniques that handle ties are usually referred to as "imputation solutions."

OPs are usually considered useful for time series of at least length $n > 100D$.

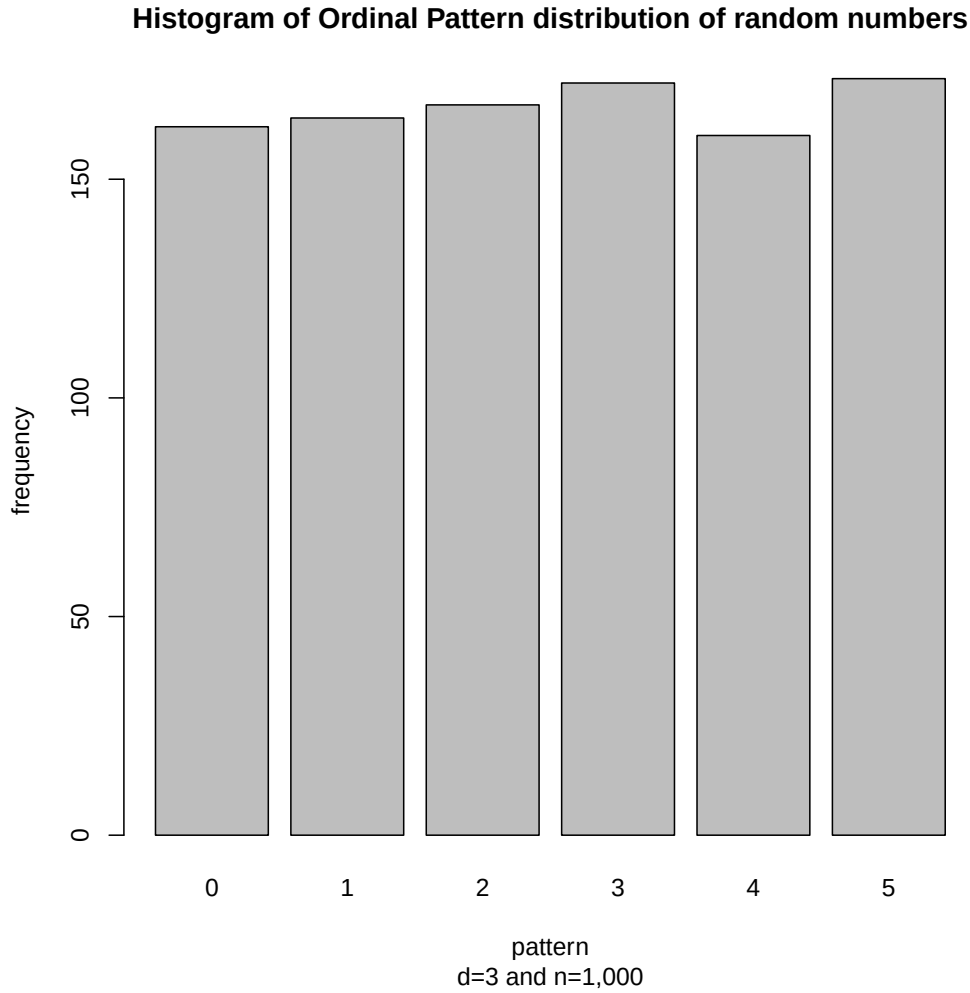

Figure 2.1: Simple histogram of the ordinal pattern distribution of random numbers

From the ordinal pattern distribution, features can be extracted. The two main used features in this field are Entropy and Complexity. In this paper, Shannon Entropy will be used, which is defined as.

$$h(n) = -\sum p(\pi)\ln(p(\pi))$$

Normalized version

$$H(n) = -\frac{\sum p(\pi)\ln(p(\pi))}{\ln(D!)}$$

There are several statistical complexity measures that can be used. They are all a product between the used entropy and a distance measure. In this paper Martin-Plastino-Rosso intensive Statistical Complexity Measure is used, where the distance measure is Jensen-Shannon divergence, and it is measured between the pattern distribution and the uniform distribution. The Q_0 is normalizing term. It is defined as

$$C[\mathscr{P}] = H[\mathscr{P}] \cdot Q_J[\mathscr{P}, \mathscr{P}_e]$$

$$Q_J[\mathscr{P}, \mathscr{P}_e] = Q_0 \cdot \mathscr{J}[\mathscr{P}, \mathscr{P}_e]$$

$$\mathscr{J}[\mathscr{P}, \mathscr{P}_e] = S[\frac{\mathscr{P} + \mathscr{P}_e e}{2}] - S[\frac{\mathscr{P}}{2}] - S[\frac{\mathscr{P}_e}{2}]$$

$$\mathscr{P}_e = \{p_j = \frac{1}{W}; j = 1..., W\}$$

$$Q_0 = -2(\frac{W+1}{W}ln(W+1) - 2ln(2W) + ln(W))^{-1}$$

[7]

Applications Going to the original Bandt and Pompe paper [1] on Semantic Scholar and getting the ten most recent citations, gives a good picture of how broadly this methodology is being used.

The topic varies from: ”Walnut crack detection...” [8], Schizophrenia [9], Analysis of Smart Drilling [10], Epileptic Seizure detection [11], mind wandering during video-based learning [12] and Random Numbers generated based on dual-channel chaotic light [13]. The rest that was found [14, 15, 16, 17]

2.3 Reproducibility

Reproducibility crisis In data science, data from many scientific fields are analysed, as shown in the application section of ordinal patterns above. It therefore increases the difficulty of reproducing the data collection stage, as most cases requires an interdisciplinary study involving multiple people. In some cases it is impossible e.g. if the equipment needed is unavailable or a time series of weather data cannot be recollected, since it is impossible to go back in time. The point of science is to eliminate trust, when sharing knowledge, however the above observations highlights the need for some degree of trust, in cases, where data cannot be reproduced. Only data that is random numbers, generated in a script, will be reproduced in this paper.

The COVID pandemic gave birth to the following quote. ”We warn against the potential misuse or misleading interpretation of public data of variable quality” [18]. Data from different governments does not have the same quality. This necessities good source criticism. In this paper data from peer-reviewed studies will be trusted, which have often either produced the data themselves or got it from reputable organizations as NASA or NOAA.

In this paper, the focus will be on direct computational reproduction.

”Computational reproducibility is most often direct (reproducing particular analysis outcomes from the same data set using the same code and software), but it can also be conceptual (analysing the same raw data set with alternative approaches, different models or statistical frameworks)”

[2]

2.4 Bibliometric analysis

31 articles were analysed for data availability and used word size. The majority of them are about climate, and they explicit have to used Permutation Entropy. Four categories were made for data availability: “Available”, “Sourced”, “On-Request” and “not Available”. Sourced is the broadest category. Some articles give a detailed description of how to pull the data from a website. Other articles give no description, expect, which organisation they got the data from. The is no differentiation between these cases in the category S. A is the best category and only used, when the data is directly downloadable. R is when the dataset has to be requested by the author and is suboptimal. N is the worst category, because there is no source to data, nor is the dataset attached to the paper in any way. Word size between 2-7 were used. Some articles used multiple word sizes, which is why the sum of “No. articles using” is higher than 31. Only 4 articles were the data readily downloadable. The data is definitely possible to get from some articles in category ”S”, but many of them were very sparse in their description of the data, to the point, where it would be hard to pull it from the website. Assuming that around half of the data from articles in ”S” can be pulled from their sources. That leaves around 50% of the article, where it is not possible to reproduce the data processing part of the article. Ideally, all papers should be reproducible and especially their data processing part, since modern technology allows for extremely efficient data sharing across vast distance to a very affordable price. According to this paper [19] more than 70% of researches have failed reproducing an article and 52% agree that there is significant reproducibility crisis. In this paper, cases occurred, where the data had been downloaded and verified to being true, but the data processing could not be reproduced.

Word size	2	3	4	5	6	7	Mutliple
No. articles using	4	13	15	9	6	1	9

A	S	R	N
4	19	3	4

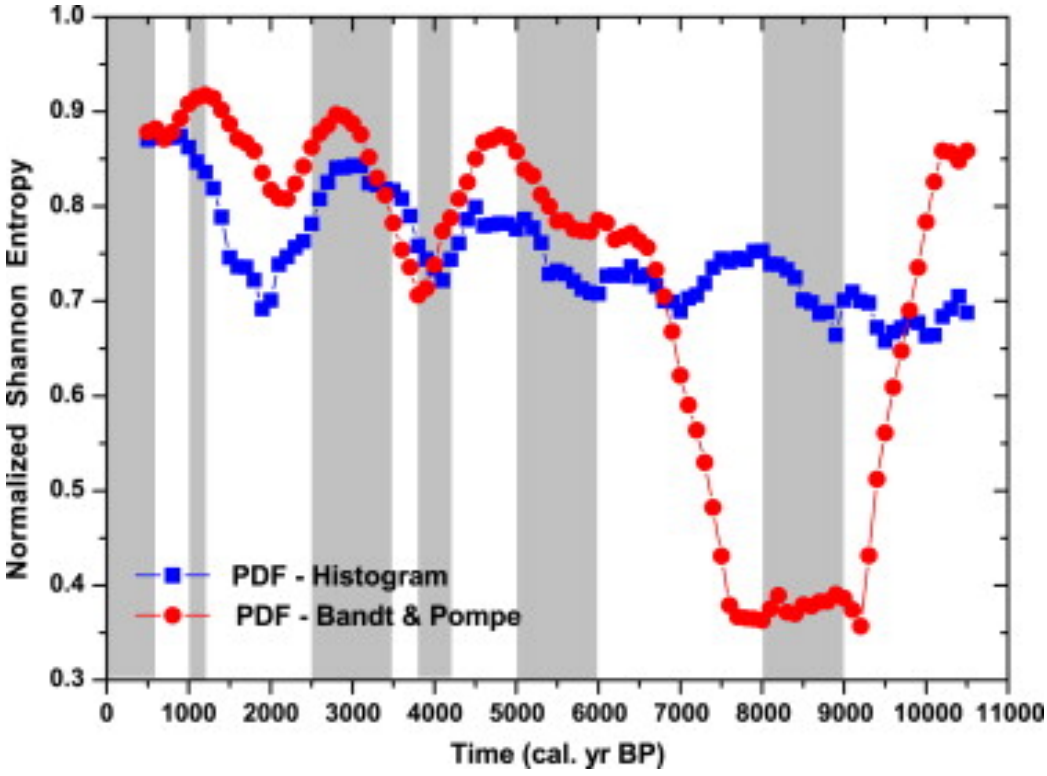

Figure 2.2: Original papers entropy plot, Red dots should be equal to the dots in plots below

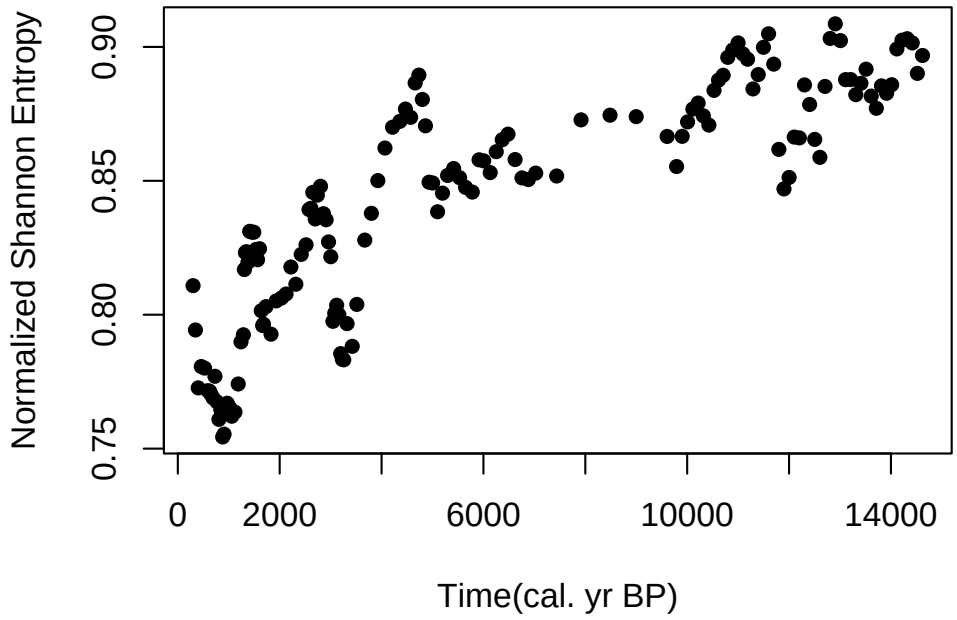

Figure 2.3: This papers reproduction. The x-axis is slightly bigger, however the data in range 0-11,000 cal. yr BP, where verified to be identical, and that is not identical in any way.

Chapter 3

Comparison of two time series

3.1 Examples of how comparisons are done in the literature,

Ordinal Patterns have been used in astrophysics to analyse geomagnetic auroras. Two articles were found to do this. They both compare the plotted points in the HxC plane to fractional Brownian motion time series. ”Many of the points are on or very near the fractional Brownian motion curve, but a single point from the Helios data lies above the fractional Brownian motion curve.” [20]. A small amount of calculation is done in this paper to rate the statistical likelihood of a point being displaced from the fractional Brownian motion(fBm), but no method is presented to evaluate, if a point is displaced. ”Figure 4 indicates that complexity-entropy values of AL overlap the fBm values for all subsampling parameters τ” [21]. Same problem in this article, where no method is presented to systemic rate if a point is overlapping or not with fBm values. Ordinal patterns is often used in climate research to detect changes in the dynamics of a weather system over time. This is done, in this paper [22], by splitting the dataset into windows and calculating the entropy for each window. No method is presented to evaluate if a change in entropy is significant.

It is assumed that a change in entropy means a change in the dynamics of the system, without considering the possibility that it might be caused by stochastic randomness of sampling.

3.2 Statistical Test and Confidence Intervals

Confidence intervals describes the range within a sampling of a distribution has a certain percentage of being in. Statistical Test can be used to reject a wide range of hypothesises. P value is used as rejection criteria, where the lower the p values is the lower is the chance of making a type 1 error, which is rejecting a true hypothesis.[23]

3.3 Confidence Interval of an Ordinal Pattern Distribution Entropy

By assuming that $\mathbf{X}_n = (X_{1,n}, X_{2,n}, ..., X_{K,n})$ with $n \in \mathbb{N}$ is a sequence of independent and identically distributed K-variate vectors of random variables. Furthermore, assuming as n tends to infinity,

$$\sqrt{n}(X_{1,n} - \theta_1, X_{2,n} - \theta_2, ..., X_{K,n} - \theta_K)$$

converges in distribution to the multivariate normal law $\mathscr{N}(\mathbf{0}, S_\mathbf{X})$, where $S_\mathbf{X}$ is the covariance matrix. The following terms are defined: m is the embedding dimension. $\mathbf{q} = (q_1, q_2, ..., q_{m!})$, where q_i is the probability of observing the ordinal pattern π_i. $\mathbf{D_q} = Diag(q_1, q_2, ..., q_{m!})$ Diagonal matrix. $\mathbf{Q}^{(\ell)}$, which is the transition matrix with elements $q_{ij}^{(\ell)} = Pr(\psi = \pi_i \wedge \psi_{t+\ell} = \pi_j)$ for $\ell = 1, 2, ..., m-1$

From this starting point it is derived that

$$\sqrt{n}[S(\hat{\mathbf{q}}) - S(\mathbf{q})] \xrightarrow[n\to\infty]{\mathscr{D}} \mathscr{N}(0, \sigma_\mathbf{q}^2)$$

$$\sigma^2_{\mathbf{q}} = \sum_{i=1}^{m!} (\mathbf{\Sigma_q})_{ii} + 2 \sum_{i=1}^{m!-1} \sum_{j=i+1}^{m!} (\mathbf{\Sigma_q})_{ij}$$

$$(\mathbf{\Sigma_q})_{ij} = \begin{cases} (ln(q_i)+1)^2 \mathbf{\Sigma}_{ii} & \text{if } i = j \\ (ln(q_i)+1)(ln(q_j)+1)\mathbf{\Sigma}_{ij} & \text{if } i \neq j \end{cases}$$

$$\mathbf{\Sigma} = \mathbf{D_q} - (2m-1)\mathbf{q}\mathbf{q}^T + \Sigma_{\ell=1}^{m-1}(\mathbf{Q}^{(\ell)} + \mathbf{Q}^{(\ell)^T})$$

[24]It is the variance that is of interest, as it will be used for statistical tests. The test goes as follows. Let $x = (x_1, x_2, ..., x_{n_x})$ and $y = (y_1, y_2, ..., y_{n_y})$ be two independent time series of length $n_x = T_x + D_x - 1$ and $n_y = T_y + D_y - 1$. p_x and p_y is the ordinal distributions of the time series. $H(\hat{p}_x)$ and $H(\hat{p}_y)$ is their entropies. A new distribution W is constructed. $W \xrightarrow{D} \mathcal{N}(\mu_W, \sigma^2_W)$, with $\mu_W = \mu_{n_x,p_x} - \mu_{n_y,p_y}$ and $\sigma^2_W = \sigma^2_{n_x,p_x} + \sigma^2_{n_y,p_y}$. The p value ends up being $2(1 - \Phi(\xi))$, where $\xi = \frac{H(\hat{p}_x) - H(\hat{p}_u)}{\sigma_W}$. [25]

3.4 Implementation

The three main libraries that will be used specific to the field of ordinal patterns are: statcomp[26], pdc[27] and StatOrdPattHxC, which is a library developed by the supervisor of this project. Slight modifications are made to StatOrdPattHxC. As will be shown later, these three libraries perform quite differently in a lot of cases. Solutions for this are proposed. StatOrdPattHxC has the variance and statistical test implemented, which statcomp and pdc do not. A wide range of standard R libraries is utilized as well.

Chapter 4

Applications

4.1 Power-law noise

The statistical test mentioned in the previous chapter is tested on computer-generated power-law noise series.[28] the β term in the formula $\frac{1}{f^\beta}$ will be referenced ask for the rest of the paper. Three values of k are chosen: $k_1 = (1, 3, 5)$. For each k_1 value, another set of k values is generated using the formula $k_2 = (k_1 - e^0, k_1 - e^{-1}, k_1 - e^{-2}, k_1 - e^{-3}, k_1 + 0, k_1 + e^0, k_1 + e^{-1}, k_1 + e^{-2}, k_1 + e^{-3})$. The statistical test is used on every possible pair of a value from k_1 and a value from k_2. Two power-law series are generated using each value in the pair. The series has 1,000 elements. The hypothesis will be rejected if the p-value is less than 0.05. 10 iterations are done for each pair. A plot will be made for each k_1 value of the rejection rate of the statistical test. A NaN plot is included for each plot, which shows the rate, at which the statistical test produces NaN values. NaN values are only produced in the last graph. Power-law series with k=5 have entropy around 0.4, the rest of the time series that will be examined in the paper have entropy above 0.9, so this should not be a problem. The Rejection plot for $k_1 = 1$ is quite surprising in the sense that many values around k_1 has a high rejection rate. $k_1 = 3$ looks the most as, what was expected since the further away a point gets from k_1 the larger is the rejection rate.

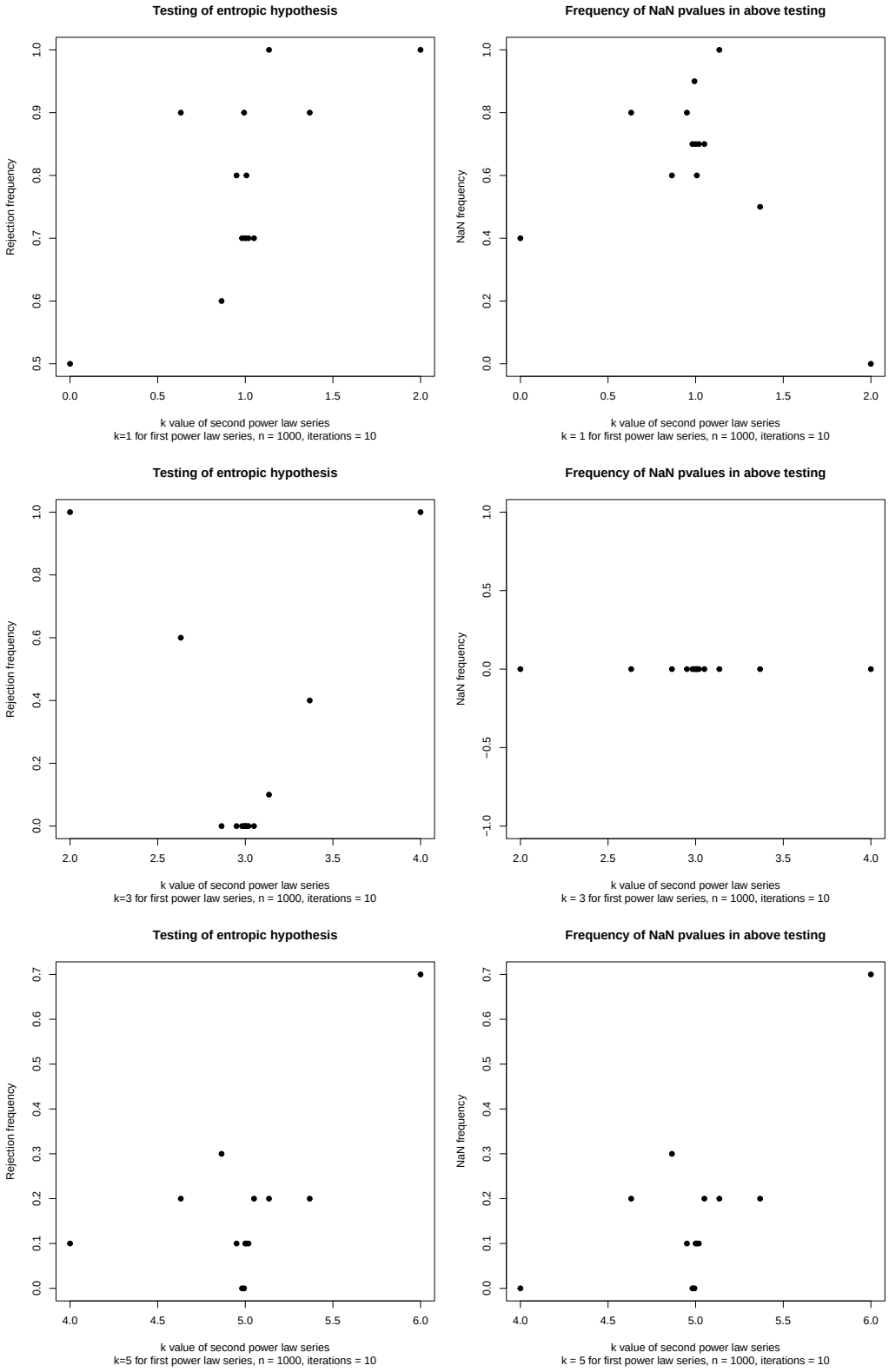

Figure 4.1: Power-law experiment

4.2 Idea behind new tie breaking

In the next section, a new way to break ties will be implemented and used. The idea behind it is when having a tie, instead of either randomly assigning it to any of the possible patterns, which was proposed by Bandt and Pompe in their original paper [1], or always assigning a type of tie to a given pattern, which is done in the article being used in the next section [25], the new idea is to assign an equally large weight to all the possible patterns in case of a tie. In a tie, with k possible patterns, each of these k patterns gets a weight of $\frac{1}{k}$. The amount of possible patterns is the product of the occurrence of each unique value. To calculate the ordinal pattern distribution, simply sum up all the weights for each pattern and divide by the total amount of weight to get the frequency. Ordinal patterns are often used in fields that do science on real-world phenomena. In cases where the measuring equipment has a low precision, ties will often occur, e.g. a weight measurement. It is commonly known that objects have an atomic weight, since all everyday objects are made of particles that have an atomic weight. The atomic weight unit is $1\mu = 1.66.. \cdot 10^{-27}kg$, so the theoretical weight of an object has a lot of decimals, which could theoretically be measured, however putting a person on a normal household scale will only give a result in kilograms with one decimal. A person measuring themselves on a scale at different times and weighing the same does not mean they had the same atomic weight at both times. Their data would result in a tie, but based on the above argumentation it is fair to assume that in reality, the weight has either gone up or down, however, it is impossible with the given instruments to measure, which it is, so the only fair assumption is that both cases are equally likely. This can almost be seen as superposition[29] of the change of weight. The weight has gone both up and down until a more precise measurement is made, but until that, it is only fair to think both cases are possible and in this case equally possible. This type of tie-breaking should work, where it can be argued that a theo-

retical measurement has a higher precision than the actual measurement. It does, however, not make sense to use it, when the measured value can be argued to be of the same precision as a theoretical measurement, e.g. "How many cows are in front of me at a certain type?". This implementation will briefly be referenced as "My implementation", but mostly as a "Theoretical Split". Note that the theoretical split method does not need noise added. Adding noise removes all ties, which means it should perform identically to "article implementation", which is the implementation made in the article[25], since it is built upon that code. The theoretical split and the second tie-breaking solution are deterministic, whereas the Bandt and Pompe solution is stochastic.

4.3 Temperature Data

[25] Will be partly reproduced and additional plots and tables will be made. Only the maximum temperature part of the climate data in section 6 of the article will be used. As can be seen in Figure 3.2 the percentage of ties in this dataset is extremely high. It is therefore quite important, how ties are handled since they make up a bulk of the dataset.

The libraries mentioned earlier all handle ties differently, as seen in Figure 3.3. The first three columns are the setting of each experiment. The start date column is included, because the paper, where the data is from, accidentally started their data on 1992-08-14, instead of 1992-08-08 as they said the data started from, it does however not make a big difference, which can be seen between the top three rows and bottom three

location	d=3	d=4	d=5	d=6
Miami	41.9	61.6	75.9	85.5
Edinburgh	27.2	45.1	61.4	74.5
Dublin	26.5	44.5	61.1	74.8

Figure 4.2: Table of ties in percentage

noiseAdded	naOmitted	StartDate	location	statcomp	pdc	article implementation	StatOrdPattHxC	my implementation
FALSE	FALSE	1992-08-14	Miami	0.97187	0.91806	0.91806		0.97768
			Edinburgh	0.98067	0.97145	0.97159		0.98565
			Dublin	0.98298	0.95095	0.96755		0.98432
TRUE	TRUE	1992-08-08	Miami	0.97847	0.97847	0.97847	0.97847	0.97847
			Edinburgh	0.98690	0.98690	0.98690	0.98690	0.98690
			Dublin	0.98625	0.98625	0.98625	0.98625	0.98625
TRUE	FALSE	1992-08-08	Miami	0.97811	0.97811	0.97811		0.97811
			Edinburgh	0.98719	0.98607	0.98611		0.98611
			Dublin	0.98540	0.97312	0.98396		0.98396
FALSE	TRUE	1992-08-08	Miami	0.97195	0.91804	0.91804	0.91804	0.97769
			Edinburgh	0.98068	0.97339	0.97339	0.97339	0.98676
			Dublin	0.98313	0.97007	0.97007	0.97007	0.98596
FALSE	FALSE	1992-08-08	Miami	0.97195	0.91804	0.91804		0.97769
			Edinburgh	0.98069	0.97148	0.97162		0.98567
			Dublin	0.98295	0.95106	0.96764		0.98435

Figure 4.3: Entropy table of libraries performs on different preprocessing

rows. If all the values in a row are identically, that means the libraries perform identically. The only setting, where this happens, is when noise is added and Na values are omitted. Na omit removes Na values and the dataset is pushed together, where the Na values have been removed. In the rest of the experiment, this setting will be used.

A more in-depth analysis of the Theoretical Split vs adding noise is made. The statcomp library is the comparison library, but it does not matter, which library is chosen, since they perform identically in this case. As can be seen in Figure 3.5 the theoretical value is very close to both the mean and median of the entropy of 1,000 iterations. Figure 3.4 clearly shows the problem of adding a random sample of white noise, since the value might easily end up 0.001 off either the mean, median, or theoretical value, which all seem to be quite precise values of the entropy since the theoretical value is calculated quite differently then the mean and median, but they still end up much closer, then a random sample do.

The above experiment is repeated, but where the theoretical split is fed a constant dataset. The constant values here represent very imprecise measurements, where the true values of the observed phenomenon are assumed to be different. E.g. trying to measure white noise, with a bad instrument. It is important to note that if a dataset contains just a single constant value, where the measurement is precise. The entropy would naturally be 0, since there is full predictability, e.g. “how many cows are on the field at a given time”. The iterations are calculated on random

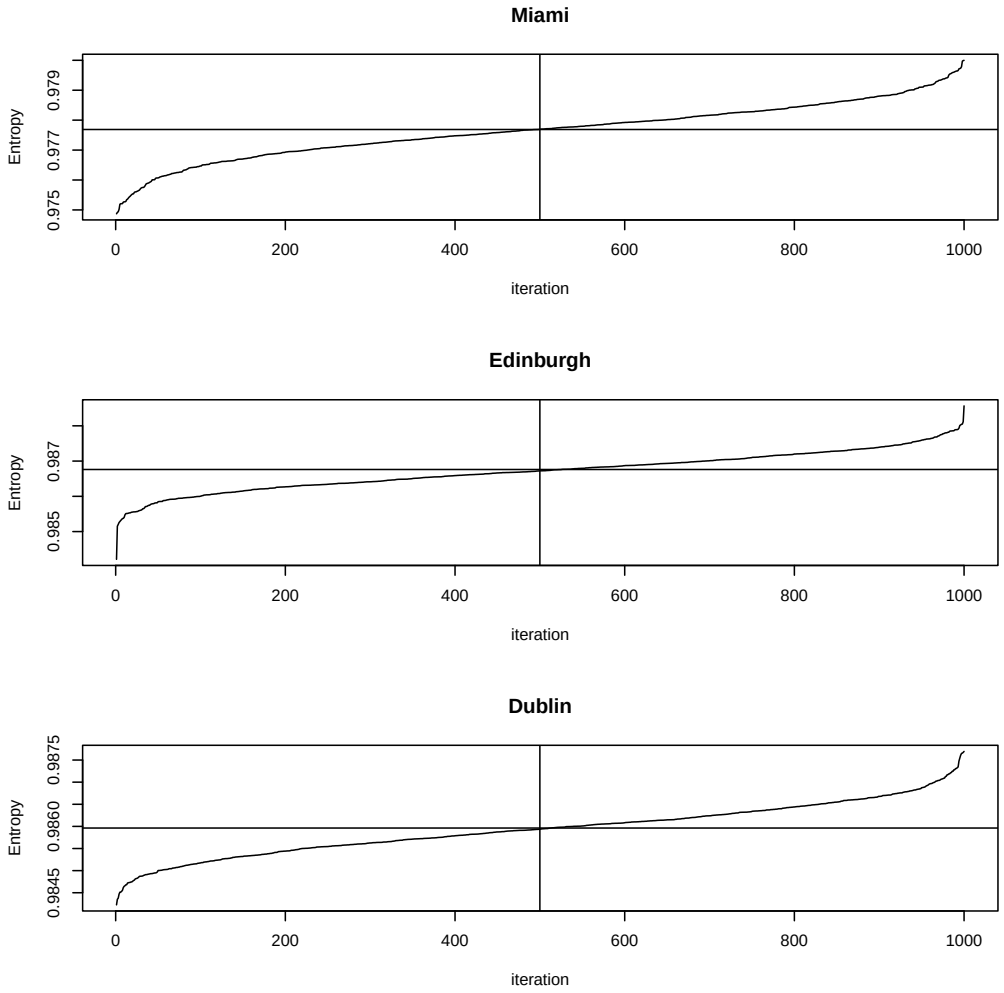

Figure 4.4: Iterations of adding noise vs theoretical split in sorted order, the vertical line is median and horizontal is theoretical split value.

location	meanRandom	medianRandom	Theoretical Split
Miami	0.9776571	0.9776953	0.9776874
Edinburgh	0.9867158	0.9867235	0.9867614
Dublin	0.9859386	0.9859376	0.9859631

Figure 4.5: Table of mean and median of adding noise and theoretical split. Temperature dataset

numbers, which ideally should be white noise. Figure 3.6 rarely has the entropy of ideal white noise, which is 1, where the theoretical split can correctly calculate that the poorly measured white noise has entropy 1. Figure 3.7 shows that the median measurement is closer than the mean to the theoretical split, so it might be a better measurement. In Figure 3.5 the median is generally also closer to the theoretical split. The median has another benefit over the mean, in the fact that it might not always be mathematically true to take the mean of an approximation, where the median is an entropy generated by an actual time series.

Figure 3.8 is a plot of the theoretical split values as the entropy of the three locations on the HxC plane, with confidence intervals on the entropy. At first glance, it looks like the confidence interval sticks out of the boundaries, however it is important to remember that a change in entropy leads to a change in complexity, so the confidence intervals are not breaking the boundaries.

P values are calculated. 10 iterations are done on adding noise for breaking ties and compared with the p-value of the theoretical split. The theoretical p value is only larger than one iteration for Miami-Edinburgh, two for Miami-Dublin and four for Edinburg-Dublin, which is OK. Ideally, it should be between iteration 5 and 6. Most importantly, it is the range of the iterations, which definitely confirms it is implemented so what correctly.

Adding noise has a couple of problems in the sense that it is much more computer-intensive to calculate just 10 iterations compared with the theoretical value once. ’

’

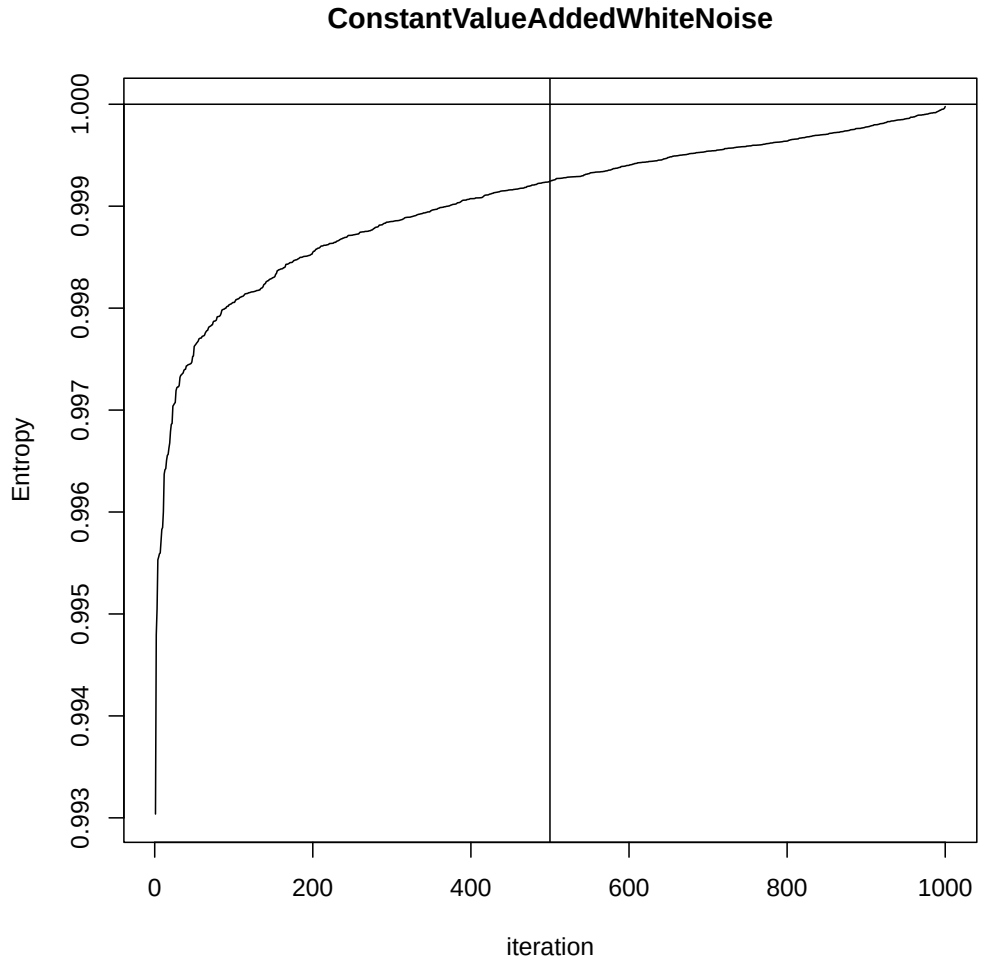

Figure 4.6: Constant dataset

meanRandom	medianRandom	Theoretical Split
0.99904	0.9992463	1

Figure 4.7: Mean and median of adding noise and theoretical split, constant dataset.

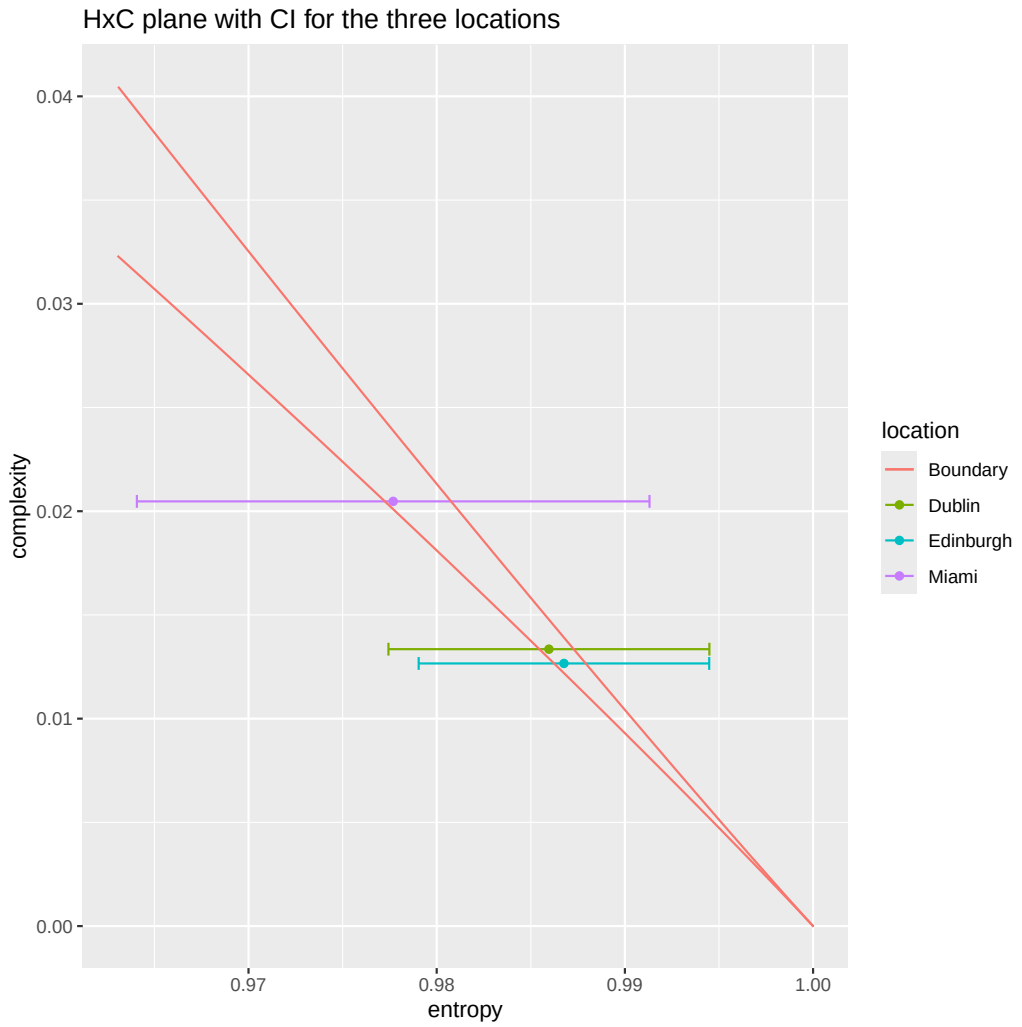

Figure 4.8: HxC plane of locations with confidence interval

Iteration	Miami-Edinburgh	Miami-Dublin	Edinburgh-Dublin
5	0.0013274374	0.0239569279	0.3638758
10	0.0003244188	0.0069446469	0.3927209
7	0.0036630440	0.0363118439	0.4463358
9	0.0022075899	0.0117689594	0.6680660
6	0.0029797057	0.0013769505	0.7810984
1	0.0048836351	0.0127590321	0.8032568
2	0.0065975420	0.0162918041	0.8188427
4	0.0018596130	0.0039581132	0.8870523
8	0.0020127362	0.0040575392	0.8873406
3	0.0008503533	0.0009603385	0.9572742

Figure 4.9: Noise added, 10 iterations, sorted by column "Edinburgh-Dublin"

Iteration	Miami-Edinburgh	Miami-Dublin	Edinburgh-Dublin
StatComplexity.R	0.000566	0.002065	0.723044

Figure 4.10: p value, when using theoretical split

Chapter 5

Conclusion

The main findings of the paper is the analysis of tie breaking. To ensure stable results, across current implementations of the libraries in the field, it is necessary to both add noise and remove Na values. A new method is proposed that breaks ties evenly among possible patterns. It has the benefit over adding noise that it is a deterministic method, where adding noise is a stochastic method. For the entropy it performed, very well and should definitely be considered, when dealing with dataset, wit ha large degree of ties. For the p-values, it was slightly less impressive, since its p values were not as close to the median of the 10 iterations. It would be interesting to run a larger amount of iterations for adding noise in the p value section, however, that would require quite a lot of computation power. The last benefit of the implementation is that it runs fast, since only one iteration needs to be run.

Bibliography

[1] C. Bandt and B. Pompe, "Permutation Entropy: A Natural Complexity Measure for Time Series," *Physical Review Letters*, vol. 88, pp. 174102–1–174102–4, Apr. 2002.

[2] F. Fidler and J. Wilcox, "Reproducibility of Scientific Results," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, winter 2018 ed., 2018.

[3] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, July 1948.

[4] R. López-Ruiz, H. L. Mancini, and X. Calbet, "A statistical measure of complexity," *Physics Letters A*, vol. 209, pp. 321–326, Dec. 1995.

[5] M. T. Martin, A. Plastino, and O. A. Rosso, "Statistical complexity and disequilibrium," *Physics Letters A*, vol. 311, pp. 126–132, May 2003.

[6] P. W. Lamberti, M. T. Martin, A. Plastino, and O. A. Rosso, "Intensive entropic non-triviality measure," *Physica A: Statistical Mechanics and its Applications*, vol. 334, pp. 119–131, Mar. 2004.

[7] J. M. Amigó and O. A. Rosso, "Ordinal methods: Concepts, applications, new developments, and challenges—In memory of Karsten Keller (1961–2022)," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 33, p. 080401, Aug. 2023.

[8] H. Zhang, F. Zhang, X. Jia, Q. Jiao, Z. Zhan, and L. Li, "Walnut crack detection based on EEMD and acoustic feature optimization," *Postharvest Biology and Technology*, vol. 212, p. 112874, June 2024.

[9] Q. Wang, X. Yang, W. Yan, J. Yu, and J. Wang, "Networked information interactions in schizophrenia magnetoencephalograms based on permutation transfer entropy," *Biomedical Signal Processing and Control*, vol. 91, p. 105977, May 2024.

[10] K. Szwajka, J. Zielińska-Szwajka, and T. Trzepieciński, "Experimental Analysis of Smart Drilling for the Furniture Industry in the Era of Industry 4.0," *Materials*, vol. 17, p. 2033, Apr. 2024.

[11] Abhishek Parikh, "EEG Sensor-Based Frequency Domain Analysis for Epileptic Seizure Detection," *Journal of Electrical Systems*, vol. 20, pp. 1033–1040, Apr. 2024.

[12] S. Tang and Z. Li, "EEG complexity measures for detecting mind wandering during video-based learning," *Scientific Reports*, vol. 14, p. 8209, Apr. 2024.

[13] G. Liu, P. Mu, K. Wang, G. Guo, X. Liu, and P. He, "Random Numbers Generated Based on Dual-Channel Chaotic Light," *Electronics*, vol. 13, p. 1603, Apr. 2024.

[14] c. C. Demirel, J. Gott, K. Appel, K. Lüth, C. Fischer, C. Raffaelli, B. Westner, Z. Zavecz, A. Steiger, S. Mota-Rolim, S. Ribeiro, M. Zeising, N. Adelhöfer, and M. Dresler, "Electrophysiological correlates of lucid dreaming," Apr. 2024.

[15] Z. Du, L. Yang, Y. Hao, J. Wang, Y. Li, S. Chen, and C. Zhou, "Randomness of scalar and vector random distributed soliton bunch in mode-locked fiber lasers," *Optical Fiber Technology*, vol. 84, p. 103782, May 2024.

[16] Y.-C. Sun, C.-Y. Ni, K.-N. Ying, A.-H. Xiong, T. Shuai, and Z.-H. Shen, "Laser ultrasonic spatially resolved acoustic spectroscopy for grain size study based on Improved Variational Mode Decomposition (IVMD)," *NDT & E International*, vol. 144, p. 103090, June 2024.

[17] M.-J. Li, X.-F. Zhou, F. Wang, M.-H. Bi, G.-W. Yang, M.-M. Xu, M. Hu, and H.-Z. Li, "Research on chaotic secure optical communication system based on dispersion keying with time delay signatures concealment," *Chaos, Solitons & Fractals*, vol. 182, p. 114720, May 2024.

[18] M. J. Struelens and P. Vineis, "COVID-19 Research: Challenges to Interpret Numbers and Propose Solutions," *Frontiers in Public Health*, vol. 9, p. 651089, Apr. 2021.

[19] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, pp. 452–454, May 2016.

[20] J. M. Weygand and M. G. Kivelson, "Jensen-Shannon Complexity Measurements in Solar Wind Magnetic Field Fluctuations," *ASTROPHYSICAL JOURNAL*, vol. 872, pp. 7–8, Feb. 2019. Web of Science ID: WOS:000458369900006.

[21] A. Osmane, A. P. Dimmock, and T. Pulkkinen, "Jensen-Shannon Complexity and Permutation Entropy Analysis of Geomagnetic Auroral Currents," *JOURNAL OF GEOPHYSICAL RESEARCH-SPACE PHYSICS*, vol. 124, pp. 2546–2547, Apr. 2019. Web of Science ID: WOS:000477707800013.

[22] P. M. Saco, L. C. Carpi, A. Figliola, E. Serrano, and O. A. Rosso, "Entropy analysis of the dynamics of El Nino/Southern Oscillation during the Holocene," *PHYSICA A-STATISTICAL MECHANICS AND ITS APPLICATIONS*, vol. 389, p. 5026, Nov. 2010. Web of Science ID: WOS:000282241600058.

[23] M. Smithson, *Confidence Intervals*. SAGE Publications, Inc., 2003.

[24] A. A. Rey, A. C. Frery, J. Gambini, and M. M. Lucini, "The asymptotic distribution of the permutation entropy," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 33, p. 113108, Nov. 2023.

[25] E. T. C. Chagas, A. C. Frery, J. Gambini, M. M. Lucini, H. S. Ramos, and A. A. Rey, "Statistical properties of the entropy from ordinal patterns," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 32, p. 113118, Nov. 2022.

[26] "statcomp: Statistical Complexity and Information Measures for Time Series analysis," Oct. 2019.

[27] "pdc: Permutation Distribution Clustering," Sept. 2015.

[28] J. Timmer and M. König, "On generating power law noise.," *Astronomy and Astrophysics*, vol. 300, p. 707, Aug. 1995. ADS Bibcode: 1995A&A...300..707T.

[29] E. Schrödinger, "An Undulatory Theory of the Mechanics of Atoms and Molecules," *Physical Review*, vol. 28, pp. 1049–1070, Dec. 1926.