

Entropic Measures of Time Series Data in a Biological Agent Based Model to Classify System Health

Robert Kramer¹

Abstract

Entropic measures have been shown to be useful in the classification of overall dynamics in biological systems [1]. Entropic measures were used to attempt to classify the dynamics in a simple biological agent based NetLogo model of lateral gene transfer in early phagocytic prokaryotic life. Within the model a set of parameters were previously found where the probability of the emergence of the super organisms was high. Several experiments were run recording the number of each organism present in the model. Time series data of this activity was recorded and used to measure multiscale multivariate sample entropy, $MSMV SampEn$. First the entropy measures were used to infer complexity using only total population data. Then, The entropy values were then used to predict the emergence of super-organisms at t time steps in the future. Analysis was inconclusive. The analysis could be improved with different measures than simply population and more labeled data.

Keywords

Artificial Life — Multivariate Sample Entropy — Agent Based Model

¹Systems Science, Portland State University

*Professor: Joe Fusion

Contents

1	Introduction	1
1.1	Complexity and Nonlinear Dynamics	1
	Introduction	1
1.2	Measures of Entropy	2
	MSE	
1.3	Nonlinear Measures	2
1.4	System Health	3
1.5	Hypothesis and Objectives	3
2	Materials and Methods	3
2.1	ABM Model	3
	"Summary of Key Results	
2.2	Dynamical Modes in ABM model	4
2.3	Analysis in R	4
3	Results and Discussion	4
4	Conclusion	6
	References	6

1. Introduction

1.1 Complexity and Nonlinear Dynamics

The study of nonlinear dynamics and statistical mechanics has given insight into the general properties of complex adaptive systems with applications in Alife. Complexity is not

precisely defined, but complex systems can be thought of as having a set of typical properties[2].

1. Complex systems contain many constituents interacting nonlinearly
2. The constituents of a complex system are interdependent
3. A complex system possesses a structure spanning several scales.
4. A complex system is capable of emergent behavior
5. Complex systems have an interplay between chaos and certainty.
6. There is an interplay between cooperation and competition.

It has been hypothesized that living systems exist in a region of dynamics known as complex. These systems are thought of as complex adaptive systems. Summary statistics like Langton's λ are used to roughly categorize the dynamics of simple systems like Cellular Automata[3, 4]. This categorization can be used to infer the possibility of computational complexity in the model. Complex adaptive systems may arise at the "Edge of Chaos" or at "Critical" states similar to phase transitions[5]. These phase transitions accompany a transition in entropy and highlight the interplay between predictability and uncertainty necessary for the existence of complex systems. Entropic measures of complex systems may serve as similar summary statistics for the dynamics of real-world biological and complex adaptive systems. Knowl-

edge of the general dynamics and information generation or entropy of systems may provide a computational tool for determining the health of complex systems. The theory of maximum entropy production has been proposed as a guiding principle of ecosystems and the Gaia hypothesis [6]. General dynamical knowledge could be used as a diagnostic tool for intervention[7, 1, 8], for robust control of complex systems[9], to help better classify systemic risk, or serve as a guideline for the design of environments conducive to evolutionary forces.

1.2 Measures of Entropy

Background

Information entropy is a measure of information, disorder, or uncertainty. Shannon's concept of compressibility has been used as a method of discerning the complexity of a system[10]. Dynamic entropy is the rate of information production[1]. ApEn or approximate entropy was first developed as a means of estimating the Kolmogorov-Sinai entropy for real world signals with shorter N lengths. The Kolmogorov-Sinai entropies are based on the limit as the number of samples goes to infinity. With m representing the dimension of the sample vector, roughly $10^m - 20^m$ data points are needed to accurately classify the dynamic system[11]. Pincus developed approximate entropy *ApEn* to accurately estimate an entropic statistic for real world signals. *ApEn* was found to be biased. "This bias causes *ApEn* to lack two important expected properties. First, *ApEn* is heavily dependent on the record length and is uniformly lower than expected for short records. Second, it lacks relative consistency." [1] The Sample Entropy *SampEn* statistic was developed to improve performance on shorter sample sizes and be less dependent on the parameters chosen[1].

$$\text{SampEn}(m, r, N) = -\ln \frac{A^m(r)}{B^m(r)} \quad (1)$$

SampEn (1) is a measure of the conditional probability a sequence is similar in dimension $m + 1$ given the sequence was similar in dimension m . The tolerance for similarity is an element wise parameter r . For example, consider the sequence $(1, 2, 3, 1, 2, 3)$ where $m = 2$ and $r = 1$. *ApEn* would first compare $(1, 2)$ with the overlapping sequences $(2, 3), (3, 1), (1, 2),$ and $(2, 3)$. Four fifths of the pairs are within $r = 1$. Then the same process would sum all similar 2 dimensional vectors in the sequence to all others. The difference between the log of this calculation and the log of the same calculation with 3 dimensional vectors, or $m + 1$, is the *ApEn*. *SampEn* is very similar except self similar sequences are not counted. However, *SampEn* always assigns a higher value to higher stochasticity. If our entropic statistic is a measure of complexity, purely random signal should be considered less complex than chaotic systems. Multiscale entropy was introduced to compare the information generated across multiple time scales. *MSE* of $1/f$ noise is roughly constant as N increases, while white noise steadily decreases as better estimates are found[11].

1.2.1 MSE

A is the number of matches for dimension $m + 1$ within the tolerance r . B is the number of matches for dimension m within r . *MSE* can be used on independent signals of the same phenomenon by combining them into one sequence, but if the signals are correlated, the statistic produces a biased result[12]. However, we would like to find a measure of complexity integrating multiple correlated signals of the same system. Multivariate Multiscale entropy was developed to overcome this challenge.

$$\text{MSE}(M, \tau, r, N) = -\ln \left[\frac{B^{m+1}(r)}{B^m(r)} \right] \quad (2)$$

Cross-apEntropy is used to compare two distinct signals " Importantly, the *ApEn* algorithm counts each sequence as matching itself, a practice carried over from the work of Eckmann and Ruelle (5) to avoid the occurrence of $\ln(0)$ in the calculations. This step has led to discussion of the bias of *ApEn* (22, 23, 27). However, further studies revealed it was a biased statistic. Later, *SampEn* or sample entropy was introduced to correct some of the biases. *SampEn* did not account for complexity on different scales. *MSE* or multiscale entropy was introduced by partitioning the data into disjoint scales.[13, 14] Development is still underway, and there exist several entropic measures under study. I choose a version known as multiscale multivariate sample entropy or *MSMV SampEn* [12]. I used *MSMV SampEn* as my primary summary statistic for complexity of my simulation of early life. I did not utilize the multivariate feature, instead I used the package to measure single variate multiscale sample entropy.

1.3 Nonlinear Measures

Sample entropy is the basis of the multiscale multivariate entropy measure used for analysis. The sample entropy is closely related to a class of statistics typically used in the study of dynamical systems known as invariants. Lyapunov exponent, correlation dimension, and information dimension are examples of these invariants[15]. Practical use of entropic measures in the literature include electroencephalograms, human postural sway, eeg brain activity, and electrocardiograms among others [16, 17, 7, 14]. The literature surrounding these practical applications is rooted in more general nonlinear techniques. I do not know of examples where these techniques have been applied to heterogeneous systems with quasi-non hierarchical control like ecosystems. It remains unclear if there exist the kind of underlying noiseless function implied by dynamical analysis. I would like to be able to infer the deterministic nature of my system with correlation in phase space and the minimum number of parameters needed with correlation dimension. Pincus specifically warns against this and gives counter examples to discourage these myths with general examples of the proper use of *ApEn* with real data [18]. However, chaotic models have been used to successfully model complex systems and a general analysis can be fruitful

in practice. A fundamental assumption during these analyses is the existence of a deterministic chaotic function governing the underlying behavior. The simulation studied contains both deterministic mappings (rules for behavior) and stochastic elements (mutations).

Deterministic chaos can be thought of as a folding of phase space between two starting points. Even if the starting points are very close, folding can quickly take these points away from each other [2]. If the entropy creation due to folding increases the uncertainty at the next time step greater than the precision of measurement, the system will seem stochastic to the observer.

Phase space is the space of the parameters and trajectories are the paths from time t to $t + 1$ through phase space. With full knowledge of phase space the potential attractors can be visualized. In practice, the phase space is commonly visualized on a 2d plane where $y(t)$ vs $y(t + \tau)$ is plotted. The tau is estimated using autocorrelation or the mutual information statistic.

$$I(\tau) = \sum P(x(k), x(k + \tau)) \ln \frac{P(x(k), x(k + \tau))}{P(x(k)) \cdot P(x(k + \tau))}$$

The mutual information difference between the data and the independence model for a given tau. When this is minimized our data is decoupled. The autocorrelation function could also have been used.

The phase space of the simulation is made combination of the environment variables. The dna completely determines behavior given a the environment factors.. There are 32 possible, but several are not viable. Therefore I could conceive of a phase space on the order of 20.

1.4 System Health

System health can be thought of in many different ways. Overall I would like a measure that indicates the need to pay special attention, or intervene within a complex system I would seek to regulate, optimize, or control. I focused on how health is defined for ecosystems because of their complexity and focus on overall health instead of optimization. There is much debate on what ecosystem health is, but it is generally accepted that system organization, resilience and vigor, as well as the absence of signs of ecosystem distress describe the essential elements[19]. Vigor is a measurement of metabolism or primary productivity. Organization is the relation and number of parts. Resilience is the ability of the system to maintain structure. I would also add for my purposes the fostering of emergent behavior and signs of evolutionary forces in my simulation. I attempted to approximate a statistical measure for overall system health.

$$E_h[\text{hamming}] * H(\text{sequencetag}) * \frac{\sum \text{patchSugar}}{\max(\text{patchSugar})}$$

The expected value of the hamming distance is a proxy measure of the resilience because it gives an indication of the

length genealogies in the system. The lengths are larger when more long term gene structures are present. If the ecosystem is unable to support these structures, I conclude the resilience is lower. The entropy of the sequence tags give an indication of the uncertainty in the types of species present. I am using it as a proxy for system organization. Sequence tag entropy is also a proxy for the diversity. Finally I use the utilization of sugar as a proxy measurement for vigor or metabolism. These proxies are obviously not perfect, but together they seem to give a reasonable first approximation. I watched the measure while running simulations to gauge whether the statistic seemed correlated with what looked healthy.

Entropy has not, to my knowledge, been applied to ecosystem health. However, *ApEn* has been used to quantify the health of biological signals as referenced previously. I would like to adopt a similar definition of health in relation to regularity as those described in bio-signal literature, but in relation to diversity and population. The system health is thought of as not being too regular or too chaotic. I want an entropy which is neither too high or too low[20].

1.5 Hypothesis and Objectives

The hypothesis is entropic measures indicate the region of dynamics, or complexity, of a simulated complex system. Furthermore, I hypothesize the measure can be used to classify the health of the system. Entropic measures were used to attempt to classify the dynamics in a simple biological netlogo model of lateral gene transfer in early phagocytic prokaryotic life. An attempt will be made to infer the health of the system and the emergence of “super” prokaryotes hypothesised as the precursors to modern eukaryotes from these measures. If a measure can be found, conditions under which this measurement could be used in a real world system will be explored.

2. Materials and Methods

2.1 ABM Model

I began with a netlogo ABM previously developed to study lateral gene transfer in early prokaryotic life. The summary is listed below

2.1.1 "Summary of Key Results

Perfect protection against infection would block evolution by lateral genetic transfer. Plausible genealogies of eukaryotic origin were obtained. In a “rough” environment, increased motility despite increased metabolic expenditure is favored. Population dynamics are unanticipated and complex. It remains unclear whether the original aerobic-anaerobic bacterial interactions were parasitic, commensal or symbiotic, but if they were parasitic, initial virulence was low. Extensions of this model may be used to model eukaryotic nucleus formation. The addition of other DNA elements to model more realistic genomes may be fruitful. The model could be applied to the spread of culture among other domains.”

This model is a decent general model of competing agents capable of undergoing evolution through selection. It demon-

strates several different states of dynamics. The model shows how interesting behavior happens at the boundaries or edges of resources as well as at the limit for survival with sugar regrowth levels set to the minimum to enable long term behavior. DNA is a 32 bit sequence. Each possible sequence is then mapped to a sequence tag attached to each agent enabling easy tracking of genealogies. The hamming distance is the number of times the DNA changed since its first ancestor arrived.

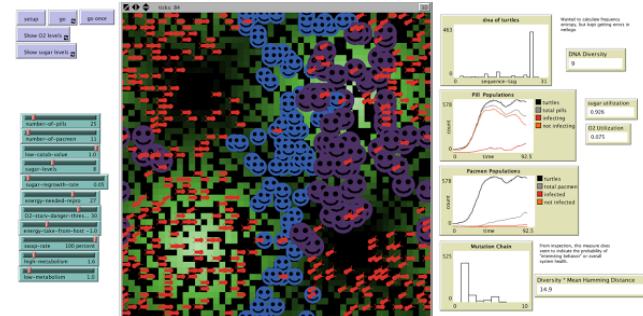


Figure 1. NetLogo Simulation Screenshot

The model was modified to enable visual exploration of health measures, utilization of resources, and regions of dynamics. Agents can have a low and high catabolism, a low or high metabolism tied to mobility, oxygen tolerance, and the ability to phagocytose. The ‘pills’ look like ants and can be phagocytized by the smiling faces or ‘packmen’. Red pills are oxygen intolerant and have low metabolism. Pink pills are oxygen tolerant and have a high metabolism. The simulation begins with blue packmen, but if a blue pacman eats a pink pill and that pill survives then the blue packman may have purple (super) packmen due to lateral gene transfer. The supers have the ability to use O₂ in their catabolism and have higher mobility. We posited this was the method of the emergence of mitochondria in an earlier study.

My primary initial study was to visually surface validate the idea that the health measure and diversity were meaningful.

2.2 Dynamical Modes in ABM model

Previous work with the abm simulation identified initial conditions conducive to the emergence of super eukaryotes with O₂ tolerance and a higher catabolism. The supers also lead to interesting dynamics. Figure (2) shows 20 runs with initial conditions conducive to emergence. There are several basic modes. The most common dynamical mode occurs when the O₂ tolerant pills die out and blue packmen do not phagocytize pills with dna capable of catabolism evolution. These events lead to utilization of separate niches within the different O₂ zones of the environment. There is stability, but also low diversity and little change. The middle ending populations in fig (2) represent this mode. The highest population modes occur when the phagocytic packmen die out from overconsumption and the smaller pills take over the niche. This normally occurs when supers take over because of their increased appetites and red O₂ tolerant pills take over the O₂ niches. The two highest ending populations show this behavior. The third basic mode

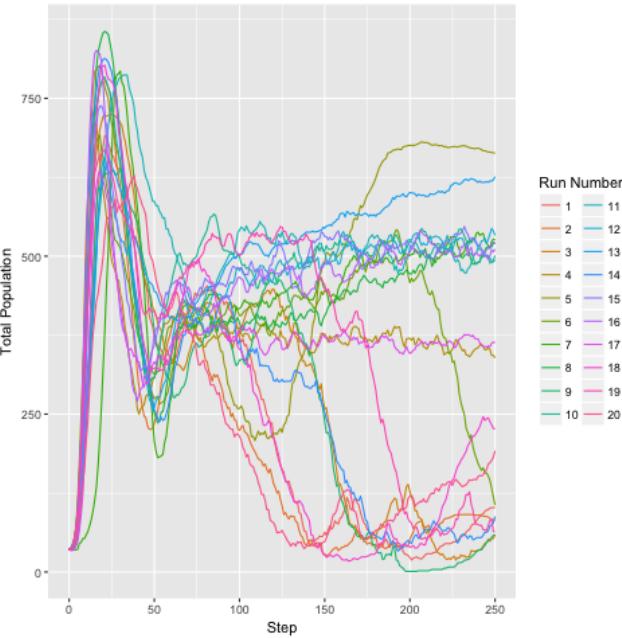


Figure 2. total population signals of test set

is the emergence of supers with stable Lotka–Volterra type population fluctuations due to resource limits. This mode is found in the lower fluctuating populations. This mode is the least stable, but the most diverse and contains the longest genealogical structures. I would like to classify these regions as the most complex. In my measure of system health, this mode is favored.

2.3 Analysis in R

I ran experiments using the behaviorspace function in netlogo. I initially ran experiments without collecting utilization data. I then ran two experiments with the utilization data. The R environment allowed me to explore the data in further detail. I performed nonlinear analysis using the package nonlinearTseries [21]. I also made use of a developmental package for simple *MSMV SampEn* [22]. I explored several of the relations between data and the overall behavior. I also trained a classifier on the “health” based on *MSMV SampEn*, but my classifier did not predict any true instances and I believe the method was flawed. Explicit details on the methods I used for analysis can be found in the R markdown document accompanying this article.

3. Results and Discussion

I found the *MSMV entropy* measure did not correlate with the *health measure* given the simulated overall population time series data.

Figure (3) shows a comparison of health and entropy. I calculated a Pearson’s product-moment correlation of -0.2949 with a $p - \text{value} = 2.2e - 05$.

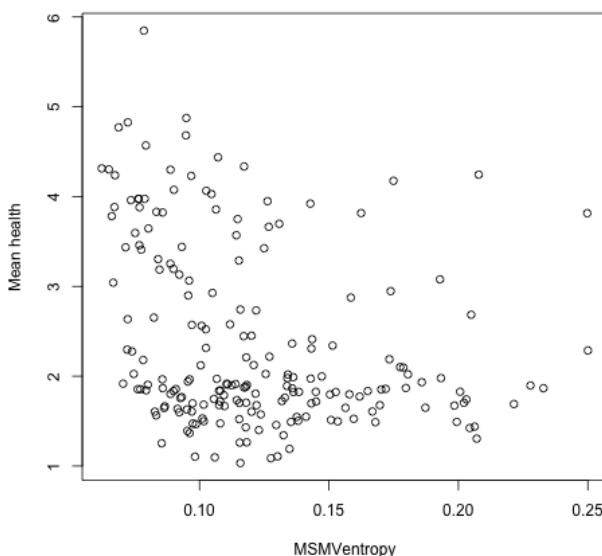


Figure 3. MSMVentropy Vs Mean health of a run

Classification

I used classification tree and regression analysis (CART) to attempt to classify the health at time step 150 using the initial 100 time steps. The *MSMVentropy* statistic did not infer any additional information. Figure (4) shows the CART predictor found the highest accuracy by simply predicting good health. The inability to infer system health from the measure is expected given the near independence of the measure to overall health.

Nonlinear Analysis

I used nonlinear analysis methods to compare the population signal to a random walk with similar mean and variance. I treated the training set of 200 simulated runs as one large population signal. The methods I used for analysis were limited to single variate analysis and I wanted to get an overall perspective on the population data. The population data is focused on because I assume that overall population could be estimated for real ecological systems.

I estimated the time lag τ using mutual information [23, 21]. I found a $\tau = 7$ for the population signal and $\tau = 115$ for the random walk. I do not know why the random walk was correlated for so long. The random walk was specifically chosen to have a mean and variance similar to that of the population signal. Perhaps the non random selection criteria enforced an artificial correlation.

Figure (5) shows the approximate phase space of the population signal with a time lag of 7. There is some interesting behavior. The empty middle space show a jump indicating times of population explosion and collapse. From watching the simulations I can infer the source of these fluctuations is the emergence of expanding and contracting super-species populations.

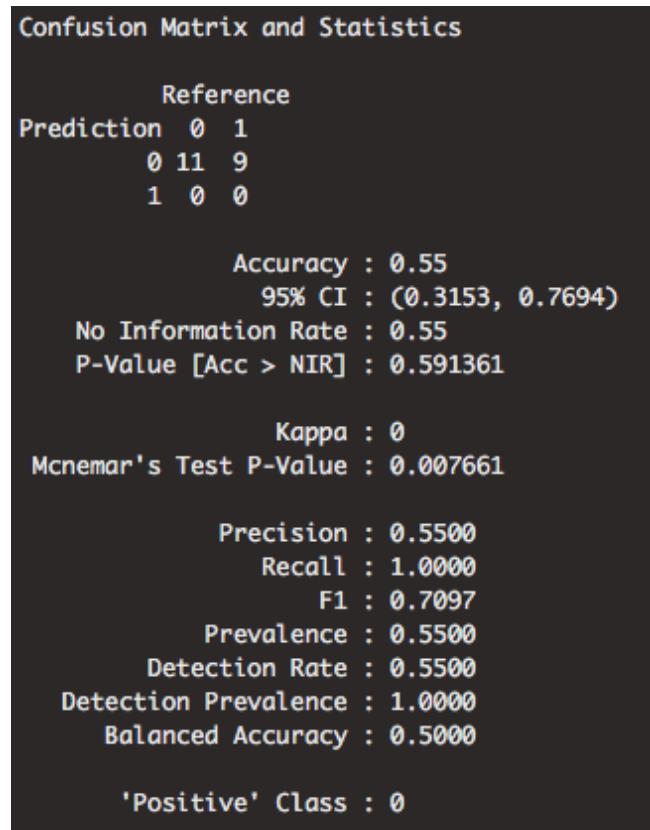


Figure 4. Confusion matrix of CART analysis

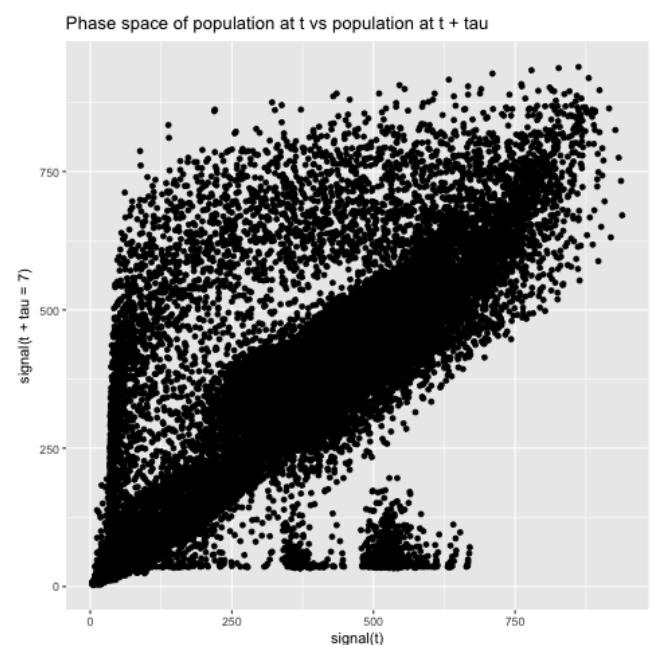


Figure 5. Phase space approximation

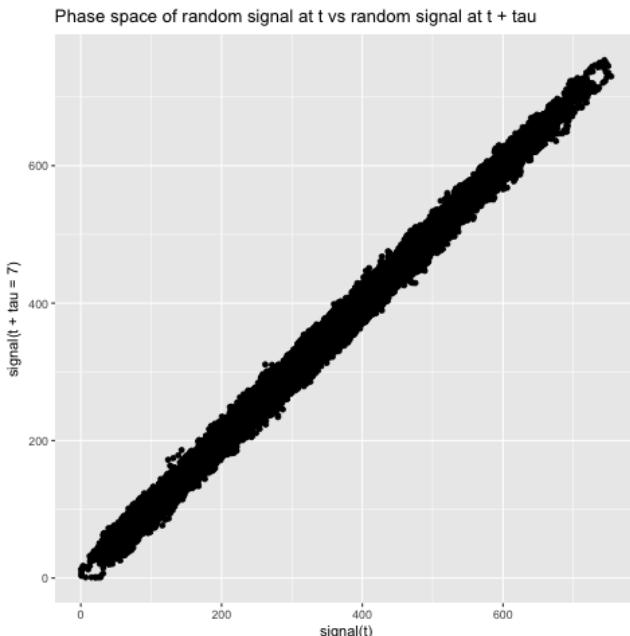


Figure 6. Phase space of random walk

The random phase space in figure (6) shows what we would expect. The next time is simply some random amount from the mean. The line is formed because the next time step in a random walk is dependent on the last.

I calculated the embedding dimension using Cao's method [24]. The sample entropies were calculated with nonlinearTseries package[21]. For further details see the R markdown file [25].

	random walk	population signal
tau	115.00	7.00
embedding dimension	7.00	11.00
sample entropy	0.39	0.42

The table shows the results. I found a similar complexity for the random walk signal and the population signal.

4. Conclusion

Within the context of the simulation and health measure chosen, neither sample entropy or multiscale entropy was a useful measure. I found the simulated data was not entirely appropriate for use with *MSE* or *SampEn* with a focus on progressing to real world ecosystem measures. I could better classify the signals with simpler statistics like the mean, variance, or simply the population after a certain amount of time. The health measure's ability to differentiate between the complex super behavior and less complex, but higher population dynamical modes was useful. Overall results are inconclusive and future work should include sequence analysis commonly found in natural language processing [9]. Future simulations should include relatively stable signals with the capacity to qualitatively change dynamics to be fully appropriate for use in this

analysis. An agent based model of a healthy aquarium moving to an algae bloom would be appropriate.

References

- [1] J S Richman and J R Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.*, 278(6):H2039–49, June 2000.
- [2] M Baranger. Chaos, complexity and entropy: a physics talk for non-physicists. 2000.
- [3] Life at the edge of chaos (langton 1991).pdf.
- [4] Melanie Mitchell, Peter Hraber, and James P Crutchfield. Revisiting the edge of chaos: Evolving cellular automata to perform computations. 31 March 1993.
- [5] P Bak, C Tang, and K Wiesenfeld. Self-organized criticality: An explanation of the 1/f noise. *Phys. Rev. Lett.*, 59(4):381–384, 27 July 1987.
- [6] Axel Kleidon, Yadvinder Malhi, and Peter M Cox. Maximum entropy production in environmental and ecological systems. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 365(1545):1297–1302, 12 May 2010.
- [7] Temujin Gautama, Danilo P Mandic, and Marc M Van Hulle. Indications of nonlinear structures in brain electrical activity. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 67(4 Pt 2):046204, April 2003.
- [8] Raúl Alcaraz, Daniel Abásolo, Roberto Hornero, and José J Rieta. Optimal parameters study for sample entropy-based atrial fibrillation organization analysis. *Comput. Methods Programs Biomed.*, 99(1):124–132, July 2010.
- [9] Ming Dong. A tutorial on nonlinear Time-Series data mining in engineering asset health and reliability prediction: Concepts, models, and algorithms. *Math. Probl. Eng.*, 2010, 8 June 2010.
- [10] S M Pincus. Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. U. S. A.*, 88(6):2297–2301, 15 March 1991.
- [11] Anne Humeau-Heurtier. The multiscale entropy algorithm and its variants: A review. *Entropy*, 17(5):3110–3123, 12 May 2015.
- [12] Mosabber Uddin Ahmed and Danilo P Mandic. Multivariate multiscale entropy: a tool for complexity analysis of multichannel data. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 84(6 Pt 1):061918, December 2011.
- [13] R B Govindan, J D Wilson, H Eswaran, C L Lowery, and H Preißl. Revisiting sample entropy analysis. *Physica A: Statistical Mechanics and its Applications*, 376:158–164, 15 March 2007.

- [14] R Alcaraz and J J Rieta. A review on sample entropy applications for the non-invasive analysis of atrial fibrillation electrocardiograms. *Biomed. Signal Process. Control*, 5(1):1–14, 2010.
- [15] Schreiber Kantz. *Nonlinear Time Series Analysis*. Cambridge, 2004.
- [16] Nicholas Rohrbacker. Analysis of electroencephogram data using Time-Delay embeddings to reconstruct phase space.
- [17] Sofiane Ramdani, Benoît Seigle, Julien Lagarde, Frédéric Bouchara, and Pierre Louis Bernard. On the use of sample entropy to analyze human postural sway data. *Med. Eng. Phys.*, 31(8):1023–1031, October 2009.
- [18] S M Pincus and A L Goldberger. Physiological time-series analysis: what does regularity quantify? *Am. J. Physiol.*, 266(4 Pt 2):H1643–56, April 1994.
- [19] D J Rapport, R Costanza, and A J McMichael. Assessing ecosystem health. *Trends Ecol. Evol.*, 13(10):397–402, 1 October 1998.
- [20] Jennifer M Yentes, Nathaniel Hunt, Kendra K Schmid, Jeffrey P Kaipust, Denise McGrath, and Nicholas Stergiou. The appropriate use of approximate entropy and sample entropy with short data sets. *Ann. Biomed. Eng.*, 41(2):349–365, February 2013.
- [21] cran. cran/nonlinearTseries. <https://github.com/cran/nonlinearTseries>. Accessed: 2017-4-14.
- [22] areshenk. areshenk/MSMVSampEn. <https://github.com/areshenk/MSMVSampEn>. Accessed: 2016-12-10.
- [23] Ewelina Brzozowska and Marta Borowska. Selection of phase space reconstruction parameters for EMG signals of the uterus. *Studies in Logic, Grammar and Rhetoric*, 47(1).
- [24] Liangyue Cao. Practical method for determining the minimum embedding dimension of a scalar time series. *ELSEVIER N Physica D*, 110:43–50, 1997.
- [25] Robert Kramer. Multiscale multivariate sample entropy as an indicator of system health, 5 May 2017.