Surrogate Testing
ePortfolio Entry 5 - Module 7

Group 6: Djourdan Johnson, Jacuqot Qiu, Lotte Michels, Nawat Nawati Azhati, Nuo Xu, Xuelin Wei

1 Checking for Result Significance

Throughout this project, we apply a variety of methods to characterize the dynamics of a honeybee colony as a complex system. In general, complex systems often produce data that look complicated, but sometimes simple stochastic processes or linear correlations can mimic complexity measures. Surrogate testing guards against mistaking such processes for complexity.

Specifically, surrogate testing shows whether that metric is statistically significant against a well-defined null model. The core idea of is to generate proxy data and calculate a metric of interest on this data. These steps are typically repeated 20 times to generate a distribution of surrogate metric scores (Kantz and Schreiber, 2004). The true metric is than compared against this distribution using inferential statistics to assess its significance. This is fundamental in complexity science because it helps distinguish genuine complex dynamics from simpler or random processes.

Moulder et al. (2018) propose various ways to generate surrogate time-series, each depending on its own null hypothesis (H0):

1. **Data Shuffling**: randomizing the order of values in a signal (without replacement). H0 = there is no temporal structure within a time-series/there is no time dependency between two or more time-series.

2. **Amplitude Adjusted Fourier Transform (AAFT)**: randomizing time-series while preserving the original amplitude distribution and linear autocorrelation. This method was originally proposed by Theiler et al. (1992). It tests the same null hypothesis as data shuffling.

3. **Segment Shuffling**: cutting the data into a number of segments and randomizing the order of these segments. H0 = there is no time dependency between randomly paired segments (of a certain size) for two or more time series.

4. **Data Sliding**: splitting the data in two and appending the latter part to the beginning of the time-series. H0 = Long lags between two time series do not influence their time dependency.

5. **Participant Shuffling**: interchanging time-series between interactive and non-interactive signal pairs, for instance over different subjects in the data. H0 = the observed measure between series is no different between interactive and non-interactive pairs.

For the beehive within the MSPB data (Zhu et al., 2024), the first four surrogate analysis methods are applicable. Figure 1 roughly visualizes how these methods create surrogate signals, applied to the raw humidity signal. The methods will be further explained and demonstrated by assessing the significance of the recurrence metrics found in the previous ePortfolio entry (entry 4/module 6). The exact pre-processing steps and parameter settings from this module will be re-used to ensure fair surrogate tests. The code used for these analyses can be found here. To keep this module concise, we will restrict ourselves to three metrics of interest: the recurrence rate (RR), determinism (DET) and laminarity (LAM).

2 Univariate Surrogate Testing for RQA

In module 6, the univariate recurrence within the hive power, humidity and temperature signals were examined using recurrence quantification analysis (RQA). A summary of the original results is provided in Table 1. The robustness of these results was assessed using data shuffling and AAFT. These approaches were implemented as Python functions and applied to all three signals.

Signal	**Original RQA Metrics**		
	RR	DET	LAM
Temperature	0.180	0.982	0.990
Humidity	0.099	0.925	0.955
Hive Power	0.109	0.718	0.809

Table 1: Original Recurrence Quantification Analysis (RQA) results per signal, computed in Entry 4/Module6. RR = Recurrence Rate, DET = Determinism, LAM = Laminarity.

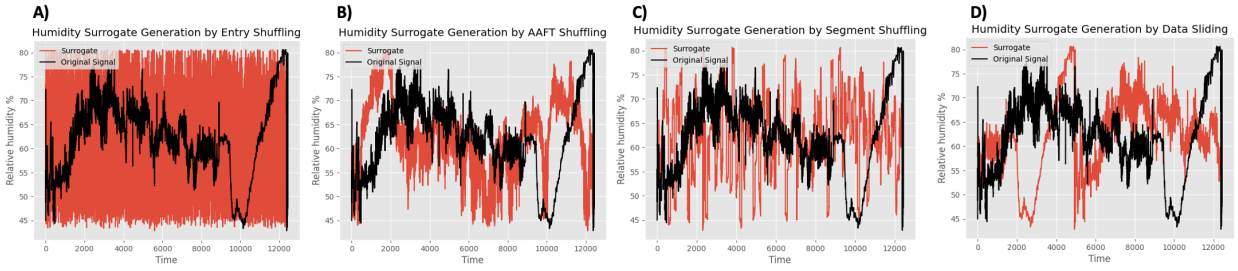

Figure 1: Example surrogate signals for the raw humidity time-series. Different methods are displayed. A: Data shuffling B: AAFT shuffling. C: Segment shuffling, with a segment size of 96 timepoints. D: Data sliding, with a cut-off at 60% of the data.

2.1 Data Shuffling

Surrogate time-series were created for all three signals by randomizing the order of the constituent values (Moulder et al., 2018). As such, temporal properties within the sequences are destroyed. Next, RQA was performed on the surrogate signals, using the same parameters as for the original signals (see previous ePortfolio entry). This process was repeated 20 times, as recommended by Kantz and Schreiber (2004), to create a distribution of surrogate RQA results.

Based on these acquired distributions, a mean and 95% confidence interval (CI) for the recurrence metrics of interest were computed. The results are displayed in Table 2. Data shuffling, in this case, is conducted to test the null hypothesis that *there is no temporally recurrent structure within the time-series of interest*. If the original metric falls outside of the surrogate CI, that null hypothesis can be rejected. Alternatively, if the original metric falls within the surrogate CI, the null hypothesis cannot be rejected and the recurrence metric may not be significant.

Combined, tables 1 and 2 show that the true RQA metrics fall outside of the surrogate confidence intervals. Data shuffling surrogates thus reject the null hypotheses for all variables and conclude that the RQA metrics in Table 1 are significant.

2.2 Amplitude Adjusted Fourier Transform

Data shuffling suggest the original metrics to be significant. However, to make our findings even more robust, we can apply multiple surrogate generation techniques and compare their outcomes. In contrast to the data shuffling technique, Amplitude Adjusted Fourier Transform (AAFT) is designed to generate surrogate sequences that ensure a Gaussian distribution and maintain some original signal features like the mean and autocorrelation (Moulder et al., 2018; Theiler et al., 1992). As such, AAFT generated sequences may be more plausible surrogates for a signal. This can also be viewed in Figure 1 when comparing plots A and B.

To implement AAFT, functionalities were used from the pyunicorn package (Donges et al., 2016). Like before, 20 surrogates were created and submitted to RQA, using the original parameters. Again, metric means and 95% confidence intervals (CIs) were computed, as displayed in Table 3.

AAFT surrogates test the same null hypothesis as data shuffling. Hence, the same decision rules apply regarding the statistical significance of the original metrics. Thus, juxtaposed to Table 1, the AAFT results show that the true RQA metrics for temperature and humidity are significantly higher than expected under the null hypothesis. For temperature and humidity, AAFT surrogates thus reject the null hypothesis that these signals feature no temporally recurrent structure.

This is not the case for the power signal: for all three recurrence metrics, tables 1 and 3 indicate that the original power signal scores significantly lower than the AAFT surrogate results. The power signal may thus be less recurrent, deterministic and laminar than expected under the null hypothesis. The observed RQA metrics in the power signal are not robust.

3 Bivariate Surrogate Testing for CRQA

So far, it was demonstrated that data shuffling and AAFT can be applied to effectively assess the significance of univariate RQA. These methods can also be applied to check the robustness of bi- or multivariate findings. However, it should be noted that data shuffling and AAFT largely get rid of any temporal properties (Moulder et al., 2018). As a result, surrogate signals may not adhere to certain contextual aspects. For instance, regarding beehives, Zhu et al. (2024) describe that the dynamics in honeybees show typical daily cycles (following a 'biological clock'). Such biological constraints are lost in AAFT and data shuffling, which can cause null hypotheses to be rejected too easily in a bi- or multivariate scenario.

To avoid such spurious findings, Moulder et al. (2018) define two alternative and more plausible surrogate meth-

Metric	Temperature			Humidity			Power		
	Mean	Lower CI	Upper CI	Mean	Lower CI	Upper CI	Mean	Lower CI	Upper CI
RR	0.0043	0.0042	0.0044	0.00267	0.0026	0.0027	0.0069	0.0068	0.0070
DET	0.0080	0.0073	0.0087	0.0051	0.0049	0.00522	0.0135	0.0131	0.0139
LAM	0.0532	0.0490	0.0575	0.0119	0.0109	0.0129	0.0346	0.0332	0.0361

Table 2: **Data shuffling** surrogate RQA metrics per signal. Each metric is characterized by the mean and 95% confidence interval of its surrogate distribution. For each variable, the original metrics are significantly higher than the surrogate mean.

Metric	Temperature			Humidity			Power		
	Mean	Lower CI	Upper CI	Mean	Lower CI	Upper CI	Mean	Lower CI	Upper CI
RR	0.1159	0.1077	0.1242	0.0533	0.0504	0.0562	0.1716	0.1686	0.1746
DET	0.9555	0.9512	0.9597	0.8410	0.8322	0.8497	0.9365	0.9342	0.9387
LAM	0.9721	0.9695	0.9747	0.8961	0.8907	0.9015	0.9569	0.9553	0.9585

Table 3: **AAFT** Surrogate RQA metrics per signal. Each metric is characterized by the mean and 95% confidence interval of its surrogate distribution. For temperature and humidity, all original metrics are significantly higher than the surrogate mean. For hive power, all original metrics are significantly lower than the surrogate mean.

ods: segment shuffling and data sliding. These techniques can be employed to check the significance of the cross-RQA (CRQA) results described in the previous ePortfolio entry on module 6. For that module, the pairwise cross-recurrence between the detrended temperature and hive power signals was explored. With regards to the metrics of interest, the following results of were found: RR = 0.023, DET = 0.537, LAM = 0.537 (also reported in tables 4 and 5).

Bi-variate surrogate tests can further assess these results. For each of those tests, a surrogate signal is created for the detrended temperature, while the un-shuffled version of the detrended hive power timeseries was used.

3.1 Segment Shuffling

Segment shuffling entails that timeseries are divided into smaller, separate chunks (Moulder et al., 2018). These chunks are randomly shuffled among each other, while the temporal order within chunks is retained. For the detrended temperature signal, a segment size of 199 timepoints was selected: segments represented almost 50 hours (a bit more than two days) in the data. This segment size could evenly divide the timeseries into 57 chunks and takes the day cycles described by Zhu et al. (2024) into account. The chunks were cut, put in random order and concatenated again to create a segment shuffled surrogate signal. This process was iterated 20 times, to create 20 of such surrogates for the detrended temperature signal.

Next, the surrogate temperature signals were coupled with the power signal and inputted to pairwise CRQA, using the same parameters as for the original computation (see previous ePortfolio entry). This resulted in a distribution of 20 CRQA measurements. Subsequently, means and 95% confidence intervals (CIs) for each surrogate recurrence metric of interest were computed. The results are displayed in Table 4. Segment shuffling, in our case, is employed to test the null hypothesis that *there is no time dependency between randomly ordered days for two time-series*. Whether the original metric falls out- or inside of the surrogate CI determines whether, respectively, this null hypothesis is rejected or not.

The segment surrogate results indicate that all the original recurrence metrics are significant. That is, the coupling of temperature and hive power shows higher recurrence, determinism and laminarity than expected under the null hypothesis. Segment surrogates thus reject this null hypothesis.

Metric	**Original**	**Surrogate Results**		
		Lower CI	Upper CI	Mean
RR	0.09989	0.057	0.055	0.059
DET	0.700	0.590	0.576	0.602
LAM	0.701	0.599	0.586	0.611

Table 4: **Segment shuffling** surrogate results for the CRQA between the (detrended) temperature and hive power signals.

3.2 Data Sliding

Instead of dividing signals into multiple chunks, data sliding entails that a time-series is split in two and that the order of the two parts is switched (Moulder et al., 2018). Whereas segment shuffling thus maintains small and local temporal trends, data sliding can be applied to take larger and more global temporal lags into account.

Moulder et al. (2018) suggest that signals are split around 60%. In addition, Kantz and Schreiber (2004) propose that surrogate testing is repeated 20 times to create a surrogate distribution of outcomes as a statistical benchmarking baseline. We combined both recommendations. First, 20 cut-off points were randomly sampled between 50% and 70%. The signal parts created by these cut-offs were then swapped, creating 20 data sliding surrogate timeseries. This process was conducted for the detrended temperature signal, while again keeping the hive power time-series fixed.

The resulting sequences were submitted to the pairwise CRQA computation from module 6, again using the same parameters as for the original analysis (see previous ePortfolio entry). Once more, means and 95% CIs were computed for each surrogate recurrence metric. The results are displayed in Table 5. Data sliding tests the null hypothesis that *long lags between two time series do not influence their time dependency*. Whether the original metric falls out- or inside of the surrogate CI statistically determines whether, respectively, this null hypothesis is rejected or not.

Contrasted to the original results, Table 5 indicates that the original recurrence rate and determinism results for CRQA are significantly higher than expected under the null hypothesis. This suggests that the recurrence and determinism between the (detrended) temperature and power signals is not influenced by long lags. Moreover, laminarity was not found to be as robust as the other two metrics. Table 5 indicates that the original LAM result is significantly lower than the surrogate outcomes. The coupling between temperature and hive power may thus be less laminar than expected under the null hypothesis.

To slightly nuance these findings, it should be noted that the original and surrogate results for data sliding (Table 5) are very close to each other, especially compared to the segment surrogate tests (Table 4).

4 Surrogate Testing for Windowed CRQA

The previous ePortfolio entry further explored the cross-recurrence between the detrended temperature and power signals by applying windowed CRQA. Specifically, a 1000 timepoint window size and 200 timepoint step size were used to obtain CRQA results for smaller, local patches in the signals. The windowed results were visualized to show the changes in cross-recurrence over time. This plot revealed several peaks in cross-recurrence between detrended temperature and power around.

Metric	**Original**	**Surrogate Results**		
		Lower CI	Upper CI	Mean
RR	0.09989	0.09977	0.09968	0.09986
DET	0.700	0.693	0.687	0.699
LAM	0.7007	0.701	0.7009	0.7019

Table 5: **Data sliding** surrogate results for the CRQA between the (detrended) temperature and hive power signals.

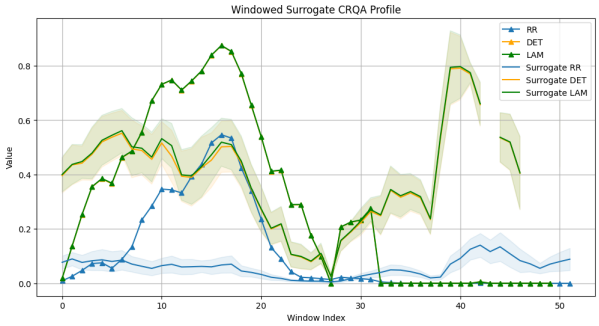

Figure 2: Windowed surrogate CRQA results against the original CRQA results (dotted). Surrogate sequences are wrapped by 95% CI areas.

To assess the robustness of these findings, windowed surrogate testing was applied. The segment shuffled surrogate signals generated in the previous section were used were used for this task. For each surrogate, the original sliding window (described before) was applied to compute the surrogate CRQA over time. Thus, 20 sequences of windowed surrogate CRQA results were obtained. The mean and 95% CI of these sequences are displayed in Figure 2, plotted against the original windowed CRQA profile.

This figure shows that small recurrence peaks in the original profile may not be robust. They either fall within the surrogate CI or are surpassed by the surrogate outcomes. The large peak on the left of the plot, however, shows significantly higher values for the original recurrence metrics compared to the surrogate signals. The underlying windows correspond to the time period ranging from around the end of June to the beginning of August. This peak in temperature-power cross-recurrence may thus forebode thermoregulatory behavior shown by beehives as they transition from colder to warmer weather (Research and Consortium, 2010). Additional analyses in our ePortfolio, such as the Dynamic Time Warping approach in Entry 6/Module 8, will further delineate this phenomenon.

5 Conclusion

Surrogate testing is important in complexity science to avoid false positive findings by creating a null model that complex measures can be benchmarked against. Various surrogate testing approaches were applied and compared to check the robustness of RQA findings from Entry 4/Module 6 of our ePortfolio. Data shuffling and AAFT surrogates suggested that temperature and humidity exhibit significant recurrence. In contrast, AAFT indicated that recurrence in beehive power was not as robust.

Additionally, segment shuffling and data sliding surrogates demonstrated the robustness of the cross-recurrence between the temperature and hive power. Solely the laminarity metric was found not to be resistant to long time lags in the time-series. Lastly, the CRQA was further validated with a windowed surrogate approach. The results highlighted a cross-recurrence peak between temperature and hive power around ...

References

[Donges et al.2016] Jonathan Donges, Jobst Heitzig, Boyan Beronov, Marc Wiedermann, Jakob Runge, Qing Yi Feng, Liubov Tupikina, Veronika Stolbova, Reik Donner, Norbert Marwan, Henk Dijkstra, and Jürgen Kurths. 2016. Unified functional network and nonlinear time series analysis for complex systems science: The pyunicorn package. *EGU General Assembly Conference Abstracts*, 4.

[Kantz and Schreiber2004] Holger Kantz and Thomas Schreiber. 2004. *Nonlinear time series analysis*. Cambridge University Press, 1.

[Moulder et al.2018] Robert G Moulder, Steven M Boker, Fabian Ramseyer, and Wolfgang Tschacher. 2018. Determining synchrony between behavioral time series: An application of surrogate data generation for establishing falsifiable null-hypotheses. *Psychological Methods*, 23(4):757–773, 3.

[Research and Consortium2010] Mid-Atlantic Apiculture Research and Extension Consortium. 2010. Seasonal Cycles of Activities in Colonies, 6.

[Theiler et al.1992] James Theiler, Stephen Eubank, André Longtin, Bryan Galdrikian, and J. Doyne Farmer. 1992. Testing for nonlinearity in time series: the method of surrogate data. *Physica D Nonlinear Phenomena*, 58(1-4):77–94, 9.

[Zhu et al.2024] Yi Zhu, Mahsa Abdollahi, Ségolène Maucourt, Nico Coallier, Heitor R. Guimarães, Pierre Giovenazzo, and Tiago H. Falk. 2024. MSPB: a longitudinal multi-sensor dataset with phenotypic trait measurements from honey bees. *Scientific Data*, 11(1), 8.