Wavelet Analysis of Nonstationary Circadian Time Series

Jessica Kate Hargreaves

PhD

University of York

Mathematics

July 2018

Abstract

Rhythmic data are ubiquitous in the life sciences, with biologists needing reliable statistical tools for the analysis of such data. When these signals display rhythmic yet nonstationary behaviour, common in many biological systems, the established methodologies are often misleading.

Chapter 2 develops and tests a new method for clustering nonstationary rhythmic biological data. The method combines locally stationary wavelet time series modelling with functional principal components analysis and thus extracts time—scale patterns useful for identifying common characteristics. We demonstrate the advantages of our methodology over alternative approaches by means of a simulation study and for real circadian data applications.

Motivated by three complementary applications in circadian biology, Chapter 3 develops new reliable statistical tests to identify whether a particular experimental treatment has caused a significant change in a rhythmic signal that displays nonstationary characteristics. As circadian behaviour is best understood in the spectral domain, we develop novel hypothesis testing procedures in the (wavelet) spectral domain, which facilitate the identification of three specific types of spectral difference. We demonstrate the advantages of our methodology over alternative approaches by means of a comprehensive simulation study and for real data applications, involving both plant and animal signals.

Chapter 4 investigates the effect of industrial and agricultural pollutants on the plant circadian clock. We examine the impact of exposure to a comprehensive range of environmentally relevant pollutants by utilising the methodologies developed in Chapters 2 and 3. Our findings indicate that many of the tested chemicals have an effect on the plant circadian clock, most of which would have remained undetected by classical methods overlooking nonstationarity. The results of Chapter 4 demonstrate the additional insight gained by using the appropriate methodologies, as developed in Chapters 2 and 3, and also have important implications for understanding environmental ramifications associated with soil pollution.

Contents

Abstract **2**

Contents **3**

List of Tables **6**

List of Figures **10**

Introduction **17**

Acknowledgements **20**

Declarations **21**

0.1 Chapter 2 . . . 21
0.2 Chapter 3 . . . 21
0.3 Chapter 4 . . . 21

1 Literature Review **23**

1.1 Basis Representations . . . 23
1.1.1 Fourier Analysis . . . 23
1.1.2 Wavelet Representations . . . 27
1.2 Wavelet Theory . . . 29
1.2.1 Multiresolution Analysis . . . 29
1.2.2 The Discrete Wavelet Transform (DWT) . . . 31
1.2.3 Matrix Representation of the Discrete Wavelet Transform . . . 35
1.2.4 The Nondecimated Wavelet Transform . . . 36
1.3 Stationary Time Series Analysis . . . 40
1.3.1 Fourier Analysis of Stationary Time Series . . . 41
1.3.2 Stationary Time Series Analysis of Circadian Data . . . 42
1.3.3 Wavelet Analysis of Stationary Time Series . . . 47
1.4 Nonstationary Time Series Analysis . . . 47
1.4.1 Locally Stationary Time Series . . . 48
1.4.2 Locally Stationary Wavelet Model . . . 50
1.4.3 Nonstationary Time Series Analysis of Circadian Data . . . 52

2 Clustering Nonstationary Circadian Rhythms Using Locally Stationary Wavelet Representations **55**

2.1 Introduction . . . 55
2.2 Motivation . . . 56
2.2.1 Experimental Details . . . 57
2.2.2 BRASS Analysis . . . 57
2.2.3 Nonstationarity in Circadian Rhythms . . . 58
2.2.4 Individual-level Variability in Circadian Rhythms . . . 58
2.3 Proposed Clustering Method . . . 59
2.3.1 Modelling Nonstationary Time Series . . . 59

2.3.2 Overview of Current Clustering/Classification Techniques that Account for Nonstationarity . . . 60
2.3.3 Proposed Functional Principal Components Analysis for the Wavelet Spectral Content . . . 62
2.3.4 Proposed Clustering Method . . . 63
2.4 Simulation Study . . . 67
2.4.1 Simulated Data . . . 67
2.5 Results . . . 70
2.6 Real Data Analysis . . . 72
2.6.1 Previously Published Circadian Data . . . 72
2.6.2 Novel Circadian Plant Data . . . 77
2.7 Conclusions and Further Work . . . 81
2.8 Appendix: Supplementary Figures . . . 83
2.9 Appendix: Experimental Details: Novel Circadian Plant Data . . . 86
2.10 Appendix: Results of Simulation Study Cases 1 and 4 . . . 87
2.11 Appendix: Experimental Details: Previously Published Circadian Data . . . 88

3 Wavelet Spectral Testing: Application to Nonstationary Circadian Rhythms 89
3.1 Introduction and Motivation . . . 89
3.1.1 Motivating Datasets . . . 89
3.1.2 Aims and Structure of this Chapter . . . 92
3.2 Overview: Nonstationary Processes and Hypothesis Testing in the Spectral Domain 92
3.2.1 Modelling Nonstationary Processes . . . 92
3.2.2 Existing Spectral Domain Hypothesis Testing . . . 94
3.3 Proposed Spectral Domain Hypothesis Tests . . . 96
3.3.1 Lead Dataset: Hypothesis Testing for Spectral Equality ('WST' and 'FT') . . . 96
3.3.2 Ultradian Dataset: Hypothesis Testing for Spectral Equality Across Scales ('HFT') . . . 99
3.3.3 Nematode Dataset: Hypothesis Testing for 'Same Shape' Spectra ('HT') . . . 100
3.3.4 Summary . . . 101
3.4 Simulation Studies . . . 102
3.4.1 Power Comparisons . . . 103
3.4.2 Size Comparisons . . . 105
3.4.3 Sensitivity Analysis . . . 107
3.4.4 Summary of Findings . . . 108
3.5 Real Data Analysis: Back to the Motivating Circadian Datasets . . . 108
3.5.1 Lead Dataset . . . 109
3.5.2 Ultradian Dataset . . . 109
3.5.3 Nematode Dataset . . . 110
3.6 Conclusions and Further Work . . . 112
3.7 Appendix: Experimental Details . . . 114
3.7.1 Experimental Overview: Lead and Ultradian Datasets . . . 114
3.7.2 Lead Nitrate Dataset . . . 114
3.7.3 Ultradian Dataset . . . 114
3.7.4 Nematode Dataset . . . 114

3.8 Appendix: Real Data Analysis: Supplementary Material . . . 116
3.9 Appendix: Tenability of the Normality Assumption . . . 117
3.10 Appendix: Haar-Fisz Transform . . . 118
3.11 Appendix: Detailed Description of Simulation Studies . . . 119
3.11.1 Detailed Description of Adaptive Neyman Test . . . 119
3.11.2 Basic Structure of Hypothesis Tests and Model Details . . . 121
3.11.3 Supplementary Tables . . . 127
3.12 Appendix: Summary Table . . . 133

4 Investigating the Effect of Soil Pollution on the Plant Circadian Clock 134
4.1 Introduction and Motivation . . . 134
4.2 Experimental Details . . . 135
4.3 Traditional Fourier Analysis . . . 136
4.3.1 Discussion of Findings . . . 136
4.3.2 Testing for Stationarity . . . 139
4.4 Wavelet Spectral Testing Using the Methodology Developed in Chapter 3 . . . 141
4.4.1 Discussion of Findings . . . 141
4.4.2 Conclusions . . . 147
4.5 Extension to Other Chemicals . . . 148
4.5.1 Discussion of Findings . . . 148
4.5.2 Conclusions . . . 154
4.6 Cluster Analysis Using the Methodology Developed in Chapter 2 . . . 154
4.6.1 Clustering DEFRA Chemicals . . . 155
4.6.2 Discussion of Findings . . . 155
4.6.3 Example: Clustering Within Individual Microtiter Plates . . . 155
4.6.4 Discussion of Findings . . . 156
4.7 Conclusions and Further Work . . . 159
4.8 Appendix: Supplementary Tables . . . 161
4.9 Appendix: Additional Clustering Example . . . 164
4.9.1 Discussion of Findings . . . 164

5 Conclusions and Further Work 167

References 170

List of Tables

1 Summary of the output of the analysis of the circadian dataset in BRASS. The 'number of plants excluded by BRASS' is the number of time series for which BRASS was not able to return a period estimate. 'RAE' (Relative Amplitude Error) is a value between 0 and 1 and gives information about the goodness of fit of the model (a value of 0 indicates a perfect fit). Results with an RAE over 0.4 are discarded. Recall: there are 24 plants in each of the groups. 58

2 Results for the Priestley-Subba Rao test of stationarity, implemented in the `fractal` package in R and available from the CRAN package repository. Number of non-stationary plants indicates the number of time series (in each group) with enough evidence to reject the null hypothesis of stationarity at the 1% significance level. Recall: there are 24 plants in each of the groups. 59

3 **Case 4.** The abruptly changing parameters of two nonstationary autoregressive processes. 70

4 Comparison of the proposed LSW-PCA clustering method with the methods proposed by Rouyer et al. (2008) and Antoniadis et al. (2013) for the simulation studies. Percentages show correct clustering rates. 72

5 Results of clustering the copper dataset into two clusters using the proposed LSW-PCA method. The modal cluster for each copper regime is highlighted in bold. . 74

6 Results of clustering the (normalised, truncated) cerium dataset into three groups using the proposed LSW-PCA method. The modal cluster for each concentration is highlighted in bold. 77

7 Distance measure (Section 2.3.4.1) comparison for the proposed LSW-PCA method for Cases 1 and 4. 87

8 **Case 1:** Comparison for selection of principal components for proposed LSW-PCA clustering method. Percentages show correct clustering rates. 87

9 **Case 4:** Comparison for selection of principal components for proposed LSW-PCA clustering method. Percentages show correct clustering rates. 87

10 Wavelet information comparison for the proposed LSW-PCA method for Cases 1 and 4. Percentages show correct clustering rates. 87

11 A summary of the hypothesis tests developed in this chapter. 102

12 Simulated power estimates (%) for models P1-P7 with nominal size of 5% with $N_1 = N_2 = 25$ realisations from each group. Highest empirical power estimates are highlighted in bold. 104

13 **Performance Comparison:** Simulated power estimates (%) for models P8-P12 with nominal size of 5% with $N_1 = N_2 = 25$ realisations from each group and using the false discovery rate procedure (FDR). Note: Control group period is 24 hours in each model. 105

14 Simulated size estimates (%) for models M1-M4 with nominal size of 5% and $N_1 = N_2 = 25$ realisations from each group. Empirical size estimates over the nominal size of 5% are highlighted in bold. 107

15 The number of rejections (as a percentage in brackets) for each relevant proposed test and multiple-hypothesis testing procedure for the motivating example datasets. 109

16 A summary of the output of the analysis of the motivating example datasets in BRASS: the mean period estimate for the control and test groups in hours (obtained using FFT-NLLS analysis (Plautz et al., 1997)), the difference between the period estimates and the corresponding p–value. 116

17 Results for the Priestley-Subba Rao test of stationarity, implemented in the `fractal` package in R and available from the CRAN package repository. Number of non-stationary plants indicates the number of time series (in each motivating example dataset) with enough evidence to reject the null hypothesis of stationarity at the 5% significance level (as a percentage in brackets). 116

18 **P6: AR Processes with Abruptly Changing Parameters.** The abruptly changing parameters of two nonstationary autoregressive processes. 124

19 **P7: AR Processes With Slowly Changing Parameters.** The slowly changing parameters of two nonstationary autoregressive processes. 125

20 Simulated power and size estimates (%) for the HFT for models P1-P7 and M1-M4 with nominal size of 5% and $N_1 = N_2 = 1$ realisation from each group. 127

21 Simulated power estimates (%) for models P1-P7 with nominal size of 5%. $N = N_1 = N_2$ is the number of realisations in each group. Highest empirical power estimates are highlighted in bold. 127

22 Simulated size and power estimates (%) for models P8-P12 and M5 with nominal size of 5% and using the false discovery rate procedure (FDR). $N = N_1 = N_2$ is the number of realisations in each group. Note: Control group period is 24 hours in each model. 128

23 Simulated size estimates (%) for models M1-M4 with nominal size of 5%. $N = N_1 = N_2$ is the number of realisations in each group. Empirical size estimates over the nominal size of 5% are highlighted in bold. 128

24 **M4: AR Process with Slowly Changing Parameters.** Numbers of rejections in empirical size estimates for the **Raw Periodogram F-Test** (FT), with Bonferroni Correction (Bon.) and false discovery rate (FDR) and with nominal size of 5%. "Modified Empirical Size Estimate" is calculated by examining only cases with more than one significant coefficient. 128

25 **Potential Non-Gaussian Innovations:** Simulated size and power estimates (%) for models P1-P5 and M1, M2 with nominal size of 5% and $N_1 = N_2 = 25$ realisations from each group. Innovations are distributed as: standard normal (denoted N(0,1)) or t-distribution with 5 or 3 degrees of freedom (denoted t_5, t_3 respectively). For the FT, the modified size and power estimates are recorded (i.e. only consider cases when more than 5 rejections are reported– see Section 3.4.2). Empirical size estimates over the nominal size of 5% are highlighted in bold. 129

26 **Potential Non-Gaussian Errors:** Simulated size and power estimates (%) for models P8-P12 and M5 with nominal size of 5% and $N_1 = N_2 = 25$ realisations from each group. The noise term in equation (105) is distributed as: standard normal (denoted N(0,1)) or t-distribution with 5 or 3 degrees of freedom (denoted t_5, t_3 respectively). For the FT, the modified size and power estimates are recorded (i.e. only consider cases when more than 5 rejections are reported– see Section 3.4.2). Empirical size estimates over the nominal size of 5% are highlighted in bold. . . 130

27 **Sensitivity to Generation and Estimation Wavelet Mismatch:** Simulated size and power estimates (%) for models P1-P5 and M1, M2 with nominal size of 5% and $N_1 = N_2 = 25$ realisations from each group. In all settings, the Haar wavelet is used for spectral estimation, but the following wavelets are used to generate the true spectra: Haar wavelets, Daubechies' least-asymmetric wavelets with 4 vanishing moments (V.M.) and Daubechies' extremal phase wavelets with 10 vanishing moments, respectively. 131

28 **Sensitivity to the Change of Modelling Wavelet:** Simulated power estimates (%) for models P6-P12 with nominal size of 5% and $N_1 = N_2 = 25$ realisations from each group. Different wavelets are used for the wavelet spectral estimation: Haar wavelets, Daubechies' least-asymmetric wavelets with 4 vanishing moments (V.M.) and Daubechies' extremal phase wavelets with 10 vanishing moments, respectively. 132

29 A summary of the hypothesis tests developed in this chapter. 133

30 **BRASS Results– DEFRA Chemicals**. Summary of the output of the analysis of the DEFRA chemicals in BRASS. "Treatment" represents the element under investigation within the chemical compound. $*$ indicates a significant change in period from the respective control group. † denotes an RAE value above the 0.4 threshold. "Number Analysed" is the number of time series for which BRASS was able to return a period estimate. There are 24 plants in each treatment group. ‡ Note that the Lead (Max) treatment group coincides with the 'Lead dataset' from Chapter 3. 137

31 Results for the Priestley-Subba Rao test of stationarity, implemented in the `fractal` package in R and available from the CRAN package repository. Number of nonstationary time series indicates the number of time series (in each treatment group) with enough evidence to reject the null hypothesis of stationarity at the 5% significance level (as a percentage in brackets). 141

32 **FT (FDR) results– DEFRA Chemicals**. The number of rejections (as a percentage in brackets) for the FT with FDR (at the 5% significance level) for the DEFRA Chemicals with † denoting 0% rejections. "Treatment" represents the element under investigation within the chemical compound. The estimated mean difference in period (using FFT–NLLS) is also shown for reference with $*$ indicating a significant change in period from the respective control group. 142

33 Results of clustering plate 0953 into two clusters using the proposed LSW-PCA method. The modal cluster for each treatment group is highlighted in bold. . . . 156

34 **FT (FDR) results– DEFRA Chemicals (plate 0953)**. The number of rejections (as a percentage in brackets) for the FT with FDR (at the 5% significance level) for the DEFRA Chemicals (plate 0953). 158

35 **Extension chemicals Part 1** (atomic numbers 3–27). The chemicals and concentrations used in the salt stress experiment (Section 4.5), where "Treatment" represents the element under investigation within the chemical compound (corresponding to the periodic table representation used in Figure 52) and "AN" represents the associated atomic number. For each chemical, the number of rejections (as a percentage in brackets) for the FT with FDR (at the 5% significance level) and the estimated mean difference in period (using FFT–NLLS), with ∗ indicating a significant change in period from the respective control group. ‡ indicates time series and a barcode plot for the chemical are shown in Figures 53 or 54. 161

36 **Extension chemicals Part 2** (atomic numbers 37–83). The chemicals and concentrations used in the salt stress experiment (Section 4.5), where "Treatment" represents the element under investigation within the chemical compound (corresponding to the periodic table representation used in Figure 52) and "AN" represents the associated atomic number. For each chemical, the number of rejections (as a percentage in brackets) for the FT with FDR (at the 5% significance level) and the estimated mean difference in period (using FFT–NLLS), with ∗ indicating a significant change in period from the respective control group. ‡ indicates time series and a barcode plot for the chemical are shown in Figures 53 or 54. 162

37 **Results of clustering the 12 DEFRA Chemicals** in Figures 47 and 48 and their respective controls into 2 groups using the LSW-PCA method. There are 24 plants in each treatment group. ∗ indicates a treatment with 0 plants in cluster 2. 163

38 Results of clustering plate 0952 into three clusters using the proposed LSW-PCA method. The modal cluster for each treatment group is highlighted in bold. . . . 164

39 **FT (FDR) results– Comparing DEFRA Chemicals (plate 0952)**. The number of rejections (as a percentage in brackets) for the FT with FDR (at the 5% significance level) for the DEFRA Chemicals (plate 0952). 166

List of Figures

1 **Example 1.1.3**. Dashed black line: Underlying cosine curve with a frequency of $1/(2\Delta t) = 0.5$ (i.e. the Nyquist frequency) and an amplitude of 2; Dashed blue line: Underlying cosine curve with a frequency of $1 = 2 \times 1/(2\Delta t)$ (i.e. double the Nyquist frequency) and an amplitude of 2; Black circles: Observed value of underlying functions at $t = 1, 2, \ldots, 5$. 25

2 **Example 1.1.6**. Top left: First underlying cosine curve with a frequency of 6/128 and an amplitude of 2; Top right: Second underlying cosine curve with a frequency of 10/128 and an amplitude of 4; Bottom left: The time series is a linear combination of two underlying cosine curves (see equation (3)); Bottom right: raw periodogram of the series with the frequencies 6/128 and 10/128 indicated by vertical red lines and horizontal green lines indicating values of 4 and 16 (which correspond to the amplitudes of the underlying cosine components squared– see equation (3)). 26

3 **Example 1.1.7**. Top left: First underlying cosine curve with a frequency of 6/128 and an amplitude of 2; Top right: Second underlying cosine curve with a frequency of 10/128 and an amplitude of 4; Bottom left: The time series is the concatenation of the two cosine curves (see equation (4)); Bottom right: raw periodogram of the series with the frequencies 6/128 and 10/128 indicated by vertical red lines and horizontal green lines indicating values of 4 and 16 (which correspond to the amplitudes of the underlying cosine components squared– see equation (4)). 27

4 Panel (a): Haar mother wavelet. Panels (b), (c), (d): translations and dilations of the Haar mother wavelet (using equation (10) for various combinations of $j = 1, 2$ and $k = 1, 2$). 29

5 Successive approximations of the Doppler test function introduced by Donoho and Johnstone (1994) using the Haar wavelet basis. Plot (a) shows the original function, plots (b), (c), (d), (e) and (f) display successively finer scale approximations (where $j = 5, 6, 7, 8$ and 9 respectively). 32

6 Flow diagram of the discrete wavelet transform of an observed dataset, $\mathbf{c_j}$, using successive applications of the low and high pass filters g and h. The orange boxes (below) give the number of coefficients at each level. 34

7 Top row: left and right: identical copies of the Doppler function. Bottom left: Haar discrete wavelet coefficients, $\{d_{j,k}\}$, of Doppler function (plotted with a different scale for each resolution level). Bottom right: as left but with Daubechies 'extremal-phase' with 8 vanishing moments. Note the smoother wavelet with a higher number of vanishing moments, has resulted in a sparser representation of the Doppler signal than the Haar wavelet 35

8 **Graphical depiction of the DWT.** The dotted arrows represent applying the filter $\mathscr{G}$ and the solid arrows represent applying the filter $\mathscr{H}$ (i.e. the application of the relations in (26)). This figure is reproduced following Figure 2.2 in Nason (2010). 36

9 **Example 1.2.5: the DWT is not translational invariant.** Figure (a) depicts the original data sequence whilst (b) depicts the same sequence rotated by a simple unit shift. Figures (c) and(d) depict the detail coefficients of the Haar DWT for the original and shifted data respectively. Note that the coefficients in Figure (d) do not correspond to a simple shift of the coefficients displayed in Figure (c). 38

10 **Example 1.2.5 of the translational invariance of the NDWT.** Figure (a) depicts the NDWT Haar wavelet detail coefficients of the original data. Figure (b) depicts the NDWT Haar wavelet detail coefficients of the shifted data. Observe that the coefficients in Figure (b) are a unit shift of the coefficients displayed in Figure (a). 38

11 **Stationary processes**. Top: An example realisation of a white noise process (Example 1.3.3) of length $T = 1000$. Bottom: An example realisation of a stationary ARMA(2, 1) process (Example 1.3.4) of length $T = 1000$ with AR parameters $(\alpha_1, \alpha_2) = (0.9, -0.2)$ and MA parameter of 0.5. 41

12 **Example 1.3.8: Spectral Estimation** for the realisation of an ARMA(2,1) process (Example 1.3.4) in Figure 11. Top: Raw periodogram. Bottom: Smoothed periodogram (using the Daniell kernel with parameter $m = 10$). 43

13 The defined rhythmic parameters: periodicity, phase, amplitude and clock precision (based on an image from Hanano et al. (2006)). 44

14 **Example 1.3.9: Implementation of FFT–NLLS**. Black line: A time series from the control group (Chapter 2); Blue line: cosine curve with period 27.03 hours (the period estimate obtained using FFT–NLLS). 46

15 **Example 1.4.1**: A time series for each of the four groups (see Chapter 2) is shown as an example– Group 1, a time series from the 100μM group; Group 2, a time series from the 150μM group; Group 3, a time series from the 200μM group. Red arrows: Plots of the estimated locations of the nonstationarities in the circadian plant signals in response to differing quantities of ammonium cerium nitrate, using the wavelet spectrum test (Nason, 2013), implemented in the `locits` package in R which is available on CRAN. 49

16 **Example 1.4.3.** Figure (a) depicts the spectrum defined in equation (57); (b) depicts a realisation generated from the spectrum shown in (a); (c) shows the mean of 100 uncorrected periodogram estimations computed on realisations from the spectrum shown in (a) and (d) shows the mean of 100 corrected periodogram estimations computed on realisations from the spectrum shown in (a). Note that the spectral estimate in (d) is much closer to the true underlying spectrum than (c). 53

17 Luminescence evolution over time for plants subjected to a control and 3 different ammonium cerium nitrate concentrations. Time is measured in hours relative to *zeitgeber* time (time of last external temporal cue: the dawn signal of lights-on). Top left: Each plant signal from the control group (in grey) along with the group average (dashed black). Other panels: Each realisation from the groups (in grey) along with the group average and the control group average (dashed black). Group 1: 100μM ammonium cerium nitrate with average in blue. Group 2: 150μM ammonium cerium nitrate with average in green. Group 3: 200μM ammonium cerium nitrate with average in red. (Each time series has been normalised to have mean zero.) Note: the free run started from time 24; shaded bars below each graph indicate the subjective darkness that plants expected to experience during the 'normal' day. 56

18 **Case 1.** Top left: Group 1 wavelet spectrum; Top right: Group 2 wavelet spectrum; Bottom left: Group 1 realisation and Bottom right: Group 2 realisation. 68

19 **Case 2.** Left: Group 1 wavelet spectrum (gradual period change from 24 to 25 hours); Centre: Group 2 wavelet spectrum (gradual period change from 24 to 26 hours); Right: Group 3 wavelet spectrum (gradual period change from 24 to 27 hours). 69

20 **Case 3.** Left: Group 1 wavelet spectrum (2-day transition); Centre: Group 2 wavelet spectrum (3-day transition); Right: Group 3 wavelet spectrum (5-day transition). 69

21 **Case 4.** Nonstationary autoregressive processes. Top left: Estimated wavelet spectrum of Group 1; Top right: Estimated wavelet spectrum of Group 2; Bottom left: Group 1 realisation; Bottom right: Group 2 realisation. 70

22 Luminescence evolution over time for plants subjected to a control and 2 different copper regimes. Time is measured in hours relative to *zeitgeber* time (time of last external temporal cue: the dawn signal of lights-on). Centre: Each plant signal from the **'Control'** group (in grey) along with the group average (dashed black). Other panels: Each realisation from the groups (in grey) along with the group average (in blue) and the control group average (dashed black). Left: **'Deficiency'** Group (1/2 MS). Right: **'Excess'** group (10 μM CuSO4). (Each time series has been normalised to have mean zero.) The grey and white bars indicate the subjective night and day, respectively. 73

23 Results of clustering the copper dataset into two clusters using the proposed LSW-PCA method. For each treatment group the individual signals are plotted in: red for Cluster 1 and blue for Cluster 2. The average of each treatment group is shown in black. Within each treatment group, the Cluster 1 average is shown in bold red and the Cluster 2 average in bold blue. 75

24 Results of clustering the copper dataset into two clusters using the proposed LSW-PCA method. The individual signals (grey) along with the cluster average in: red for Cluster 1 and (dashed) blue for Cluster 2. 76

25 Cluster average estimated spectra on the copper dataset using the proposed LSW-PCA method. 76

26 The results of clustering the cerium dataset into three groups using the proposed LSW-PCA method. The individual signals (grey) along with the cluster average in: (dashed) black for Cluster 1; blue for Cluster 2 and red for Cluster 3. The average of Cluster 1 (conceptualised as essentially 'Control') is shown (in dashed black) in all plots for reference. 79

27 Cluster average estimated spectra on the cerium dataset using the proposed LSW-PCA method. Cluster 1 approximately corresponds to the 'Control' group; Cluster 2 depicts 'Low concentration' behaviour (100 μM) and Cluster 3 the 'Higher concentration' (150 μM and 200 μM). 79

28 First two principal components obtained using the proposed LSW-PCA method on the cerium dataset. 80

29 The cerium dataset projected onto the first two principal components obtained from the LSW-PCA clustering method. The colours represent the clusters: black for Cluster 1, blue for Cluster 2 and red for Cluster 3. The symbols represent the plant treatments. 80

30 Summary of the BRASS analysis of the circadian plant signals in response to differing quantities of ammonium cerium nitrate, represented by plots of period estimates plotted against the respective relative amplitude errors (RAE). The colours and symbols represent the plant treatment groups: blue squares for the Control Group; green circles for Group 1 (100μM); red triangles for Group 2 (150μM) and purple stars for Group 3 (200μM). 83

31 Plots of the estimated locations of the nonstationarities in the circadian plant signals in response to differing quantities of ammonium cerium nitrate, using the wavelet spectrum test (Nason, 2013), implemented in the `locits` package in R which is available on CRAN. A time series for each of the four groups is shown as an example– Group 1, a time series from the 100μM group; Group 2, a time series from the 150μM group; Group 3, a time series from the 200μM group. 84

32 The screeplot used to inform the selection of the number of principal components to retain for the cerium dataset. Note 2 or 3 components could potentially be used, but for ease of interpretation (see Section 2.3.4.2), 2 were selected for clustering. 85

33 **Lead dataset:** Luminescence profiles over time for untreated *A. thaliana* plants (Control) and those exposed to lead nitrate (Lead). Left: Individuals in the control group (in grey) along with the group average (blue). Right: Individuals in the lead treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero. 90

34 **Ultradian dataset:** Luminescence profiles over time for control and mutant *A. thaliana* plants. Left: Individuals in the control group (in grey) along with the group average (blue). Right: Individuals in the mutant group (in grey) along with the mutant group average (red) and the control group average (blue). Each time series has been standardised to have mean zero. 91

35 **Nematode dataset:** Luminescence profiles over time for untreated *C. elegans* (Control) and those subjected to a pharmacological treatment (Treatment). Left: Individuals in the control group (in grey) along with the group average (blue). Right: Individuals in the treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero. 91

36 **Lead dataset.** Left: Average estimated spectrum of the 'Control' group; Centre: Average estimated spectrum of the 'Lead' group; Right: 'Barcode' plot for FT (with FDR). 110

37 **Ultradian dataset.** Left: Average estimated spectrum of the 'Control' group; Centre: Average estimated spectrum of the 'Mutant' group; Right: 'Barcode' plot for HFT (with FDR). 111

38 **Nematode dataset.** Left: Average estimated spectrum of the 'Control' group; Centre: Average estimated spectrum of the 'Treatment' group; Right: 'Barcode' plot for FT (with FDR). 111

39 Q–Q plots for a representative series from the control (Plots A, C, E) and test groups (Plots B, D, F) of each of our motivating datasets. Lead Dataset: Plots A and B. Ultradian Dataset: C and D. Nematode Dataset: E and F. 117

40 **P1:Fixed Spectra.** Top left: Group 1 wavelet spectrum; Top right: Group 2 wavelet spectrum; Bottom left: Group 1 realisation; Bottom right: Group 2 realisation. . . 122

41 **P2:Fixed Spectra-Fine Difference.** Top left: Group 1 wavelet spectrum; Top right: Group 2 wavelet spectrum; Bottom left: Group 1 realisation; Bottom right: Group 2 realisation. 122

42 **P3:Fixed Spectra-Plus Constant.** Top left: Group 1 wavelet spectrum; Top right: Group 2 wavelet spectrum; Bottom left: Group 1 realisation; Bottom right: Group 2 realisation. 123

43 **P4/P5: Gradual Period Change.** Left: Group 1 wavelet spectrum (gradual period change from 24 to 25 hours); Centre: Group 2 wavelet spectrum (gradual period change from 24 to 26 hours); Right: Group 3 wavelet spectrum (gradual period change from 24 to 27 hours). 124

44 **P6: AR Processes with Abruptly Changing Parameters.** Nonstationary autoregressive processes. Top left: Estimated wavelet spectrum of Group 1; Top right: Estimated wavelet spectrum of Group 2; Bottom left: Group 1 realisation; Bottom right: Group 2 realisation. 125

45 **P7: AR Processes with Slowly Changing Parameters.** Top left: Estimated wavelet spectrum of Group 1; Top right: Estimated wavelet spectrum of Group 2; Bottom left: Group 1 realisation; Bottom right: Group 2 realisation. 126

46 **P10: 'Function Plus Noise' Time Series with Constant Period.** Top left: Estimated wavelet spectrum of Group 1 (24 hour period); Top right: Estimated wavelet spectrum of Group 4 (23 hour period); Bottom left: Group 1 realisation; Bottom right: Group 4 realisation. Grey lines indicate a 24 hour period. 126

47 **DEFRA Chemicals:** Luminescence profiles over time for *A. thaliana* plants exposed to a selection of the DEFRA chemicals. Each Panel: Individuals in the chemical treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero. 138

48 **DEFRA Chemicals:** Luminescence profiles over time for *A. thaliana* plants exposed to a selection of the DEFRA chemicals. Each Panel: Individuals in the chemical treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero. 140

49 **Mercury (Max):** Luminescence profiles over time for untreated *A. thaliana* plants (denoted 'Control') and those exposed to mercuric chloride ($HgCl_2$) at a concentration of 5μM (denoted 'Mercury (Max)'). Left: Individuals in the control group (in grey) along with the group average (blue). Right: Individuals in the Mercury (Max) treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero. 143

50 **'Barcode' plots** for FT (with FDR) for the time series shown in Figure 47. 145

51 **'Barcode' plots** for FT (with FDR) for the time series shown in Figure 48. 146

52 **Periodic tables**, coloured by effect on the circadian clock of *A. thaliana* (Oakenfull et al., 2018). **A**: Coloured by FFT-NLLS period estimates (red outlines indicate a statistically significant change in period for all compounds tested). **B**: Coloured by percentage change from control using FT (FDR) analysis. **A** and **B**: Green elements are essential to life and were not tested individually; White elements were not tested due to safety or solubility. The actinoids and group 7 elements have been omitted as they were not tested. 149

53 **Time series and Barcode plots for Strontium, Platinum and Rubidium.** Time series (left panels): Blue lines indicate the control average for each chemical; grey lines indicate individual time series within each chemical treatment group and red lines indicate the average time series for the chemical treatment group. Barcode plots (right panels): Barcode plots for FT (with FDR) at the 5% significance level. 150

54 **Time series and Barcode plots for Gold, Tungsten and Lutetium.** Time series (left panels): Blue lines indicate the control group average for each chemical; grey lines indicate individual time series within each chemical treatment group and red lines indicate the average time series for the chemical treatment group. Barcode plots (right panels): Barcode plots for FT (with FDR) at the 5% significance level. 151

55 **Ruthenium:** Luminescence profiles over time for untreated *A. thaliana* plants (denoted 'Control') and those exposed to ruthenium chloride ($RuCl_3$) at a concentration of 2mM (denoted 'Ruthenium'). Left: Individuals in the control group (in grey) along with the group average (blue). Right: Individuals in the Ruthenium treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero. 153
56 **DEFRA Chemicals (plate 0953):** Luminescence profiles over time for *A. thaliana* plants exposed to a selection of the DEFRA chemicals. Each Panel: Individuals in the chemical treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero. 156
57 The results of clustering the DEFRA Chemicals (plate 0953) into 2 groups using the LSW-PCA method. The individual signals (grey) along with the cluster average in: red for Cluster 1 and blue for Cluster 2. The individual signals of the Lead (Max) treatment group in Cluster 1 are plotted in green. 157
58 **DEFRA Chemicals (plate 0952):** Luminescence profiles over time for *A. thaliana* plants exposed to a selection of the DEFRA chemicals. Each Panel: Individuals in the chemical treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero. 164
59 The results of clustering the DEFRA Chemicals (plate 0952) into 3 groups using the LSW-PCA method. The cluster average time series in: red for Cluster 1 (conceptualised as 'Cadmium (Max)'); blue for Cluster 2 (conceptualised as 'Arsenic (Max)') and green for Cluster 3 (conceptualised as 'Arsenic (Half)'). 165
60 The results of clustering the DEFRA Chemicals (plate 0952) into 3 groups using the LSW-PCA method. The individual signals of the modal treatment group (grey) along with the cluster average in: red for Cluster 1; blue for Cluster 2 and green for Cluster 3. For each cluster, the individual signals in the non–modal treatment group are plotted in: red for Cadmium (Max); blue for Arsenic (Max) and green for Arsenic (Half). 166

Introduction

The earth rotates on its axis every 24 hours resulting in a day and night cycle. Correspondingly, almost all species exhibit changes in their behaviour between day and night (Bell-Pedersen et al., 2005). These daily rhythms are not only caused by a response to daily changes in the physical environment, but are also the result of an internal timekeeping system or 'biological clock' within the organism (Vitaterna et al., 2001; Minors and Waterhouse, 2013). In particular, most plants are able to anticipate dawn and adjust their biochemistry accordingly. The mechanisms underlying the biological timekeeping systems, and the potential consequences of their failure, are among the issues addressed by researchers in the field of circadian biology (McClung, 2006; Bujdoso and Davis, 2013).

Circadian rhythms are a subset of biological rhythms with a period of approximately 24 hours. The term 'circadian' (derived from the Latin words "circa" (about) and "dies"(day)) was first used by Franz Halberg in the 1950s (McClung, 2006). Furthermore, a defining attribute of circadian rhythms is that they are "endogenously generated and self-sustaining" (McClung, 2006). In other words, they are the result of an internal timekeeping system–"endogenously generated"– and the period remains approximately 24 hours under constant environmental conditions, such as constant light (or dark) and constant temperature (i.e. when deprived of any external time cues)– "self-sustaining".

The first recorded observations (in western literature) of circadian rhythms appeared in the fourth century BC, when Androsthenes described the daily leaf movements of the tamarind tree (McClung, 2006). However, at the time it was assumed that these movements were due to the plant reacting to the day-night cycle (not the result of an internal clock) and it took over 2000 years for these observations to be experimentally tested. The first instance of scientific literature on circadian rhythms was in 1729 when the French astronomer de Mairan discovered that the daily leaf movements of certain plants persisted in constant darkness. This demonstrated for the first time that the plant could not be reacting to the external cues associated with a light–dark cycle, potentially indicating the existence of an internal timekeeping system. However, these experiments did not take temperature into account and it took a further 30 years before de Mairan's observations were independently repeated (in constant darkness) with constant ambient temperature (McClung, 2006). Almost 100 years later, the period length of these leaf movements was accurately measured and shown to be only *approximately* 24 hours. The result that the rhythms were not exactly 24 hours was crucial as it provided evidence that these rhythms were driven by an internal timekeeping system and not simply responses to an undetected geophysical cue associated with the rotation of the earth on its axis (such as light leaking into the laboratory darkroom!)

However, leaf movement is only one among many circadian rhythms in plants that include: germination; growth; enzyme activity; stomatal movement and gas exchange; photosynthetic activity; flower opening and fragrance emission (McClung, 2006). Therefore, in the 1970s, researchers began using genetic analysis with the intention of: identifying components of circadian clocks and elucidating the oscillator mechanism central to the circadian clock in a number of organisms, including the laboratory model plant species *Arabidopsis thaliana*. These early experiments were quite labour intensive, but advances in experimental methods in the 1990s meant that relative gene expression could be quantified *in vivo* (Plautz et al., 1997; Southern

and Millar, 2005; Perea-García et al., 2016a). Experiments recording plant response to light entrainment (constant light) result in datasets that, from a statistical point of view, can be considered as time series realisations.

Time series are ubiquitous and their analysis has found important applications in, for example, economics, climatology and, of course, circadian biology. For series that satisfy certain properties, such as stationarity (i.e. statistical properties such as the mean and variance are assumed constant over time) there are well—established methods of statistical analysis which are classically based on Fourier representations (see for example Priestley (1982); Shumway and Stoffer (2000); Brillinger (2001); Percival and Walden (2006) for an introduction to the topic). This thesis is concerned with analysis methods for nonstationary time series. In particular, we address a number of applied problems in the field of circadian biology, where nonstationarity is common (Zielinski et al., 2014) and replicate information is available. Access to replicate information, though standard in many biological applications, is atypical for time series data. Consequently, there is a gap in the current time series literature. In this thesis, we are primarily interested in clustering nonstationary time series and also determining if two (groups of) time series differ in terms of their spectral structure, and, if so, how?

Wavelets can be thought of as localised, oscillatory basis functions with several attractive properties for function representation. They are localised in both time and frequency, providing sparse multiscale representations for many signals. Due to their time localisation, wavelets provide natural 'building blocks' for nonstationary series. In this thesis, we develop clustering and hypothesis testing procedures based on wavelets.

This thesis is structured as follows: Chapter 1 provides an overview of aspects of the literature which are essential to the work subsequently developed. In particular, we give an overview of basis representations and an introduction to wavelet theory including the discrete wavelet transform (DWT). We then introduce the topic of stationary time series analysis, and its relevant applications in circadian biology. We also review the current state-of-the-art period estimation methods for circadian data. We then describe various approaches to nonstationary time series analysis and, in particular, the locally stationary wavelet (LSW) model of Nason et al. (2000), which provides the modelling framework for the methodology developed in Chapters 2 and 3 .

The work in Chapter 2 is motivated by the phenomenon of individual-level variability in plant response to stimuli, despite their sharing identical genetic characteristics (Doyle et al., 2002). The presence of multiple nonstationary behaviours within the same experimental treatment group motivates the development of a clustering procedure that can detect these different characteristics and analyse them separately, whilst accounting for nonstationarity. Hence, in Chapter 2, we develop and test (both through an extensive simulation study and application to a previously published circadian dataset) a new method for clustering rhythmic biological data. The proposed methodology combines locally stationary wavelet time series modelling with functional principal components analysis and thus extracts the time-scale patterns arising in a range of rhythmic data. Interesting and encouraging results are obtained by applying the clustering methodology to a newly–generated circadian dataset. Nevertheless, the developed methodology has wider applicability; it can be applied to other circadian datasets, as well as to data originating in other fields.

Chapter 3 addresses the problem of identifying whether a particular experimental treat-

ment has caused a significant change in a rhythmic biological signal. When these signals display nonstationary behaviour, the established methodologies may be misleading. Therefore, in this chapter, we develop new methodology that enables the formal comparison of nonstationary processes. As circadian behaviour is best understood in the spectral domain (Hargreaves et al., 2018), we develop novel hypothesis testing procedures in the (wavelet) spectral domain, embedding replicate information when available. Motivated by three complementary applications in circadian biology, our new methodology allows the identification of three specific types of spectral difference. We demonstrate the advantages of our methodology over alternative approaches, by means of a comprehensive simulation study and real data applications, using both published and newly generated circadian datasets. In contrast to the current standard methodologies, our proposed method successfully identifies differences within the motivating circadian datasets, and facilitates wider ranging analyses of rhythmic biological data. This demonstrates the utility of the proposed methodology, which again is not restricted to these applications.

Throughout this thesis, our work is motivated by a specific application in the field of circadian biology– the effect of industrial and agricultural pollutants on the plant circadian clock (Foley et al., 2005; Senesil et al., 1998; Hargreaves et al., 2018; Nicholson et al., 2003). Specifically, the Department for Environment, Food and Rural Affairs (DEFRA) developed 'Soil Guideline Values' (SGVs) that can be used to determine appropriate concentrations of certain chemicals in soil. Therefore, in Chapter 4, we apply the wavelet spectral testing and clustering methodologies developed within this thesis to investigate the impact of exposure to the chemicals at the concentrations outlined in the DEFRA report, as well as to chemicals not included in the report, on the plant circadian clock. Our findings indicate that many of the tested chemicals have an effect on the plant circadian clock. Therefore, the results of Chapter 4 could be used to inform a revision of the SGVs. Thus, the results of Chapter 4 not only have important implications for understanding environmental ramifications associated with soil pollution, but also demonstrate the additional insight gained by using the appropriate methodologies, as developed in Chapters 2 and 3.

Finally, Chapter 5 concludes with a summary of our work and some interesting ideas for future research.

Acknowledgements

First and foremost my thanks go to my team of supervisors: Marina Knight, Jon Pitchford and Seth Davis. A few people questioned the wisdom of having three voices, but I think we proved them wrong! Thank you Jon for being the maths–biology translator! Also thank you for your support and kindness and sense of humour (which was so often needed)! Thank you Seth for your generosity with your time, all your useful/ useless (delete as appropriate) facts and all the conversations about American sports and Eurovision! And thank you to Marina Knight (always right). I have nominated her for supervisor of the year every year for the last four years! But, due to the fact that I never threatened to quit my PhD (which of course is down to their excellent supervision), she is yet to win the award! Marina, you are an inspiration on a personal and professional level. And look– I went a whole paragraph without writing "in particular" or "therefore"! You have taught me so much!

Thanks also go to Agostino Nobile for being an excellent and very helpful Thesis Advisory Panel. I cannot apologise enough for the 40 page TAP reports that arrived 2 days before meetings! Also, thank you for all your time and patience when we worked together lecturing Stats 2. I wouldn't be where I am without that experience and you were there every step of the way.

Also thank you to all members of the Davis Lab. The lessons I have learnt through presenting and listening at the weekly lab meetings have been invaluable. Special mentions to: Jack; Kayla (one of the kindest people I have ever met); Mandi (Go Tribe!) and Rachael (without which none of the work in this thesis would have been possible). I wish all of you the best in your future endeavours.

Finally, I want to thank my family for much more than words can ever say. Mainly, proof reading and counselling! And generally putting up with me in work–mode for the past four years.

Declaration

I declare that this thesis is a presentation of original work and that I am the sole author, under the supervision of Dr. Marina Knight, Dr. Jon Pitchford and Prof. Seth Davis. Unless specified otherwise (see below), I performed all literature research, programming, analysis and writing of the thesis chapters. Under the supervision of Dr. Marina Knight, Dr. Jon Pitchford and Prof. Seth Davis, I developed and implemented the statistical methodology in Chapters 2 and 3. The novel circadian datasets analysed in this thesis were obtained by the Davis and Chawla Labs (Biology, University of York).

This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

The following chapters of the thesis have been published in or submitted to peer–reviewed journals. I am the first author on two of the three papers, but have received feedback and corrections from my co–authors.

0.1 Chapter 2

The novel circadian dataset analysed in this chapter was obtained by the Davis Lab (Biology, University of York). The BRASS analysis of this dataset was performed by R. Oakenfull.

This chapter has been published as:

Hargreaves, J. K., Knight, M. I., Pitchford, J. W., Oakenfull, R. and Davis, S. J. (2018). Clustering nonstationary circadian plant rhythms using locally stationary wavelet representations. *SIAM Multiscale modeling and simulation*, 16(1):184–214.

0.2 Chapter 3

The novel 'Lead dataset' analysed in this chapter was obtained by the Davis Lab (Biology, University of York). The BRASS analysis of this dataset was performed by R. Oakenfull. The novel 'Nematode dataset' analysed in this chapter was obtained by the Chawla Lab (Biology, University of York). The BRASS analysis of this dataset was performed by J. Munns.

This chapter has been submitted for publication to the Annals of Applied Statistics as:

Hargreaves, J. K., Knight, M. I., Pitchford, J. W., Oakenfull, R., Chawla, S., Munns, J. and Davis, S. J. (2018). Wavelet spectral testing: application to nonstationary circadian rhythms. *arXiv preprint arXiv:1803.09507*.

0.3 Chapter 4

The novel circadian dataset analysed in this chapter was obtained by the Davis Lab (Biology, University of York). The BRASS analysis of this dataset was performed by R. Oakenfull. I performed the wavelet spectral testing and produced the resulting figures.

The results and discussion of the BRASS analysis and wavelet spectral testing are in preparation for publication (with R. Oakenfull as the lead author) as:

Oakenfull, R., Hargreaves, J. K., Knight, M. I., Pitchford, J. W. and Davis, S. J. (In Preparation). Out of the sewage rises new maths...

In Chapter 4, I present the results of the above analyses. However, I selected which of the above results to discuss in detail and wrote the text of the thesis chapter (with feedback and

guidance from my supervisors). I also performed the cluster analysis (which is not included in the manuscript in preparation).

1 Literature Review

This chapter provides an overview of aspects of the literature which are essential to the work presented in this thesis. Section 1.1 gives an overview of basis representations and Section 1.2 gives a more detailed introduction to wavelet theory. Section 1.3 introduces the topic of stationary time series analysis, and, in particular, Section 1.3.2 its applications in circadian biology (which motivated the work in this thesis), as well as reviewing the current state-of-the-art period estimation methods for circadian data. Finally, Section 1.4 describes approaches to nonstationary time series analysis and, in particular, the locally stationary wavelet model.

1.1 Basis Representations

We begin by first reviewing some relevant concepts from Fourier analysis. An understanding of these methods provides the motivation for the use of wavelets, since certain signals cannot be represented efficiently using the trigonometric functions which form the basis of Fourier analysis. Fourier methods also underpin some of the commonly used period estimation methods for circadian data (see Section 1.3.2) and provide the benchmark for comparison with the (wavelet-based) methodology we develop in later chapters.

Our review of Fourier analysis follows the description in Priestley (1982) and the review of wavelet theory synthesises the descriptions in Daubechies (1992), Vidakovic (1999) and Nason (2010). We refer the reader to these texts for a more detailed discussion.

1.1.1 Fourier Analysis

In classical Fourier analysis, trigonometric functions (i.e. sine and cosine waves) are used to form the bases for functions in the space of *square integrable functions*,

$$L^2(\mathbb{R}) = \left\{ f \mid \int_{-\infty}^{\infty} |f(t)|^2 \, dt < \infty \right\}.$$

We define the Fourier series representation of a function, f, as follows.

Definition 1.1.1. *Let f be periodic (with period 2π) and square integrable over the interval $[0, 2\pi)$. Then the **Fourier series** representation of f is:*

$$f(x) = \frac{a_0}{2} + \sum_{n \in \mathbb{Z}_+} \Big(a_n \cos(nx) + b_n \sin(nx) \Big),$$

where the Fourier coefficients are calculated from

$$a_n = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos(nx) \, dx, \quad b_n = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin(nx) \, dx.$$

The Fourier coefficients, a_n and b_n in Definition 1.1.1, are calculated using the L^2 inner product. The magnitudes of the Fourier coefficients provide information about the frequency composition of the signal. The Fourier functions, $\{\cos(nx), \sin(nx)\}_{n \in \mathbb{N}}$, form an orthonormal basis and can be thought of as the "building blocks" from which certain periodic functions can be constructed.

However, most functions are not periodic. The Fourier transform is an extension of the Fourier series in that it provides a representation of non-periodic functions in the space of

absolutely integrable functions,

$$L^1(\mathbb{R}) = \{g \mid \int_{-\infty}^{\infty} |g(t)|dt < \infty\}.$$

The trigonometric "building blocks" of the Fourier series in Definition 1.1.1 are replaced by complex exponentials in the definition of the Fourier (and inverse Fourier) transform.

Definition 1.1.2. *The **Fourier transform** of a function $g \in L^1(\mathbb{R})$ is given by*

$$\hat{g}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g(x) \exp^{-i\omega x} dx.$$

*If $\hat{g}$ is the Fourier transform of g and $\hat{g}, g \in L^1(\mathbb{R})$, then the **inverse Fourier transform** is given by*

$$g(x) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \hat{g}(\omega) \exp^{i\omega x} d\omega. \tag{1}$$

Note that in the Fourier integral representation, frequency varies on a continuous scale, as opposed to the Fourier series decomposition which involves a discrete set of frequencies.

1.1.1.1 Sampling and Aliasing

In many practical applications, a discrete series is obtained by sampling a continuous function at equal intervals, Δt. For a sampling interval $\Delta t > 0$ and an arbitrary time offset t_0, we can define a discrete process through

$$X_t \equiv X(t_0 + t\Delta t),$$

for $t = 0, \pm 1, \pm 2, \ldots$. The frequency $1/(2\Delta t)$ is called the **Nyquist frequency** (or folding frequency) and defines the highest frequency that can be seen in discrete sampling. Higher frequencies sampled in this way will appear at lower frequencies called **aliases** (Shumway and Stoffer, 2000).

Example 1.1.3. In this example, we demonstrate the effect of aliasing by sampling from two different cosine curves (one at the Nyquist frequency and one over this value) at equal intervals $\Delta t = 1$. The results can be seen in Figure 1. In Figure 1, the dashed lines represent the underlying (continuous) functions from which we are sampling. The dashed black line represents a cosine curve with a frequency of $1/(2\Delta t) = 0.5$ (i.e. the Nyquist frequency) and an amplitude of 2. This function makes a cycle every two time units, therefore, the value of each observation of this function is zero (the black circles, Figure 1). The dashed blue line represents a cosine curve with a frequency of $1 = 2 \times 1/(2\Delta t)$ (i.e. double the Nyquist frequency) and an amplitude of 2. This function makes a cycle every time unit, therefore, the value of each observation of this function is also zero. This demonstrates how sampling the function with the higher frequency in this way would give the same results as sampling the function at the Nyquist frequency (known as aliasing).

Example 1.1.4. In Chapter 2, we analyse a dataset taken from a broad investigation of the effect of various salt stresses on the plant circadian clock. In this experiment, measurements were taken at intervals of approximately 45 minutes. Therefore, the Nyquist frequency (the highest

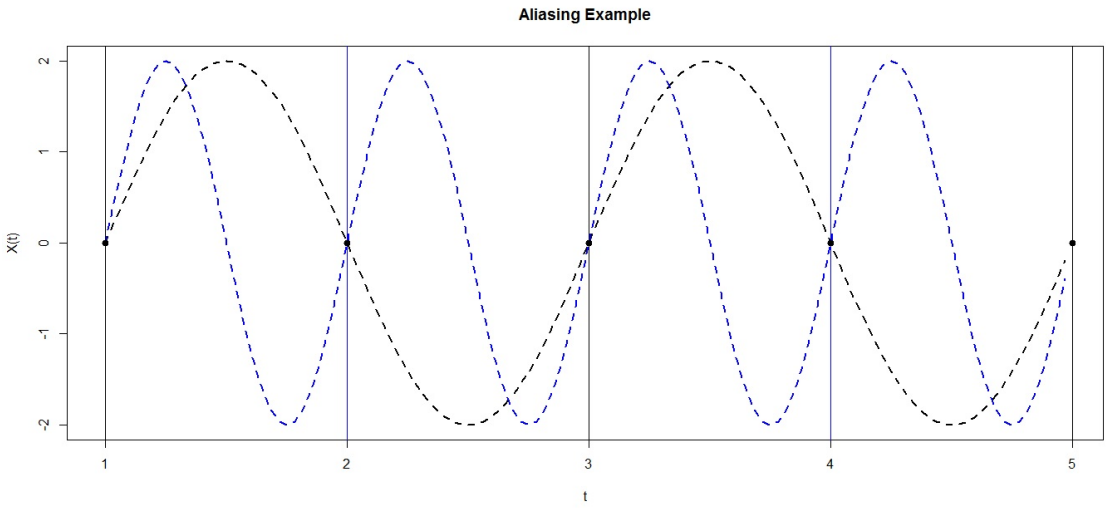

Figure 1: **Example 1.1.3**. Dashed black line: Underlying cosine curve with a frequency of $1/(2\Delta t) = 0.5$ (i.e. the Nyquist frequency) and an amplitude of 2; Dashed blue line: Underlying cosine curve with a frequency of $1 = 2 \times 1/(2\Delta t)$ (i.e. double the Nyquist frequency) and an amplitude of 2; Black circles: Observed value of underlying functions at $t = 1, 2, \ldots, 5$.

frequency that can be seen in discrete sampling) is

$$\frac{1}{2\Delta t} = \frac{1}{2 \times 0.75} = \frac{2}{3},$$

which is equivalent to a period of 1.5 hours.

1.1.1.2 Discrete Fourier Transform

We can now define the discrete Fourier transform as follows.

Definition 1.1.5. *Given data $X_1, \ldots, X_n$ we define the **discrete Fourier transform (DFT)** to be*

$$d(\omega_j) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_t \exp^{-2\pi i \omega_j t} \tag{2}$$

*for $j = 0, 1, \ldots, n-1$, where frequencies $\omega_j = j/n$ are called the **Fourier** or **fundamental frequencies**.*

Example 1.1.6. In this example, we create a simple time series, and then demonstrate how we can extract the frequency information using Fourier analysis. The time series is the sum of two underlying cosine curves: the first at a frequency of 6/128 with an amplitude of 2 and the second at a frequency of 10/128 with an amplitude of 4:

$$X(t) = 2\cos\left(2\pi t \frac{6}{128}\right) + 4\cos\left(2\pi t \frac{10}{128}\right). \tag{3}$$

The underlying cosine curves and resulting time series (sampled at $t = 1, \ldots, 128$) are shown in Figure 2. The (squared) DFT (called the periodogram– also see Section 1.3.1 later) is also plotted in Figure 2. Note that the periodogram is only non–zero at the frequencies 6/128 and 10/128 and the value of the periodogram at these values is equal to the amplitude of the corresponding

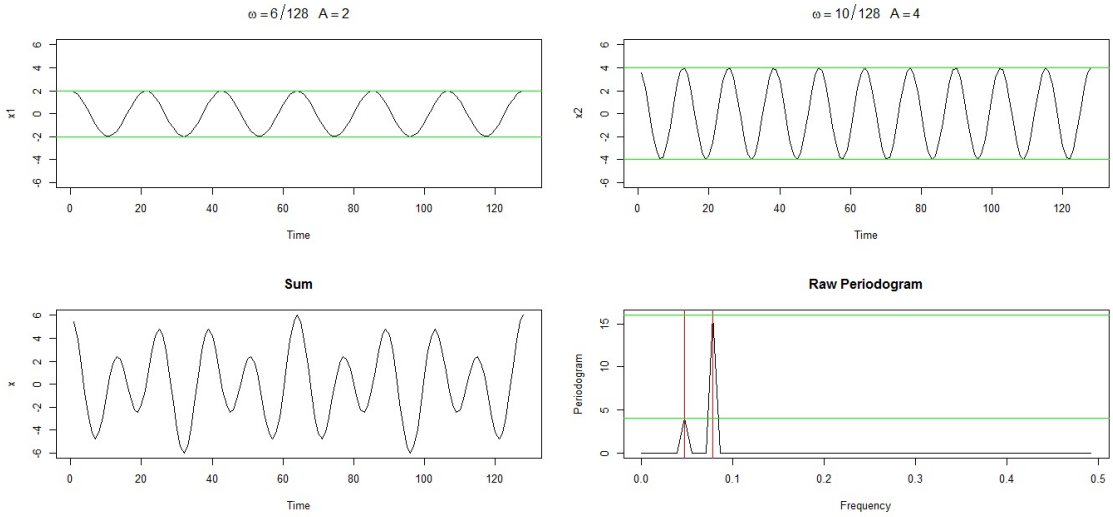

Figure 2: **Example 1.1.6**. Top left: First underlying cosine curve with a frequency of 6/128 and an amplitude of 2; Top right: Second underlying cosine curve with a frequency of 10/128 and an amplitude of 4; Bottom left: The time series is a linear combination of two underlying cosine curves (see equation (3)); Bottom right: raw periodogram of the series with the frequencies 6/128 and 10/128 indicated by vertical red lines and horizontal green lines indicating values of 4 and 16 (which correspond to the amplitudes of the underlying cosine components squared– see equation (3)).

underlying cosine curve squared. Therefore, the periodogram has correctly determined the underlying frequencies of our time series.

Example 1.1.7. In this example, we modify the time series from Example 1.1.6 such that the period of the series abruptly changes. The time series is now the concatenation of the above two underlying cosine curves:

$$X(t) = \begin{cases} 2\cos\left(2\pi t \frac{6}{128}\right), & t \in [1,128]. \\ 4\cos\left(2\pi t \frac{10}{128}\right), & t \in (128,256]. \end{cases} \tag{4}$$

The underlying cosine curves (sampled at $t = 1,\ldots,128$), resulting time series (sampled at $t = 1,\ldots,256$) and periodogram are shown in Figure 3. Note that the periodogram is almost identical to the periodogram in Figure 2. Therefore, this analysis has identified the periodicity of the data, however, it cannot detect changes of period through time. Such changes are common in many biological systems (see Section 1.3.2) and will call for more sophisticated methodology, able to cope with time–varying periods and amplitudes.

When representing a series by a combination of basis functions, it is usually desirable that the representation is sparse (i.e. there are only a small number of non-zero coefficients). This is because a sparse representation can aid understanding of the signal structure. Furthermore, a sparse representation also leads to better signal compression. In other words, we would prefer to represent a large number of data points by a much smaller number of basis coefficients instead. Sparse decompositions can be achieved by using basis functions with similar properties to the function that is being represented. Fourier functions are localised in frequency but not in time. Therefore, Fourier functions are suitable for representing smooth, periodic functions,

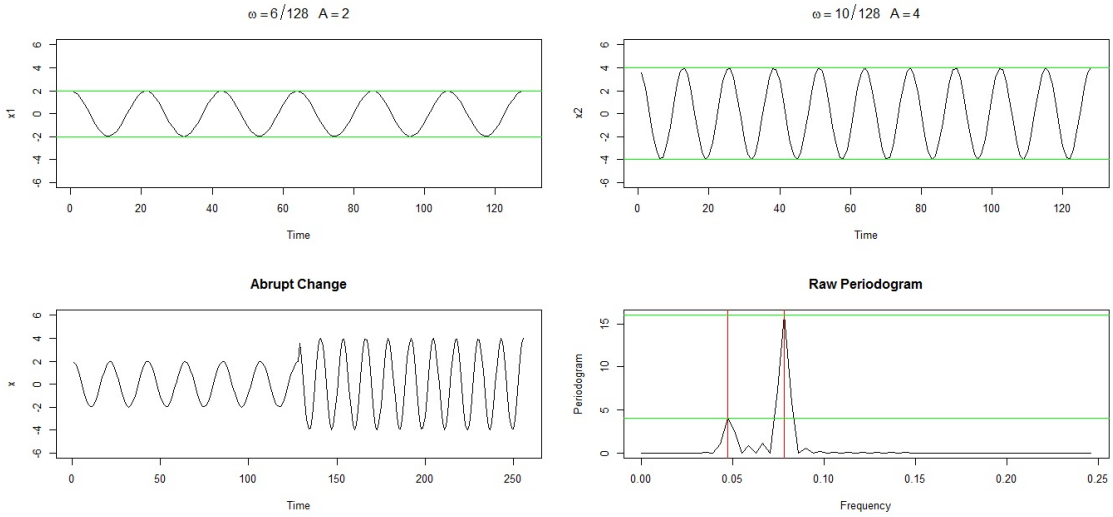

Figure 3: **Example 1.1.7**. Top left: First underlying cosine curve with a frequency of 6/128 and an amplitude of 2; Top right: Second underlying cosine curve with a frequency of 10/128 and an amplitude of 4; Bottom left: The time series is the concatenation of the two cosine curves (see equation (4)); Bottom right: raw periodogram of the series with the frequencies 6/128 and 10/128 indicated by vertical red lines and horizontal green lines indicating values of 4 and 16 (which correspond to the amplitudes of the underlying cosine components squared– see equation (4)).

but are not as suitable for functions with local features such as sharp changes and discontinuities. In order to represent such functions, we would prefer basis functions that have short support i.e. are localised in time. One solution is to use wavelets, which are described in the next section.

1.1.2 Wavelet Representations

The name "wavelet" gives us a clue as to two important properties of wavelets: this word appears to describe a "little wave" (as opposed to a "big wave" such as the trigonometric functions in Fourier theory). A "wavelet" can thus be thought of as a small, localised wave. This property makes it an ideal candidate to represent functions with local features (which proved problematic for the Fourier functions above). Formally, as in Daubechies (1992) we define a wavelet as follows.

Definition 1.1.8. *A **wavelet** is any square integrable function, $\psi \in L^2(\mathbb{R})$ which satisfies the **admissibility condition**,*

$$C_\psi = \int_{\mathbb{R}} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty, \tag{5}$$

where $\hat{\psi}(\omega)$ is the Fourier transform of $\psi(x)$ (see definition 1.1.2).

The admissibility condition (5) implies

$$\int_{-\infty}^{\infty} \psi(x)\, dx = 0, \tag{6}$$

which ensures its oscillatory behaviour (Vidakovic, 1999).

A wavelet basis can be formed by *translating* and *dilating* a basis function called the **mother wavelet**, which we will denote $\psi(x)$. In this thesis, we focus on wavelet functions whose dyadic

dilations and translations form an orthonormal basis of $\mathbb{L}^2(\mathbb{R})$. Formally, the collection of functions $\{\psi_{j,k}\}_{j,k\in\mathbb{Z}}$, defined by:

$$\psi_{j,k}(x) = 2^{\frac{j}{2}}\psi(2^j x - k), \tag{7}$$

known as a **discrete (decimated) wavelet family**, forms an orthonormal basis of $\mathbb{L}^2(\mathbb{R})$. The functions $\psi_{j,k}(x)$ in equation (7) are the wavelets generated by the mother ψ. Informally, this demonstrates that once the "type" of wavelet has been chosen and fixed (in this case, the ψ function) we can now generate other wavelets by transforming the mother wavelet. In particular, we can generate wavelets (in our case, the $\psi_{j,k}$'s) by dilating and translating the mother wavelet. In fact, the parameters j and k in equation (7) are known as the **dilation** and **translation parameters** respectively. The dilation parameter indicates the wavelet scale (see section 1.2.1) and the translation parameter indicates the location. The wavelet family then forms an orthonormal basis of $\mathbb{L}^2(\mathbb{R})$ and is analogous to the sine and cosine functions used in Fourier analysis.

When for each $k \in \{0,\dots,m\}$, we have

$$\int_{-\infty}^{\infty} x^k \psi(x)\,dx = 0, \tag{8}$$

the wavelet ψ in equation (8) is said to have $m+1$ **vanishing moments**. The vanishing moments property of a wavelet implies that the wavelet coefficients of polynomials of degree m or less are zero in a decomposition on such a wavelet basis. Therefore, this property has important implications when selecting a wavelet basis that would give a sparse representation of a given function.

Example 1.1.9. The Haar Basis. The simplest wavelet is the *Haar wavelet* (see Figure 4) and we discuss it as an introductory example throughout this review. The Haar wavelet is commonly used to introduce the topic of wavelets due to its simplicity, yet it displays many characteristic features of wavelets.

The Haar mother wavelet is a mathematical function, $\psi^H : \mathbb{R} \to \{\pm 1, 0\}$, defined by

$$\psi^H(x) = \begin{cases} 1, & \text{if } x \in [0, 1/2). \\ -1, & \text{if } x \in [1/2, 1). \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Using equation (7), the translations for $j,k \in \mathbb{Z}$ of the Haar mother wavelet are given by

$$\psi^H_{j,k}(x) = \begin{cases} 2^{\frac{j}{2}}, & \text{if } x \in \left[\frac{k}{2^j}, \frac{k}{2^j} + \frac{1}{2^{j+1}}\right). \\ -2^{\frac{j}{2}}, & \text{if } x \in \left[\frac{k}{2^j} + \frac{1}{2^{j+1}}, \frac{k}{2^j} + \frac{1}{2^j}\right). \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

and example plots for various values of j and k are given in Figure 4.

As in Fourier analysis, we can use wavelet functions as a basis to represent other functions. Recall (above) that dilations and translations of a mother wavelet function $\psi(x)$ (ie. $\psi_{j,k}$) define an orthonormal basis in $L^2(\mathbb{R})$. Throughout this thesis, we will consider only real–valued wavelet functions, therefore, we define the wavelet representation of a function f as follows.

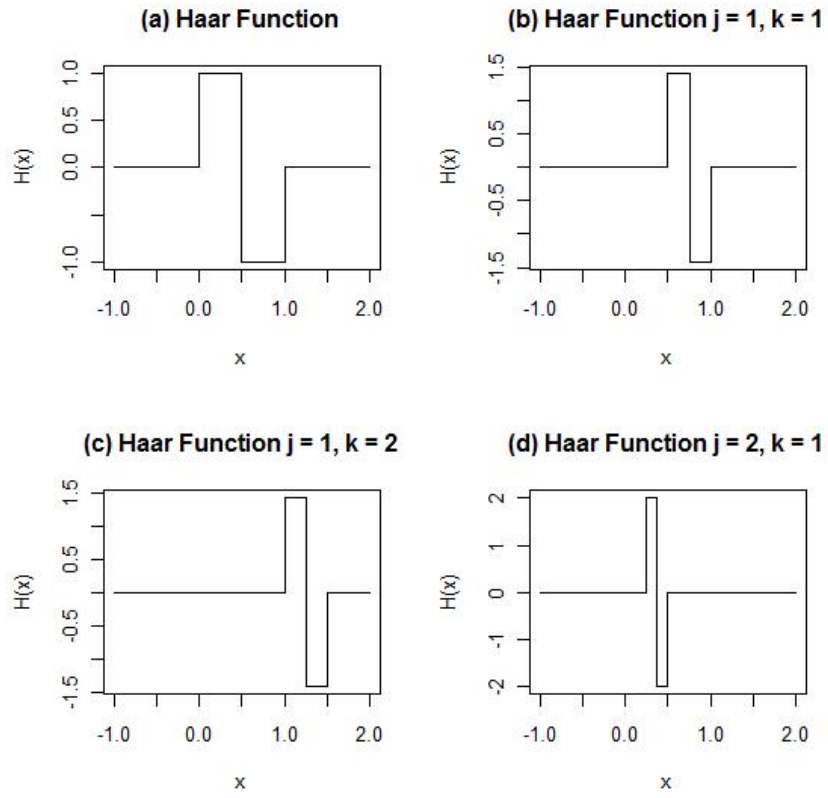

Figure 4: Panel (a): Haar mother wavelet. Panels (b), (c), (d): translations and dilations of the Haar mother wavelet (using equation (10) for various combinations of $j = 1,2$ and $k = 1,2$).

Definition 1.1.10. *Given a function $f \in L^2(\mathbb{R})$, its **wavelet representation** is given by*

$$f(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} d_{j,k} \psi_{j,k}(x), \tag{11}$$

where, due to the orthogonality of wavelets, for $j,k \in \mathbb{Z}$:

$$d_{j,k} = \int_{-\infty}^{\infty} f(x) \psi_{j,k}(x) dx = < f, \psi_{j,k} >, \tag{12}$$

where $< \cdot, \cdot >$ is the L^2-inner product.

The numbers $\{d_{j,k}\}_{j,k \in \mathbb{Z}}$ are referred to as the **wavelet coefficients** of f. As for the Fourier coefficients discussed in Section 1.1.1, the wavelet coefficients also provide information about the structure of the function, f. However, we note that the Fourier coefficients only provide information about the amplitude associated with each frequency, whereas the wavelet coefficients provide information about the amplitude of the wavelet at both a given (time) location and scale (associated with frequency, see Section 1.2.1).

1.2 Wavelet Theory

1.2.1 Multiresolution Analysis

A common way of introducing wavelet bases and demonstrating their properties is to construct them within the framework of a multiresolution analysis (MRA), introduced by Mallat

(1989a,b). An MRA provides a mathematical framework for looking at functions at different resolution levels or scales. Essentially, an MRA of, for example, the space of square integrable functions, $L^2(\mathbb{R})$, allows for the approximation of any function $f \in L^2(\mathbb{R})$, at different resolutions by projecting the function f onto a sequence of approximation spaces. Informally, we can think of the approximations at different resolution levels in terms of a camera "zooming" in and out: a higher resolution level is equivalent to zooming in and obtaining a fine detailed representation, whereas a lower resolution level is equivalent to zooming out and obtaining a coarse representation. In this section, we will briefly discuss some of the important features of an MRA as presented in Mallat (1989a,b), Fan and Gijbels (1996) and Nason (2010).

Definition 1.2.1. *A **multiresolution analysis** of $L^2(\mathbb{R})$ is a chain of nested closed subspaces, $\{V_j\}_{j\in\mathbb{Z}}$ of $L^2(\mathbb{R})$,*

$$\ldots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \ldots \tag{13}$$

satisfying the following conditions:

1. *The spaces have trivial intersection:*

$$\bigcap_{j\in\mathbb{Z}} V_j = \{0\}.$$

2. *The union is dense in $L^2(\mathbb{R})$:*

$$\overline{\bigcup_{j\in\mathbb{Z}} V_j} = L^2(\mathbb{R}).$$

3. *The following scale relations exist:*

$$\begin{aligned} f(x) \in V_j &\iff f(2x) \in V_{j+1}, \forall x \in \mathbb{R}, \quad \text{and} \\ f(x) \in V_0 &\iff f(x-k) \in V_0, \forall k \in \mathbb{Z}, x \in \mathbb{R}. \end{aligned} \tag{14}$$

4. *There exists a **scaling function** $\phi(x) \in V_0$, with $\int_{-\infty}^{\infty} \phi(x)\,dx = 1$, such that $\{\phi(x-k), k \in \mathbb{Z}\}$ constitutes an orthonormal basis of V_0.*

Equations (14) of condition 3 along with condition 4, imply that $\{\phi_{j,k} := 2^{j/2}\phi(2^j x - k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis of $V_j, \forall\, j \in \mathbb{Z}$ (Vidakovic, 1999). Furthermore, since $\phi \in V_0 \subset V_1$, and $\{\phi_{1,k}\}_{k\in\mathbb{Z}}$ is an orthonormal basis of V_1, the function $\phi(x) \in V_0$ can be represented as a linear combination of functions from V_1:

$$\phi(x) = \sum_{k\in\mathbb{Z}} h_k \phi_{1,k}(x) = \sum_{k\in\mathbb{Z}} h_k 2^{\frac{1}{2}} \phi(2x-k) \tag{15}$$

for some coefficients $h_k, k \in \mathbb{Z}$, which form a vector that is referred to as a **low-pass filter**. Equation (15) is known as the **scaling equation** and is fundamental in the construction of wavelets.

This theoretical framework allows us to develop the mother wavelet function, $\psi(x)$, in terms of an MRA. We can think of the mother wavelet function, $\psi(x)$, as explaining the detail at each level j. In other words, it represents the information that is lost when moving from one approximation space, V_{j+1}, to the next (coarser) space, V_j. Now, consider the detail space, which we will denote W_j, to be the orthogonal complement of V_j in V_{j+1}, so that:

$$V_{j+1} = V_j \oplus W_j, \forall\, j \in \mathbb{Z}, \tag{16}$$

(where $\oplus$ denotes the direct sum of spaces). Repeated application of the relationship in equation (16) gives

$$V_{j+1} = V_0 \oplus \bigoplus_{i=0}^{j} W_i. \tag{17}$$

Furthermore, condition 2 states that the union, $\bigcup_{j\in\mathbb{Z}} V_j$, is dense in $L^2(\mathbb{R})$, therefore, taking the limit and using condition 1, we obtain

$$L^2(\mathbb{R}) = \bigoplus_{j\in\mathbb{Z}} W_i. \tag{18}$$

Therefore, an orthonormal basis for $L^2(\mathbb{R})$ could be obtained from the orthonormal bases for $W_j, \forall\, j \in \mathbb{Z}$. In particular, the spaces W_j inherit the scaling property (condition 5) from the V_j. Therefore, if $\psi(x)$ is a function such that its integer translations form an orthonormal basis of W_0, then through dyadic dilations and translations, $\{\psi_{j,k}(x)\}_{k\in\mathbb{Z}}$ is an orthonormal basis for the space W_j. Hence, $\{\psi_{j,k}(x)\}_{j,k\in\mathbb{Z}}$ provides an orthonormal basis for $L^2(\mathbb{R})$.

As in the derivation of the scaling equation (15), since $\psi(x) \in W_0 \subset V_1$, the function $\psi(x)$ can similarly be represented as a linear combination of the functions from V_1:

$$\psi(x) = \sum_{k\in\mathbb{Z}} g_k 2^{\frac{1}{2}} \phi(2x-k) \tag{19}$$

for some coefficients $g_k, k \in \mathbb{Z}$, which form a vector that is referred to as a **high-pass filter**.

Informally, we can think of the space V_j as the collection of functions with detail up to some finest resolution scale. This space can contain functions with less detail, but there is some maximum level of detail allowed in this collection. Here, larger values of j indicate V_j contains functions with finer detail. Therefore, if a function is in V_j then it is also in V_k if $k > j$. Intuitively, we can think of V_{j+1} as being "V_j plus some detail (W_j)" (see equation (16)). Therefore, an approximation of a function, f, at resolution level j is given by:

$$f_j(x) = \sum_{k\in\mathbb{Z}} c_{j,k}\phi_{j,k}(x) = P_j f \tag{20}$$

where P_j is the projection operator onto V_j. Essentially, in equation (20), we approximate the function at resolution level j by not including any of the detail from the finer scales. Note that as $\{\phi_{j,k}, k \in \mathbb{Z}\}$ are orthonormal, the $\{c_{j,k}\}$ may be obtained using:

$$c_{j,k} = < f, \phi_{j,k} > = \int_{-\infty}^{\infty} f(x)\phi_{j,k}(x)dx \tag{21}$$

as in equation (12). Intuitively, we start with a low-resolution function, f_j, and then add finer and finer detail by including a new layer of detail coefficients (the "zooming in" of our camera). Figure 5 illustrates this concept for successive resolution levels, j. We can see that the finer-scale approximations (with larger values of j) capture more and more of the detail of the original function.

1.2.2 The Discrete Wavelet Transform (DWT)

In many practical situations, functions or data sets are observed at a finite number of discrete time points. In such cases, the representation of a continuous function in Definition 1.1.10

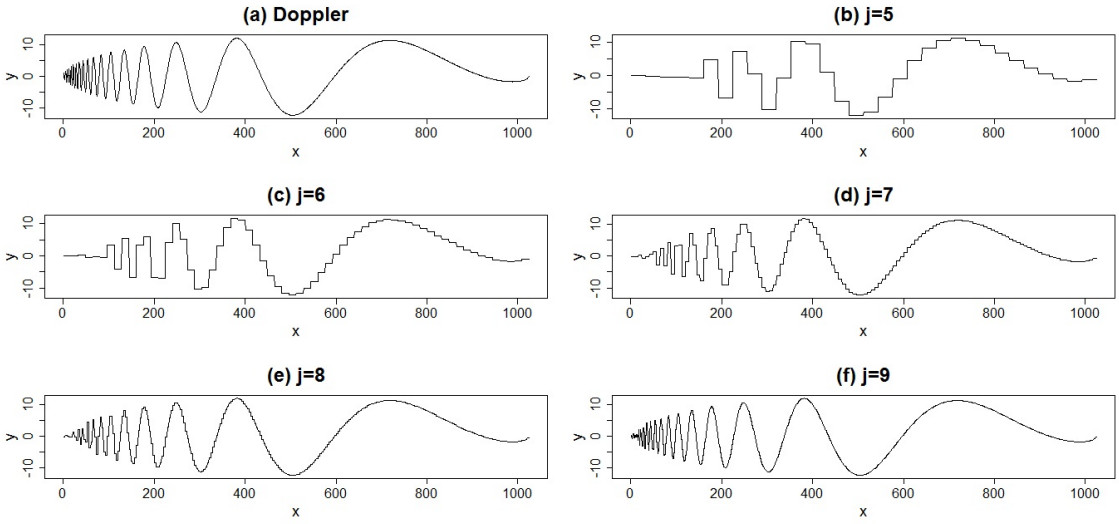

Figure 5: Successive approximations of the Doppler test function introduced by Donoho and Johnstone (1994) using the Haar wavelet basis. Plot (a) shows the original function, plots (b), (c), (d), (e) and (f) display successively finer scale approximations (where $j = 5, 6, 7, 8$ and 9 respectively).

would not be suitable. In this section, we introduce the discrete equivalent of equation (11) and discuss an efficient scheme for performing the discrete wavelet transform, **Mallat's Pyramid Algorithm** (Mallat, 1989a,b). Our description of the DWT is based largely upon Vidakovic (1999) and Nason (2010).

The basic premise of this method is to filter the data sequence using the low pass filter, $\mathcal{H} = \{h_k\}$, and high pass filter, $\mathcal{G} = \{g_k\}$, associated with the scaling equations (15) and (19) in Section 1.2.1, to obtain the wavelet coefficients at different levels. Essentially, we start with a data sequence and compute coarser level wavelet coefficients using a relation which we derive next.

Assume a function, f, is observed at $N = 2^J$ equally spaced locations $\{x_i, i = 0, \ldots, N-1\}$. First, interpolate the observations by using the basis of scaling functions from the space V_J. Set $c_{J,i} = f(x_i)$ for $i = 0, \ldots, N-1$, then a function $\bar{f}$ can be constructed using $\{\phi_{J,k}(x)\}_{k\in\mathbb{Z}}$ as follows:

$$\bar{f}(x) = \sum_k c_{J,k}\phi_{J,k}(x). \tag{22}$$

The function $\bar{f}$ can be used as an approximation of the observed function f. Consequently, the wavelet coefficients of $\bar{f}$ are actually an approximation of the wavelet coefficients $d_{j,k} =< f, \psi_{j,k} >$ of the observed function f, and are sometimes referred to as the **empirical** wavelet coefficients of f. The empirical wavelet coefficients are approximately proportional to their continuous counterparts (see e.g. Abramovich et al. (2000)).

To obtain the empirical wavelet coefficients, note that $\bar{f}$ is an element of V_J (since $\{\phi_{j,k}, k \in \mathbb{Z}\}$ is a basis of V_j). Equation (16) implies that any function $v_j \in V_j$ may be represented uniquely as:

$$v_j(x) = v_{j-1}(x) + w_{j-1}(x)$$

where $v_{j-1} \in V_{j-1}$ and $w_{j-1} \in W_{j-1}$. Recall: $\{\phi_{j,k}, k \in \mathbb{Z}\}$ is a basis of V_j and $\{\psi_{j,k}, k \in \mathbb{Z}\}$ is a

basis of W_j. Therefore:

$$v_j(x) = v_{j-1}(x) + w_{j-1}(x) = \sum_l c_{j-1,l}\phi_{j-1,l}(x) + \sum_l d_{j-1,l}\psi_{j-1,l}(x) \tag{23}$$

for some coefficients $\{c_{j,l}\}$ and $\{d_{j,l}\}$ known as the **smooth** and **detail** coefficients of the transformation respectively. This is because $\{c_{j,l}\}$ provides a coarser description of the original function and $\{d_{j,l}\}$ extracts the features lost when representing the function in a coarser version.

To obtain the smooth coefficients of the transform, equation (23) together with the orthogonality of the $w_{j-1}(x)$ and $\phi_{j-1,l}(x)$ imply that:

$$c_{j-1,l} =< v_j, \phi_{j-1,l} >, \tag{24}$$

and by equations (7) and (15):

$$\phi_{j-1,l}(x) = \sum_k h_{k-2l}\phi_{j,k}(x). \tag{25}$$

Therefore, substituting (25) into equation (24), we obtain:

$$\begin{aligned} c_{j-1,l} &=< v_j, \sum_k h_{k-2l}\phi_{j,k} > \\ &= \sum_k h_{k-2l} < v_j, \phi_{j,k} > \\ &= \sum_k h_{k-2l} c_{j,k}, \end{aligned}$$

where the last line follows from equation (24). An equation to obtain the detail coefficients can be developed in a similar way. To summarise, the DWT of the sequence is then obtained recursively using the relations:

$$c_{j-1,l} = \sum_k h_{k-2l}c_{j,k} \quad \text{and} \quad d_{j-1,l} = \sum_k g_{k-2l}c_{j,k} \tag{26}$$

to obtain

$$\mathbf{d} = (c_{0,0}, \mathbf{d_{J-1}}, \mathbf{d_{J-2}}, \ldots, \mathbf{d_1}, d_{0,0}), \tag{27}$$

where $\mathbf{d_j}$ is the vector of coefficients, $\mathbf{d_j} = (d_{j,0}, \ldots, d_{j,2^j-1})$.

By examining the relations in (26), we can see that the coarser level coefficients are given by multiplying the data sequence by the coefficients g_k and h_k given in the scaling equations (15) and (19). These coefficients are specific to the wavelet selected to perform the decomposition.

Figure 6 gives a visual representation the implementation of this algorithm. This figure illustrates that at each step of the algorithm, an input vector, $\mathbf{c_j}$, is transformed into two output vectors, $\mathbf{c_{j-1}}$ and $\mathbf{d_{j-1}}$, using the filters defined in (26). Furthermore, note that the output of each step, $\mathbf{c_{j-1}}$, becomes the input for the next step of the algorithm, producing vectors $\mathbf{c_{j-2}}$ and $\mathbf{d_{j-2}}$ and so on. The resulting wavelet transform in (27) is the collection of detail coefficients at each level together with the smooth or father coefficient at the zero level. Also, note the $2l$ term in the relations in (26). This represents the **decimation** step in the DWT. In other words, it ensures the number of coefficients is halved at each level. To illustrate this, the number of coefficients at each level is displayed in the orange boxes in Figure 6.

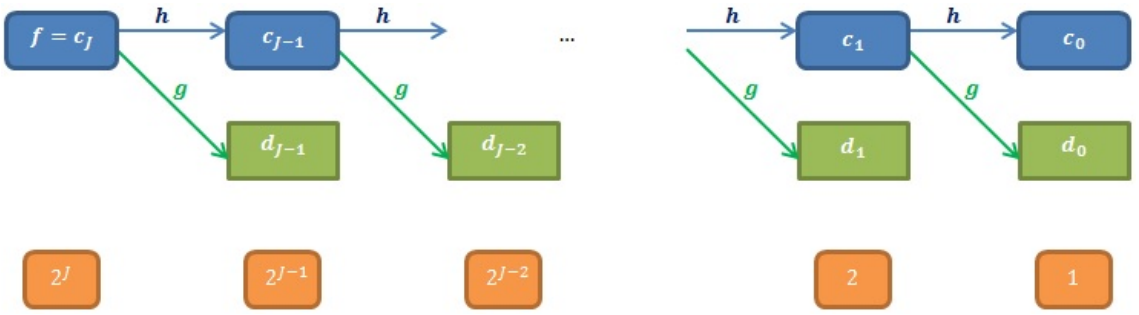

Figure 6: Flow diagram of the discrete wavelet transform of an observed dataset, $\mathbf{c_j}$, using successive applications of the low and high pass filters g and h. The orange boxes (below) give the number of coefficients at each level.

Finally, we note that it is possible to reconstruct the original data series from the output coefficients in equation (27). In order to do this we perform the **inverse (discrete) wavelet transform** (IWT). The inverse relation is given by:

$$c_{j,n} = \sum_k h_{n-2k} c_{j-1,k} + \sum_k g_{n-2k} d_{j-1,k}, \tag{28}$$

where h_n and g_n are known as the **quadrature mirror filters** defined by (15) and (19) (Mallat, 1989b). Note that the filters associated with the inverse transform have the same structure as those that computed the forward transform in (26).

To summarise, the IWT takes the coarsest level father and mother coefficients and uses them to reconstruct the next finer level using equation (28). The reconstruction of the original data sequence is then achieved by iterating this process and climbing the resolution levels back to the original data.

Example 1.2.2. Figure 7 shows a plot of the "Doppler" test function introduced by Donoho and Johnstone (1994) along with a plot of the Haar wavelet transform (the detail coefficients at each level) of the Doppler function. Each coefficient is depicted by a small vertical line (the bigger the vertical line, the larger the wavelet coefficient). The coefficients $d_{j,k}$, corresponding to the same resolution level j, are arranged along an imaginary horizontal line. Note that the number of coefficients is halved at each resolution level.

The oscillatory nature of the Doppler signal is clearly visible in the wavelet coefficients, especially at the finer scales (resolution levels 6–9). Large variation in the fine–scale coefficients corresponds with the high frequencies in the Doppler function whereas large variation in coarser–level coefficients corresponds with lower frequencies. Thus, the plot of wavelet coefficients can be thought of as a time-frequency display of the varying frequency information contained with the Doppler signal.

Finally, the Daubechies 'extremal-phase' (with eight vanishing moments) wavelet coefficents are also plotted in the bottom right subplot of Figure 7. As discussed in Section 1.1.1, sparse decompositions can be achieved by using basis functions with similar properties to the function that is being represented. Therefore, the smoother wavelet with a higher number of vanishing moments, has resulted in a sparser representation of the Doppler signal than the Haar wavelet.

Example 1.2.3. A Numerical Example of the DWT. Suppose that we begin with the following

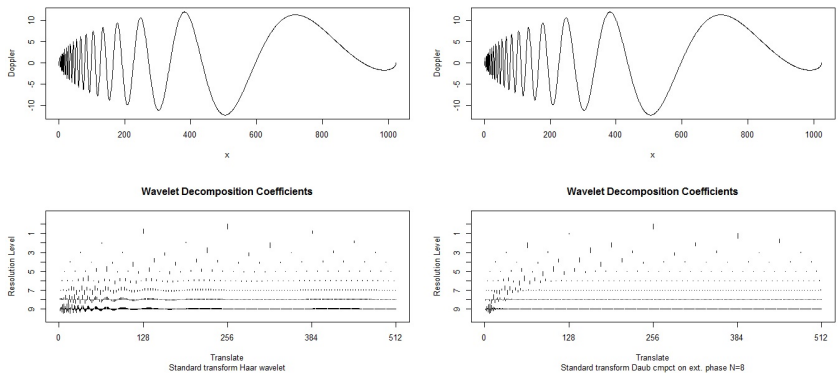

Figure 7: Top row: left and right: identical copies of the Doppler function. Bottom left: Haar discrete wavelet coefficients, $\{d_{j,k}\}$, of Doppler function (plotted with a different scale for each resolution level). Bottom right: as left but with Daubechies 'extremal-phase' with 8 vanishing moments. Note the smoother wavelet with a higher number of vanishing moments, has resulted in a sparser representation of the Doppler signal than the Haar wavelet

data sequence (from Nason (2010)):

$$\mathbf{y} = (y_0, \ldots, y_{N-1}) = (1,1,7,9,2,8,8,6).$$

In this example, we will find the wavelet decomposition using the Haar basis from Example 1.1.9. The low and high pass filters for the Haar basis are:

$$\mathscr{H} = (h_0, h_1) = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) \quad \text{and} \quad \mathscr{G} = (g_0, g_1) = \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right). \tag{29}$$

Since there are eight elements of $\mathbf{y}$, $N = 8 = 2^J$ and hence $J = 3$. Recall: we set $\mathbf{c_J}$ equal to our original data sequence. Therefore, $\mathbf{c_3} = \mathbf{y}$. Repeatedly applying equation (26), we obtain the output, as in (27):

$$(21\sqrt{2}/2, 0, -\sqrt{2}, -3\sqrt{2}, \sqrt{2}, -7, -2, -3\sqrt{2}/2).$$

The computations are displayed in a graphical form in Figure 8. On examining Figure 8, we note that the coefficients can be visualized as an inverted pyramid (hence the name "Pyramid Algorithm"). It is also useful to note the decimation step of the DWT whilst examining this representation. We can see here that we use two coefficients from the previous level to calculate the next coefficient, and then move on to the next (non-overlapping) pair. This will be useful to bear in mind when we discuss the nondecimated wavelet transform in Section 1.2.4.

1.2.3 Matrix Representation of the Discrete Wavelet Transform

Example 1.2.3 above illustrates that the DWT takes a vector input and produces a set of output coefficients that can also be represented as a vector, as in equation (27). Furthermore, we note that since the output vector has been obtained from the input using a series of summations and scalings, we can alternatively compute the output from the input using matrix multiplication. Therefore, an alternative way to formulate the DWT is to construct an orthogonal matrix W associated with the particular wavelet being used.

More formally, note that at each step of the DWT, the input signal is represented on two

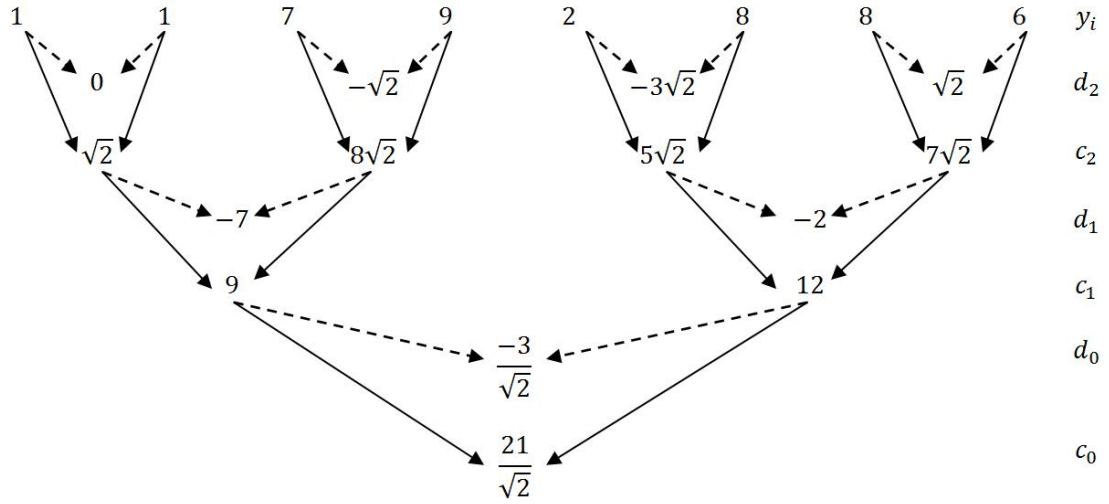

Figure 8: **Graphical depiction of the DWT.** The dotted arrows represent applying the filter $\mathscr{G}$ and the solid arrows represent applying the filter $\mathscr{H}$ (i.e. the application of the relations in (26)). This figure is reproduced following Figure 2.2 in Nason (2010).

different bases (see equation 23). Since any change of basis of this type can be represented by matrix multiplication, it follows that the DWT can also be represented in this manner. Furthermore, since the bases used for representing the signal at each step are orthonormal, the matrix W is an orthogonal matrix (i.e. $W^T W = I_{2^J}$, where I_n is the identity matrix of order n). The DWT can then be formally defined as the matrix multiplication of the orthogonal matrix W with a vector of data points, $\mathbf{y}$:

$$\mathbf{d} = W\mathbf{y}, \tag{30}$$

where $\mathbf{d}$ is the output vector comprising both the discrete mother and father wavelet coefficients defined in (27).

Finally, recall from Section 1.2.2 that it is possible to reconstruct the original data series from the output coefficients using the IWT. We can also develop the inverse discrete wavelet transform in matrix notation. In particular, multiplying both sides of equation (30) by the inverse of the matrix W gives:

$$W^{-1}\mathbf{d} = \mathbf{y}. \tag{31}$$

Therefore, the original data is obtained by pre–multiplying the output vector of coefficients by the inverse of the matrix W. Finally, recall that matrix W was orthogonal (so $W^{-1} = W^T$), which implies that the inverse discrete wavelet transform in matrix notation is W^T.

1.2.4 The Nondecimated Wavelet Transform

In Section 1.2.2, we noted that the $2l$ term in the relations in (26) represents the decimation step in the DWT. Furthermore, recall the discussion of Figure 8 in Example 1.2.3– to calculate one coefficient at a particular level, we use two coefficients from the previous level and then move on to the next non-overlapping pair for the next coefficient to be calculated. Hence, the $2l$ in the index of the summations in (26) essentially picks every even element from a vector.

For example, in Example 1.2.3 we calculated:

$$d_{2,0} = (y_0 - y_1)/\sqrt{2}$$
$$d_{2,1} = (y_2 - y_3)/\sqrt{2}.$$

The first two coefficients encode the difference between (y_0, y_1) and (y_2, y_3) respectively. But what about the information that might be contained in the difference between y_2 and y_1? One of the motivations behind the **nondecimated wavelet transform** (NDWT) is to "fill in the gaps" caused by the decimation step in the discrete wavelet transform (Nason and Silverman, 1995).

Example 1.2.4. We begin by returning to Example 1.2.3 in Section 1.2.2. If we shifted the original sequence cyclically by one position, we would obtain the sequence:

$$(y_7, y_0, \ldots, y_6). \tag{32}$$

Then, taking the Haar wavelet transform as before gives:

$$d_{2,1} = (y_1 - y_2)/\sqrt{2},$$

i.e. the "missing information" outlined above. Applying the transform to the shifted sequence in (32) obtains the "missing" odd elements of the filter vector.

Therefore, to obtain more information about the data, we could calculate both the original set of (even) wavelet coefficients and the coefficients that resulted after shifting and transforming the sequence (the odd coefficients). However, as a result, the orthogonal structure of the DWT is lost. Furthermore, the extra transformation is redundant. In particular, we could use either the original or the shifted coefficients to reconstruct the original sequence using the IWT.

Another undesirable property of the DWT is that it is not **translation invariant**. In particular, an undesirable consequence of the decimation step is that a shift in the data leads to a non-trivial change in the wavelet transform. Thus, the DWT of a shifted data set is not a shift of the DWT of the original data. However, the NDWT of a shifted data set is a shift of the NDWT of the original data.

Example 1.2.5. In this example, we return to the example dataset from Section 1.2.2, which is plotted in Figure 9(a). Figure 9(b) depicts the same data sequence rotated by a simple unit shift (as in (32)), and the detail coefficients associated with the original and shifted sequences. Note how the detail coefficients associated with the shifted sequence (Figure 9(d)) do not correspond to a simple shift of the detail coefficients associated with the original sequence. However, the NDWT of a shifted data set is a shift of the DWT of the original data (see Figure 10). Note how the coefficients in Figure 10(b) are a unit shift of the coefficients displayed in Figure 10(a).

In order to describe the NDWT, we formally introduce some notation. Firstly, define the action of a filter $\mathscr{P}$ on a sequence (or vector) $\mathbf{x} = \{x_n\}$ by

$$(\mathscr{P}\mathbf{x})_j = \sum_n p_{n-j} x_n.$$

Now define the **even dyadic decimation operator** $\mathscr{D}_0$ by:

$$(\mathscr{D}_0\mathbf{x})_l = x_{2l}. \tag{33}$$

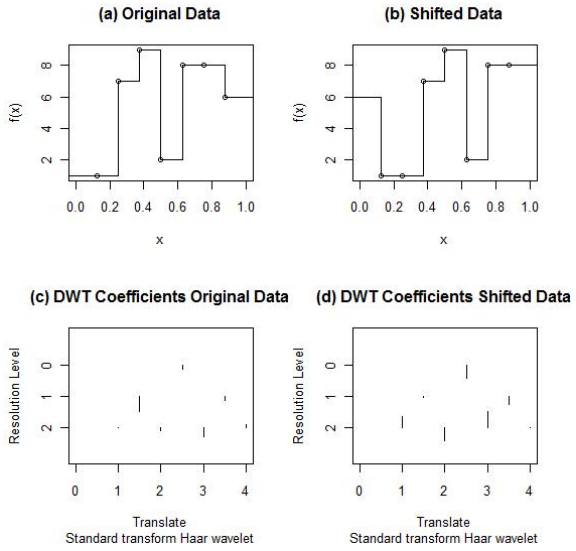

Figure 9: **Example 1.2.5: the DWT is not translational invariant.** Figure (a) depicts the original data sequence whilst (b) depicts the same sequence rotated by a simple unit shift. Figures (c) and(d) depict the detail coefficients of the Haar DWT for the original and shifted data respectively. Note that the coefficients in Figure (d) do not correspond to a simple shift of the coefficients displayed in Figure (c).

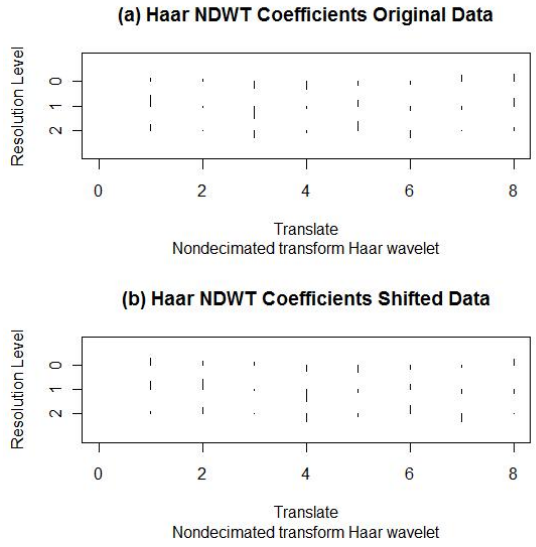

Figure 10: **Example 1.2.5 of the translational invariance of the NDWT.** Figure (a) depicts the NDWT Haar wavelet detail coefficients of the original data. Figure (b) depicts the NDWT Haar wavelet detail coefficients of the shifted data. Observe that the coefficients in Figure (b) are a unit shift of the coefficients displayed in Figure (a).

Therefore, $\mathscr{D}_0$ represents selecting every other element of the filter vector (in this case, the even elements). Using this notation, we can write the operations described by (26) more succinctly as:

$$\mathbf{c_{j-1}} = \mathscr{D}_0\mathscr{H}\mathbf{c_j} \quad \text{and} \quad \mathbf{d_{j-1}} = \mathscr{D}_0\mathscr{G}\mathbf{c_j}, \tag{34}$$

where $\mathscr{H}$ and $\mathscr{G}$ denote the low and high pass filters respectively (see Section 1.2.2). Note that in (34) we have denoted the inputs and outputs of these operations using vector notation $\mathbf{c_J}, \mathbf{c_{J-1}}, \mathbf{d_{J-1}}$ rather than indexed sequences. Similarly, define the **odd dyadic decimation operator** $\mathscr{D}_1$ by:

$$(\mathscr{D}_1\mathbf{x})_l = x_{2l+1}. \tag{35}$$

Therefore, $\mathscr{D}_1$ does exactly the same as $\mathscr{D}_0$, except it takes the odd elements of the filter vector instead.

We will now describe the NDWT as in Nason and Silverman (1995) and Nason (2010). The basic idea of the NDWT is to retain both the odd and even decimations at each scale and continue to do the same at each subsequent scale.

Definition 1.2.6. The nondecimated wavelet transform.

1. *Given the input vector* $\mathbf{y} = (y_0, \ldots, y_{N-1})$, *apply and retain both* $\mathscr{D}_0\mathscr{G}\mathbf{y}$ *and* $\mathscr{D}_1\mathscr{G}\mathbf{y}$ *(the odd and even indexed filtered observations).*

2. *Perform a similar operation to obtain the finest-scale father wavelet coefficients and compute* $\mathscr{D}_0\mathscr{H}\mathbf{y}$ *and* $\mathscr{D}_1\mathscr{H}\mathbf{y}$.

3. *For the next level wavelet coefficients, apply both* $\mathscr{D}_0\mathscr{G}$ *and* $\mathscr{D}_1\mathscr{G}$ *to both* $\mathscr{D}_0\mathscr{H}\mathbf{y}$ *and* $\mathscr{D}_1\mathscr{H}\mathbf{y}$.

4. *Similarly, to obtain the father wavelet coefficients at this level, apply both* $\mathscr{D}_0\mathscr{H}$ *and* $\mathscr{D}_1\mathscr{H}$ *to both* $\mathscr{D}_0\mathscr{H}\mathbf{y}$ *and* $\mathscr{D}_1\mathscr{H}\mathbf{y}$.

5. *Continue in this manner, applying* $\mathscr{D}_0\mathscr{G}$ *and* $\mathscr{D}_1\mathscr{G}$ *and* $\mathscr{D}_0\mathscr{H}$ *and* $\mathscr{D}_1\mathscr{H}$ *to each father wavelet coefficient in the previous level.*

The NDWT is useful for studying (nonstationary) time series, as discussed in Section 1.3.

Example 1.2.7. NDWT using the Haar Basis. To summarise, when performing the DWT using the Haar basis, to calculate one coefficient at a particular level, we use two adjacent coefficients from the previous level and then move on to the next non-overlapping pair for the next coefficient to be calculated. However, when performing the NDWT for the Haar basis, to calculate one coefficient at a particular level, we use two adjacent coefficients from the previous level but then move on to the next pair for the next coefficient to be calculated.

For example, given a data sequence (y_0, y_1, y_2, y_3), we would calculate:

$$\begin{aligned} c_{1,0} &= (y_0 - y_1)/\sqrt{2} \\ c_{1,1} &= (y_1 - y_2)/\sqrt{2} \\ c_{1,2} &= (y_2 - y_3)/\sqrt{2} \\ c_{1,3} &= (y_3 - y_0)/\sqrt{2}. \end{aligned}$$

1.3 Stationary Time Series Analysis

A time series is a set of random variables recorded sequentially through time. The analysis of experimental data that have been observed at different points in time leads to specific challenges in statistical modelling and inference. This is because successive time series observations are (generally) not independent. The correlation introduced by the sampling of adjacent points in time means that many conventional statistical methods (traditionally dependent on the assumption that adjacent observations are independent and identically distributed) are not applicable. The systematic approach by which one goes about answering the mathematical and statistical questions posed by these time correlations is commonly referred to as time series analysis.

The impact of time series analysis in many different applications is highlighted by listing the diverse fields in which important time series problems may arise. For example, economics (e.g. daily stock market quotations or monthly unemployment figures); meteorology (e.g. measurements of rainfall or temperature) and medicine (e.g. blood pressure measurements or magnetic resonance imaging of brain activity). In particular, this thesis shall consider the application of time series analysis to data originating from various experiments in the field of circadian biology.

In this section, we begin by stating some key results in stationary time series analysis following Priestley (1982), Shumway and Stoffer (2000), Brillinger (2001), and Percival and Walden (2006). Intuitively, a time series is stationary if its statistical characteristics are assumed constant over time. This means that parameters such as the mean and variance (if they exist) do not change over time. This foundation will then allow us to describe and contrast how wavelets can be used to analyse nonstationary time series in Section 1.4.

Definition 1.3.1. *A stochastic process $\{X_t, t \in \mathcal{T}\}$ is said to be **strictly stationary** if, for all $n \geq 1$, for any $t_1, \ldots, t_n \in \mathcal{T}$, and for any τ such that $t_1 + \tau, ..., t_n + \tau \in \mathcal{T}$ are also contained in the index set, $\mathcal{T}$, the joint distribution function of $\{X_{t_1}, \ldots, X_{t_n}\}$ is the same as that of $\{X_{t_1+\tau}, \ldots, X_{t_n+\tau}\}$.*

Often this assumption is relaxed to that of weak or second-order stationarity:

Definition 1.3.2. *A stochastic process $\{X_t, t \in \mathcal{T}\}$ is said to be **weakly stationary** or **second–order stationary** if, for all $n \geq 1$, for any $t_1, \ldots, t_n \in \mathcal{T}$, and for any τ such that $t_1 + \tau, ..., t_n + \tau \in \mathcal{T}$ are also contained in the index set, $\mathcal{T}$, all the joint moments of orders 1 and 2 of $\{X_{t_1}, \ldots, X_{t_n}\}$ exist, are finite and are equal to the corresponding joint moments of $\{X_{t_1+\tau}, \ldots, X_{t_n+\tau}\}$.*

Hence,

$$\mathbb{E}(X_t) = \mu_X \quad \text{and} \quad \text{Cov}(X_t, X_{t+\tau}) = \gamma(\tau), \tag{36}$$

where $\mu_X \in \mathbb{R}$. Therefore, the autocovariance of a weakly stationary time series is dependent only on the time lag, τ, and not the value of time.

Example 1.3.3. A sequence $\{Z_t, t \in \mathbb{Z}\}$ of uncorrelated random variables with mean zero and finite variance σ_Z^2 (often called a **purely random process** or **white noise**) is a weakly stationary process (Shumway and Stoffer, 2000).

A sequence of independent and identically distributed (iid) random variables, W_t, with mean zero and finite variance σ_W^2 (known as **white independent noise**) is both a strictly and

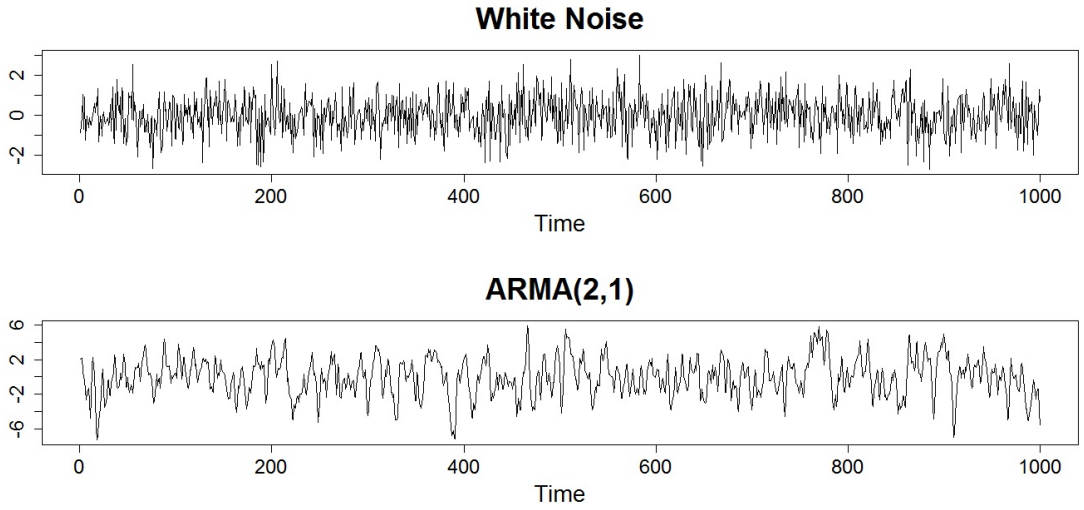

Figure 11: **Stationary processes**. Top: An example realisation of a white noise process (Example 1.3.3) of length $T = 1000$. Bottom: An example realisation of a stationary ARMA(2, 1) process (Example 1.3.4) of length $T = 1000$ with AR parameters $(\alpha_1, \alpha_2) = (0.9, -0.2)$ and MA parameter of 0.5.

weakly stationary process (Shumway and Stoffer, 2000). A common example of a white (independent) noise series is **Gaussian white noise**, wherein the W_t are independent normal random variables, with mean 0 and variance σ_W^2. An example realisation of a Gaussian white noise process (with variance $\sigma_W^2 = 1$) can be found in Figure 11.

Example 1.3.4. Autoregressive moving average (ARMA) processes are one of the most commonly used time series models. An ARMA(p, q) process X_t is defined as

$$X_t = \sum_{j=1}^{p} \alpha_j X_{t-j} + Z_t + \sum_{i=1}^{q} \beta_i Z_{t-i}, \tag{37}$$

where Z_t is a white noise process (see Example 1.3.3). An ARMA(p, q) process is stationary if the polynomial

$$\alpha(\lambda) = 1 - \alpha_1 \lambda - \cdots - \alpha_p \lambda^p \tag{38}$$

has no roots inside the unit circle (Shumway and Stoffer, 2000). An example realisation of a stationary ARMA(2, 1) process can be found in Figure 11.

1.3.1 Fourier Analysis of Stationary Time Series

The Cramér-Rao representation of stationary processes (Priestley, 1982) states that all zero-mean discrete time second-order stationary time series $\{X_t\}_{t \in \mathbb{Z}}$ can be written as

$$X_t = \int_{-\pi}^{\pi} A(\omega) \exp(i\omega t) d\xi(\omega), \tag{39}$$

where $A(\omega)$ is the amplitude of the process and $\{\xi(\omega)\}_\omega$ is a stochastic process with orthonormal increments (i.e. $\mathbb{E}(d\xi(\omega)) = 0$ and $\mathrm{Cov}(d\xi(\omega_1), d\xi(\omega_2)) = d\omega_1 \delta_{\{\omega_1 = \omega_2\}}(\omega_1)$, where $\delta(\cdot)$ is the Kronecker delta function). The representation in (39) implies that a stationary process can be represented by a "Fourier-type" expansion (see Section 1.1.1). In other words, a stationary time series can be thought of as a linear combination of Fourier sinusoids of various frequen-

cies with an associated amplitude. However, in equation (39), for each frequency, ω, $d\xi(\omega)$ is a random quantity and the integral is a stochastic integral (unlike the representation in (1) for deterministic series).

Definition 1.3.5. *The quantity*

$$f(\omega) = |A(\omega)|^2 \tag{40}$$

is called the ***spectrum*** *or* ***spectral density function****.*

The spectral density function quantifies the contribution of a frequency, ω, to the process variance.

Example 1.3.6. The spectral density of an ARMA(p,q) process X_t (see equation (37)) is given by

$$f_X(\omega) = \frac{\sigma^2}{2\pi}\left|\frac{\beta(e^{-i\omega})}{\alpha(e^{-i\omega})}\right|^2, \tag{41}$$

where $\beta(\lambda) = 1 + \beta_1\lambda + \cdots + \beta_q\lambda^q$ and $\sigma^2 = \text{Var}(Z_t)$.

The periodogram is an estimator of the spectral density, and is defined as the squared modulus of the discrete Fourier transform (see Section 1.1.1):

Definition 1.3.7. *Given data $x_1, \ldots, x_n$ we define the* ***periodogram*** *to be*

$$I(\omega_j) = |d(\omega_j)|^2 \tag{42}$$

for $j = 0, 1, 2, \ldots, n-1$, where $d(\omega_j) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t \exp^{-2\pi i\omega_j t}$ and $\omega_j = j/n$ (see equation (2)).

The periodogram is asymptotically unbiased for the spectral density, but it is not a consistent estimator of the spectral density. Therefore, a common approach to obtain a consistent estimator of the spectral density is to smooth the periodogram by averaging in the spectral domain. There are many different approaches to smoothing the periodogram, for a detailed description see Priestley (1982).

Example 1.3.8. Figure 12 depicts the spectral estimate (the periodogram) and the smoothed periodogram for the realisation of an ARMA(2,1) process (Example 1.3.4) in Figure 11. In this example, we used kernel smoothing. In particular, we used a centred moving average procedure, the Daniell kernel with parameter m, which is defined as follows

$$\hat{x}_t = \frac{x_{t-m} + \cdots + x_{t-1} + x_t + x_{t+1} + \ldots x_{t+m}}{2m+1}.$$

In this particular example, we used $m = 10$.

1.3.2 Stationary Time Series Analysis of Circadian Data

Almost all species exhibit changes in their behaviour between day and night (Bell-Pedersen et al., 2005). These circadian rhythms are not only caused by a response to daily changes in the physical environment, but are also the result of an internal timekeeping system or 'biological clock' within the organism (Vitaterna et al., 2001; Minors and Waterhouse, 2013). For many species, a circadian clock is believed to enhance survival by directing anticipatory changes in physiology in tune with environmental fluctuations. When an organism is deprived of external

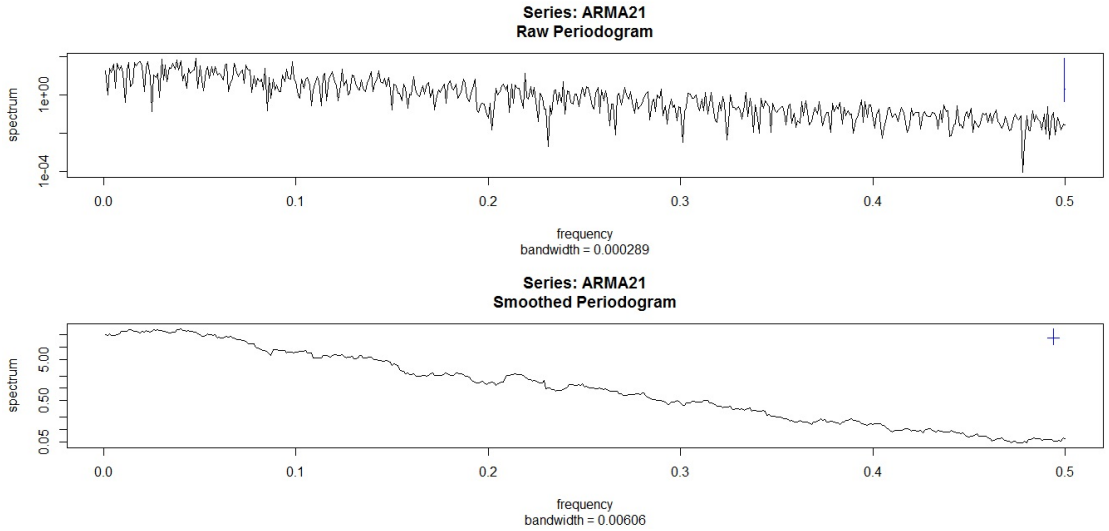

Figure 12: **Example 1.3.8: Spectral Estimation** for the realisation of an ARMA(2,1) process (Example 1.3.4) in Figure 11. Top: Raw periodogram. Bottom: Smoothed periodogram (using the Daniell kernel with parameter $m = 10$).

time cues, its circadian rhythms typically persist qualitatively but may change in detail; the study of these changes can reveal the biochemical reactions underpinning the circadian clock (McClung, 2006; Bujdoso and Davis, 2013).

Period and phase estimation (see Figure 13 for a visual interpretation of this terminology) are the fundamental elements of most circadian analyses. There are many different techniques for estimating period, all with different advantages and disadvantages, different assumptions and different levels of complexity. The current standard estimates period via software packages such as BRASS (Biological Rhythm Analysis Software System (Edwards et al., 2010)) or BioDare (Moore et al., 2014). BioDare and BRASS implement six of the most commonly used methods to estimate period: Enright periodogram (EPR) (Enright, 1965); Lomb-Scargle periodogram (Lomb, 1976); Fast Fourier Transform Non–Linear Least Squares (FFT-NLLS) (Plautz et al., 1997); mFourfit (Edwards et al., 2010); Maximum Entropy Spectral Analysis (MESA) (Burg, 1972) and Spectrum Resampling (Costa et al., 2011).

The six methods above represent the range of the approaches to period estimation for circadian time series in the literature. In particular, these six methods can be categorised as one of the following three approaches to period estimation: intuitive algorithms; curve fitting methods and spectrum-based methods. EPR is an example of one of the more intuitive approaches to analysing rhythmic biological data. mFourfit and FFT-NLLS are examples of curve fitting methods where we use a function (with known period) to represent our data and then report the period of the modelling function as the estimate. Lomb-Scargle Periodogram, MESA and spectrum resampling represent spectrum-based methods, where the fact that the data is a time series means that the theory and methods of stationary time series analysis (Section 1.3) can be used to produce the period estimate.

In the remainder of this section, we briefly introduce and review one method pertaining to each category (intuitive algorithms: EPR, curve fitting methods: FFT–NLLS and spectrum–based methods: MESA). We refer the interested reader to the original papers for more detailed descriptions of the above techniques. Alternatively, Zielinski et al. (2014) conducted an exten-

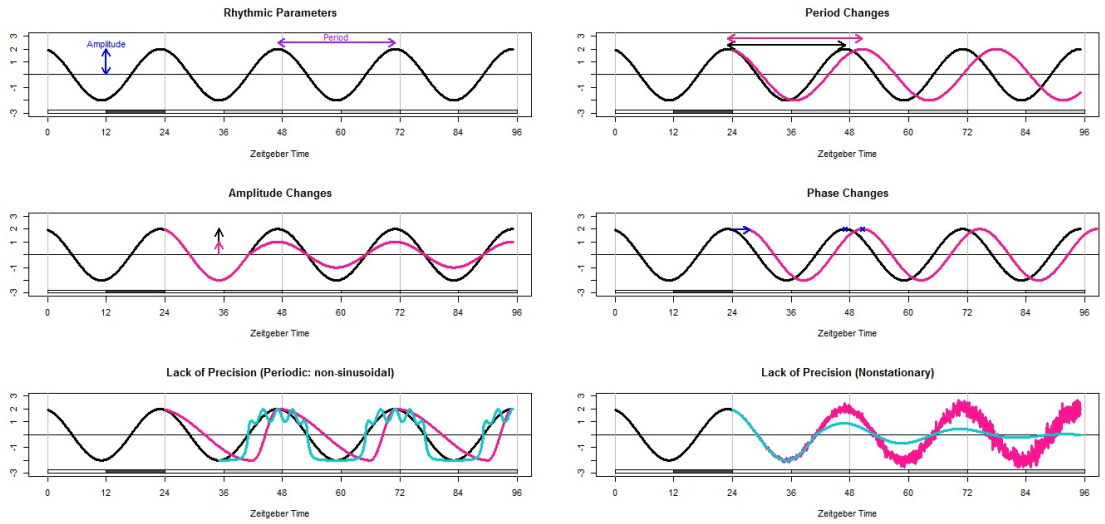

Figure 13: The defined rhythmic parameters: periodicity, phase, amplitude and clock precision (based on an image from Hanano et al. (2006)).

sive review of the six period estimation methods above in the context of analysing circadian data.

1.3.2.1 Intuitive Algorithms: EPR

The EPR is one of the more intuitive algorithms for the analysis of rhythmic biological data. The concept behind EPR is that, if the period of the data was known, the data could be split into sections where the length of the section was the same as the underlying period. Then each of the sections should contain similar data, since rhythmic data should exhibit some form of repeating pattern. Furthermore, overlaying the sections should give a clear waveform (with peak and trough) where the troughs align and give a low sum across the sections and similarly the peaks align and give a large sum. Therefore, the resulting waveform should have a large amplitude. However, if the data were not split exactly into sections whose length is equal to the period, then the peaks and troughs would not align and summing the sections together would result in a lower amplitude. Therefore, to analyse data with unknown period, the algorithm iterates through a series of test period values, implements the above procedure (for each test period) and selects the period that gives the average waveform with the highest amplitude. More recently, the calculation of the resulting average waveform has been improved. Therefore, it is the modified version of the Enright Periodogram which is actually implemented in BRASS and BioDare and used for circadian data with constant period and equally spaced observations. The main advantage of EPR is that it is intuitively accessible and computationally simple. The main limitation of EPR (for stationary data) is that the step size between test periods is constrained by the duration and sampling frequency of the collected data (see Section 1.1.1.1).

1.3.2.2 Curve Fitting Methods: FFT–NLLS

One general approach to period estimation is based on the idea of curve fitting. The motivation is that, if the data can be represented by a function of known period, then the period of the data can be assumed to be the same as that of the function. As an example of curve fitting methods, we outline the FFT–NLLS method since it is the most commonly used period analysis technique

in the field of circadian biology (see Costa et al. (2011); Perea-García et al. (2016a)). Hence, we will also use FFT–NLLS as the benchmark to assess the performance of the methodologies we develop in later chapters.

In FFT–NLLS, a model-based approach is adopted. That is, a function is chosen to represent the data that depends on parameters that determine its period and shape. The next stage is to estimate optimal parameters for this function (in other words, parameters that define the function that best fits the data) using non-linear least-squares fitting.

The FFT-NLLS algorithm was developed to analyse data that has constant period (Zielinski et al., 2014). The data are modelled by a sum of (up to 25) cosine functions. More formally the function used to represent the data for FFT-NLLS, $\widetilde{f}_{FFT}$, is given by:

$$\widetilde{f}_{FFT}(t) = \sum_{i=1}^{N} \alpha_i \cos\left[\frac{2\pi(t-\phi_i)}{\tau_i}\right], \tag{43}$$

where α_i is the amplitude of each cosine; ϕ_i its phase and τ_i its period and $N \leq 25$.

FFT–NLLS is a two–step procedure, in which a Fast Fourier Transform (FFT) is coupled with a non–linear least squares (NLLS) fitting of cosine functions to the data (Plautz et al., 1997) in the following way:

1. Remove long–term trends in the time series by fitting a linear regression model to the data and then subtracting the estimate from the original series.

2. Calculate the FFT of the transformed series.

3. Use FFT peak frequencies to sequentially (in order of descending power, up to a maximum of 25 frequencies) initialise NLLS cosine fitting (using a modified Gauss–Newton minimisation algorithm) which estimates the parameters $(\tau_i, \phi_i, \alpha_i)$.

4. Output confidence intervals for the estimated parameters of the fitted curves.

5. Stop when the latest period estimate, $\hat{\tau}_i$ is not statistically significant or the maximum number of frequencies was reached.

6. Report all estimated significant periodicities, $\hat{\tau}_1, \ldots, \hat{\tau}_l, l \leq 25$.

Under the assumption of constant period for the circadian component, the period estimate is taken to be the period of the cosine component lying within a user–defined range of likely circadian periods (typically between 15 and 35 hours). If more than one cosine component is within this range, it is up to the user to decide which period to select.

In Step 3 of the FFT–NLLS algorithm described above, the non-linear least squares procedure (NLLS) is used to find parameter estimates by iteratively improving initial values via numerical search. However, it only works well when given sensible starting values. Thus, a Fast Fourier Transform (FFT) is performed on the circadian time series to obtain good period and amplitude estimates using the data (as opposed to using user-defined or default values as the initial guess). In Example 1.1.6 (Section 1.3.1), we found that the FFT, when used to compute the DFT and thus the periodogram, was an effective method to identify the frequency components of a linear combination of cosine curves and their respective amplitudes. However, in Example 1.1.7 (Section 1.3.1), we found that although the periodogram effectively identified the frequency components of a concatenation of cosine curves and their respective amplitudes, it

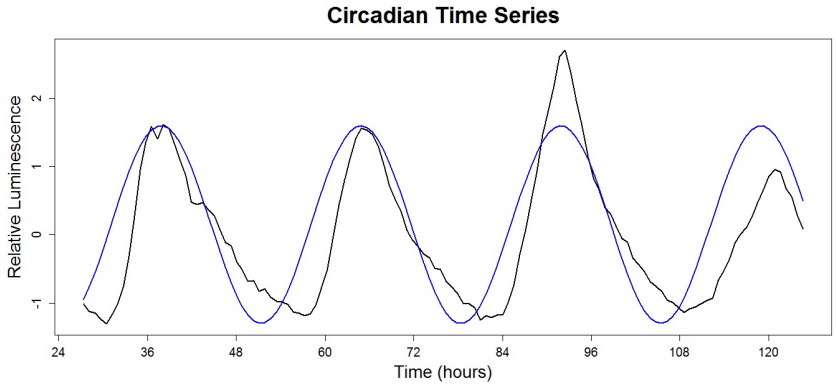

Figure 14: **Example 1.3.9: Implementation of FFT–NLLS**. Black line: A time series from the control group (Chapter 2); Blue line: cosine curve with period 27.03 hours (the period estimate obtained using FFT–NLLS).

could not identify changes of period and thus could not differentiate between the signals in examples 1.1.6 and 1.1.7. This illustrates a disadvantage of FFT-NLLS– because the technique utilises the FFT, it is limited to modelling (linear combinations of) sinusoidal waveforms with constant period and does not perform well on data that are not of this type (Zielinski et al., 2014; Hargreaves et al., 2018).

Example 1.3.9. In Chapter 2, we analyse a dataset taken from a broad investigation of the effect of various salt stresses on the plant circadian clock. An example time series from the control group of this dataset is shown in Figure 14. For reference, we also plot a cosine curve with the estimated period obtained using FFT–NLLS (via BRASS). Note that the period estimate appears to approximately describe the data. However, we also note a 'lack of precision' and changes in period and amplitude (see Figure 13 for a visual interpretation of this terminology). These features are not captured by this method, demonstrating the limitations of this analysis.

1.3.2.3 Spectrum-based Methods: MESA

Another general approach to period estimation is spectrum-based methods, based on stochastic modelling (e.g. MESA). MESA first fits an autoregressive model to the data (see Section 1.3). Various methods (e.g. examination of the autocorrelation and partial autocorrelation functions, or an information criterion) can be used to determine the order of the AR process, p. The associated parameters can then be estimated using, for example, the method of moments, least squares estimation or maximum likelihood estimation. The estimated coefficients can then be used to obtain an estimate of the spectrum of the data (see Section 1.3.1). Recall: the (frequency) spectrum quantifies the contribution of a frequency, ω, to the process variance. Therefore, since frequency is the inverse of period, finding the maximum of a frequency spectrum is equivalent to finding the strongest period of the data. For the MESA approach, an estimate of the spectrum is constructed using the following formula (where the scaling constant has been removed):

$$\hat{f}(\omega) = \frac{1}{\left|1 - \sum_{k=1}^{\hat{p}} \hat{\alpha}_k e^{-i\omega k}\right|^2} \tag{44}$$

where $\hat{p}$ is an estimate of the order of the AR process, $\hat{\alpha}_k$ are the estimated model coefficients, ω is the circular frequency: $\omega = \frac{2\pi}{\tau}$ and τ is the period. The period value corresponding to the

maximum of $\hat{f}$ is then returned as the period estimate for the MESA approach.

The main advantage of MESA is that although it is still Fourier-based, it does not assume any pre-determined shape of the data (as opposed to FFT-NLLS which assumed the shape of the circadian component of the underlying function was sinusoidal). However, MESA does assume that the data can be modelled as an AR process, which often may not be appropriate. Furthermore, its performance is also dependent on the estimation of the order of the AR process, p, and the associated parameters.

1.3.3 Wavelet Analysis of Stationary Time Series

The wavelet methods introduced in Section 1.2 can also be a useful tool for stationary time series analysis. In this section, we briefly introduce (wavelet) scale analysis of stationary time series. For an introduction to this topic see Nason (2010), for a comprehensive review of the field see Percival and Walden (2006) or Chiann and Morettin (1998).

The **wavelet variance** is the process variance represented in the wavelet domain and is represented by the wavelet spectrum. Since the wavelet basis is orthogonal, energy is preserved in the wavelet domain. The wavelet variance can be estimated using, for example, the discrete wavelet transform or the nondecimated wavelet transform. Since the wavelet variance decomposes the variance of certain stochastic processes by scale, it is useful in applications such as signal processing, where the process can conceptualised as variations operating over a range of different scales. However, in general a wavelet spectrum is less informative than the Fourier spectrum since it has a much lower frequency resolution. In such cases, Fourier analysis would be advisable. For example, the theoretical model of a circadian rhythm assumes that the data can be represented by a function of known period (see Section 1.3.2). In this situation, it is of interest to estimate the period of the function to a high degree of accuracy and, thus, Fourier analysis would be preferable.

1.4 Nonstationary Time Series Analysis

In the representation in equation (39) (Section 1.3), note that for stationary processes the amplitude $A(\omega)$ does not depend on time (i.e. the frequency behaviour is the same across time). However, for many real time series, including the motivating circadian datasets we analyse in later chapters, this assumption is not realistic (Zielinski et al., 2014). Price et al. (2008) asserted that data arising from circadian experiments is nonstationary and discussed the features which support this claim, namely a progressively dampened signal with a changing period. A modelling framework for time series where the frequency behaviour can vary with time would therefore be preferable in such applications.

Example 1.4.1. In Chapter 2, we analyse a dataset taken from a broad investigation of the effect of various salt stresses on plants. Four time series from this dataset are shown in Figure 15 and nonstationary behaviour such as changes in amplitude can easily be noted in each.

Furthermore, we investigated whether the individual plant signals in Figure 15 are (second-order) stationary via hypothesis testing. We employed a wavelet-based test of stationarity, the wavelet spectrum test (Nason, 2013), implemented in the `locits` package in R, which is available on CRAN. All four plant signals provided enough evidence to reject the null hypothesis of stationarity.

Additionally, this test also indicates where the nonstationarities are located in the series and these are also plotted for reference (as red double–headed arrows) in Figure 15. Each arrow corresponds to one of the nonstationarities identified by the test. The span of the arrow indicates the time period over which the nonstationarity has been detected. In the time domain, the estimated nonstationarities appear to coincide with the changes in amplitude previously noted (in Figure 15) within this example. (Note: the right–hand axis in Figure 15 indicates the scale of the (time–varying) wavelet spectrum (see Section 1.4.2) that contains the nonstationarity-further details are given in Section 3.3.3.)

1.4.1 Locally Stationary Time Series

If the stationarity assumption is dropped, other (less restrictive) assumptions still have to be imposed on the process to enable inferences on the process characteristics. Throughout this chapter, we will focus on trend–free processes with a second order structure that varies slowly with time. Such time series are called **locally stationary** (Dahlhaus, 1997; Nason et al., 2000), since they appear to have stationary behaviour over short periods of time. This ensures that their statistical characteristics (such as the autocovariance function) can be (locally) estimated by pooling the observed data over regions of local stationarity.

One way of introducing time dependence into a model is by replacing the amplitudes $A(\omega)$ in equation (39) with a time-dependent form. Priestley (1965) introduced a time-frequency model with the amplitude replaced by $A_t(\omega)$, leading to a class of nonstationary processes called **oscillatory processes**. The amplitude variation as a function of time was assumed to have a degree of regularity which ensured the locally stationary character of the process. Priestley (1965) also defines a time–dependent evolutionary spectrum, which describes the frequency content of the process over regions of time.

Dahlhaus (1997) developed the **locally stationary Fourier** (LSF) model where the process X_t is modelled as a triangular stochastic array $\{X_{t;T}\}_{t=0}^{T-1}$ such that

$$X_{t;T} = \int_{-\pi}^{\pi} A^0_{t;T}(\omega) \exp(i\omega t) d\xi(\omega), \tag{45}$$

where there exists K such that

$$\sup_{t,w} |A^0_{t;T}(\omega) - A(t/T, \omega)| \leq K/T, \tag{46}$$

$\forall T$, and $\{\xi(\omega)\}_\omega$ is a random process satisfying certain specific properties (see Dahlhaus (1997) for a detailed description). As discussed in Section 1.4, asymptotic considerations are more difficult in a nonstationary setting as any future observations may not contain any information on the structure of the process at the current time. Therefore, in the LSF setting, the evolution of the individual time-dependent amplitudes, $A_t(\omega)$, is controlled through a function, $A(z,\omega)$, dependent on **rescaled time**, $z = t/T$, $t = 0, \ldots, T-1$ (see equation (46)), known as the asymptotic transfer function. The asymptotic transfer function regulates the behaviour of the time–varying individual amplitudes, $A_t(\omega)$. The smoothness of $A(z,\omega)$ with respect to z, tunes the degree of local stationarity of the process. As the length of the time series T increases, there is more information about the local behaviour of the function $A(z,\omega), z = t/T \in (0,1)$, which thus paves the way to estimation.

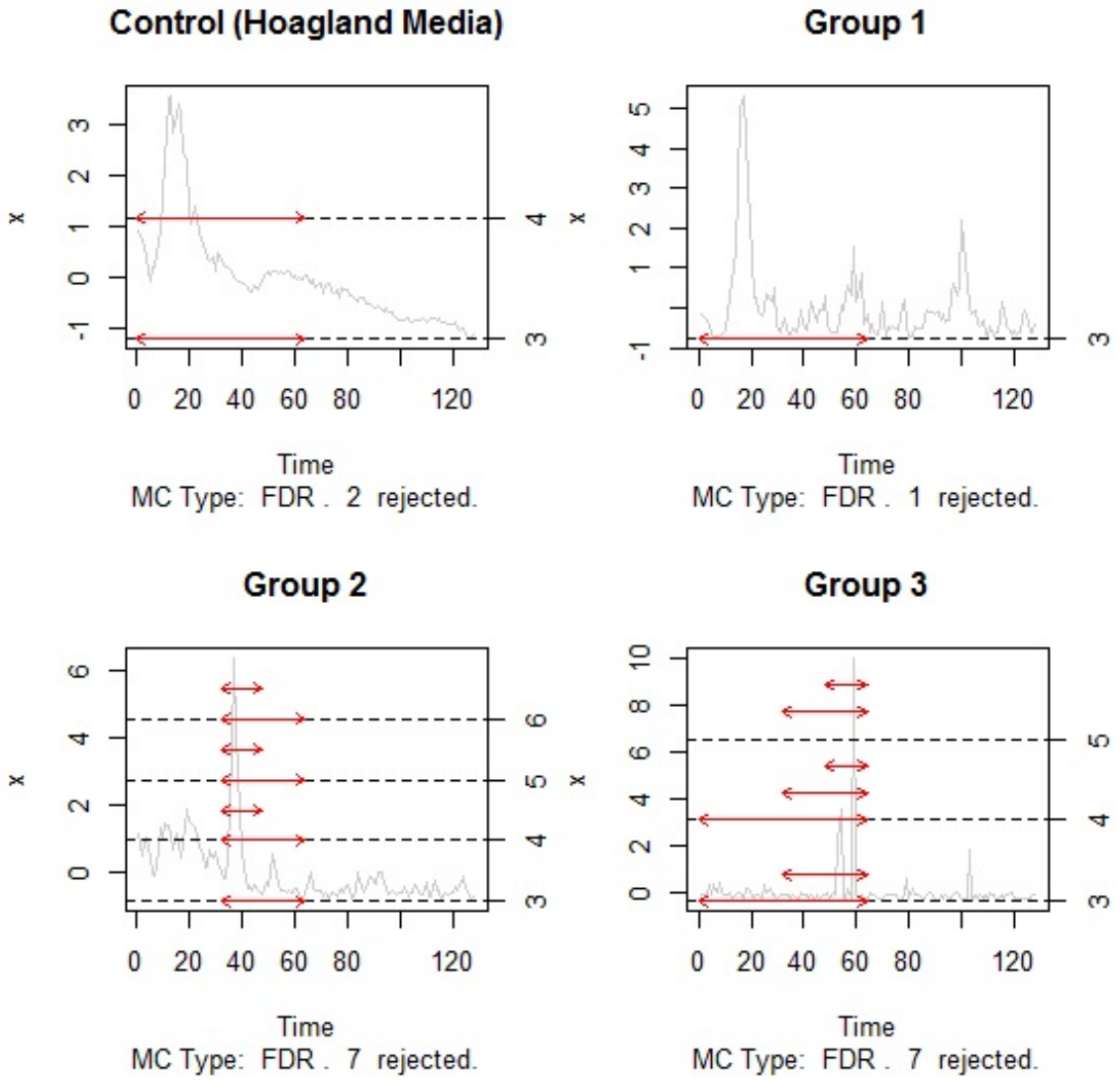

Figure 15: **Example 1.4.1**: A time series for each of the four groups (see Chapter 2) is shown as an example– Group 1, a time series from the 100μM group; Group 2, a time series from the 150μM group; Group 3, a time series from the 200μM group. Red arrows: Plots of the estimated locations of the nonstationarities in the circadian plant signals in response to differing quantities of ammonium cerium nitrate, using the wavelet spectrum test (Nason, 2013), implemented in the `locits` package in R which is available on CRAN.

Dahlhaus (1997) also defined an associated evolutionary spectral function which is also defined in terms of rescaled time, $z = t/T$,

$$f_X(z,\omega) = |A(z,\omega)|^2.$$

This spectrum has the advantage of being uniquely defined (Dahlhaus, 1997), as opposed to the time–dependent evolutionary spectrum of the oscillatory processes (Priestley, 1965) discussed above.

1.4.2 Locally Stationary Wavelet Model

Later, Nason et al. (2000) introduced a locally stationary wavelet (LSW) model, where the Fourier building blocks (present in the LSF model) are replaced by families of discrete nondecimated wavelets. The LSW model forms the basis of the methodology we develop in this thesis. Therefore, for the remainder of this section, we introduce the definition of an LSW process as well as several related quantities. We begin by describing nondecimated discrete wavelets, the building blocks of the LSW model.

Let $\{h_k\}$ and $\{g_k\}$ be the low– and high–pass quadrature mirror filters as defined in Section 1.2.2. Following Nason et al. (2000), the compactly supported **discrete wavelet vectors** $\psi_j = (\psi_{j,0},\ldots,\psi_{j,(N_j-1)})$ of length N_j for scale $j > 0$, are obtained using the following formulae:

$$\begin{aligned} \psi_{1,n} &= \sum_k g_{n-2k}\delta_{0,k} = g_n, \quad \text{for} \quad n = 0,\ldots,N_1-1, \\ \psi_{j+1,n} &= \sum_k h_{n-2k}\psi_{j,k}, \quad \text{for} \quad n = 0,\ldots,N_{j+1}-1, \\ N_j &= (2^j-1)(N_h-1)+1, \end{aligned} \tag{47}$$

where $\delta_{0,k}$ is the Kronecker delta and N_h is the number of non-zero elements of $\{h_k\}$. The notation $j = 1$ denotes the finest scale wavelet, $j = 2$ the next finest scale and so on.

The collection of (discrete) **nondecimated wavelet vectors**, $\psi_{j,k}(t)$ for $t = 0,1,\ldots,T-1$, is formed by translations of the discrete wavelet vectors ψ_j to all (discrete) integer locations k as:

$$\psi_{j,k}(t) \coloneqq \psi_{j,k-t}. \tag{48}$$

Note the notation in equation (48), established in Nason et al. (2000), will be used throughout this thesis.

Definition 1.4.2. *A **Locally Stationary Wavelet** (LSW) process (Nason et al., 2000), $\{X_{t;T}\}_{t=0}^{T-1}$, $T = 2^J \geq 1$, is a sequence of doubly indexed stochastic process with the following representation:*

$$X_{t;T} = \sum_{j=1}^{J}\sum_{k\in\mathbb{Z}} w_{j,k;T}\psi_{j,k}(t)\xi_{j,k}, \tag{49}$$

where $\{\xi_{j,k}\}$ is a random orthonormal sequence of increments, $\{\psi_{j,k}(t) = \psi_{j,k-t}\}_{j,k}$ is a set of discrete non-decimated wavelets and $\{w_{j,k:T}\}$ is a set of amplitudes, each of which at a scale j and time k. The quantities in representation (49) *posses the following properties:*

1. $\mathbb{E}(\xi_{j,k}) = 0$. *Hence,* $\mathbb{E}\left(X_{t,T}\right) = 0$ *for all t and T.*
2. $cov(\xi_{j,k}\xi_{l,m}) = \delta_{j,l}\delta_{k,m}$, *where $\delta_{j,l}$ is the Kronecker delta.*

3. *There exists for each $j \geq 1$ a Liphschitz continuous function $W_j(z)$ for $z \in (0,1)$ which satisfies the following properties:*

   - $\sum_{j=1}^{\infty} |W_j(z)|^2 < \infty$ *uniformly in* $z \in (0,1)$.
   - *The Lipschitz constants L_j are uniformly bounded in j and*

$$\sum_{j=1}^{\infty} 2^j L_j < \infty.$$

   - *There exists a sequence of constants C_j such that, for each T*

$$\sup_k |w_{j,k;T} - W_j(k/T)| \leq C_j/T,$$

   where, for each $j = 1, \ldots, J$, the sup *is over $k = 0, \ldots, T-1$, and where $\{C_j\}$ fulfils*

$$\sum_{j=1}^{\infty} C_j < \infty.$$

Intuitively, the representation in (49) can be thought of as building a time series model $\{X_{t;T}\}$ out of a linear combination of oscillating functions $(\psi_{j,k})$ with random amplitudes $(w_{j,k;T}\xi_{j,k})$. Therefore, it is simply the multiscale version of the representation for stationary processes in (39).

Property 3 of the quantities in the LSW representation (49) states that the amplitudes $\{w_{j,k;T}\}$ are not allowed to evolve too rapidly through not deviating too much from a "control" function $W_j(z)$, which itself has certain constraints to prevent it from oscillating too wildly. Intuitively, this condition sets a limit on how "nonstationary" a time series can be, in order to allow estimation (as discussed for the alternative locally stationary models outlined in Section 1.4.1).

An analogous quantity to the spectrum of a stationary process (equation (40)), which quantifies the contribution of a frequency ω to the process variance, is introduced in the LSW setting. This quantity, commonly referred to as the **evolutionary wavelet spectrum** (EWS), quantifies the power distribution in an LSW process over time and scale and is formally defined as:

$$S_j(z) = |W_j(z)|^2, \tag{50}$$

for $j = 1, \ldots, J$, and rescaled time $z \in (0,1)$.

As in Nason et al. (2000), define the **autocorrelation wavelets**, $\Psi_j(\tau)$, of the discrete wavelets as:

$$\Psi_j(\tau) = \sum_{k \in \mathbb{Z}} \psi_{j,k}(0)\psi_{j,k}(\tau), \tag{51}$$

for all $j = 1, \ldots, J$ and $\tau \in \mathbb{Z}$.

The process **autocovariance function** of an LSW process $X_{t,T}$ at lag τ and rescaled time location z is defined as

$$c_T(z,\tau) = \mathrm{cov}(X_{\lfloor zT \rfloor,T}, X_{\lfloor zT \rfloor+\tau,T}). \tag{52}$$

Nason et al. (2000) show that $c_T(z,\tau) \to c(z,\tau)$ as $T \to \infty$, where

$$c(z,\tau) = \sum_{j=1}^{J} S_j(z)\Psi_j(\tau) \tag{53}$$

is the local autocovariance function and $\lfloor x \rfloor$ is the largest integer less than or equal to x.

An asymptotically unbiased estimator of the EWS $\{S_j(z)\}$ is obtained by correcting the **raw wavelet periodogram**

$$I^j_{k,T} = |d_{j,k;T}|^2, \tag{54}$$

where

$$d_{j,k;T} = \sum_{t=0}^{T} X_{t,T}\psi_{j,k}(t) \tag{55}$$

are the **empirical nondecimated wavelet coefficients**. The correction is attained by premultiplying the raw wavelet periodogram vector $\mathbf{I}(z) := (I^j_{\lfloor zT \rfloor,T})_{j=1}^J$ by the inverse of the $J \times J$ **autocorrelation wavelet inner product matrix**,

$$A_J = (\sum_{\tau} \Psi_j(\tau)\Psi_l(\tau))_{j,l},$$

where $\Psi_j(\tau)$ is the autocorrelation wavelet. Thus, the **corrected wavelet periodogram** is defined as

$$\mathbf{L}(z) = A_J^{-1}\mathbf{I}(z), \text{ for all } z \in (0,1). \tag{56}$$

Example 1.4.3. Let $T = 256$ and specify a wavelet spectrum $S_j(z)$ as follows:

$$S_j(z) = \begin{cases} 4\cos^2(2\pi z), & \text{for } j = 3, z \in (0,1) \\ 1, & \text{for } j = 7, z \in (1/256, 56/256) \\ 0, & \text{otherwise.} \end{cases} \tag{57}$$

Figure 16(a) provides a visualisation of the wavelet spectrum in equation (57) and an example of a signal realisation generated from equation (57) can be found in 16(b). (A realisation can be generated from a spectrum using the `locits` R package; for more information on how to generate an LSW process from a defined spectrum see Nason (2010).)

To demonstrate the importance of the bias correction of the wavelet periodogram, we begin with the spectrum in equation (57) and simulate a realisation (as outlined above). We then compute the raw wavelet periodogram (equation (54)) and the corrected periodogram (equation (56)). We repeat this process for 100 realisations and then average the respective periodograms to produce Figures 16(c) and (d). On examining Figures 16(c) and (d), note that the (corrected) spectral estimate in (d) is much closer to the true underlying spectrum than the raw wavelet periodogram in (c).

As in the stationary setting, the wavelet periodogram is not a consistent estimator of the wavelet spectrum (Nason, 2010). One method to overcome this is to smooth the raw wavelet periodogram as a function of (rescaled) time within each scale j, and then to apply the correction above. Various smoothing approaches have been proposed in the literature, see e.g. smoothing using variance stabilisation of Fryzlewicz and Nason (2006).

1.4.3 Nonstationary Time Series Analysis of Circadian Data

As discussed in Section 1.3.2, Zielinski et al. (2014) conducted an extensive review of period estimation methods for circadian data. Zielinski et al. (2014) omitted wavelet-based methods because they have not been shown to be better than the six methods discussed in 1.3.2 for

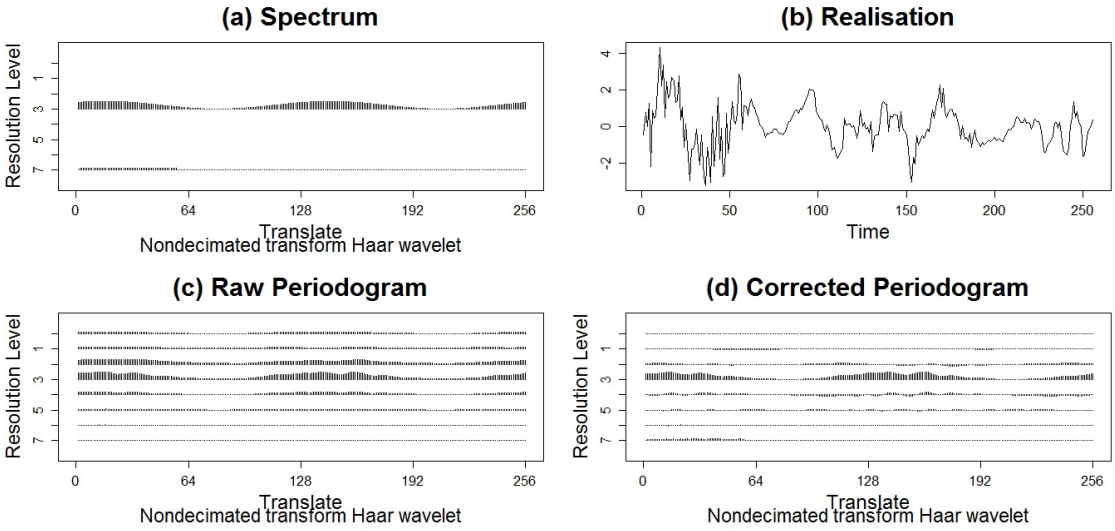

Figure 16: **Example 1.4.3.** Figure (a) depicts the spectrum defined in equation (57); (b) depicts a realisation generated from the spectrum shown in (a); (c) shows the mean of 100 uncorrected periodogram estimations computed on realisations from the spectrum shown in (a) and (d) shows the mean of 100 corrected periodogram estimations computed on realisations from the spectrum shown in (a). Note that the spectral estimate in (d) is much closer to the true underlying spectrum than (c).

stationary data with constant period, which was the focus of the paper. However, the authors assert that the wavelet transform can be performed to extract changes of period over time and, therefore, wavelet-based methods are particularly useful for analysing nonstationary time series. Zielinski et al. (2014) also states that nonstationarity is common in many biological systems and this was evidenced by the circadian time series in Example 1.4.1. Therefore, in this section we briefly review wavelet-based circadian data analysis tools present in the current literature.

Price et al. (2008) asserted that data arising from circadian experiments is nonstationary and discussed the features which support this claim, namely a progressively dampened signal with changing period. Therefore, Price et al. (2008) advocated the use of wavelets to analyse circadian data and developed a technique for characterising the modal periods present in circadian data using a continuous wavelet decomposition (this is disseminated in the `waveclock` package in R, currently on CRAN archive). Later, Harang et al. (2012) also supported the circadian data nonstationarity view, and furthermore claimed that circadian analysis under nonstationary behaviour by means of traditional Fourier methods can lead to inaccurate results. Harang et al. (2012) thus recommended the use of wavelets to allow for the changes in period to be tracked through time; the authors developed 'WAVOS'- a wavelet-based MATLAB toolkit that allows for analysis of nonstationary circadian data.

Leise et al. (2013) discussed the appropriateness of traditional methods to determine period length from experimental datasets that assume a rhythm of fixed period and amplitude, proposing that most biological rhythms exhibit changes in both period and amplitude (see Example 1.4.1). The authors extended wavelet methods to measure how biological rhythms vary over time and developed MATLAB scripts to implement their analysis using both continuous and discrete wavelet transforms.

The methodology we develop in Chapters 2 and 3 is different, as it combines the use of

wavelets (ideal for analysing nonstationary behaviour due to their time localisation) with the rigorous statistical (process) modelling introduced in Section 1.4.2. Using this statistical modelling framework will, of course, be advantageous when the LSW modelling assumption is correct. For example, unbiased estimators of the EWS can be calculated (see Example 1.4.3). However, there may be times when the data is nonstationary but the underlying model is not an LSW process. In such circumstances, the added computational burden of utilising the LSW methodology may be a disadvantage. However, in the simulation studies in Chapter 2, we demonstrate the advantages of utilising the LSW methodology over standard wavelet–based approaches in a range of different scenarios (both when the LSW modelling assumption is correct and when the data consists of nonstationary AR processes (see Section 1.3.1)).

2 Clustering Nonstationary Circadian Rhythms Using Locally Stationary Wavelet Representations

In this chapter we develop and test a new method for clustering rhythmic biological data. The proposed method is the result of joint work with M.I. Knight, J. W. Pitchford, R. Oakenfull and S. J. Davis, and corresponds to the publication Hargreaves et al. (2018). Please see page 21 for details of author contributions.

2.1 Introduction

The earth rotates on its axis every 24 hours resulting in a day and night cycle. Correspondingly, almost all species exhibit changes in their behaviour between day and night (Bell-Pedersen et al., 2005). These daily rhythms are not only caused by a response to daily changes in the physical environment, but are also the result of an internal timekeeping system or 'biological clock' within the organism (Vitaterna et al., 2001; Minors and Waterhouse, 2013). In particular, most plants are able to anticipate dawn and adjust their biochemistry accordingly. When an organism is deprived of external time cues, these rhythms typically persist qualitatively but may change in detail; the study of these changes can reveal the biochemical reactions underpinning the circadian clock and, at a larger scale, can provide valuable insight into the possible consequences of environmental change (McClung, 2006; Bujdoso and Davis, 2013).

Experiments recording plant response to light entrainment result in datasets that, from a statistical point of view, can be considered as time series realisations. Period and phase estimation (see Figure 13 in Chapter 1 for a visual interpretation of this terminology) are the fundamental elements of most circadian analyses. The current standard uses BRASS (Biological Rhythm Analysis Software System (Edwards et al., 2010)) to estimate the period of each time series using Fourier analysis (see Moore et al. (2014) or Zielinski et al. (2014) for a complete description of the underlying period analysis methods). Data stationarity is an implicit assumption within the underlying methodology – put simply, its statistical characteristics are assumed constant over time. However, in reality, nonstationary behaviour is common in biological systems (Zielinski et al., 2014). Here we propose, develop and test methods that are capable of detecting changes of period over time by drawing on the plant time-frequency signature as quantified by its spectrum.

The methodology developed here is general, but our concrete example concerns (i) identifying if a plant's clock is affected under exposure to different concentrations of ammonium cerium nitrate, (ii) establishing which concentrations produce similar effects and (iii) subsequently characterising these effects. The answers to these questions have important implications, not only for the understanding of the mechanism of the plant's circadian clock, but also for the environmental impact associated with soil pollution (Yang et al., 2016).

In order to answer the above questions, we propose to estimate the spectral behaviour of our time series under the formal framework of locally stationary wavelet processes (Nason et al., 2000), introduced in Section 1.4.2, which are able to account for data nonstationarity. Wavelets (introduced in Section 1.1.2) are ideal for identifying discriminant local time and scale (frequency) features, and time-frequency (scale) patterns are known to be indicative of the plant response to various stimuli (Zielinski et al., 2014). A functional principal components analysis on the spectral data treated as an 'image' (as suggested in a Fourier context by Holan

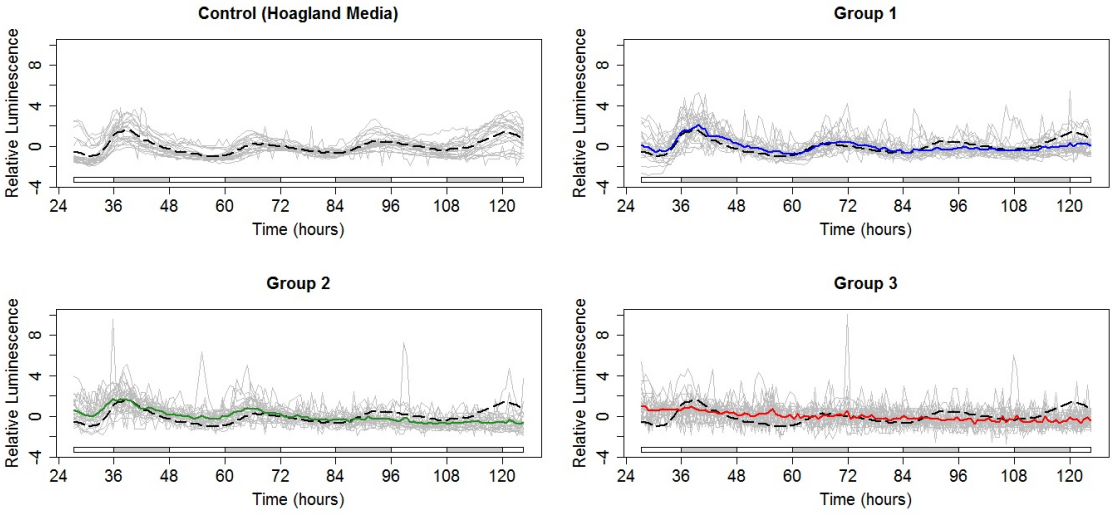

Figure 17: Luminescence evolution over time for plants subjected to a control and 3 different ammonium cerium nitrate concentrations. Time is measured in hours relative to *zeitgeber* time (time of last external temporal cue: the dawn signal of lights-on). Top left: Each plant signal from the control group (in grey) along with the group average (dashed black). Other panels: Each realisation from the groups (in grey) along with the group average and the control group average (dashed black). Group 1: 100μM ammonium cerium nitrate with average in blue. Group 2: 150μM ammonium cerium nitrate with average in green. Group 3: 200μM ammonium cerium nitrate with average in red. (Each time series has been normalised to have mean zero.) Note: the free run started from time 24; shaded bars below each graph indicate the subjective darkness that plants expected to experience during the 'normal' day.

et al. (2010)) is then used to reduce the data dimensionality and allows the extraction of important behavioural features. Furthermore, this functional representation is also used to inform a clustering method that facilitates quantifying the effects induced by different concentrations of ammonium cerium nitrate.

This chapter is organised as follows. Section 2.2 outlines the novel circadian dataset and establishes its nonstationary behaviour. Section 2.3 develops our proposed novel locally stationary wavelet-based clustering method. The findings of an extensive simulation study are presented in Section 2.4. Section 2.6.1 demonstrates the additional insight our clustering method can provide when applied to a published circadian plant dataset. Section 2.6.2 presents the results of clustering the novel circadian plant dataset using the proposed methodology and examines them in the context of several relevant biological questions. Section 2.7 concludes with a brief discussion and suggests topics for further investigation.

2.2 Motivation

In this section we briefly outline the experimental details that led to the novel circadian dataset and assess the prominent features of the plant rhythms under analysis, namely their lack of stationarity. This result, along with several others recorded in the literature (e.g. Price et al. (2008), Leise et al. (2013)) motivates the development of analysis techniques that can account for nonstationarity. Furthermore, we also discuss the phenomenon of individual-level variability in plant response to stimuli, despite their sharing identical genetic characteristics (Doyle et al., 2002). The presence of multiple behaviours within the same treatment group motivates our

development of a clustering procedure that can detect these different characteristics and analyse them separately. For completeness, we also report the results of the analysis a circadian biologist would typically use.

2.2.1 Experimental Details

The novel circadian dataset (henceforth referred to as the cerium dataset) was obtained by the Davis Lab (Biology, University of York) following a similar method to Hanano et al. (2006). For a detailed description of these methods see Appendix 2.9. Briefly, for each plant, gene expression levels are measured (using a firefly luciferase reporter system) at regular intervals resulting in an individual time series. In this experiment, the gene of interest was 'cold and circadian regulated and RNA binding 2', known as CCR2 (Doyle et al., 2002).

The cerium dataset consists of a total 96 plant signals (time series) recorded at 128 time points, with the control and groups 1–3 (each corresponding to a different concentration of ammonium cerium nitrate) all containing 24 plants. The control group is grown in Hoagland's media (Hoagland et al., 1950), which contains essential nutrients required for plant growth, and is not exposed to any additional levels of ammonium cerium nitrate. To examine the effects of cerium on the circadian clock, the other three groups, while also grown in Hoagland's media, were additionally exposed to varying additional concentrations of ammonium cerium nitrate– 100μM for Group 1, 150μM for Group 2 and 200μM for Group 3. A plot of individual luminescence time series, the average expression at each time point, for each of the treatment groups, is shown in Figure 17. Note that time is measured in hours relative to *zeitgeber* time, which is the time of the last external temporal cue: the dawn signal of lights-on.

2.2.2 BRASS Analysis

In the circadian community, analysis of this data would typically be performed by the Microsoft Excel macro BRASS (introduced in Section 1.3.2). Table 1 provides a summary of the output of the analysis of the cerium dataset in BRASS. In particular, it shows the mean period estimate (obtained using FFT-NLLS analysis (Plautz et al., 1997) considering only period estimates between 15 and 40 hours), the number of plants that could not be analysed by BRASS and the mean Relative Amplitude Error (RAE) for each of the 4 groups. RAE is a value between 0 and 1 and gives information about the goodness of fit of the model (a value of 0 indicates a perfect fit). In the circadian community, standard practice dictates that results with an RAE value above the threshold of 0.4 are discarded (Doyle et al., 2002). Circadian biologists often visualise the results in a scatter plot of relative amplitude error against period length for the plants analysed by BRASS (see e.g. Hanano et al. (2006)) and such a plot for this dataset is given in Figure 30, Appendix 2.8.

On examining Table 1, note that not all data is used to produce the period estimate reported by BRASS– in particular, the 'number of plants excluded by BRASS' is the number of time series for which the FFT–NLLS algorithm (Plautz et al., 1997) was not able to return a period estimate, possibly due to a loss of rhythmicity. Thus, under the assumption of stationarity (and the above constraints), these methods are not able to analyse all data produced by this experiment, indicating that this dataset is not suitably modelled using Fourier methods.

Furthermore, by just reporting the results of this analysis, the biologist would conclude that adding 100μM or 150μM ammonium cerium nitrate produces no detectable effect on the

Group	Hoagland's	Group 1 (100μM)	Group 2 (150μM)	Group 3 (200μM)
Average period estimate (in hours)	27	27	26	24
Number of plants excluded by BRASS	7	10	12	21
Average RAE	0.23	0.44	0.41	0.74

Table 1: Summary of the output of the analysis of the circadian dataset in BRASS. The 'number of plants excluded by BRASS' is the number of time series for which BRASS was not able to return a period estimate. 'RAE' (Relative Amplitude Error) is a value between 0 and 1 and gives information about the goodness of fit of the model (a value of 0 indicates a perfect fit). Results with an RAE over 0.4 are discarded. Recall: there are 24 plants in each of the groups.

circadian clock (as these period estimates are similar to the control, see Table 1). Moreover, within the circadian community, the results from adding 200μM ammonium cerium nitrate would not be considered, since they produce an RAE value of 0.74 (which is over the threshold of 0.4). Therefore, using the current methodology, the circadian biologist would not be able to conclude that exposure to ammonium cerium nitrate (at any of the tested concentrations) has an effect on the circadian clock of *A. thaliana*. However, visual examination of Figure 17 shows that this chemical appears to have a strong effect on these plants, providing further evidence that more statistically advanced approaches are needed.

2.2.3 Nonstationarity in Circadian Rhythms

In Section 1.4.3 we reviewed the literature that asserts that data arising from circadian experiments is nonstationary and also discussed a number of wavelet–based methods for nonstationary time series analysis of circadian data (Price et al., 2008; Harang et al., 2012; Leise et al., 2013). Therefore, for our novel circadian dataset, we investigated whether the individual plant signals are (second-order) stationary via hypothesis testing.

We employed two tests for stationarity– a Fourier-based test (Priestley and Rao, 1969) and a wavelet-based test (Nason, 2013). The Fourier-based test we used was the Priestley-Subba Rao (PSR) test. The results, which can be found in Table 2, show that over 70% of the plant signals provided enough evidence to reject the null hypothesis of stationarity. This conclusion is backed-up by the wavelet-based spectrum test for stationarity. Additionally, this test also indicates where the nonstationarities are located in the series. (A visual representation for each group can be found in Figure 31, Appendix 2.8.)

Therefore, in agreement with previous observations in circadian literature (see Section 1.4.3), both tests suggest that our circadian data also displays nonstationary features. In order to assess the impact of different concentrations of ammonium cerium nitrate, we propose a novel clustering technique that combines the use of wavelets (ideal for analysing nonstationary behaviour) with rigorous statistical (process) modelling. Additionally, to mitigate against individual plant variability, our technique proposes the use of time-scale patterns as explained next.

2.2.4 Individual-level Variability in Circadian Rhythms

We noticed in our dataset the presence of individual-level variability in plant responses to the same stimuli, despite their sharing identical genetic characteristics (Doyle et al., 2002). For

Group	Hoagland's	Group 1 (100μM)	Group 2 (150μM)	Group 3 (200μM)
Number of nonstationary plants	22	19	19	8

Table 2: Results for the Priestley-Subba Rao test of stationarity, implemented in the `fractal` package in R and available from the CRAN package repository. Number of nonstationary plants indicates the number of time series (in each group) with enough evidence to reject the null hypothesis of stationarity at the 1% significance level. Recall: there are 24 plants in each of the groups.

example, different types of behaviour can be seen in the control group of Figure 17. This is particularly noticeable at the beginning (prior to time $T = 36$) and end (after time $T = 96$) of the experiment where the plant signals displayed one of two different amplitudes. This variability highlights the issues caused by taking an average period estimate for each group and comparing the results, or comparing the average raw time series for each group. Although all plants in each treatment group share identical genetic characteristics and have been treated in identical conditions, they respond differently. In such situations, looking at average behaviour masks the individual differences and is conducive to misleading conclusions, as also acknowledged in other fields (Fiecas and Ombao, 2016). This motivates our choice to cluster the circadian plant data using their time-frequency (scale) patterns and further accounts for their proven (see Section 2.2.3) nonstationary features.

2.3 Proposed Clustering Method

Our proposed methodology combines the use of wavelets, as recommended (but not implemented) by Zielinski et al. (2014) in their review of period estimation methods for circadian data, with rigorous stochastic nonstationary time series modelling. We exploit the locally stationary wavelet processes of Nason et al. (2000), arriving at a novel and general approach for clustering circadian signals according to their leading time-scale spectral patterns, as extracted by functional principal components analysis.

2.3.1 Modelling Nonstationary Time Series

In Section 1.4 we introduced a number of statistically rigorous approaches to modelling nonstationary time series. In our work we adopt the locally stationary wavelet (LSW) model (Nason et al., 2000). Recall (Section 1.4.3) that the advantage of wavelets is that they are localised in both time and scale (frequency) and are therefore well-suited to modelling second-order characteristics that evolve over time. Therefore, the locally stationary wavelet model combines the advantages of a wavelet analysis with rigorous stochastic nonstationary time series modelling.

Under the locally stationary wavelet (LSW) process framework, a time series $\{X_{t;T}\}_{t=0}^{T-1}$, $T = 2^J \geq 1$ is defined to be a sequence of (doubly-indexed) stochastic processes with the following representation

$$X_{t,T} = \sum_{j=1}^{J} \sum_{k \in \mathbb{Z}} w_{j,k;T} \psi_{j,k}(t) \xi_{j,k}, \tag{58}$$

where $\{\xi_{j,k}\}$ is a random orthonormal increment sequence, $\{\psi_{j,k}(t) = \psi_{j,t-k}\}_{j,k}$ is a set of discrete non-decimated wavelets and $\{w_{j,k;T}\}$ is a set of amplitudes, each of which at a scale j and

time k.

The properties of the random increment sequence $\{\xi_{j,k}\}$ ensure that $\{X_{t,T}\}$ is a zero-mean process (see definition 1.4.2 in Section 1.4.2). In practice, for a process with non-zero mean, it is customary to re-centre it around zero (Nason, 2010) and this is our approach here, as the quantity of our primary interest is the process spectral signature.

The definition of the LSW process in equation (58) requires the data to be of dyadic length ($T = 2^J$). In many practical applications, this is not realistic and there are a number of approaches to address this situation. For example, the practitioner could truncate the time series and analyse a segment of the data (of length $T = 2^J$), and this is our approach here. Alternatively, it is possible to extend the data to the next greater power of two by artificially appending values. In particular, common approaches include padding the data with zeros, replicating a data value (such as the final value) or reflecting the dataset about an end point. Another approach is to interpolate data values to produce a new data set of the required length (Ogden, 1997). However, preconditioning the data could lead to misleading results. Therefore, we do not artificially extend the data in this thesis.

In Section 1.4.2, we formally defined the evolutionary wavelet spectrum (EWS) as

$$S_j(z) = |W_j(z)|^2, \tag{59}$$

at each scale $j \in \overline{1,J}$ and rescaled time $z = k/T \in (0,1)$. An unbiased estimator of the EWS $\{S_j(z)\}$ is obtained by correcting the raw wavelet periodogram

$$I^j_{k,T} = |d_{j,k;T}|^2, \tag{60}$$

where $d_{j,k;T} = \sum_{t=0}^{T} X_{t,T}\psi_{j,k}(t)$ are the empirical nondecimated wavelet coefficients. Thus, the corrected wavelet periodogram is

$$\mathbf{L}(z) = A_J^{-1}\mathbf{I}(z), \text{ for all } z \in (0,1), \tag{61}$$

where $A_J = (\sum_\tau \Psi_j(\tau)\Psi_l(\tau))_{j,l}$ is the autocorrelation wavelet inner product ($J \times J$) matrix and $\Psi_j(\tau) = \sum_k \psi_{j,k}(0)\psi_{j,k}(\tau)$ is the autocorrelation wavelet. For the remainder of this chapter, let us denote the corrected and smoothed periodogram of a time series (plant signal) $\{X_{t,T}\}_{t=0}^{T-1}$ as $\{\hat{S}_j(z)\}_j$, for rescaled time $z \in (0,1)$.

2.3.2 Overview of Current Clustering/Classification Techniques that Account for Nonstationarity

The problem of clustering and classification for nonstationary data has received a good deal of attention in the statistical literature, thanks to its relevance in many applied fields. In the context of monitoring potential nuclear testing, Shumway (2003) considered the use of time-varying spectra for the classification and clustering of nonstationary time series by means of locally stationary Fourier models and Kullback-Leibler discrimination measures. Also in this context, Fryzlewicz and Ombao (2009) developed a procedure for the *classification* of nonstationary time series. The observed data were modelled as realisations of locally stationary wavelet processes and their corresponding wavelet spectra were estimated and used as the signal classification signature. In the context of an industrial experiment, Krzemieniewska et al.

(2014) further developed this method by proposing an alternative divergence index to the simple squared quadratic distance of Fryzlewicz and Ombao (2009) for comparing the spectra of two time series. Note that the above techniques are underpinned by rigorous process modelling but the focus is on classification into known groups, rather than on clustering. When classifying animal communication signals, known to have a nonstationary character, Holan et al. (2010) achieved dimension reduction by treating each windowed Fourier spectrum as an 'image' and performing a functional principal components analysis. In this context, the authors proposed to classify nonstationary time series by means of a generalised linear model that incorporated the (dimension-reduced) spectrogram of a short-time Fourier transform into the model as a predictor.

For clustering applications, the maximum covariance analysis (MCA) on wavelet representations of *two series* has been proposed in previous works. MCA has the advantage of extracting common time-scale (frequency) patterns while also reducing the dimension of the data. Rouyer et al. (2008) used MCA to yield a quantitative measure of the common time-scale content in squared wavelet coefficients for pairs of time series. This subsequently yields a distance matrix used to obtain a cluster tree that groups signals according to their spectral time-scale patterns. In the context of an energy application, Antoniadis et al. (2013) also used an MCA over the wavelet coefficients obtained via a continuous wavelet transform and quantify signal similarity by comparing the evolution in time of each pair of leading patterns. This builds a distance matrix which is then used within classical clustering algorithms to differentiate among high dimensional populations.

Formally, consider two time series, $\{X_t^{(i)}\}$ and $\{X_t^{(j)}\}$. Both Antoniadis et al. (2013) and Rouyer et al. (2008) obtained a time-scale decomposition of each time series (the wavelet transform and its squared version, respectively). Regardless of the usage of wavelet coefficients or their squared version, denote these new quantities in the wavelet domain by $Q^{(i)}$ and $Q^{(j)}$, for the $\{X_t^{(i)}\}$ and $\{X_t^{(j)}\}$ signals respectively, and define the time-scale covariance matrix by

$$R^{(i,j)} = Q^{(i)} Q^{(j)H}, \tag{62}$$

where $Q^{(j)H}$ denotes the conjugate transpose and $R^{(i,j)}$ is a $J \times J$ matrix with possibly complex values. Performing a singular value decomposition of $R^{(i,j)}$ gives the following decomposition:

$$R^{(i,j)} = U^{(i)} \Lambda^{(i,j)} V^{(j)H} \tag{63}$$

where the columns of $U^{(i)}$ and $V^{(j)}$ are the orthonormal singular vectors of $Q^{(i)}$ and $Q^{(j)}$ respectively, and $\Lambda^{(i,j)}$ is a diagonal matrix with the singular values of the decomposition arranged in decreasing order. Denote the k-th pair of the singular vectors of $U^{(i)}$ and $V^{(j)}$ as u_k and v_k respectively. We can then define the k-th leading pattern as the projections of $Q^{(i)}$ and $Q^{(j)}$ over their respective k-th singular vectors:

$$P_k^{(i)} = u_k^H Q^{(i)} \text{ and } P_k^{(j)} = v_k^H Q^{(j)}. \tag{64}$$

This process is then repeated for each pair of time series to produce the leading patterns and singular vectors which are then used with various distance measures (described in Section 2.3.4.1) to obtain the dissimilarity matrix which forms the input of classical clustering algo-

rithms.

Contrasting with the classification techniques described above, these clustering approaches are not underpinned by rigorous statistical modelling, and while they propose respectively the usage of wavelet coefficients or their squares, the reasoning that should drive this choice is not discussed by either Rouyer et al. (2008) or Antoniadis et al. (2013).

2.3.3 Proposed Functional Principal Components Analysis for the Wavelet Spectral Content

In this work we propose to combine the rigorous modelling framework provided by the locally stationary wavelet (LSW) processes that allows for the reliable (unbiased and consistent) estimation of the spectral time-scale features specific to each plant, with the dimension reduction afforded through the use of a functional principal components analysis (FPCA).

In our biological problem of interest, the time-scale representation of the signal is high-dimensional. Since any useful biological information is likely to relate to the low-dimensional mechanisms known to regulate the clock (Bujdoso and Davis, 2013), this motivates our proposal to use a FPCA to perform dimension reduction over the spectral content. In the spirit of Holan et al. (2010), we treat our LSW spectral estimate as an 'image' and the spectral coefficients as time-scale 'pixels'. The pixels are not independent– in fact, the spectrum presents coherent patterns that should be accounted for. This motivates the use of the Karhunen-Loéve representation (at the heart of FPCA) which, in our context, for a continuous spectrum $\{S(\mathbf{v}) : \mathbf{v} = (j,z), \mathbf{v} \in \mathbb{R}\times(0,1)\}$ allows for its covariance function $C_S(\mathbf{v},\mathbf{v}')$ to be decomposed via an eigen-decomposition (Ramsay and Silverman, 2005). Consequently, the spectra may be decomposed as $S(\mathbf{v}) = \sum_{m\geq 1} \alpha_m \phi_m(\mathbf{v})$, with scores $(\alpha_m)_m$ independent random variables whose variance is given by the corresponding eigenvalues ($\mathrm{Var}(\alpha_m) = \lambda_m$) and $\phi_m(\mathbf{v})$ orthonormal eigenvectors that capture the variability in the spectral domain.

Assuming we observed N plant signals at $T = 128$ equally spaced time points, we model the i-th plant signal as an LSW process $\{X^{(i)}_{t,T}\}_{t=0}^{T-1}$ for each $i = 1,\ldots,N$. As biological evidence points towards the relevance of the plant spectral signature in understanding its response to stimuli, we estimate the wavelet spectrum by means of its corresponding corrected and smoothed periodogram, $\{\hat{S}^{(i)}_j(t/T)\}_{j=1}^{J}$ for each time series $i = 1,\ldots,N$, where $t = 0,\ldots,T-1$ and $J = \log_2(T)$. The estimated spectra, viewed as continuous functions $\{\hat{S}^{(i)}(\mathbf{v})\}$ with $\mathbf{v} = (j, z = t/T) \in \mathbb{R}\times(0,1)$, are then treated as input observations in a FPCA. Their corresponding estimated covariance function $\hat{C}(\mathbf{v},\mathbf{v}')$ thus summarises the dependence of plants across time *and* scale.

Although the continuous Karhunen-Loéve representation is often the most realistic from the point of view of modelling a biological process, due to the discrete nature of observations resulting from most experiments, it is rarely considered in applications. In practice, we use its empirical version, also known as empirical orthogonal function analysis, as is common in e.g. spatial statistics and geophysics (Cressie and Wikle, 2015). In particular, the estimated spectral coefficients can be arranged in N matrices, each of size $J\times T$, which we denote $\hat{S}^{(1)},\ldots,\hat{S}^{(N)}$. In

particular, for each time series $i = 1, \ldots, N$,

$$\hat{S}^{(i)} = \begin{bmatrix} \hat{S}_1^{(i)}\left(\frac{0}{T}\right) & \hat{S}_1^{(i)}\left(\frac{1}{T}\right) & \ldots & \hat{S}_1^{(i)}\left(\frac{T-2}{T}\right) & \hat{S}_1^{(i)}\left(\frac{T-1}{T}\right) \\ \hat{S}_2^{(i)}\left(\frac{0}{T}\right) & \hat{S}_2^{(i)}\left(\frac{1}{T}\right) & \ldots & \hat{S}_2^{(i)}\left(\frac{T-2}{T}\right) & \hat{S}_2^{(i)}\left(\frac{T-1}{T}\right) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{S}_J^{(i)}\left(\frac{0}{T}\right) & \hat{S}_J^{(i)}\left(\frac{1}{T}\right) & \ldots & \hat{S}_J^{(i)}\left(\frac{T-2}{T}\right) & \hat{S}_J^{(i)}\left(\frac{T-1}{T}\right) \end{bmatrix}$$

For each plant signal (each $i = 1, \ldots, N$), vectorise the matrix $\hat{S}^{(i)}$, i.e. concatenate the rows of the matrix $\hat{S}^{(i)}$ to produce a vector $\hat{\mathbf{s}}^{(i)}$ with length $J \times T = n$:

$$\hat{\mathbf{s}}^{(i)} = \left[\left(\hat{S}_1^{(i)}\left(\frac{0}{T}\right), \ldots, \hat{S}_1^{(i)}\left(\frac{T-1}{T}\right)\right), \ldots, \left(\hat{S}_J^{(i)}\left(\frac{0}{T}\right), \ldots, \hat{S}_J^{(i)}\left(\frac{T-1}{T}\right)\right)\right]^T.$$

These N vectors are combined to form a data matrix Q of size $N \times n$, where each row of Q represents the spectral content of a plant. Formally,

$$Q = \left[\hat{\mathbf{s}}^{(1)}, \ldots, \hat{\mathbf{s}}^{(N)}\right]^T. \tag{65}$$

Note that in practice, this analysis is equivalent to performing a classical principal components analysis on the mean centred data, which we still denote by Q in order not to further clutter the notation. The spectral decomposition of the sample covariance matrix $R = Q^T Q$ is given by

$$R = U \Lambda U^T, \tag{66}$$

where U is an orthonormal matrix whose columns are the eigenvectors of R (also known as the principal directions of the data; here, we can conceptualise these as representing 'images') and Λ is a diagonal matrix whose diagonal elements are eigenvalues of R (positive real numbers arranged in decreasing order of magnitude; these are proportional to the variance accounted for by each direction).

We can achieve size reduction by choosing to represent our data in fewer dimensions. The usual practice is to use the set of $p < n$ eigenvectors of R corresponding to the p largest eigenvalues and aggregate these in an $n \times p$ matrix, U_{PCA}, which performs the PCA projection. Therefore, for each eigenvector, we can find a corresponding projection in the principal component space by computing

$$QU_{\text{PCA}}.$$

In this transformed space, each process is now represented by a p-dimensional vector, i.e. the principal co-ordinates of the i-th process are given by the i-th row of the matrix QU_{PCA}, denoted from now on as $\text{Score}^{(i)}$ (p-dimensional vector). Therefore,

$$\text{Score}^{(i),T} = \left[\text{Score}_1^{(i)}, \ldots, \text{Score}_p^{(i)}\right] = \hat{\mathbf{s}}^{(i),T} U_{\text{PCA}}. \tag{67}$$

2.3.4 Proposed Clustering Method

Our proposal is to construct a clustering method that assesses time series similarity/ dissimilarity on the basis of their spectral content as distilled in the scores developed in Section 2.3.3 above. Next we shall introduce potential distance measure candidates and assess various methods to determine the number of principal components to retain and the optimal number of

clusters.

2.3.4.1 Distance Measures

The success of any clustering algorithm depends on the adopted dissimilarity measure. In this section, we propose four possible distance measures and discuss their advantages and disadvantages. The proposed distance measures consist of developments of those adopted in the work reviewed in Section 2.3.2. The distance measures are then utilised to form an $N \times N$ matrix, D, which we will refer to as the dissimilarity matrix. In particular, the (i, j)th entry of the dissimilarity matrix is defined as the value of a chosen distance measure between the two time series $\{X_{t,T}^{(i)}\}_{t=0}^{T-1}$ and $\{X_{t,T}^{(j)}\}_{t=0}^{T-1}$, for each $i = 1,\ldots,N$ and $j = 1,\ldots,N$. In our simulation studies (Section 2.4), we compare the performance of clustering algorithms embedding the different distance measures outlined below.

The simplest choice for the dissimilarity measure is the squared quadratic (SQ) distance between two time series, $\{X_{t,T}^{(i)}\}_{t=0}^{T-1}$ and $\{X_{t,T}^{(j)}\}_{t=0}^{T-1}$. This distance measure is adopted by Fryzlewicz and Ombao (2009) who quote its advantages of good practical performance and computational ease. In our context it is defined as the sum of the squared differences between the scores relating to the p principal components retained

$$SQ(X_{t,T}^{(i)}, X_{t,T}^{(j)}) = \sum_{k=1}^{p} \left[\text{Score}_k^{(i)} - \text{Score}_k^{(j)}\right]^2, \tag{68}$$

where $\text{Score}_k^{(i)}$ denotes the score associated to the k-th principal component of time series $\{X_{t,T}^{(i)}\}$, as defined in equation (67). The value $SQ(i, j)$ is the (i, j)th entry of the dissimilarity matrix, D.

Our proposal is to develop this simplistic measure by aggregating the scores in the most significant p directions using a *weighted* combination with weights given by the squared singular values. We refer to this measure as the weighted squared quadratic (WSQ) distance and define the WSQ distance between two time series, $\{X_{t,T}^{(i)}\}_{t=0}^{T-1}$ and $\{X_{t,T}^{(j)}\}_{t=0}^{T-1}$ as the weighted sum of the squared differences between their scores in p directions. Formally

$$WSQ(X_{t,T}^{(i)}, X_{t,T}^{(j)}) = \frac{\sum_{k=1}^{p} \lambda_k \left[\text{Score}_k^{(i)} - \text{Score}_k^{(j)}\right]^2}{\sum_{k=1}^{p} \lambda_k}, \tag{69}$$

where $\text{Score}_k^{(i)}$ is as in equation (67) and λ_k denotes the corresponding k-th squared singular value. The value $WSQ(i, j)$ is the (i, j)th entry of the dissimilarity matrix, D.

We now outline the distance measures as adopted in Antoniadis et al. (2013) and Rouyer et al. (2008). Both approaches hinge on the singular vectors and leading patterns for each time series pair. Specifically, Antoniadis et al. (2013) compared the time evolution of each pair of leading patterns. In particular, for the k-th pair of leading patterns corresponding to time series $\{X_{t,T}^{(i)}\}_{t=0}^{T-1}$ and $\{X_{t,T}^{(j)}\}_{t=0}^{T-1}$, the authors take the first difference (Δ) and measure energy by means of its modulus

$$d_k(i, j) = |\Delta(P_k^{(i)} - P_k^{(j)})|. \tag{70}$$

Finally, the most significant p directions are aggregated using a weighted combination with

weights given by the squared singular values:

$$D(i,j) = \frac{\sum_{k=1}^{p} \lambda_k d_k^2(i,j)}{\sum_{k=1}^{p} \lambda_k}. \tag{71}$$

The last comparison metric is

$$DT(i,j) = \frac{\sum_{k=1}^{p} \lambda_k (RD(P_k^{(i)}, P_k^{(j)}) + RD(\mathbf{u}_k^{(i)}, \mathbf{u}_k^{(j)}))}{\sum_{j=1}^{p} \lambda_k}, \tag{72}$$

where $\mathbf{u}_k^{(i)}$ and $\mathbf{u}_k^{(j)}$ are the k-th singular vectors of $X_{t,T}^{(i)}$ and $X_{t,T}^{(j)}$ respectively, and RD denotes the measure from Rouyer et al. (2008), adapted from Keogh and Pazzani (1998). Formally, for two vectors $\mathbf{u} = [u_1, \ldots, u_n]^T$ and $\mathbf{v} = [v_1, \ldots, v_n]^T$ of length n,

$$RD(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{n-1} atan[|(u_i - v_i)) - (u_{i+1} - v_{i+1})|]. \tag{73}$$

The metric in equation (73) compares two vectors by measuring the angle between each pair of corresponding segments (a segment is defined as a pair of consecutive points of a vector) and is a method for measuring parallelism between curves. The overall distance is then computed as a weighted mean of the distance for each of the p pairs of leading patterns and singular vectors retained (with the weights being equal to the amount of covariance explained by each axis), see equation (72).

Note that in the simulation study (Section 2.4), when comparing our method with the methods outlined in Antoniadis et al. (2013) and Rouyer et al. (2008), we cluster the data using their specified time-scale decomposition and distance measure.

2.3.4.2 Determining the number of principal components to retain

Recall the aim to reduce the dimensionality of our problem; for each of the distance metrics above, we must decide how many axes, p, to retain. Antoniadis et al. (2013) and Rouyer et al. (2008) both decided to use the number of axes that correspond to a fixed percentage of the total covariance (as is common in principal components analysis). A different approach is to select the number of components based on a screeplot. This displays the proportion of variance explained by the (ordered) eigenvalues, and p is then selected by looking for an elbow in the screeplot. Finally, our proposed methodology is motivated by an applied problem in the field of circadian biology. In order to interpret the results of our proposed clustering algorithm and potentially to gain biological insight, practitioners expressed a desire for a method of visualising the clusters. In particular, if two principal components were retained, the scores could be plotted as a (colour–coded) two-dimensional scatter plot (see Figure 28). Therefore, we also investigated the impact on our proposed methodology of always retaining two principal components.

2.3.4.3 Determining the Number of Clusters

One of the most difficult tasks in clustering is determining the number of clusters (Antoniadis et al., 2013). This can be informed through a number of statistical techniques (Kaufman and

Rousseeuw, 2009) as well as by scientific expert knowledge. For example, the 'elbow method' examines the percentage of variance explained as a function of the number of clusters; the number of clusters is then chosen by looking for an elbow in the plot of this function. Tibshirani et al. (2001) developed this methodology by estimating the number of clusters in a dataset via the 'gap statistic'. This technique uses the output of any clustering algorithm and compares the change in within–cluster dispersion (e.g. the pooled within–cluster sum of squares around the cluster mean) with that expected under an appropriate reference null distribution. Tibshirani et al. (2001) provide two choices for the reference distribution (see the original manuscript for further details). Alternatively, the 'silhouette method' (Rousseeuw, 1987) can be used. The 'silhouette' of a data point is a number between -1 and 1, with values of 1 indicating correct clustering. Briefly, the silhouette of an observation compares the average distance of that observation to all other elements in the cluster to which it has been assigned with the average distance between the observation and the "closest" alternative cluster. Optimization techniques are then used to determine the number of clusters that gives rise to the largest 'silhouette' (Kaufman and Rousseeuw, 2009).

2.3.4.4 Proposed LSW-PCA Clustering Algorithm

Our proposed clustering method, which we shall refer to as LSW-PCA clustering, is outlined in Algorithm 1 below. We perform a partitioning around medoids (PAM). The motivation behind this choice was that this method (implemented in R) admits a general dissimilarity matrix as input (as opposed to the raw data). Therefore, this method permitted the comparison of the proposed distance measures (outlined in Section 2.3.4.1). Furthermore, PAM is known to be more robust than other alternatives such as k-means (Antoniadis et al., 2013). Each of the proposed choices, i.e. spectral information, number of principal components retained (p) and distance measure, are informed by the findings of the simulation study (see Section 2.4 and Appendix 2.10).

Algorithm 1 Proposed LSW-PCA clustering algorithm

Assume that each of the N observed (e.g. circadian) signals is a realisation of a locally stationary LSW process $\{X_{t,T}^{(i)}\}_{t=0}^{T-1}$, with $i = 1, 2, \ldots, N$.

1. *Spectral estimation*: estimate the spectral content of each process by using a model-based LSW corrected estimator and aggregate all information in a matrix (see Section 2.3.3).

2. *Dimension reduction*: achieve dimension reduction by projecting the spectral information of each process in a functional principal component space and obtain the scores associated to each signal. The number of principal components retained (p) is decided by means of the screeplot of percentage variance explained (see Section 2.3.4.2).

3. *Spectral distance matrix*: quantify the spectral differences between two signals by using the (weighted) squared quadratic distance measure (see Section 2.3.4.1).

4. *Cluster the data*: by performing a partitioning around medoids (PAM) with the distance matrix above as input.

2.4 Simulation Study

The goals of our simulation study are twofold. First, we investigate the impact of the wavelet information choice (e.g. wavelet coefficients versus model-based spectral estimate), distance measure choice and methods to determine the number of principal components to retain. Secondly, we assess the comparative performance of our proposed procedure with other methods. Since our work is motivated by an application in the field of circadian biology, we have designed our simulated scenarios to display typical characteristics of circadian rhythms and also to reflect the limitations of empirical work in the life sciences, where the resolution and length of the time series would be limited in practice.

2.4.1 Simulated Data

The basic structure of each simulated experiment can be described as follows. A dataset of $N = 100$ (50 simulations from each of the two groups) was generated. For cases 1, 2 and 3, the data was generated using the LSW representation (see equation (58)) with Daubechies' extremal phase wavelet with one vanishing moment and a Gaussian orthonormal increment sequence with mean zero and unit variance (the `locits` R package was used). For Case 4, the data was generated from an AR process (see Section 1.3) with time–varying coefficients. For the proposed methodology, each periodogram was level smoothed by log transform, followed by translation invariant global universal thresholding and then the inverse transform was applied. For each scale of the wavelet periodogram, only levels 3 and finer were thresholded. For all methods, using the appropriate estimated spectral information, we obtained a dissimilarity matrix for each of the methods under investigation. This matrix was the input of a PAM algorithm (performed in the `cluster` R package) which clustered the data into two groups. We then compared the clusters with the known group memberships and recorded the correctly clustered percentage. The above procedure was then repeated 100 times and the results for each method were averaged.

Case 1: Defined spectra. For this study, we assume each time series is a realisation from one of $g = 1, 2$ possible groups, each with different spectral characteristics. Define the evolutionary wavelet spectrum of each group $\{S_j^{(g)}(z)\}_{j=1}^{J}$ with $J = \log_2(T)$ for all $z \in (0,1)$ and $T = 64$ by

$$S_j^{(1)}(z) = \begin{cases} 4\cos^2(4\pi z), & \text{for } j = 2, z \in (1/64, 16/64) \\ 4\cos^2(2\pi z), & \text{for } j = 3, z \in (17/64, 1) \\ 0, & \text{otherwise;} \end{cases} \tag{74}$$

and

$$S_j^{(2)}(z) = \begin{cases} 4\cos^2(2\pi z), & \text{for } j = 2, z \in (17/64, 1) \\ 4\cos^2(4\pi z), & \text{for } j = 3, z \in (1/64, 1/2) \\ 0, & \text{otherwise;} \end{cases} \tag{75}$$

The choice above encompasses changes in amplitude and period through time, akin to those of interest to the circadian biologist. Figure 18 provides a visualisation of the wavelet spectra above (top row) and an example of a signal realisation from each of the two groups (bottom row).

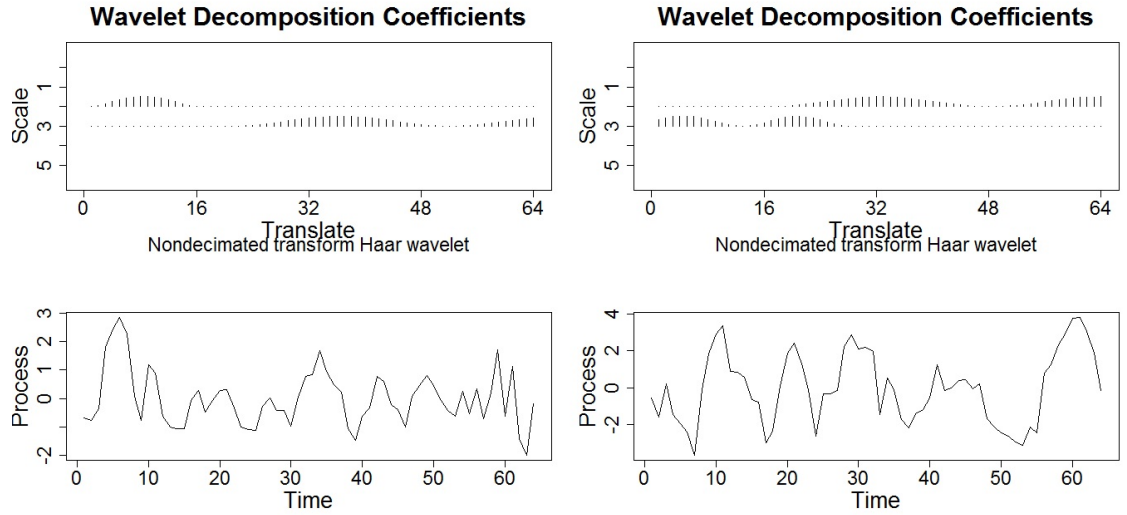

Figure 18: **Case 1.** Top left: Group 1 wavelet spectrum; Top right: Group 2 wavelet spectrum; Bottom left: Group 1 realisation and Bottom right: Group 2 realisation.

Case 2: Gradual period change. For our second study, we assume each time series is a realisation from one of 3 possible groups, each with different spectral characteristics. In particular, each group represents a time series that gradually changes period from 24 to: 25 (Group 1), 26 (Group 2) and 27 (Group 3) over (approximately) two days, before continuing with the relevant period for a further two days. The purpose of this simulation study is to replicate a typical circadian experiment with changes that could not be captured by standard analyses that assume stationarity and report an average period value. Therefore, we will take $T = 256$ which is equivalent to a free-running period of 4 days with equally spaced observations every 22.5 minutes. Figure 19 shows the wavelet spectra which represent the gradually changing periods that define each of the 3 groups above. Notice that the increased period is shown by the movement up through the resolution levels and the gradual increase in period of the wavelet coefficients. To determine which changes can be discriminated by the methods, we perform two studies within this setting (i) Case 2A: simulations from Group 1 and Group 2, and (ii) Case 2B: simulations from Group 1 and Group 3.

Case 3: Different rates of change. For our next study, let us assume each time series is a realisation from one of 3 possible groups, each with different spectral characteristics. In particular, each group represents a time series that gradually changes period from 24 to period 27 over 2 days (Group 1), 3 days (Group 2), 5 days (Group 3) and then continues with period 27 for the remainder of the experiment. The purpose of this simulation study is to replicate a circadian experiment with changes that could not be captured by standard analyses that assume stationarity and report an average period value. Therefore, we also take $T = 256$ which is equivalent to a free-running period of 4 days with equally spaced observations every 22.5 minutes. Figure 20 shows the wavelet spectra which represent the characteristics that define each of the 3 groups above. To determine which changes can be discriminated by the methods, we perform three studies within this setting: (i) Case 3A: simulations from Group 1 and Group 2, (ii) Case 3B: simulations from Group 1 and Group 3, and (iii) Case 3C: simulations from Group 2 and Group

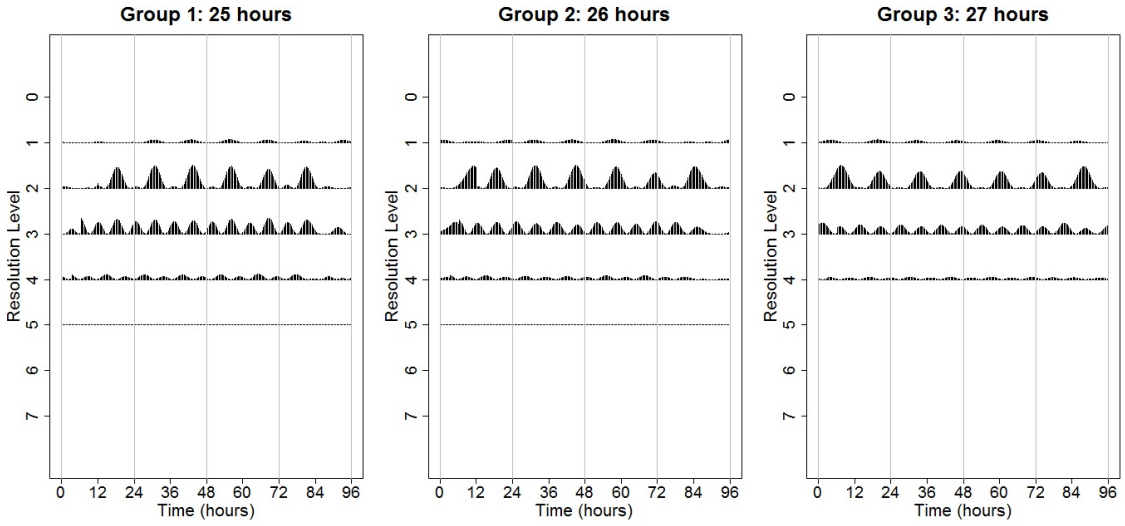

Figure 19: **Case 2.** Left: Group 1 wavelet spectrum (gradual period change from 24 to 25 hours); Centre: Group 2 wavelet spectrum (gradual period change from 24 to 26 hours); Right: Group 3 wavelet spectrum (gradual period change from 24 to 27 hours).

3.

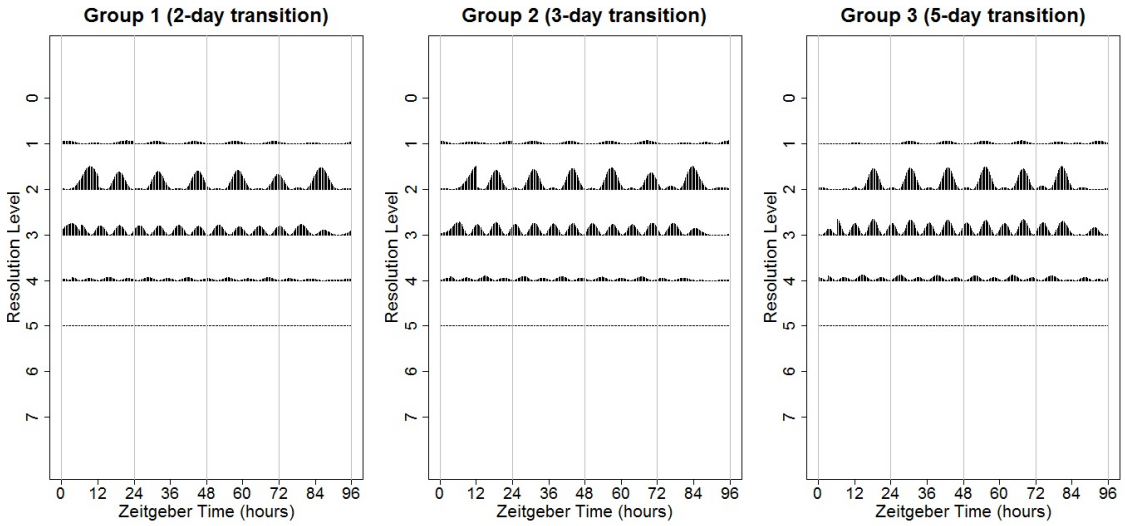

Figure 20: **Case 3.** Left: Group 1 wavelet spectrum (2-day transition); Centre: Group 2 wavelet spectrum (3-day transition); Right: Group 3 wavelet spectrum (5-day transition).

Case 4: Nonstationary AR process. The signals in cases 1, 2, and 3 are generated from a defined group spectrum, satisfying the underlying LSW modelling assumptions of our proposed methodology. The purpose of this study is to asses the performance of our tests when this assumption is not met. Therefore, we simulate from an important class of nonstationary processes– AR processes with time-varying coefficients. We propose a simulation study in a setting as described in Fryzlewicz and Ombao (2009) Section 4.1 Case 1 (AR processes with abruptly changing parameters). The r_i-th time series from group $i = 1,2$, denoted $X_{n,t}^{(i),r_i}$ is generated from the process defined by:

$$X_t^{(i),r_i} = \phi_1^{(i)}(t) X_{t-1}^{(i),r_i} + \phi_2^{(i)}(t) X_{t-2}^{(i),r_i} + \epsilon_t^{(i),r_i}, \tag{76}$$

Time-varying parameters	Time Index	Group $i=1$	Group $i=2$
$\phi_1^{(i)}(t)$	$t=1,\ldots,53$	0.8	0.8
	$t=54,\ldots,128$	-0.9	0.6
	$t=129,\ldots,256$	0.8	0.8
$\phi_2^{(i)}(t)$	$t=1,\ldots,256$	-0.81	-0.81

Table 3: **Case 4.** The abruptly changing parameters of two nonstationary autoregressive processes.

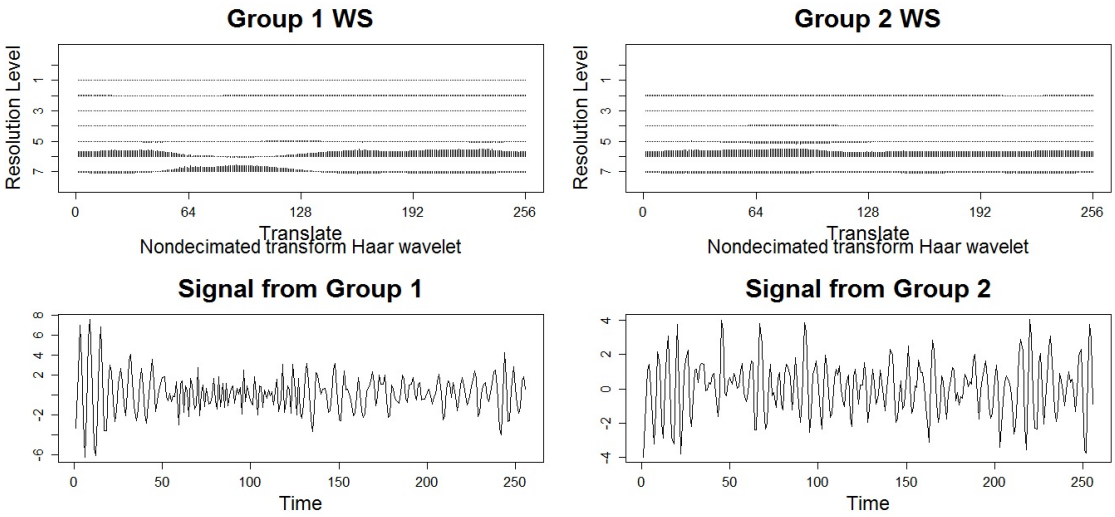

Figure 21: **Case 4.** Nonstationary autoregressive processes. Top left: Estimated wavelet spectrum of Group 1; Top right: Estimated wavelet spectrum of Group 2; Bottom left: Group 1 realisation; Bottom right: Group 2 realisation.

where the innovations $\epsilon_t^{(i),r_i}$ are independent and identically distributed (iid) Gaussian with zero mean and unit variance. In this study, the squared difference between the group spectra is relatively large and the abruptly changing parameters for the two groups are shown in Table 3. Representative time series plots from each group and the estimated spectra are shown in Figure 21.

2.5 Results

For each of our simulation studies outlined above, we investigate the impact of the wavelet information choice (e.g. wavelet coefficients versus model-based spectral estimate), distance measure choice and methods to determine the number of principal components to retain. We report our findings next, with detailed results for Cases 1 and 4 presented in Appendix 2.10.

Distance measure choice. To examine the effect of the choice of distance measure on our proposed clustering method, we performed the simulation studies as outlined above using all four distance measures defined in Section 2.3.4.1. We found that our method is fairly robust to the choice of distance measure, although the squared and weighted squared quadratic distances (SQ, respectively WSQ), appear to give superior results to the distance choices in Antoniadis et al. (2013) and Rouyer et al. (2008).

Dimension choice. We also examined the different methods outlined in Section 2.3.4.2 to select the number of principal components to retain for our LSW-PCA clustering method. We thus compared determining the number of principal components to retain by examining the screeplot with the situation where we retain the minimal number of components that correspond to 90% of the total covariance. Once again we found that the LSW-PCA clustering method is robust to the way in which we choose the number of principal components to retain. Based on these results, we suggest using the LSW-PCA clustering method with the squared quadratic distance (see equation (68)), and retaining principal components by examining the screeplot. However, note that our algorithm is robust to an automatic choice based on a set percentage of the total covariance.

Furthermore, recall in Section 2.3.4.2 we outlined that in certain practical situations (such as our motivating example), retaining two principal components could aid visualisation and hence interpretation of the results of our clustering algorithm. Therefore, we also compared the above methods with this situation. We found that, in these settings, our proposed methodology is also fairly robust to this choice. For example, in Case 4 (AR processes), we found that this method had a correct clustering rate of 99% compared with 98% for the total covariance explained method (detailed results can be found in Table 9). However, this is potentially due to the other methods also choosing similar numbers of components (typically less than 5). Therefore, in certain practical situations, to aid ease of interpretation, we would permit the choice of retaining two components, if this was justifiable using the screeplot or total covariance explained methods.

Wavelet information choice. In Section 2.3.2 we noted that other wavelet-based clustering approaches in the literature, while non-model based techniques (unlike our proposed LSW-PCA), extract the information by means of wavelet coefficients (Antoniadis et al., 2013) or squared wavelet coefficients (Rouyer et al., 2008). Therefore, to justify our decision to formulate our proposed methodology using within the LSW framework, we performed two simulation studies (using the Case 1 and Case 4 settings). This allows us to compare utilising the LSW methodology over standard wavelet–based approaches in a range of different scenarios, both when the LSW modelling assumption is correct (Case 1) and when the data consists of nonstationary AR processes (Case 4). Therefore, to investigate the impact of wavelet information choice, we performed each simulation study with the following input data: original signals (thus extracting time-dependent information only), wavelet coefficients (time-scale information), squared wavelet coefficients (second-order time scale information) and finally the LSW corrected wavelet periodogram (to consistently estimate the spectrum under the LSW modelling framework, but without the FPCA stage). The results can be found in Table 10 in Appendix 2.10.

For Case 1, we found that clustering based on the raw data and the raw wavelet transform gave poor results (54% correctly clustered compared to 63% for squared wavelet coefficients and 69% for the corrected periodogram) which supports the assertion that clustering based on the second-moment information is preferable. Also note that using the FPCA approach further improves the results, from 69% correctly clustered to 76% (see Table 4). Similar results are also obtained for the Case 4 setting (nonstationary AR processes), see Table 10 in Appendix 2.10. These results demonstrate the advantages of utilising the LSW methodology over standard wavelet–based approaches in a range of different scenarios, both when the LSW modelling as-

sumption is correct (Case 1) and when the data consists of nonstationary AR processes (Case 4).

Performance comparison. Finally, we compare the LSW-PCA method with the competitor methods proposed by Rouyer et al. (2008) and Antoniadis et al. (2013) (outlined in Section 2.3.2). Both of these benchmark methods do well in practice and represent the state-of-the-art among procedures for clustering nonstationary time series. The results are summarised in Table 4. These simulation studies provide empirical evidence that our proposed LSW-PCA method works very well and outperforms its competitors for clustering nonstationary time series. Again we see that (for this particular application) methods based on the second-order information (our LSW-PCA method and the Rouyer et al. (2008) method) perform better than the method based on the wavelet transform (Antoniadis et al., 2013). Moreover, our method, which utilises an LSW model to obtain an unbiased, consistent estimator of the underlying spectral information, performs considerably better still than the method which uses the raw wavelet periodogram. These results also show that our proposed method, which performs an FPCA on the estimated spectral coefficients of the entire dataset, outperforms the pairwise methods of Rouyer et al. (2008) and Antoniadis et al. (2013). However, note that in Cases 2A, 3A and 3C, the LSW-PCA method also has difficulty discriminating between the defined groups. These results may be due to the resolution of the data. Therefore, if the analyst predicted that a treatment effect would be characterised by this behaviour, we would recommend increasing the length of the experiment and taking observations at shorter intervals which would improve the resolution of all methods.

Sim. Study	Rouyer et al. (2008)	Antoniadis et al. (2013)	LSW-PCA Method
Case 1	66%	61%	76%
Case 2A	56%	54%	65%
Case 2B	58%	55%	76%
Case 3A	54%	54%	61%
Case 3B	55%	55%	75%
Case 3C	55%	54%	63%
Case 4	54%	53%	99%

Table 4: Comparison of the proposed LSW-PCA clustering method with the methods proposed by Rouyer et al. (2008) and Antoniadis et al. (2013) for the simulation studies. Percentages show correct clustering rates.

2.6 Real Data Analysis

2.6.1 Previously Published Circadian Data

In this section, we apply our method to an already published circadian dataset, which tested the effects of copper on plants in a method similar to our cerium dataset. Our aim is to demonstrate the additional insights provided by our proposed method. The dataset from Perea-García et al. (2016a,b) examined circadian rhythms in high concentrations of copper as well as copper deficiency. This previously published circadian data will henceforth be referred to as the copper dataset.

The copper dataset was also obtained using a firefly luciferase reporter system, as described

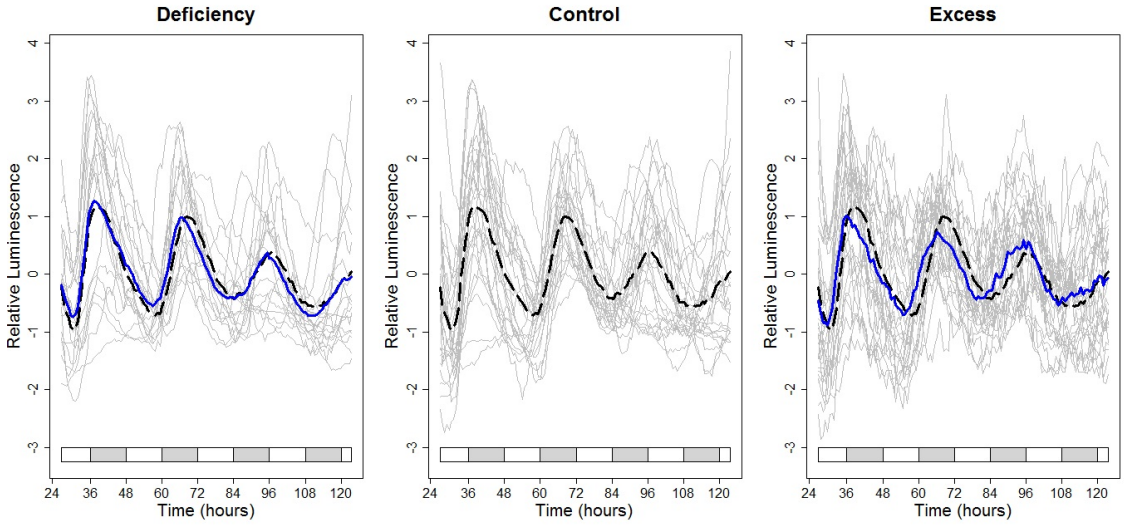

Figure 22: Luminescence evolution over time for plants subjected to a control and 2 different copper regimes. Time is measured in hours relative to *zeitgeber* time (time of last external temporal cue: the dawn signal of lights-on). Centre: Each plant signal from the **'Control'** group (in grey) along with the group average (dashed black). Other panels: Each realisation from the groups (in grey) along with the group average (in blue) and the control group average (dashed black). Left: **'Deficiency'** Group (1/2 MS). Right: **'Excess'** group (10 μM CuSO4). (Each time series has been normalised to have mean zero.) The grey and white bars indicate the subjective night and day, respectively.

in Appendix 2.9. However, this experiment used a different gene of interest, GIGANTEA (GI). For a detailed description of these experimental methods see Appendix 2.11 and Perea-García et al. (2016a,b). Briefly, plants were grown under different copper regimes: 'Deficiency' (no CuSO4), 'Sufficiency' or 'Control' (1 μM CuSO4), and 'Excess' (10 μM CuSO4). The copper dataset consists of a total of 74 plant signals (time series) recorded at 151 time points, with the 'Deficiency' group containing 19 plants; the 'Control' or 'Sufficiency' group, 26 plants and the 'Excess' group, 29 plants. Perea-García et al. (2016a) conducted an analysis in BRASS (see Section 2.2.2) and concluded that the period did not seem to be affected by copper deficiency or excess. In particular, the average period estimates for each group were reported not statistically significantly different. Therefore, it was concluded that changes in available copper were not readily detected by BRASS, even though qualitative differences were easily noted. These findings provide supportive evidence that more statistically advanced approaches are needed to analyse these types of data.

We analysed the circadian copper data using the proposed LSW-PCA clustering method (outlined in Algorithm 1) to establish and characterise the effect copper has on GI within the *Arabidopsis* circadian clock. As the LSW model is underpinned by wavelets and requires the data to be of dyadic length $\left(T = 2^J\right)$, in our analysis, we chose a segment of length $T = 128$ out of the copper dataset. This truncation was decided upon after consultation with the experimental scientists, who confirmed that the selected segments contained the times during which the plant transfered from entrained cycles into 'free-running conditions' (constant light). Figure 22 shows each individual luminescence time series from each treatment group (in grey) along with the group average (in bold) for our truncated demeaned dataset. The average of the 'Control' group is also shown in (dashed) black in each plot for comparison. For each plant

Number of plants	Deficiency	Control	Excess	Total
Cluster 1	**11**	**14**	13	38
Cluster 2	8	12	**16**	36
Total	19	26	29	74

Table 5: Results of clustering the copper dataset into two clusters using the proposed LSW-PCA method. The modal cluster for each copper regime is highlighted in bold.

we estimated the wavelet spectrum by means of the corrected wavelet periodogram estimate (with the same setting as described in the simulation study). After examining the screeplot, and for ease of interpretation, we retained two principal components to use for clustering. Using a dissimilarity matrix obtained by computing the squared quadratic distance between the first two scores of each time series, the proposed LSW-PCA clustering method yielded the results detailed in Table 5.

In determining the optimal number of clusters, we used the 'elbow method' and then validated this result via the 'silhouette method' (implemented in the `fpc` R package), as outlined in Section 2.3.4.3. Both approaches indicated that we should cluster the data into two groups. This result was also supported by consultations with experimental scientists, since clustering the data into two groups could answer the question, 'Is it the local concentration of copper, or simply the presence or absence of copper, which dictates plant–level response?' Such results would be of biological interest, as copper is an important environmental pollutant (Oakenfull et al., 2018) with guidelines governing its acceptable concentrations in soils (Environmental Protection Act, 1990). (This will be explored in more detail in Chapter 4.)

Discussion of findings. Both approaches (outlined in Section 2.3.4.3) indicated that we should cluster the data into two groups. This initial result is of biological interest, since two clusters suggests the presence of two distinct groups within this dataset, each with different time-frequency behaviour. This is in contrast to the results in Perea-García et al. (2016a), which found no detectable difference in period (even though qualitative differences were easily noted).

On examining Table 5, we can see that the LSW-PCA clustering method has clustered the behaviour of the data into the following two groups: Cluster 1 identifies similar behaviour of plants in the 'Control' and copper 'Deficiency' groups, and Cluster 2 is the modal cluster of the copper 'Excess' group. These results are biologically insightful and in agreement with Figure 22 which provides visual evidence that the plants in the copper 'Excess' group seemed to display distinct behaviour from the other groups.

However, on examining Figure 22, note the presence of two distinct types of behaviour within each treatment group. This is particularly noticeable in the 'Excess' group (where the time series appear to peak at around 36 hours or at around 40 hours). Figure 23 shows the final cluster each individual time series was assigned to: the individual signals are plotted in red for Cluster 1 and blue for Cluster 2, for each treatment group. Figure 23 highlights individual-level variability in plant response to stimuli, despite their sharing identical genetic characteristics (Doyle et al., 2002)- although all plants in each treatment group share identical genetic characteristics and have been treated in identical conditions, they respond in two different ways. Note that the treatment group averages (in black) lie between the two (within treatment group) cluster averages. This is particularly noticeable in the 'Deficiency' group. Therefore, the pres-

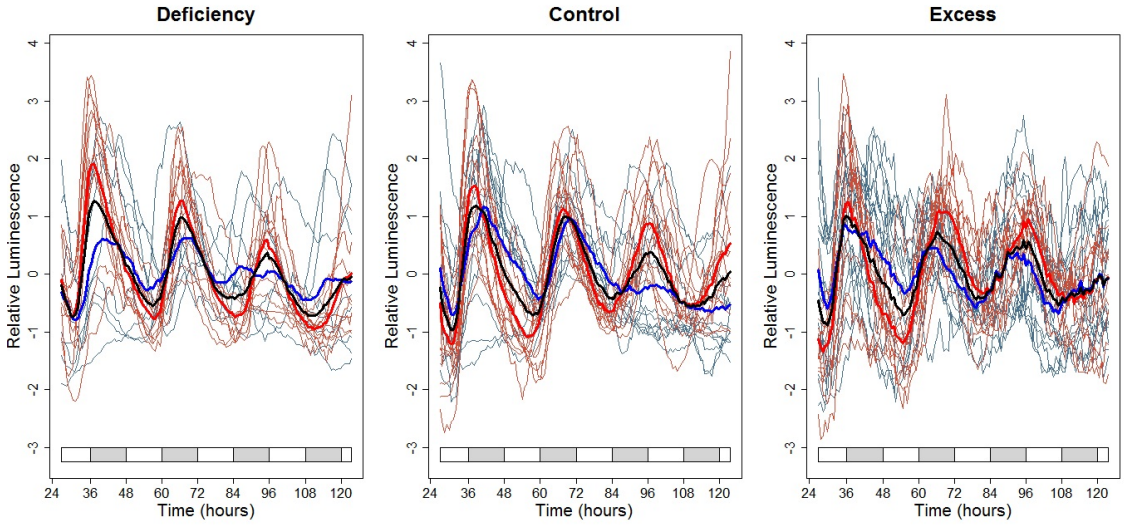

Figure 23: Results of clustering the copper dataset into two clusters using the proposed LSW-PCA method. For each treatment group the individual signals are plotted in: red for Cluster 1 and blue for Cluster 2. The average of each treatment group is shown in black. Within each treatment group, the Cluster 1 average is shown in bold red and the Cluster 2 average in bold blue.

ence of both types of behaviour in each of the original treatment groups has resulted in similar average behaviour (which could explain the misleading results of the original investigation in Perea-García et al. (2016a)). On examining Table 5 and Figure 23, we find that the Cluster 2 'Excess' behaviour can also be seen in some plants in the other two groups, particularly in the 'Control' group. The presence of 'Control' and 'Deficiency' treated plants in the cluster associated mostly with 'Excess' levels of copper may be due to the individual plants in some instances showing a general stress response, particularly those individuals from the 'Deficiency' group in Cluster 2. Alternatively, this may be due to stress induced by the experimental method itself. Thus, although both types of behaviour are present in each treatment group, we can conclude that increased levels of copper increase the likelihood of a Cluster 2-type response.

Our proposed method also allows us to characterise the behaviour associated with each cluster. The signals within each cluster are shown (in grey) along with the cluster average (in bold) in Figure 24. The cluster estimated average spectra appear in Figure 25.

Note in Figure 24 that Cluster 1 is characterised by a gradual increase in period throughout the experiment and gradual amplitude dampening with time. The amplitude dampening can also clearly be seen in the decreasing coefficients in resolution levels 2–4 (and particularly in level 2) in the average spectrum of Cluster 1 in Figure 25. The gradual increase in period can be seen as the activity in the spectrum begins in resolution level 4 and moves into levels 3 and 2 with time.

Cluster 2 is characterised by low frequency behaviour throughout the experiment (a longer period) and marked amplitude dampening with time, resulting in a rhythmicity loss. Indeed, this behaviour is also identified by the average spectrum in Figure 25. The increased period is reflected in the large coefficients at coarsest levels and the increased period of the wavelet coefficients in resolution levels 2 and 3. The dampening is apparent as the magnitude of the spectral coefficients decreases as time progresses.

Furthermore, note the nonstationary behaviour that characterises both clusters (changing

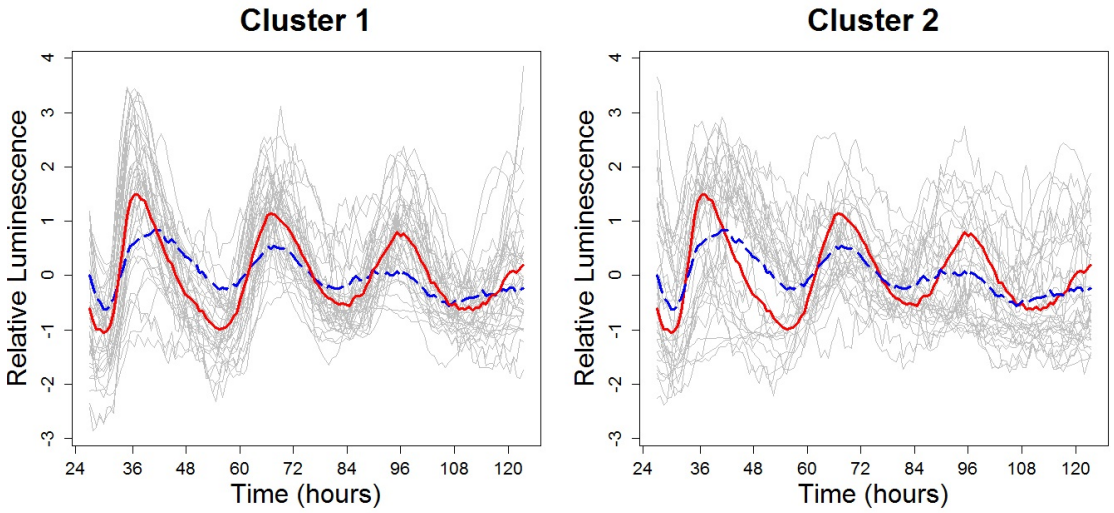

Figure 24: Results of clustering the copper dataset into two clusters using the proposed LSW-PCA method. The individual signals (grey) along with the cluster average in: red for Cluster 1 and (dashed) blue for Cluster 2.

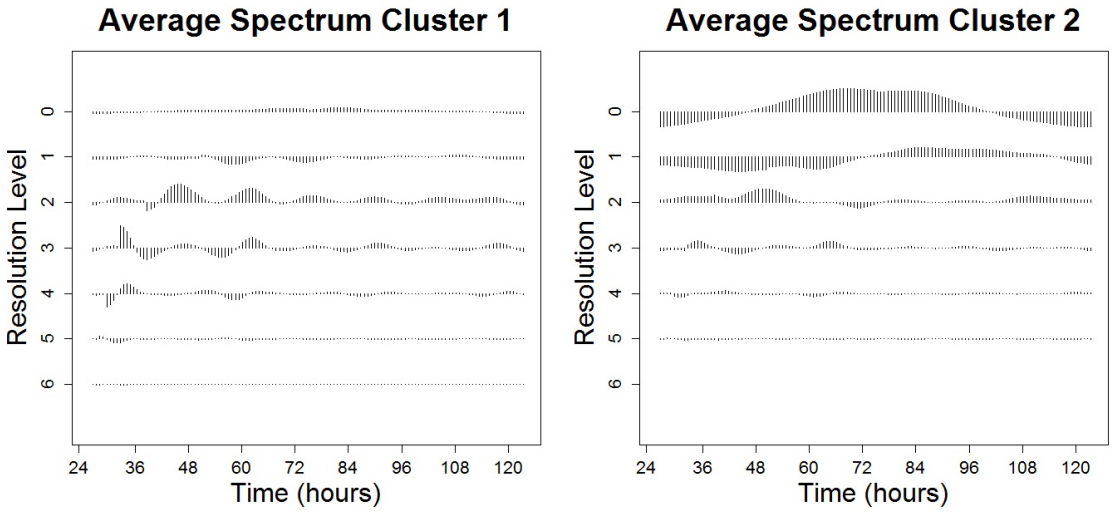

Figure 25: Cluster average estimated spectra on the copper dataset using the proposed LSW-PCA method.

Number of plants	Hoagland's	100 μM	150 μM	200 μM	Total
Cluster 1	**13**	2	3	0	18
Cluster 2	6	**14**	0	0	20
Cluster 3	5	8	**21**	**24**	58
Total	24	24	24	24	96

Table 6: Results of clustering the (normalised, truncated) cerium dataset into three groups using the proposed LSW-PCA method. The modal cluster for each concentration is highlighted in bold.

period and amplitude). The presence of these nonstationary characteristics supports our assertion that the existing methods (which assume stationarity) are inappropriate for such datasets and cannot capture this behaviour. In conclusion, our LSW-PCA clustering method has detected and characterised the interesting effects excess levels of copper have on the circadian clock, that were not detectable in the original analysis of the copper dataset (Perea-García et al., 2016a).

2.6.2 Novel Circadian Plant Data

We now return to the circadian data that motivated this work and apply our proposed LSW-PCA clustering method to analyse the novel cerium data. As the LSW model is underpinned by wavelets and requires the data to be of dyadic length ($T = 2^J$), in our analysis we chose a segment of length $T = 128$ out of the original dataset. This truncation was decided upon after consultation with the experimental scientists, as in Section 2.6.1. For each plant we estimated the wavelet spectrum by means of the corrected wavelet periodogram estimate (with the same setting as described in the simulation study in Section 2.4). For ease of interpretation we retained two principal components to cluster the data (see Section 2.3.4.2). This was justified by examining the screeplot (see Figure 32 in Appendix 2.8). The proposed LSW-PCA clustering method yielded the results detailed in Table 6.

The methods outlined in Section 2.3.4.3 were used to determine the optimal number of clusters. All methods indicated that we should cluster the data into three groups. This was supported by experimental scientists who confirmed that it would be useful to cluster the data into three groups: 'No Change' and two distinct departures from this group. In particular, we hoped to differentiate between and characterise the effects of lower and higher concentrations of cerium. This is because recent research has shown that certain compounds can produce very different effects on plant growth at low and high doses (Yang et al., 2016). Furthermore, this phenomenon seems to be present in our circadian dataset. On examining Figure 17, it appears that plants subjected to higher concentrations of cerium (150μM and 200μM) seem to exhibit similar behaviour, while the control group and concentration 100μM seem to display average behaviour which is distinct from each other and from the higher concentrations.

Discussion of findings. On examining Table 6, we can see that this method has effectively clustered the behaviour of the data into the following three groups:

1. Cluster 1: contains mostly plants in the Control dataset (Hoagland's), and very few plants subjected to lower-medium concentrations of ammonium cerium nitrate (100μM and 150μM)– conceptualised as essentially 'Control';

2. Cluster 2: contains mostly plants with lower concentration of ammonium cerium nitrate (100μM) and a few plants from the Control dataset– conceptualised as 'Low concentration';

3. Cluster 3: identifies similar behaviour to plants mostly exposed to medium-high concentrations (150μM, 200μM), but interestingly also contains a few plants from the Control and 100μM concentration.

These results are in agreement with Figure 17 (which we recall provided visual evidence that the plants subjected to higher concentrations of cerium exhibit similar behaviour, while the control group and concentration 100μM seem to display distinct behaviour). Therefore, this analysis has enabled us to achieve our first goal: to differentiate between the effects of lower and higher concentrations of cerium. Of interest to circadian biologists, however, is the presence of control and low concentration treated plants in the group associated mostly with higher concentrations. This highlights individual-level variability in plant response to stimuli, despite their sharing identical genetic characteristics (Doyle et al., 2002).

Our proposed method also allows us to characterise these groups, both in terms of first and second-order plant behaviour. The signals within each clustered group are shown (in grey) along with the cluster average (in bold) in Figure 26, while the cluster estimated average spectra appear in Figure 27.

On examining Figure 26, notice the different behaviour of Cluster 3 from the other clusters– this is characterised by high frequency behaviour throughout the experiment and a marked amplitude dampening with time, resulting in a rhythmicity loss. Indeed, this behaviour is also identified by the average spectrum in Figure 27. The high frequency behaviour is reflected in the large coefficients in resolution level 6. The dampening is apparent as the magnitude of the spectral coefficients decreases as time progresses (particularly in resolution level 2).

In contrast, Clusters 1 and 2 (approximately corresponding to the control and low concentration groups respectively) display more similar, rhythmic behaviour. On examining Figure 26, the rhythmic periods of the cluster averages seem approximately equal. However, there are also clear differences between the two groups. Firstly, there is a difference in the amplitudes of the two cluster averages. Cluster 1 has a larger peak at approximately $t = 36$ and an even larger peak at $t = 120$. This can be seen in the large coefficients around these time points in resolution levels 1-4 in the average spectrum of Cluster 1. Alternatively, Cluster 2 seems to have a very large peak at $t = 36$ followed by a distinct reduction in the amplitude of the other peaks. This can also be seen in the large coefficients in resolution levels 2-4 in the average spectrum of Cluster 2 in Figure 27.

The spectral content extracted in the first two principal components can be found in Figure 28. The projection of the original plant signals onto the principal component plane appears in Figure 29, by cluster and group membership. These indicate that the first principal component represents the departure from the control group after exposure to ammonium cerium nitrate, with larger values indicating a distinct change. The second principal component appears to reflect the spectral behaviour of the 100μM group, in particular the larger amplitude at around $t = 36$. Finally, note that Figure 29 shows that Cluster 1 has the biggest spread, while Cluster 3 is the most tightly packed. This supports biological expectations that plants behave in a similar manner when 'under stress' (Hanano et al., 2006).

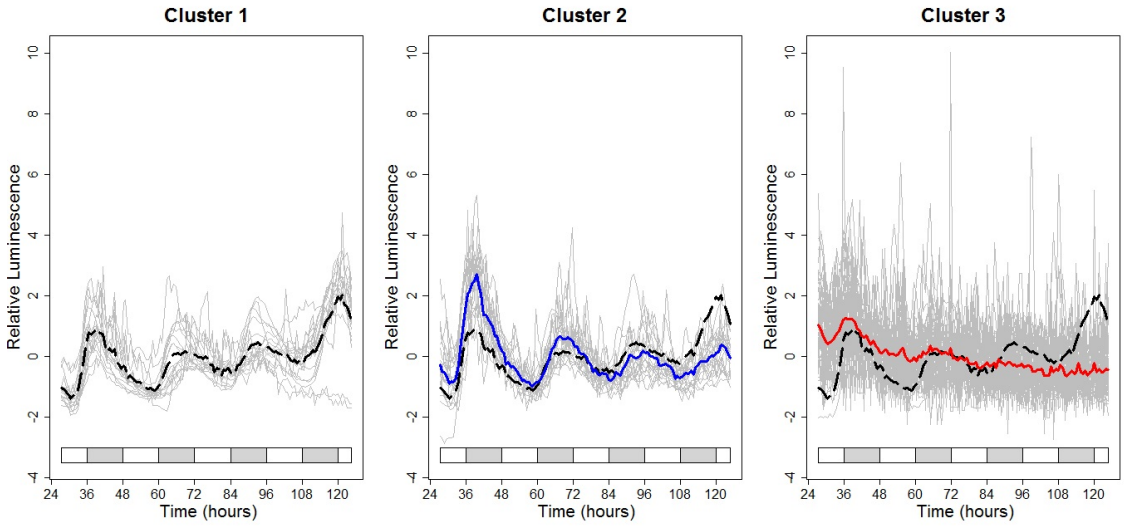

Figure 26: The results of clustering the cerium dataset into three groups using the proposed LSW-PCA method. The individual signals (grey) along with the cluster average in: (dashed) black for Cluster 1; blue for Cluster 2 and red for Cluster 3. The average of Cluster 1 (conceptualised as essentially 'Control') is shown (in dashed black) in all plots for reference.

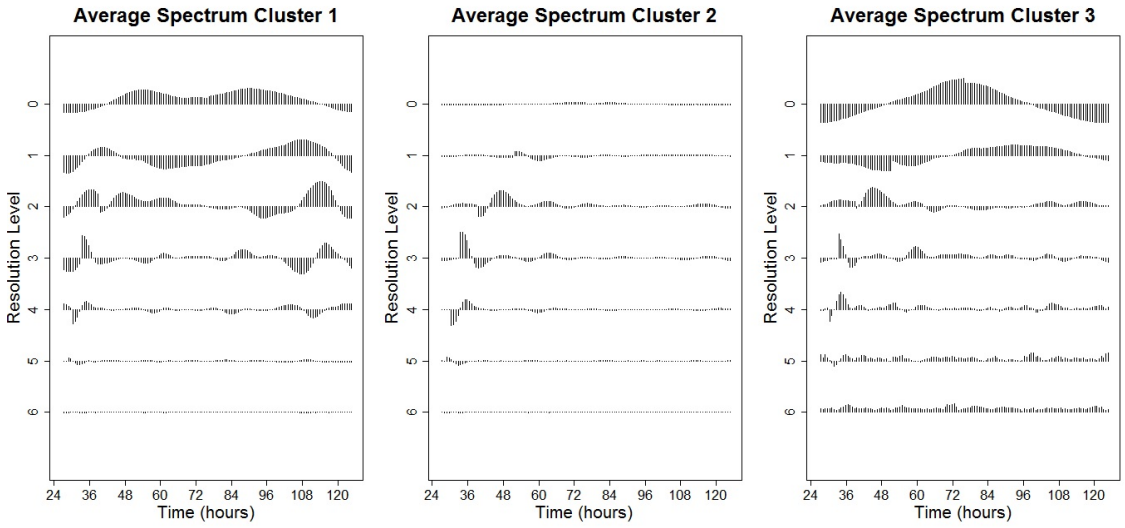

Figure 27: Cluster average estimated spectra on the cerium dataset using the proposed LSW-PCA method. Cluster 1 approximately corresponds to the 'Control' group; Cluster 2 depicts 'Low concentration' behaviour (100 μM) and Cluster 3 the 'Higher concentration' (150 μM and 200 μM).

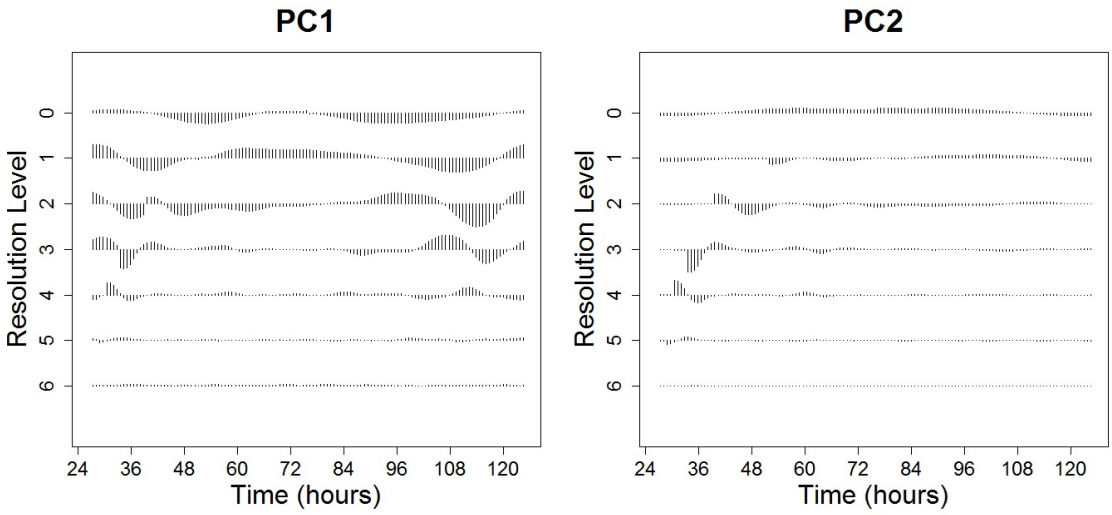

Figure 28: First two principal components obtained using the proposed LSW-PCA method on the cerium dataset.

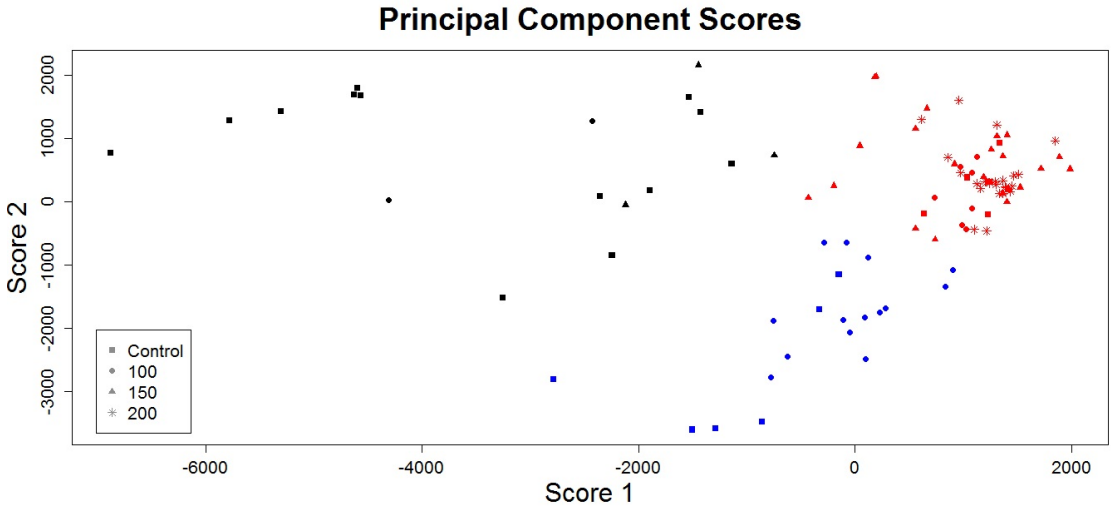

Figure 29: The cerium dataset projected onto the first two principal components obtained from the LSW-PCA clustering method. The colours represent the clusters: black for Cluster 1, blue for Cluster 2 and red for Cluster 3. The symbols represent the plant treatments.

2.7 Conclusions and Further Work

In this chapter we have developed a new procedure for clustering inherently nonstationary rhythmic data by modelling them as locally stationary wavelet processes and exploiting their local time-scale spectral properties by means of a functional principal component analysis. Our method combines the advantages of a wavelet analysis with the benefits of rigorous stochastic nonstationary time series modelling and has desirable properties, such as low sensitivity to the choice of distance measure and number of principal components to retain. These characteristics show the method's suitability in organising and understanding multiple nonstationary time series, such as the gene expression levels in our novel circadian dataset. When compared to competitor (non-model based) methods, we found that our methodology brought clear gains for simulated data (Table 4). Furthermore, when compared to existing methods (which assume stationarity), the LSW-PCA clustering method also displayed advantages for real data (Table 6).

The proposed model-based clusterings can be used to answer questions such as, 'What other concentrations of this compound produce similar effects in plants?' Our approach can also produce visualisations helpful in answering questions such as, 'What characterises the different types of reactions present in this dataset?' Such answers have important implications for understanding the mechanism of the plant's circadian clock and also environmental implications associated with soil pollution.

Also note that our proposed algorithm is not restricted to the datasets analysed in this chapter; it can be applied to other circadian datasets, as well as to data originating in other fields. The flexibility and computational efficiency of our approach allows more global analyses of plant behaviour to be undertaken which would not be possible within the stationary statistical constraints underlying traditional methods of period estimation. For example, the roles of a wide range of soil pollutants can be assessed within a single statistical framework. By extending this statistical methodology and empirical protocol to include exposure to other compounds, one could address the question, 'Which other elements in the periodic table, and at which concentrations, produce similar kinds of reactions in plants?' We can also extend the dataset to include plants with deficiencies of elements other than copper. These studies would also enable deeper understanding of the circadian clock mechanisms and its adaptations to change (Perea-García et al., 2016a).

The wavelet system gives a representation for nonstationary time series under which we estimate the wavelet spectrum and subsequently cluster the data. Ideally, we would envisage the use of the wavelet that is best suited to modelling and discriminating between the particular dataset. In simulations we found our method to be fairly robust to the wavelet choice. However, we found that Haar wavelets seemed to achieve superior results to other wavelets. This result supports the intuition that, as the scope of our work is to devise a clustering procedure that can locally identify dissimilarities (and hence discriminate) between pairs of spectra, the short support overlaps of Haar wavelets counterbalance their otherwise reduced capacity of representing smooth signals (such as certain circadian time series). However, an area of further work would be to derive a procedure for determining which wavelet system to adopt for any given dataset.

We are aware of the propensity of the recording equipment (see Appendix 2.9) to break down, resulting in gaps in the data. Such failures in hardware are an objective reality of empiri-

cal work in the life sciences, and another area of future work is to adapt current methods under the presence of missingness, or 'gappy' data, often arising in experimental data. This estimate could then be used as a classification signature or within our clustering procedure.

2.8 Appendix: Supplementary Figures

In this section we offer visual evidence to support claims in Sections 2.1, 2.2 and 2.6. All figures (30, 31 and 32) are referred to in context as part of the main body of the chapter.

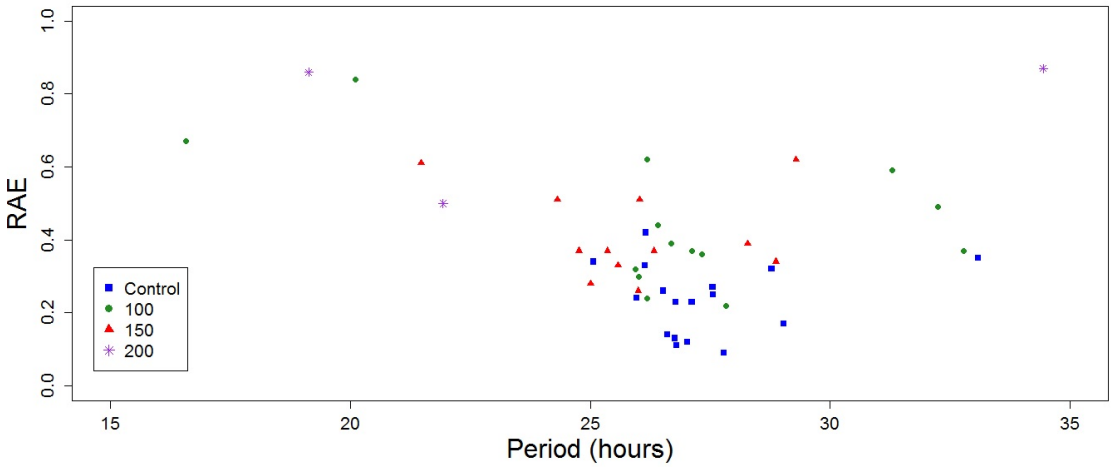

Figure 30: Summary of the BRASS analysis of the circadian plant signals in response to differing quantities of ammonium cerium nitrate, represented by plots of period estimates plotted against the respective relative amplitude errors (RAE). The colours and symbols represent the plant treatment groups: blue squares for the Control Group; green circles for Group 1 (100μM); red triangles for Group 2 (150μM) and purple stars for Group 3 (200μM).

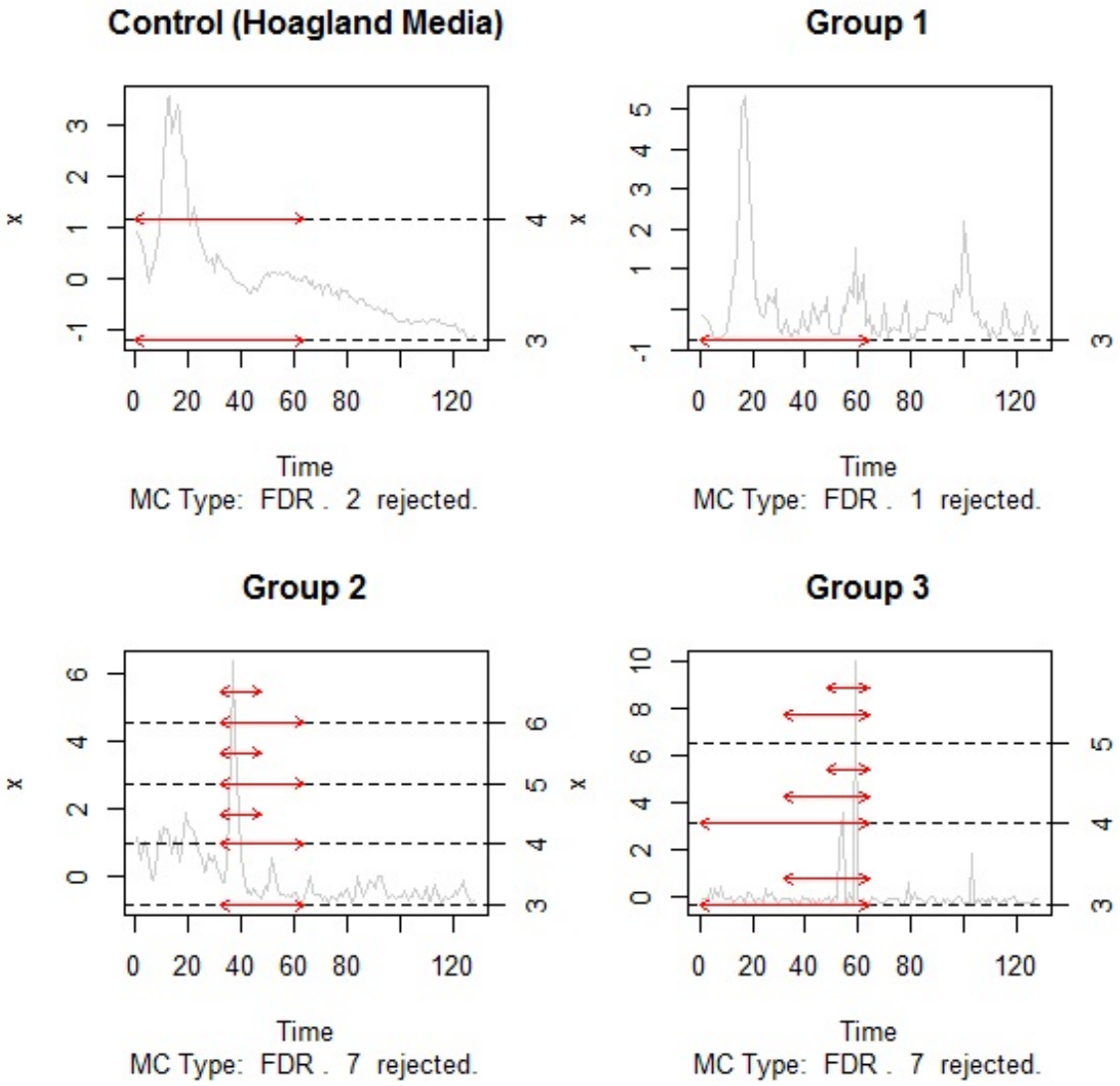

Figure 31: Plots of the estimated locations of the nonstationarities in the circadian plant signals in response to differing quantities of ammonium cerium nitrate, using the wavelet spectrum test (Nason, 2013), implemented in the `locits` package in R which is available on CRAN. A time series for each of the four groups is shown as an example– Group 1, a time series from the 100μM group; Group 2, a time series from the 150μM group; Group 3, a time series from the 200μM group.

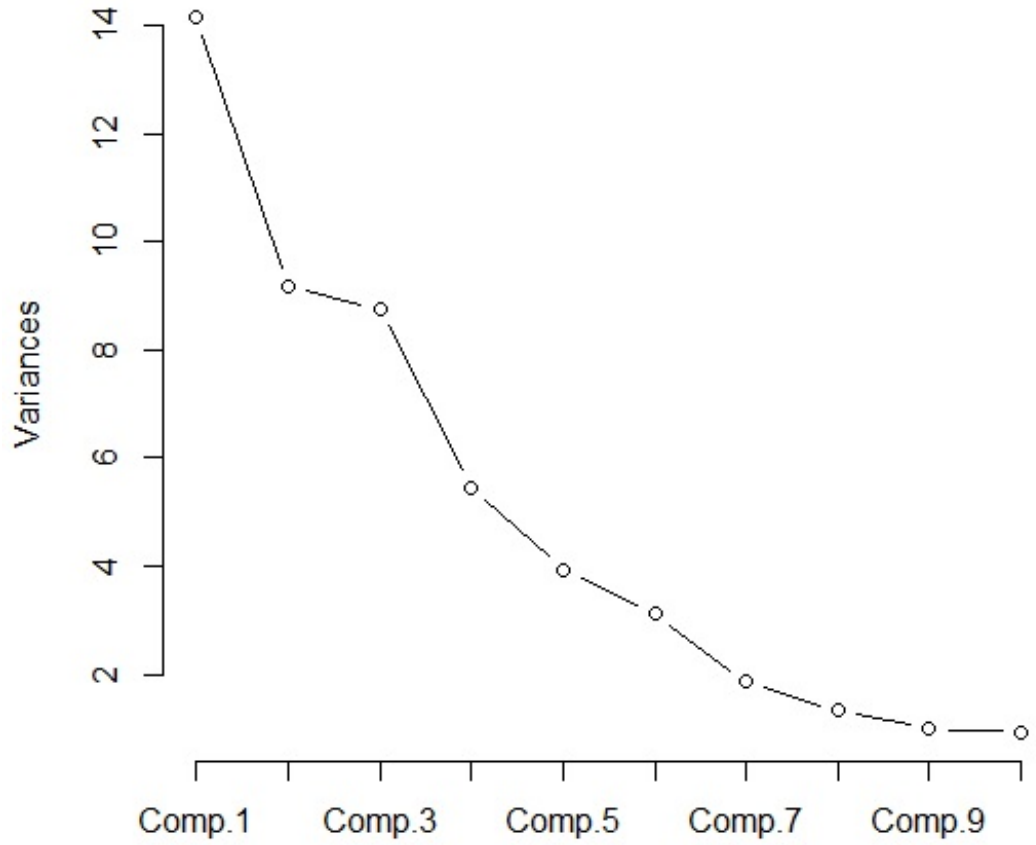

Figure 32: The screeplot used to inform the selection of the number of principal components to retain for the cerium dataset. Note 2 or 3 components could potentially be used, but for ease of interpretation (see Section 2.3.4.2), 2 were selected for clustering.

2.9 Appendix: Experimental Details: Novel Circadian Plant Data

In this section we outline the experimental details that led to the novel circadian plant rhythms under analysis (Section 2.2.1).

To obtain this dataset, the Davis Lab (Biology, University of York) used a firefly luciferase reporter system. This method uses a fusion of the gene of interest to luciferase. In this experiment, the gene of interest was 'cold and circadian regulated and RNA binding 2', known as CCR2 (further details of *CCR2:LUC* can be found in Doyle et al. (2002)). When CCR2 is expressed, luciferase is produced, causing the plant to produce quantifiable levels of light. This bioluminescence was measured using a TopCount NXT scintillation counter (Perkin Elmer), allowing relative gene expression of CCR2 to be quantified *in vivo* (Plautz et al., 1997; Southern and Millar, 2005; Perea-García et al., 2016a). These experiments were carried out using the following methods: *Arabidopsis thaliana* seeds (Ws–*CCR2:LUC*) were surface sterilised and plated onto Hoagland's media containing 1% sucrose, 1.5% phyto agar (Hoagland et al., 1950). The seeds were stratified for 2 days at 4°C and transferred to growth chambers to entrain under 12:12 light/dark cycles at a constant temperature of 20°C. These conditions were chosen to simulate the 'normal' light/dark cycles of a day. Six-day-old seedlings were transferred to 96 well microtiter plates containing Hoagland's 1% sucrose, 1.5% agar (Southern and Millar, 2005) also containing supplemental $(NH_4)_2Ce(NO_3)_6$ (ammonium cerium nitrate) at a concentration of 100μM, 150μM or 200μM. The plants were then transferred to the TOPCount machine. Measurements were taken at intervals of approximately 45 minutes. Measurement began after the transition to 12 hours of darkness (known as subjective dusk) on the seventh day of the plants' life. Therefore, the plant experiences one 'normal' day in the TOPCount machine (known as entrainment). After this, the plant was exposed to constant light (known as an LL free-run) for approximately four days. In Figure 1, the shaded bars below the graph represent the light conditions the plants would experience during the 'normal' day. The plants are under constant light throughout the experiment, however, the grey bars indicate that they would be in darkness during a 'normal' 12 hour light/12 hour dark cycle.

Our dataset therefore consists of a total 96 plant signals (time series) recorded at 128 time points, with each of the control and groups 1–3 (each corresponding to a different concentration of ammonium cerium nitrate) containing 24 plants. In particular, the control group is grown in Hoagland's media (Hoagland et al., 1950) which contains essential nutrients required for plant growth and is not exposed to any additional levels of ammonium cerium nitrate. To examine the effects of cerium on the circadian clock, the other three groups, while also grown in the Hoagland's media, were additionally exposed to varying additional concentrations of ammonium cerium nitrate– 100μM for Group 1, 150μM for Group 2 and 200μM for Group 3.

2.10 Appendix: Results of Simulation Study Cases 1 and 4

In this section we report the findings of the simulation study associated with Cases 1 and 4 in Section 2.4.1. These consist of Tables 7, 8, 9 and 10, which further justify the distance and dimension reduction choices adopted for our proposed method.

Distance Measure	**SQ**	**WSQ**	**DT**	**D**
Case 1 Correctly Clustered (%)	76%	70%	69%	65%
Case 4 Correctly Clustered (%)	99%	99%	84%	80%

Table 7: Distance measure (Section 2.3.4.1) comparison for the proposed LSW-PCA method for Cases 1 and 4.

Dimension reduction method	**90% of total covariance**	**Screeplot**
SQ distance	73%	76%
WSQ distance	69%	70%
DT distance	54%	69%

Table 8: **Case 1:** Comparison for selection of principal components for proposed LSW-PCA clustering method. Percentages show correct clustering rates.

Dimension reduction method	**90% of total covariance**	**Screeplot**	**Always retain 2 PCs**
SQ distance	98%	99%	99%
WSQ distance	99%	99%	99%
DT distance	54%	84%	80%

Table 9: **Case 4:** Comparison for selection of principal components for proposed LSW-PCA clustering method. Percentages show correct clustering rates.

Input	**Original Signals**	**Wavelet Coefficients**	**Squared Wavelet Coefficients**	**Corrected Wavelet Periodogram**
Case 1	54%	54%	63%	69%
Case 4	54%	54%	53%	87%

Table 10: Wavelet information comparison for the proposed LSW-PCA method for Cases 1 and 4. Percentages show correct clustering rates.

2.11 Appendix: Experimental Details: Previously Published Circadian Data

In this section we outline the experimental details that led to the previously published copper dataset (Section 2.6.1).

This dataset (Perea-García et al., 2016a,b) was also obtained using a firefly luciferase reporter system as described in Appendix 2.9. Experimental Details: Novel Circadian Plant Data. However, this experiment uses a different gene of interest GIGANTEA (GI). Plants were grown on plates as described in Andrés-Colás et al. (2010), incubated on MS (Murashige and Skoog) medium (Murashige and Skoog, 1962) at half concentration (1/2 MS) [phytoagar 0.8% (w/v) plus 1% sucrose (w/v) in 0.5% MES (w/v)]. WS GI:LUC seedlings were grown under different copper regimes: 'Deficiency' (1/2 MS), 'Sufficiency' or 'Control' (1 μM CuSO4), and 'Excess' (10 μM CuSO4). 96 plants were grown in total, 32 under each copper regime. The plants were entrained for 7 days under 12:12 light-dark cycles at a constant temperature of 20°C. The plants were then exposed to constant light (LL free-run) for the remainder of the experiment. Bioluminescence was then measured every hour using the same TopCount NXT system as in Appendix 2.9.

The dataset analysed in Perea-García et al. (2016a,b) consists of a total 74 plant signals (time series) recorded at 151 time points. Plants with an average luminescence of 40 or below were excluded prior to analysis as luminescence values below this are considered background noise. Therefore, the 'Deficiency' group (1/2 MS) contains 19 plants; the 'Control' or 'Sufficiency' group (1 μM CuSO4) contains 26 plants and the 'Excess' group (10 μM CuSO4) contains 29 plants.

3 Wavelet Spectral Testing: Application to Nonstationary Circadian Rhythms

In this chapter we develop and test novel hypothesis testing procedures in the (wavelet) spectral domain, embedding replicate information when available. The proposed methodology is the result of joint work with M. I. Knight, J. W. Pitchford and S. J. Davis. The novel circadian datasets analysed in this thesis were obtained by R. Oakenfull and J. Munns from the Davis and Chawla Labs (Biology, University of York), respectively. Please see page 21 for further details of author contributions. This work has been submitted for publication.

3.1 Introduction and Motivation

The 'circadian clock' enhances survival by directing anticipatory changes in physiology synchronised with environmental fluctuations. When an organism is deprived of external time cues, its circadian rhythms typically persist qualitatively but may change in detail; the study of these changes can reveal the biochemical reactions underpinning the circadian clock and, at a larger scale, can provide valuable insight into the possible consequences of environmental and ecological challenges (McClung, 2006; Bujdoso and Davis, 2013).

In many scientific applications, available data consist of signals with known group memberships and scientists are interested in establishing whether these groups display statistically different behaviour. Our work is motivated by a general problem: biologists need reliable statistical tests to identify whether a particular experimental treatment has caused a significant change in the circadian rhythm. If the changes are limited to period and/or phase then existing Fourier-based theory may be adequate. However, when the changes to the circadian clock are less straightforward, for example involving nonstationarity or changes at multiple scales (Hargreaves et al., 2018), the application of these established methods may be conducive to misleading conclusions.

3.1.1 Motivating Datasets

The potential value of our approach is illustrated by three complementary examples encompassing: the effect of various salt stresses on plants; the identification of mutations inducing rapid rhythms and the response of nematode clocks to pharmacological treatment, as described in the following sections. The biological experimental details for each dataset appear in Appendix 3.7.

3.1.1.1 Lead Nitrate Dataset (Davis Lab, Biology, University of York)

This dataset (hereafter referred to as the 'Lead dataset') is from a broad investigation of whether plant circadian clocks are affected by industrial and agricultural pollutants (Foley et al., 2005; Senesil et al., 1998; Hargreaves et al., 2018; Nicholson et al., 2003). Specifically, this experiment asks whether lead affects the *Arabidopsis thaliana* circadian clock and, if so, when and how? Figure 33 displays the luminescence profiles for both untreated *A. thaliana* plants, as well as for those exposed to lead nitrate.

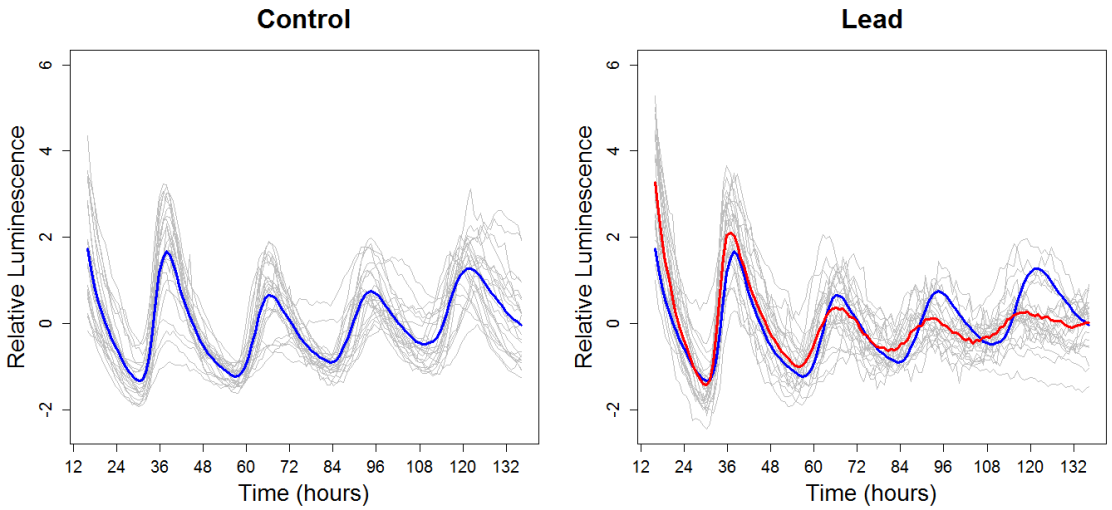

Figure 33: **Lead dataset:** Luminescence profiles over time for untreated *A. thaliana* plants (Control) and those exposed to lead nitrate (Lead). Left: Individuals in the control group (in grey) along with the group average (blue). Right: Individuals in the lead treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero.

3.1.1.2 Ultradian Dataset (Millar Lab, Biology, University of Edinburgh)

In order to understand the clock mechanism, a common approach is to mutate a gene and examine the resulting behaviour in response to a variety of stimuli. Figure 34 depicts the luminescence profiles recording plant response to light, for both the control and genetically mutated *A. thaliana* plants (Millar et al., 2015). Researchers are interested in establishing whether a specific genetic mutation induced high-frequency behaviour (known as 'ultradian rhythms') in the laboratory model plant *A. thaliana*.

3.1.1.3 Nematode Dataset (Chawla Lab, Biology, University of York)

The free-living nematode *Caenorhabditis elegans* is an animal widely used in neuroscience and genetics, but its circadian clock is still poorly understood. To increase understanding of the nematode clock, and potentially uncover rhythmicity not detected by conventional approaches, researchers applied a pharmacological treatment to *C. elegans*, based on evidence that it causes aberrant circadian rhythms in other established mammalian and insect circadian models (Kon et al., 2015; Dusik et al., 2014). Figure 35 depicts the luminescence profiles for both untreated and treated *C. elegans* and reveals apparently similar traces. In particular, the two groups seem to display similar average behaviour, but with differing intensities. Therefore, a useful research question is to ask whether these similar signals are the result of the two groups being underpinned by the same profile spectra, up to a scale-dependent additive constant.

3.1.1.4 Summary

On examining Figures 33 and 34, it is visually clear that changes in period and amplitude between the control and test groups occur in both datasets. Nevertheless, in each of our motivating examples, less easily quantified or subtle differences between these groups may also exist.

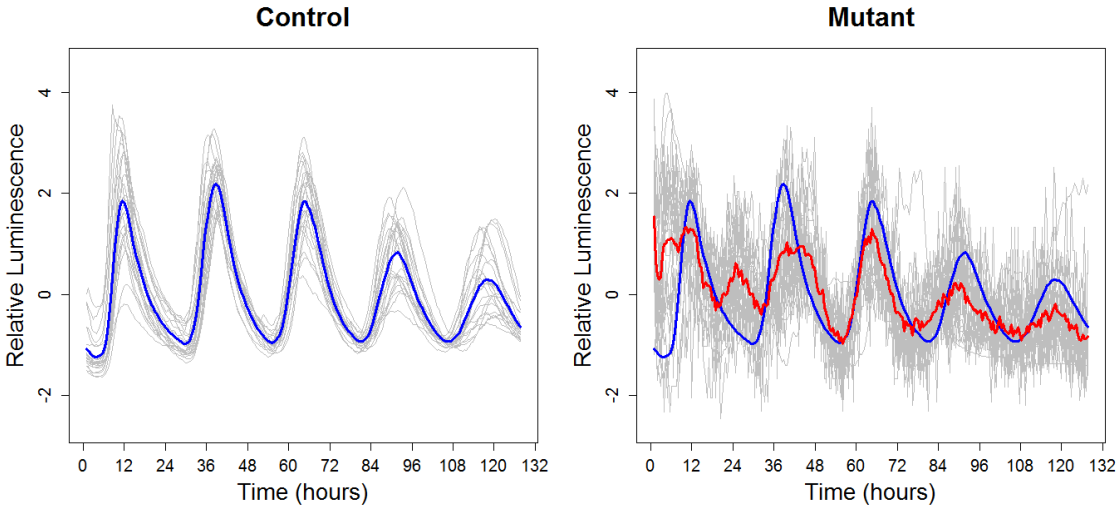

Figure 34: **Ultradian dataset:** Luminescence profiles over time for control and mutant *A. thaliana* plants. Left: Individuals in the control group (in grey) along with the group average (blue). Right: Individuals in the mutant group (in grey) along with the mutant group average (red) and the control group average (blue). Each time series has been standardised to have mean zero.

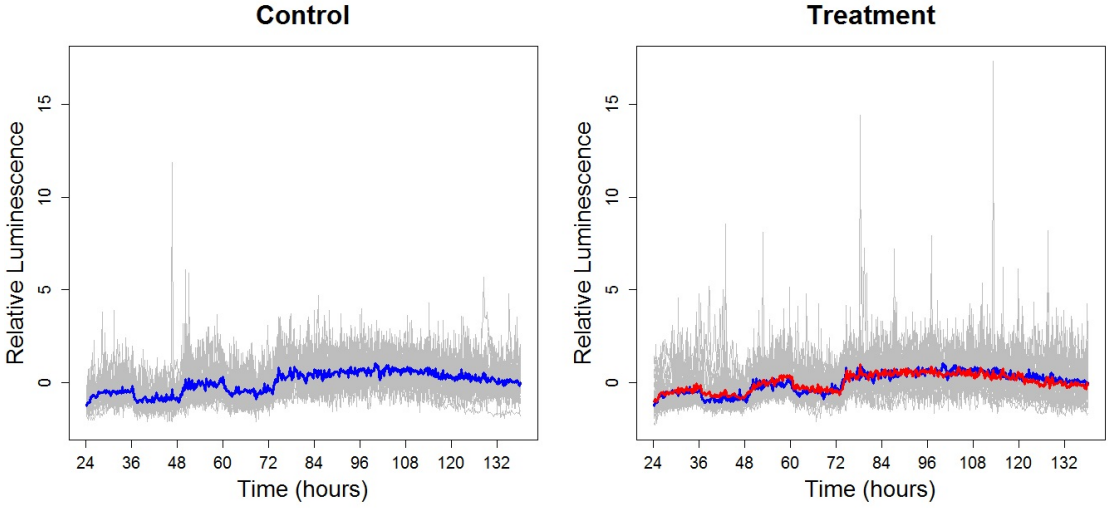

Figure 35: **Nematode dataset:** Luminescence profiles over time for untreated *C. elegans* (Control) and those subjected to a pharmacological treatment (Treatment). Left: Individuals in the control group (in grey) along with the group average (blue). Right: Individuals in the treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero.

3.1.2 Aims and Structure of this Chapter

Period estimation is central to the analysis of circadian data, with the current standard achieving this using Fourier analysis (Zielinski et al., 2014; Costa et al., 2011) via software packages, such as BRASS (Biological Rhythm Analysis Software System (Edwards et al., 2010)) or BioDare (Moore et al., 2014). The practitioner estimates the period of the control and treatment groups respectively, and then tests for statistically significant differences (see for example Perea-García et al. (2016a), Costa et al. (2011)). Crucially, in all of our motivating examples, such established Fourier-based tests found no significant difference between the groups (see Table 16 in Appendix 3.8), even though qualitative differences are easily noted (see Section 3.1.1.4).

One obvious limitation of this analysis is that the employed methodology does not typically evaluate the crucial underpinning assumption of data stationarity. In the context examined here, assuming stationarity can be inappropriate (Hargreaves et al., 2018; Leise et al., 2013), a feature shared by many biological systems (Zielinski et al., 2014). For our motivating example datasets, we investigated whether the individual time series are (second-order) stationary via hypothesis testing. We employed two tests for stationarity– a Fourier-based test (the Priestley-Subba Rao test (Priestley and Rao, 1969)) and a wavelet-based test (the wavelet–spectrum test (Nason, 2013)). The results (Table 2 in Appendix 3.8) show that, for each of our motivating example datasets, over 80% of the time series provided enough evidence to reject the null hypothesis of stationarity. This result suggests that the application of the current methodology (which assumes data stationarity) would be inappropriate for our motivating datasets and highlights the urgent need for more statistically advanced approaches.

The primary contribution of this work is the development of novel wavelet-based hypothesis tests that allow for circadian behaviour comparison while accounting for data nonstationarity. A substantial body of circadian literature advocates the use of wavelet (Price et al., 2008; Harang et al., 2012; Leise et al., 2013) and in particular spectral representations (Hargreaves et al., 2018) of circadian rhythms. This motivates our choice to formally compare circadian signals in the wavelet spectral domain by using their time-scale signature patterns, thus accounting for their proven nonstationary features.

This chapter is organised as follows. Section 3.2 reviews the theoretical wavelet-based framework we adopt for modelling nonstationary data and the relevant literature on hypothesis testing in the spectral domain. Our new hypothesis testing procedures are introduced in Section 3.3. Section 3.4 provides a comprehensive performance assessment of our new methods via simulation. Section 3.5 demonstrates the additional insight our techniques provide for the motivating circadian datasets and Section 3.6 concludes this work.

3.2 Overview: Nonstationary Processes and Hypothesis Testing in the Spectral Domain

3.2.1 Modelling Nonstationary Processes

In Section 1.4 we introduced a number of statistically rigorous approaches to modelling nonstationary time series. Motivated by literature advocating wavelets as analysis tools for circadian rhythms (Leise et al., 2013), we adopt the locally stationary wavelet (LSW) process model of Nason et al. (2000) with previously demonstrated utility for circadian analysis (Hargreaves

et al., 2018). Recall, an LSW process $\{X_{t;T}\}_{t=0}^{T-1}$, $T = 2^J \geq 1$ is represented as follows

$$X_{t,T} = \sum_{j=1}^{J} \sum_{k \in \mathbb{Z}} w_{j,k;T} \psi_{j,k}(t) \xi_{j,k}, \tag{77}$$

where $\{\xi_{j,k}\}$ is a random orthonormal increment sequence, $\{\psi_{j,k}(t) = \psi_{j,k-t}\}_{j,k}$ is a set of discrete non-decimated wavelets and $\{w_{j,k;T}\}$ is a set of amplitudes, each of which at a scale j and time k.

3.2.1.1 Practical Considerations

In this paper, we assume the innovations $\{\xi_{j,k}\}$ to be normally distributed, resulting in modelling the data $\{X_{t,T}\}$ as a Gaussian LSW process. The normality assumption is typically employed for the (Fourier) circadian testing methodology (Perea-García et al., 2016a). This assumption is also commonly made in time series analysis in general and in LSW modelling in particular (e.g. Oh et al. (2003), Van Bellegem and von Sachs (2008) and Nason and Stevens (2015)), with Nason (2013) arguing for its non-limiting character in this context. In Appendix 3.9 we show this assumption is tenable for our circadian datasets.

The properties of the random increment sequence $\{\xi_{j,k}\}$ ensure that $\{X_{t,T}\}$ is a zero-mean process. In practice, for a process with non-zero mean, it is customary to re-centre it around zero (Nason, 2010) and this is our approach here, as the quantity of our primary interest is the process spectral signature.

As is typical for wavelet representations, the data is often required to be of dyadic length, $T = 2^J$. In many practical applications, this is not realistic and there are a number of approaches to address this situation (see e.g. Ogden (1997)). Our approach is to analyse a (dyadic length) segment of the data, with the truncation decided upon careful consultation with the experimental scientists in order to ensure the time-frame of interest is represented.

3.2.1.2 The Evolutionary Wavelet Spectrum

In Section 1.4.2, we formally defined the evolutionary wavelet spectrum (EWS) as

$$S_j(z) \coloneqq |W_j(z)|^2, \tag{78}$$

at each scale $j \in \overline{1, J}$ and rescaled time $z = k/T \in (0, 1)$. We also defined the raw wavelet periodogram as

$$I_{j,k;T} \coloneqq |d_{j,k;T}|^2, \tag{79}$$

where $d_{j,k;T} = \sum_{t=0}^{T} X_{t,T} \psi_{j,k}(t)$ are the empirical nondecimated wavelet coefficients. In the remainder of this chapter we drop the explicit dependence on T for the wavelet coefficients and the periodogram.

The raw wavelet periodogram is an asymptotically unbiased estimator of the quantity $\beta_j(z)$ introduced by Fryzlewicz and Nason (2006) and defined as

$$\beta_j(z) \coloneqq \sum_{i=1}^{J} A_{i,j} S_i(z) = (AS)_j(z), \tag{80}$$

where $A = (A_{i,j})_{i,j=1}^{J} = (\sum_{\tau} \Psi_i(\tau) \Psi_j(\tau))_{i,j=1}^{J}$ is the autocorrelation wavelet inner product ma-

trix, with $\Psi_j(\tau) = \sum_k \psi_{j,k}(0)\psi_{j,k}(\tau)$ the autocorrelation wavelet (Nason et al., 2000). In other words, the expectation of the raw wavelet periodogram (computed at rescaled time z) converges pointwise to a linear combination of wavelet spectra at location z (Fryzlewicz and Nason, 2006). Recall (Section 1.4.2) that an asymptotically unbiased estimator of the EWS is the empirical wavelet spectrum (or corrected periodogram), defined as

$$\mathbf{L}(z) := A^{-1}\mathbf{I}(z), \tag{81}$$

for all $z \in (0,1)$, where $\mathbf{I}(z) := (I_{j,\lfloor zT \rfloor})_{j=1}^{J}$ is the raw wavelet periodogram vector.

The quantity $\beta_j(z)$ (equation (80)) is often easier to work with theoretically than the spectrum (see Nason (2013) and Sections 3.3.2 and 3.3.3). One immediate advantage of working with $\beta_j(k/T)$ as opposed to the spectrum $S_j(k/T)$ is the direct access to the distribution of its corresponding estimator $I_{j,k;T}$, the raw wavelet periodogram, as opposed to the distribution of the corrected periodogram $L_{j,k;T}$, needed to asymptotically estimate the spectrum. In particular, the empirical wavelet spectrum is a collection of random variables that are not independent, nor is their (joint or marginal) distribution easy to determine.

As the individual raw periodogram ordinates within each scale are correlated, Fryzlewicz and Nason (2006) model the raw wavelet periodogram as

$$I_{j,k} \sim \beta_j(z) Z_{j,k}^2,$$

where $z = k/T$ and $Z_{j,k}^2 \sim \chi_1^2$, for $j \in \mathbb{N}$, $k = 0, \ldots, 2^J - 1 = T - 1$. A way to 'correct' these undesirable features is to employ a transform that brings the raw periodogram ordinates closer to Gaussianity and decorrelates within each scale. In Section 3.3.2, we adopt the Haar-Fisz transform (denoted $\mathscr{F}$), introduced (for spectral estimation) by Fryzlewicz and Nason (2006) and apply it separately to each scale $j = 1, \ldots, J$ of the raw wavelet periodogram (see Appendix 3.10 for details), denoted $\mathscr{H}_{j,k;T} := \mathscr{F} I_{j,k;T}$. Proposition 6.1 in Fryzlewicz and Nason (2006) then suggests a potential model

$$\mathscr{H}_{j,k} \sim N(\mathscr{B}_j(z), \sigma_j^2),$$

where $\mathscr{B}_j(z) = \mathscr{F}\beta_j(z)$ with $z = k/T$ and $\mathscr{F} Z_{j,k}^2 \sim N(0, \sigma_j^2)$ and again dropping the explicit dependence on T. This model, viewed as a nonparametric additive regression model, was also employed by Nason and Stevens (2015) in the context of Bayesian spectral estimation, where its viability was demonstrated.

3.2.2 Existing Spectral Domain Hypothesis Testing

Assuming that the available data consists of multiple nonstationary time series with known group memberships, to the authors' knowledge no hypothesis tests exist to determine whether two groups are significantly different in terms of their associated (evolutionary) wavelet spectra. Wavelet spectral comparison is closest framed as a (consistent) classification method by Fryzlewicz and Ombao (2009), further improved by Krzemieniewska et al. (2014). Spectral comparison, framed as testing for spectral constancy, also appears in connection with testing for time series stationarity and white noise testing. In the Fourier domain, Priestley and Rao (1969) determined (as a hypothesis test) whether the spectrum is time-varying and, hence, whether the process is nonstationary. von Sachs and Neumann (2000) introduced the principle

of assessing the constancy of the time-varying Fourier spectrum by examining its Haar wavelet coefficients across time. In the wavelet domain, Nason (2013) developed a test for second-order stationarity which examines the constancy of a wavelet spectrum by also examining its Haar wavelet coefficients. A similar approach is adopted by Nason and Savchev (2014) in the development of white noise tests.

The problem of testing that involves curves is often posed in time series literature as a functional regression problem defined using a functional response and categorical predictors (functional ANOVA; see the monograph of Ramsay and Silverman (2005) for its introduction and the review of Morris (2015) for developments in the field). Functional regression problems are often treated by projection in the Fourier or wavelet domain, where the spectral time series representations become subject to modelling. Shumway (1988) compares groups of curves (with stationary stochastic errors) by testing whether the mean curves have the same Fourier spectrum at each given frequency. Fan and Lin (1998) developed this method by applying the adaptive Neyman test to the (Fourier or wavelet) transformed difference vector (the difference between the two group-average time series). Vidakovic (2001) introduces a wavelet-based functional data analysis, with McKay et al. (2012) developing this as an approach for comparing neurophysiological signals that are functions of time. This approach was also subsequently adopted by Atkinson et al. (2017) to develop model validation using a test statistic based on thresholded wavelet coefficients. Tavakoli and Panaretos (2016) compare pairs of stationary functional time series by developing t-tests for the equality of their (Fourier) spectral density operators. However, these approaches fail to account for potential nonstationarity in the data. This is mitigated by Guo et al. (2003), who propose a smoothing-spline ANOVA on the logarithm of the Fourier spectrum of a locally stationary process that is specifically designed to discriminate between models that contain a linear trend, modulation, time and frequency interaction terms, thus yielding global model comparisons, rather than time- and frequency- specific ones. The closest methodology for spectral comparison while allowing for a localised representation comes from Martinez et al. (2013) who identify regional differences in (the Fourier spectrograms of) bat mating chirps. The statistical modelling of windowed Fourier spectrograms as an image was first proposed by Holan et al. (2010) in a study that aimed to classify animal communication signals. Martinez et al. (2013) apply the higher-dimension functional mixed model of Morris et al. (2011) and use a Bayesian approach to fit a model that incorporates localised chirp Fourier spectrograms as the functional response and categorical regressors that identify bat location (fixed-effects) and independent bat (random)-effects. The observed data is modelled in a (projected) wavelet-domain with several distributional assumptions in place, e.g. data Gaussianity, spike Gaussian-slab prior distributions for the wavelet coefficients. However, while their windowed Fourier spectrogram does offer a time-frequency representation of the data, thus potentially capturing nonstationarity, it is sensitive to the choice of kernel and crucially of window-width (Martinez et al., 2013). In the context of clustering circadian plant rhythms, Hargreaves et al. (2018) demonstrated the superiority of a principled model-based spectral estimator that, in the spirit of Holan et al. (2010), was also used as an image in subsequent modelling. Additionally, we note that our study aims to identify not only (i) time-scale (frequency) group differences (conceptually a task close to Martinez et al. (2013)), but also (ii) to detect global scale-level differences (while still allowing for a development that incorporates potential nonstationarity) and (iii) to identify similar patterns within each scale, rather than

exact differences (the reader will find precise details in the next section).

3.3 Proposed Spectral Domain Hypothesis Tests

Aligned to our motivating examples, the key goals of our work are to develop novel hypothesis tests, each capable of detecting one of three specific types of spectral differences between two groups and to identify the scales and times (e.g. Lead and Nematode datasets– Sections 3.1.1.1 and 3.1.1.3) or scales only (e.g. Ultradian dataset– Section 3.1.1.2) at which these difference arise, as appropriate.

Formally, we model the observed nonstationary circadian rhythms using the LSW framework of Nason et al. (2000) (see Section 1.4.2 for details). Denote each individual profile by $\{X_{t,T}^{(i),r_i}\}_{t=0}^{T-1}$ with $i = 1, 2$ corresponding to one of two groups (e.g. control/ treatment) and potential replicates $r_i = 1, \ldots, N_i$ (i.e. N_i circadian traces in the ith group). Note that when $N_i = 1$ we drop the r_i index for simplicity. Assume the signals in group i are underpinned by a common wavelet spectrum and denote this by $S_j^{(i)}(t/T)$ for each group $i = 1, 2$ at scales $j \in \overline{1, J}$ ($J = \log_2 T$) and rescaled times $z = t/T \in (0, 1)$.

3.3.1 Lead Dataset: Hypothesis Testing for Spectral Equality ('WST' and 'FT')

Put simply, our soil pollutant example focussed on detecting whether the two plant groups, 'Control' and 'Lead', display significant differences in the evolution of their spectral structures, and if so, the particular scales and times at which such differences occur. Mathematically we formalise our hypotheses as

$$H_0 : S_j^{(1)}(z) = S_j^{(2)}(z), \quad \forall\, j, z \tag{82}$$

versus the alternative $H_A : S_{j^*}^{(1)}(z^*) \neq S_{j^*}^{(2)}(z^*)$ for some scale j^* and rescaled time z^*. In the time domain, we visually note that differences in the circadian rhythms of the two groups appear towards the end of the experiment (see Figure 33).

3.3.1.1 A Naive Wavelet Spectrum Test ('WST')

Since in reality we do not know the group spectrum $S_j^{(i)}(z)$, we replace it with a well-behaved estimator, denoted $\hat{S}_j^{(i)}(z)$. Assuming independent replicates are available for each group, we use the group ($i = 1, 2$) averaged spectral estimators

$$\hat{S}_j^{(i)}(k/T) = \frac{1}{N_i} \sum_{r_i=1}^{N_i} L_j^{(i),r_i}(k/T), \tag{83}$$

where $L_j^{(i),r_i}(k/T)$ is the empirical wavelet spectrum of the r_ith series in group i at scale j and time k. Assuming independence across the replicates and a Gaussian distribution for the spectral estimates, because the LSW theory constructs asymptotically unbiased spectral estimators, it follows that under the null hypothesis $\hat{S}_j^{(1)}(k/T) - \hat{S}_j^{(2)}(k/T)$ has an asymptotically normal distribution with mean zero. Hence, should our spectral estimators satisfy the classical assumptions for a t-test (which in our context amount to independence of the spectral estimates across replicates and a Gaussian distribution), we propose a naive *wavelet spectrum test* (WST), cen-

tred on a test statistic of the form

$$T_{j,k} = \frac{\hat{S}_j^{(1)}(k/T) - \hat{S}_j^{(2)}(k/T)}{\left((\hat{\sigma}_{j,k}^{(1)})^2/N_1 + (\hat{\sigma}_{j,k}^{(2)})^2/N_2\right)^{1/2}} \sim t_{df} \text{ under the null hypothesis,} \tag{84}$$

where $(\hat{\sigma}_{j,k}^{(i)})^2$ is an estimate of the variance of $\hat{S}_j^{(i)}(k/T)$ for $i = 1,2$ across the N_i observations in group i, obtained using the standard sum–of–squares sample variance formula (as in Krzemieniewska et al. (2014)). Under the null hypothesis of spectral equality, $T_{j,k}$ (asymptotically) follows a t-distribution with the number of degrees of freedom (df) directly related to the variance estimation procedure we employ. Each test statistic is then compared with a critical value derived from the t-distribution in the usual way.

When the variance of $\hat{S}_j^{(i)}(k/T)$ is unknown but common to both $i = 1,2$ groups (denoted $(\sigma_{j,k})^2 := (\sigma_{j,k}^{(1)})^2 = (\sigma_{j,k}^{(2)})^2$), it can be estimated using the pooled estimator:

$$\hat{\sigma}_{j,k}^2 = \frac{(N_1-1)(\hat{\sigma}_{j,k}^{(1)})^2 + (N_2-1)(\hat{\sigma}_{j,k}^{(2)})^2}{N_1+N_2-2}, \tag{85}$$

replacing $(\hat{\sigma}_{j,k}^{(1)})^2$ and $(\hat{\sigma}_{j,k}^{(2)})^2$ in equation (84). The number of degrees of freedom in the t-distribution of the test statistic is then $df = N_1 + N_2 - 2$.

If there is no reason to believe the group variances are equal, then use a t-distribution with degrees of freedom

$$df = \frac{\left((\hat{\sigma}_{j,k}^{(1)})^2/N_1 + (\hat{\sigma}_{j,k}^{(2)})^2/N_2\right)^2}{\frac{\left((\hat{\sigma}_{j,k}^{(1)})^2/N_1\right)^2}{N_1-1} + \frac{\left((\hat{\sigma}_{j,k}^{(2)})^2/N_2\right)^2}{N_2-1}}.$$

However, the test statistic does not exactly follow the t-distribution, since two standard deviations are estimated in the statistic. Conservative critical values may also be obtained by using the t-distribution with N degrees of freedom, where N represents the smaller of N_1 and N_2 (Moore, 2007).

Discussion. As we wish to test many hypotheses of the type $H_0 : \beta_j^{(1)}(k/T) = \beta_j^{(2)}(k/T)$ for several values of j and k, we are in the field of multiple-hypothesis testing. For all tests we develop, we use Bonferroni correction and, for a less conservative approach, the false discovery rate (FDR) procedure introduced by Benjamini and Hochberg (1995). Our simulations in Section 3.4 show that both these methods work well. However, of course the tests themselves are related to one another, but just as in Nason (2013) we do not pursue this topic further in this work.

In practice, the spectral estimators in equation (83) may breach the Gaussianity testing assumption, especially when only a low number of replicates are available. The assumption of approximate normality for individual replicate spectral estimates, cautiously used in Fryzlewicz and Ombao (2009), will be strengthened by the presence of a higher collection of group replicates (N_1, N_2) (see Section 3.4 for a discussion of WST's features and caveats).

3.3.1.2 Raw Periodogram F-Test ('FT')

We now construct a testing procedure that is not reliant on the Gaussianity assumption whose validity we challenged above. Formally, for each scale $j \in \mathbb{N}$ and rescaled time $z \in (0,1)$, the spectral equality $S_j^{(1)}(z) = S_j^{(2)}(z)$ is equivalent to $\beta_j^{(1)}(z) = \beta_j^{(2)}(z)$ as the autocorrelation wavelet

inner product matrix A that links the two (see equation (80)) is invertible. We therefore replace our initial collection of multiple hypothesis tests with equivalent re-framed versions

$$H_0 : \beta_j^{(1)}(z) = \beta_j^{(2)}(z), \forall\, j, z$$

against the alternative (H_A) that there exist a scale j^* and rescaled time z^* such that

$$\beta_{j^*}^{(1)}(z^*) \neq \beta_{j^*}^{(2)}(z^*).$$

In order to construct our test statistic, we test for spectral equality by examining the $\beta_j(z)$ quantities instead.

In reality we do not know $\beta_j^{(i)}(z)$ for $i = 1, 2$ so we replace it by an asymptotically unbiased estimator. As data are available consisting of multiple time series with known group memberships, we replace $\beta_j^{(i)}(z)$ with an estimate across the group replicates. Specifically, if we have N_i independent time series replicates from group i, we define

$$N_i \bar{I}_{j,k}^{(i)} := \sum_{r_i=1}^{N_i} I_{j,k}^{(i),r_i} \sim \beta_j^{(i)}(k/T)\chi^2_{N_i}. \tag{86}$$

The distribution above follows as the raw wavelet periodogram coefficient of each r_ith periodogram replicate $I_{j,k}^{(i),r_i}$ is (scaled) χ_1^2 distributed (e.g. Nason and Stevens (2015)) and independent of all other raw wavelet periodogram coefficients across all other replicates from the same group (also see Fryzlewicz and Ombao (2009) and the discussion in Section 3.2.1). Under the further assumption of group independence, $\bar{I}_{j,k}^{(1)}$ and $\bar{I}_{j,k}^{(2)}$ are independent and distributed as detailed in equation (86). Hence we propose the test statistic

$$F_{j,k} = \frac{\bar{I}_{j,k}^{(1)}}{\bar{I}_{j,k}^{(2)}} \sim F_{N_1,N_2} \text{ under the null hypothesis.} \tag{87}$$

Each test statistic is then compared with a critical value derived from the F_{N_1,N_2}-distribution in the usual way.

Discussion. An advantage of the FT, particularly as opposed to the WST, is that its underlying distributional assumption is theoretically, as well as practically, more reliable. We would therefore expect the FT to outperform the WST in many applications, and this is indeed validated across a variety of simulation settings (see Section 3.4).

In certain practical applications, the binary distinction provided by a hypothesis test could be seen as somewhat restrictive in terms of characterising the difference between two groups (Das and Nason, 2016). However, the WST and FT developed above both report the time-scale locations of the significant differences between the two group spectra. These can be visualised as a 'barcode' plot, where a significant difference is represented by a black line at the time-scale location of the rejection of the null hypothesis (see for example Figure 36, right). In many practical applications (such as our motivating example), such information can be extremely useful as, in contrast with the established period–estimation techniques, our proposed methodology can identify the time point at which the control and treatment groups start to have different circadian rhythms (see Section 3.5.1 for a detailed example). Alternatively, there may be practical situations when a degree of dissimilarity would be of interest (see for example Section 4.5).

For all our proposed tests, practitioners can also be informed by the number of rejections (as a coarse dissimilarity measure), with larger values potentially indicating a greater departure from the null hypothesis (as cautiously used in Nason (2013) and in Section 3.4.2). However, factors such as correlations between coefficients (see discussion in Section 3.3.1.1) mean that such numbers should be treated with caution, but just as in Nason (2013) we do not pursue this topic further in this work.

3.3.2 Ultradian Dataset: Hypothesis Testing for Spectral Equality Across Scales ('HFT')

For certain biological applications, such as the Ultradian motivating example, it is more important to identify spectral differences between groups at scale-level and the time locations of spectral differences are of less interest. For such situations, we replace the spectral comparison $H_0 : S_j^{(1)}(z) = S_j^{(2)}(z)$ of the previous section, in general equivalent to $H_0 : \beta_j^{(1)}(z) = \beta_j^{(2)}(z)$, by the comparison of the respective Haar-Fisz transforms, i.e. test for

$$H_0 : \mathscr{F}\beta_j^{(1)}(z) = \mathscr{F}\beta_j^{(2)}(z), \forall\, j,\, z.$$

Equivalently, in the notation established in Section 3.2.1 we test

$$H_0 : \mathscr{B}_j^{(1)}(z) = \mathscr{B}_j^{(2)}(z),\ \forall\, j,\, z \tag{88}$$

versus the alternative (H_A) that there exist some scale j^* and rescaled time z^* for which the equality does not hold. We shall refer to this test as the *Haar-Fisz test* (HFT).

As we do not know $\mathscr{B}_j^{(i)}(z)$, we replace it by its unbiased estimator $\mathscr{H}_{j,k}^{(i)}$ at scale j and time k (with $z = k/T$) for group $i = 1, 2$. In applications which do not provide access to replicate data, we could adopt equation (84) with $\hat{S}_j^{(i)}(k/T)$ replaced by $\mathscr{H}_{j,k}^{(i)}$ and estimate the variance across each scale as the Haar-Fisz transform stabilises variance (Nason and Stevens, 2015). When replicates are available, we use equation (83) with $\mathscr{H}_{j,k}^{(i)}$ to obtain group averaged estimators of $\mathscr{B}_j^{(i)}(z)$, denoted $\hat{\mathscr{H}}_{j,k}^{(i)}$, and propose a test statistic as in equation (84) with $\hat{S}_j^{(i)}(k/T)$ replaced by $\hat{\mathscr{H}}_{j,k}^{(i)}$. The variance estimation techniques and subsequent test statistic distribution follow as detailed in Section 3.3.1 and the results of the HFT can also be visualised as a 'barcode' plot.

Discussion. The HFT identifies both scales and times at which the null hypothesis of spectral equality in the Haar-Fisz domain does not hold. However, as the Haar-Fisz transform essentially 'averages' within each scale of the raw wavelet periodogram, potential differences 'spread' throughout the scale. This property makes it ideal for identifying scale-level differences between group wavelet spectra (see for example Figure 37, right).

In practice, due to its scale averaging construction, the HFT results in many more time-localised rejections than the actual number of differing coefficients in the original spectra. Furthermore, the HFT does sometimes have difficulty discriminating between spectra which differ by a small number of coefficients; however, the HFT does correctly identify scale-level spectral differences (see Section 3.4 for further investigations).

An additional benefit of this approach is also to bring the data (in this context, the Haar-Fisz transform of the raw wavelet periodogram) closer to Gaussianity and to break the dependencies across time. Consequently, the assumptions behind the t-test are closely adhered to and the dependencies between the multiple tests we perform are weak.

3.3.3 Nematode Dataset: Hypothesis Testing for 'Same Shape' Spectra ('HT')

In applications such as the Nematode example, the focus may be on identifying whether groups evolve according to spectra that have the same shape at each scale (up to a scale-dependent additive constant), thus indicating that the same patterns are identified in the data, albeit with potentially different magnitudes.

Mathematically, for a scale-dependent (non-zero) constant denoted by C_j, we formalise our hypotheses as

$$H_0 : S_j^{(1)}(z) = S_j^{(2)}(z) + C_j, \quad \forall\, j, z \tag{89}$$

versus the alternative $H_A : S_{j^*}^{(1)}(z^*) \neq S_{j^*}^{(2)}(z^*) + C_{j^*}$ for some scale j^* and time z^*.

Denoting by $\underline{C}$ the $J \times 1$ vector that holds C_j as its jth component and recalling equation (80), we can equivalently re-frame the problem into testing whether

$$H_0 : \beta_j^{(1)}(z) = \beta_j^{(2)}(z) + c_j, \text{ or equivalently } H_0 : \beta_j^{(D)}(z) = c_j, \quad \forall\, j, z$$

where c_j is the jth entry of the vector $\underline{c} = A\underline{C}$ and $\beta_j^{(D)}(z) := \beta_j^{(1)}(z) - \beta_j^{(2)}(z)$.

In the spirit of the tests developed in Fan and Lin (1998), and as undertaken by von Sachs and Neumann (2000) and Nason (2013), at each scale j we assess the constancy through time of $\beta_j^{(D)}(z)$ by examining its associated Haar wavelet coefficients. Although, in principle, any wavelet system could be adopted, von Sachs and Neumann (2000) note that the Haar wavelet coefficients are ideal for testing the constancy of a function. Hence we employ these wavelets and refer to the test developed in this section as the *Haar Test* (HT).

The underlying principle behind these tests is that the wavelet transform of a constant function is zero, hence under H_0 above, the wavelet coefficients of $\beta_j^{(D)}(z)$ are

$$v_{\ell,p}^j = \int_0^1 \beta_j^{(D)}(z)\psi_{\ell,p}^H(z)\,dz = c_j \int_0^1 \psi_{\ell,p}^H(z)\,dz = 0,$$

where $\{\psi_{\ell,p}^H(z)\}_{\ell,p}$ denote the usual Haar wavelets at scale ℓ and location p.

This suggests performing multiple hypothesis testing on the collection of hypotheses

$$H_0 : v_{\ell,p}^j = 0, \forall\, j, \ell \text{ and } p$$

against the alternative (H_A) that there exist j^*, ℓ^* and p^* such that $v_{\ell^*,p^*}^{j^*} \neq 0$.

As the spectral and related quantities are unknown, and since the wavelet transform is linear, we estimate each $v_{\ell,p}^j$ by $\hat{v}_{\ell,p}^j = \hat{v}_{\ell,p}^{j,(1)} - \hat{v}_{\ell,p}^{j,(2)}$, with the Haar wavelet coefficients corresponding to each group $i = 1, 2$ estimated in the spirit of Nason (2013) as

$$\hat{v}_{\ell,p}^{j,(i)} = 2^{-\ell/2}\left(\sum_{r=0}^{2^{\ell-1}-1} I_{j,2^\ell p-r}^{(i)} - \sum_{q=2^{\ell-1}}^{2^\ell-1} I_{j,2^\ell p-q}^{(i)}\right), \tag{90}$$

at each (original) scale j and Haar scale ℓ and locations p, q.

With the availability of independent replicates within each group, we estimate the group i Haar wavelet coefficients as

$$\hat{v}_{\ell,p}^{j,(i)} = \frac{1}{N_i}\sum_{r_i=1}^{N_i} \hat{v}_{\ell,p}^{j,(i),r_i}, \tag{91}$$

where each $\hat{v}_{\ell,p}^{j,(i),r_i}$ is obtained as in equation (90) for the r_i-th replicate.

Under a specific set of assumptions, Nason (2013) shows the asymptotic normality of the Haar wavelet coefficient estimator of the wavelet periodogram at scale j. Thus, in our setting, each $\hat{v}_{\ell,p}^{j,(i),r_i}$ for $i = 1, 2$ is asymptotically normal with mean $v_{\ell,p}^{j,(i),r_i}$ and variance $(\sigma_{\ell,p}^{j,(i)})^2$. Using the replicate independence, we have that $\hat{v}_{\ell,p}^{j,(i)}$ is asymptotically normally distributed with mean $v_{\ell,p}^{j,(i)}$ and variance $(\sigma_{\ell,p}^{j,(i)})^2/N_i$ and note that its distributional closeness to the normal increases via a central limit theorem argument with the increasing number of replicates.

The group independence assumption then leads to an asymptotically joint normal distribution for $(\hat{v}_{\ell,p}^{j,(1)}, \hat{v}_{\ell,p}^{j,(2)})$. Following the continuous mapping theorem, we obtain that

$$\hat{v}_{\ell,p}^{j} = \hat{v}_{\ell,p}^{j,(1)} - \hat{v}_{\ell,p}^{j,(2)}$$

has an asymptotic normal distribution with mean $v_{\ell,p}^{j,(1)} - v_{\ell,p}^{j,(2)}$ and variance

$$\frac{\left(\sigma_{\ell,p}^{j,(1)}\right)^2}{N_1} + \frac{\left(\sigma_{\ell,p}^{j,(2)}\right)^2}{N_2}.$$

In the presence of replicates, we propose a test statistic of the form discussed in equation (84)

$$T_{\ell,p}^{j} = \frac{\hat{v}_{\ell,p}^{j}}{\left((\hat{\sigma}_{\ell,p}^{j,(1)})^2/N_1 + (\hat{\sigma}_{\ell,p}^{j,(2)})^2/N_2\right)^{1/2}} \sim t_{df} \text{ under the null hypothesis,} \tag{92}$$

where $(\hat{\sigma}_{\ell,p}^{j,(i)})^2$ is an estimate of the variance of $\hat{v}_{\ell,p}^{j,(i)}$ for $i = 1, 2$ across the N_i observations in group i, obtained using the standard sum–of–squares sample variance formula and df denotes the degrees of freedom associated with the variance estimation procedure (see Section 3.3.1.1). Each test statistic is then compared with a critical value derived from the t-distribution in the usual way.

Discussion. In order to control the asymptotic bias derivation, one of the assumptions under which the distributional theory is derived consists of limiting the scales of the Haar wavelet coefficients $v_{\ell,p}^{j}$ to be sufficiently coarse, $\ell = 0, \ldots, (J - \lceil J/2 \rceil - 2)$. Furthermore, as in Nason (2013), we only consider the wavelet coefficients of the periodogram at levels $j \geq 3$ in order to avoid the effects of a region similar to the 'cone of influence' described by Torrence and Compo (1998).

To aid the visualisation of the WST, FT and HFT results, we use a 'barcode' plot that indicates the time- and scale- locations where significant differences are present (see for example Figure 38). The HT can also indicate where the significant differences are located in the series and can plot the results in a manner similar to the wavelet test of stationarity (see for example Figure 13). However, due to its construction, these locations are more difficult to interpret than for the WST, FT and HFT.

3.3.4 Summary

A summary of the hypothesis tests developed in this chapter detailing the test name, its acronym and the motivation behind its development can be found in Table 11.

Name	Acronym	Designed to ...
Wavelet Spectrum Test	WST	Detect whether two groups display significant differences in the evolution of their spectral structures, and if so, the particular scales and times at which such differences occur.
Raw periodogram F-Test	FT	Detect whether two groups display significant differences in the evolution of their spectral structures, and if so, the particular scales and times at which such differences occur.
Haar-Fisz Test	HFT	Detect differences when the total power within a scale differs between groups.
Haar Test	HT	Detect whether groups evolve according to spectra that have the same shape (up to an additive constant) at each scale.

Table 11: A summary of the hypothesis tests developed in this chapter.

3.4 Simulation Studies

The goals of the simulation studies were: (1) to evaluate the empirical power and size of our new tests; (2) to consider the effect of sample size on the accuracy of the tests; (3) to investigate two approaches to multiple-hypothesis testing: Bonferroni correction (denoted 'Bon.') and the false discovery rate procedure ('FDR'); (4) to investigate the performance of our proposed tests when certain modelling assumptions are broken and (5) to evaluate the empirical power and size of our new tests in comparison with the adaptive Neyman Test (ANT) of Fan and Lin (1998), see Section 3.2.2. This benchmark method performs well in practice when the assumption that the data can be modelled as an (unknown) underlying function plus noise (henceforth referred to as a 'function plus noise' time series) is valid. (For more details regarding the ANT see Appendix 3.11.1.)

The basic structure of each simulated experiment (a comprehensive description of the simulation studies can be found in Appendix 3.11.2) can be described as follows. In each case, we assumed that the signal was a realisation from one of $i = 1,2$ possible groups. For each group, we generated a set of $N_1 = N_2 = 1,10,25,50$ signal realisations, each of length $T = 256$, the equivalent of a free-running period of 4 days. For each realisation, we obtained the raw and corrected wavelet periodograms using (unless otherwise stated) the Haar wavelet from the `locits` software package for R (available from the CRAN package repository), although, any wavelet system can, in principle be used (see Section 3.4.3.2). The Haar–transformed and Haar–Fisz transformed raw wavelet periodogram were subsequently obtained and the spectral testing procedures carried out as described in Section 3.3. The results are compared with the known group memberships, and the procedure is then repeated 1000 times to obtain empirical size and power estimates as outlined in the following sections.

3.4.1 Power Comparisons

To explore statistical power we simulate a set of $N_1 = N_2 = 1, 10, 25, 50$ signal realisations from each group where the individual group spectra are defined such that there exists a scale j^* and time t^* such that $H_A : S^{(1)}_{j^*}(t^*/T) \neq S^{(2)}_{j^*}(t^*/T)$. The empirical power estimates are obtained by counting the number of times our tests reject the null hypothesis of spectral equality. The models we will use are denoted **P1**–**P12** respectively and are briefly described below. (Precise details can be found in Appendix 3.11.2.)

1. **P1: Fixed Spectra.** We follow Krzemieniewska et al. (2014) and design the spectra of the two groups to differ at the finest level (resolution level 7) by 100 coefficients.

2. **P2: Fixed Spectra-Fine Difference.** We modify the model **P1** such that the spectra of the two groups differ by only 6 coefficients.

3. **P3: Fixed Spectra-Plus Constant.** Modify the model **P1** such that the spectra of the two groups differ by a constant in the finest resolution level.

4. **P4/P5: Gradual Period Change.** This study replicates a typical circadian experiment with changes that cannot be captured by standard analyses assuming stationarity and only reporting an average period value. We thus define 3 possible groups, where each group represents a signal that gradually changes period from 24 to: 25 (Group 1), 26 (Group 2) and 27 (Group 3) over (approximately) two days, before continuing with the relevant period for a further two days (also see Hargreaves et al. (2018)). To determine which changes can be discriminated by the methods, we perform two studies within this setting: simulations from Groups 1 and 2 (**P4**) and simulations from Groups 1 and 3 (**P5**).

5. **P6/P7: AR Processes with time-varying coefficients.** We simulate from an important class of nonstationary processes– AR(2) processes with: abruptly (**P6**) and slowly (**P7**) changing parameters (as in Fryzlewicz and Ombao (2009)).

6. **P8–P12: 'Function Plus Noise' Time Series (Constant Period).** This study follows Zielinski et al. (2014) and generates each time series using an underlying cosine curve with additive noise, which also coincides with the theoretical assumptions of the ANT. We define time series as realisations from one of 6 possible groups, each with a different (constant) period, relevant to our circadian setting. To determine which period changes can be discriminated by the methods, we perform five studies within this setting: simulations from a group with a period of 24 hours versus a group with a period of 21, 22, 23, 23.5 and 23.75 hours (models **P8–P12** respectively).

3.4.1.1 Discussion of Findings

The empirical power values for $N_1 = N_2 = 25$ (this is the typical number of available replicates in circadian studies, see Section 3.5) for models **P1–P7** are reported in Table 12. We found that all tests perform well when the spectra differ by a large number of coefficients (model **P1**). The FT (and, to a lesser extent, the HT) are able to discriminate between spectra that differ by a small number of coefficients (model **P2**) whereas the HFT has lower empirical power. By construction, the HT cannot differentiate between spectra that differ by a constant at a particular

Model	WST (Bon.)	WST (FDR)	FT (Bon.)	FT (FDR)	HFT (Bon.)	HFT (FDR)	HT (Bon.)	HT (FDR)
P1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
P2	39.3	48.0	**100.0**	**100.0**	29.1	31.8	86.2	86.4
P3	100.0	100.0	100.0	100.0	100.0	100.0	4.3	4.4
P4	1.0	2.7	45.5	54.5	33.2	36.5	**100.0**	**100.0**
P5	5.9	14.6	97.0	99.9	**100.0**	**100.0**	**100.0**	**100.0**
P6	**100.0**	**100.0**	87.5	92.6	44.8	89.1	66.5	67.7
P7	**100.0**	**100.0**	54.3	64.5	97.4	99.9	**100.0**	**100.0**

Table 12: Simulated power estimates (%) for models P1-P7 with nominal size of 5% with $N_1 = N_2 = 25$ realisations from each group. Highest empirical power estimates are highlighted in bold.

resolution level (model **P3**), but we found that the HT performs well in our synthetic circadian example of gradual small period change across many time-scale locations (models **P4** and **P5**). Due to the higher distributional reliability of the FT, it unsurprisingly outperforms the WST when the times series are generated from a defined spectrum (models **P1**–**P5**). However, distributional properties of the time-varying AR process ensure that the WST performs best when data are generated using models **P6** and **P7**, with the HT and HFT also performing well for model **P7**.

Effect of sample size. The number of replicates in each group (N_1, N_2) are also an important factor in achieved power. The results for the HFT with $N_1 = N_2 = 1$ are shown in Table 20 (Appendix 3.11.3), since we recall that the HFT is the only proposed test which can be applied when replicate data is not available– see Section 3.3.2. The results for all tests with $N_1 = N_2 = 10$ and 50 replicates are shown in Table 21 (Appendix 3.11.3). Increasing the number of replicates should, and indeed does, increase the empirical power of all tests (with the exception of the HT for model **P3**). For example, note the increase in empirical power (particularly for models **P2** and **P4**) as the number of replicates increases from 10 to 25.

Approach to multiple-hypothesis testing. These studies show that the Bonferroni correction provides a more conservative approach. The false discovery rate gives an empirical power greater than (or equal to) that of the Bonferroni correction (see e.g. model **P6** in Table 12).

Performance comparison. We also report that the empirical power of the ANT for model **P5** (gradual period change, 25 replicates) was 10.7%, which is below the results in Table 12 for our proposed tests. This is to be expected as the underlying assumptions of the ANT are no longer met. (Similar results are obtained for models **P1**–**P7**, hence we do not provide these here.)

Table 13 presents a selection of the performance comparison results for models **P8–P12** when $N_1 = N_2 = 25$. (The results for all tests with $N_1 = N_2 = 10$ replicates are also shown in Table 22, Appendix 3.11.3.) As expected, the ANT performs extremely well in all these studies since the underlying assumptions of the methodology are adhered to. Nevertheless, it is encouraging that the WST, FT and HT also all have an empirical power over 95% (25 replicates) showing that our methodology can also be successfully applied to 'function plus noise' time series as designed for the ANT. However, the HFT had difficulty discriminating between groups when the period difference was less than 2 hours. This was no surprise as the HFT was constructed to detect differences in scale only and, due to the lower frequency resolution of the wavelet

Model	Test Group Period	WST (FDR)	FT (FDR)	HFT (FDR)	HT (FDR)	ANT
P8	21	100.0	100.0	100.0	100.0	100.0
P9	22	100.0	100.0	100.0	100.0	100.0
P10	23	100.0	100.0	92.0	100.0	100.0
P11	23.5	100.0	100.0	31.8	100.0	100.0
P12	23.75	100.0	97.9	9.1	98.3	100.0

Table 13: **Performance Comparison:** Simulated power estimates (%) for models P8-P12 with nominal size of 5% with $N_1 = N_2 = 25$ realisations from each group and using the false discovery rate procedure (FDR). Note: Control group period is 24 hours in each model.

spectrum, the total power within each scale of the wavelet spectrum will be very similar for both groups.

3.4.1.2 Power Comparisons: Conclusions

In practice, the suitability of the testing procedures is determined by a combination of factors, such as the practical problem posed by scientists, the degree to which the data adheres to the underlying theoretical assumptions and the number of available replicates. For example, models **P1-P3** all stem from a simulated LSW structure and thus would be subject to a test for time-scale equality departure, carried out through an 'FT' as its theoretical assumptions are closely adhered to. Recall that the 'WST' was proposed as a 'naive' variant and is heavily reliant on the number of replicates in order to achieve the appropriate distributional properties, thus its best results are obtained for models that have been simulated from time-varying AR processes. Meanwhile, for data following models that exhibit a gradual period change (such as **P4-P5**) one might be interested in identifying scale-dependent patterns or discrepancies, carried out through the 'HT' or 'HFT'.

3.4.2 Size Comparisons

To explore statistical size, we simulate data from a number of models and we asses how often our hypothesis tests reject the null hypothesis of spectral equality (i.e. the time series are generated in the same way for both test groups). The models are denoted **M1**–**M5** respectively and defined as follows. (Precise details can be found in Appendix 3.11.2.)

1. **M1: Fixed Spectra.** We simulate all data from the wavelet spectrum associated with Group 1 in models **P1**, **P2** and **P3**, which we define as $\{S_j^{(1)}(z)\}_{j=1}^{J}$ in equation (99).

2. **M2: Gradual Period Change.** We simulate all data from the wavelet spectrum which corresponds to a time series that gradually changes period from 24 to 25 hours over (approximately two days), before continuing with period 25 hours for a further two days (i.e. Group 1 from models **P4/P5**).

3. **M3: AR Processes With Abruptly Changing Parameters.** Each time series is generated from the process defined by equation (103) with the abruptly changing parameters as defined for group $i = 1$ in Table 18 (i.e. Group 1 from model **P6**).

4. **M4: AR Processes With Slowly Changing Parameters.** Each time series is generated from the process defined by equation (104) with the slowly changing parameters as defined for group $i = 1$ in Table 19 (i.e. Group 1 from model **P7**).

5. **M5: 'Function Plus Noise' Time Series (Constant Period).** All data are simulated (using equation (105)) from the model that corresponds to a time series with a constant period of 24 hours (i.e. Group 1 from models **P8–P12**).

3.4.2.1 Discussion of Findings

The empirical size values for models **M1–M4** with $N_1 = N_2 = 25$ (this is the typical number of available replicates in circadian experiments, see Appendix 3.7) are reported in Table 14. The results for the HFT (for models **M1–M4**) with $N_1 = N_2 = 1$ are shown in Table 20, Appendix 3.11.3 (recall: the HFT is the only proposed test which can be applied when replicate data is not available– see Section 3.3.2). The results for all tests (for models **M1–M4**) with $N_1 = N_2 = 10$ and 50 replicates are shown in Table 23 (Appendix 3.11.3). The results (for all tests) for model **M5** with $N_1 = N_2 = 10$ and 25 are shown in Table 22 (Appendix 3.11.3).

These studies show that the empirical size corresponding to all proposed tests (apart from the FT for model **M4** with $N_1 = N_2 = 10$ and 25) are less than the nominal size of 5%. A close inspection of rejections for the FT for model **M4** with $N_1 = N_2 = 10$ and 25 and both multiple-hypothesis testing methods (Table 24 in Appendix 3.11.3) reveals that, for this particular example, the number of rejections is often 1. If we disregard such situations, the empirical size of the FT also falls below the nominal size of 5% for all sample sizes and multiple-hypothesis testing procedures. In practice, circadian scientists are mostly interested in the numbers of rejections and their locations and often choose to disregard situations where very few coefficients are significantly different. Indeed, this is also our approach in Section 3.5.

Effect of sample size. Note that the tests scale well with increasing sample size, with the nominal size acting as an upper bound, a behaviour also present in other related empirical size investigations (see e.g. Cho (2016)).

Approach to multiple-hypothesis testing. These studies show that the Bonferroni correction provides a more conservative approach, whereas the false discovery rate (using the correction outlined above) is closer to the nominal size.

Performance comparison. The results for model **M5** with $N_1 = N_2 = 10$ and 25 are shown in Table 22 (Appendix 3.11.3). Note that the empirical size estimates for our proposed tests are all lower than the nominal size of 5%, whereas for 10 replicates the empirical size of the ANT is 7.9%.

3.4.2.2 Size Comparisons: Conclusions

These studies show that the empirical size corresponding to all proposed tests is less than the nominal size of 5% (apart from the FT for model **M4** with $N_1 = N_2 = 10$ and 25– where, in most cases, the number of significant coefficients was less than 5). We thus recommend using the less conservative FDR procedure (ignoring situations with very small numbers of rejections). Note this also yields better results for empirical power (see Section 3.4.1) whilst also remaining below the nominal size.

Model	WST (Bon.)	WST (FDR)	FT (Bon.)	FT (FDR)	HFT (Bon.)	HFT (FDR)	HT (Bon.)	HT (FDR)
M1	0.6	1.3	2.5	3.1	0.1	2.0	2.3	2.7
M2	0.3	0.6	3.0	3.9	0.4	3.3	2.5	2.7
M3	0.2	1.5	3.6	3.9	0.0	1.6	3.5	3.8
M4	0.4	0.9	4.6	**5.2**	1.0	2.4	3.4	3.8

Table 14: Simulated size estimates (%) for models M1-M4 with nominal size of 5% and $N_1 = N_2 = 25$ realisations from each group. Empirical size estimates over the nominal size of 5% are highlighted in bold.

3.4.3 Sensitivity Analysis

In this section we investigate the sensitivity of our proposed tests to certain modelling assumptions. We investigate: (1) departures from the normality assumption and (2) the impact of the choice of wavelet family used within the spectral estimation procedures of each of our proposed tests. Throughout this section, we use $N_1 = N_2 = 25$, since this is the typical number of available replicates in circadian experiments (see Appendix 3.7).

3.4.3.1 Departures from Normality

Recall the proposed statistical testing methodology assumes the innovations $\{\xi_{j,k}\}$ to be normally distributed. To investigate the impact of this assumption, we computationally assess the power and size of the proposed tests within the settings outlined in Section 2.4 for models **P1–P5** and **M1–M2** but simulated using non-Gaussian innovations (specifically following a t-distribution with 5, and subsequently 3, degrees of freedom). The results can be found in Table 25 (Appendix 3.11.3). Unsurprisingly, when the normality assumption is broken, the empirical power of all tests is less than (or equal to) the empirical power when the innovations follow a standard normal distribution. The increasing distributional departure from normality appears to be of little relevant influence when testing data simulated from models **P1** and **P3** (across all tests), while the empirical power drops for the HT corresponding to models **P2** and **P4/P5**. The testing procedures break for models **P4/P5** with t_3-distributed innovations, as intuitively, the presence of heavier innovations make the gradual period change structure of models **P4/P5** very difficult to discriminate. We also note that the HT is heavily reliant on the distributional assumptions (see Section 3.3.3) which explains its sensitivity. Due to its construction (see Section 3.3.1.2), the FT appears to more readily reject the null hypothesis, increasing the empirical size of the test. However, if we disregard situations where there are a very low number of rejections (see Section 3.4.2.1) the empirical size of the FT falls below the nominal size of 5% for both multiple–hypothesis testing procedures and all studies (other than M1 with FDR). We report here that the empirical power of the ANT for model **P1** (fixed spectra) with t-distributions with 5 degrees of freedom was 6.8%, which is below the results in Table 25 for all our proposed tests (which are all over 99.9%). This is to be expected since, as in Section 3.4.1, the underlying assumptions of the ANT are not valid. (Similar results are obtained for models **P2–P7**, hence we do not provide these here.)

We also investigated the power and size for models **P8–P12** and **M5** (see Section 2.4) simulated using non-Gaussian errors (specifically following t-distributions with 5, and subsequently

3, degrees of freedom). The results can be found in Table 26 (Appendix 3.11.3). The WST, FT and HT appear to share a good degree of robustness as they all have an empirical power over 99% for models **P8–P11**, showing that our methodology can also be successfully applied to 'function plus noise' time series (as designed for the ANT) with non-Gaussian error. Akin to the previous results for the gradual period change models **P4/P5**, the distribution of the noise term does appear to have an adverse effect in model **P12**, where the difference between the periods of the two underlying signals is only 15 minutes. Across this study, the HFT was most affected. A possible explanation is that the HFT was constructed to detect differences in scale only and, due to the lower frequency resolution of the wavelet spectrum, the total power within each scale of the wavelet spectrum will be very similar for both groups. This issue will have been compounded by the heavier tailed distribution of the noise term. We also report here that, in the settings of this study, the performance of ANT was sustained as its underlying assumptions are adhered to.

3.4.3.2 Choice of wavelet

The wavelet system gives a representation for nonstationary time series under which we estimate the wavelet spectrum and subsequently perform hypothesis testing. We investigated the sensitivity of our methods to the wavelet choice. For models **P1–P5**, the Haar wavelet was used for spectral estimation, but different, potentially mismatched wavelets were used to generate the processes from the spectrum: Haar wavelets, Daubechies' least-asymmetric wavelets with 4 vanishing moments and Daubechies' extremal phase wavelets with 10 vanishing moments. Models **P6–P12** were not generated from LSW spectra (see Section 2.4), hence we report the results when using a selection of wavelets for the empirical wavelet spectrum.

The results in Tables 27 and 28 (Appendix 3.11.3) show that our methodology is fairly robust to the wavelet choice. The empirical size estimates all fall below the nominal size. The results indeed support the intuition that, as the scope of our work is to devise tests that locally identify dissimilarities between pairs of spectra, the short support overlaps of Haar wavelets counterbalance their otherwise reduced capacity of representing smooth signals.

3.4.4 Summary of Findings

A summary of the hypothesis tests developed in this chapter detailing the test name, its acronym, strengths and weaknesses can be found in Table 29 (Appendix 3.12).

3.5 Real Data Analysis: Back to the Motivating Circadian Datasets

We now use our proposed methodology to analyse the motivating examples (Section 3.1). Prior to analysis, we investigate whether the normality assumption is tenable for each of our motivating datasets. The results (Appendix 3.9) show that, for each of our motivating datasets, the normality assumption is appropriate. We then model each circadian trace as an LSW process, estimate its corresponding group wavelet spectral representation and consequently construct the appropriate test statistic that aims to identify whether a departure towards a specific type of spectral difference is present or not (as described in Section 3.3). For each dataset, the corresponding number of rejections can be found in Table 15, with corresponding representative 'barcode' plots in Figures 36, 37 and 38.

Dataset (Test)	Bon.	FDR
Lead (FT)	31 (3%)	133 (15%)
Ultradian (HFT)	1102 (54%)	1538 (75%)
Nematode (HT)	0 (0%)	0 (0%)

Table 15: The number of rejections (as a percentage in brackets) for each relevant proposed test and multiple-hypothesis testing procedure for the motivating example datasets.

3.5.1 Lead Dataset

Section 3.1.1.1 outlined the scientific aims to determine if lead nitrate affects the circadian clock and, if so, to detect the times and scales at which any significant differences arise between the 'Control' and 'Lead' exposure groups. Therefore we are particularly interested in the results of the FT. Table 15 shows the results for the FT and includes both the more conservative Bonferroni correction and FDR. In order to visualise the areas of null hypothesis rejection of spectral equality between the control and lead-exposure groups, both group average estimated spectra as well as the 'barcode' plot for the FT (with FDR) appear in Figure 36. Figure 36 indicates that the differences between the two spectra lie in resolution levels 2–4, directly corresponding to a circadian rhythm, with the number of rejections increasing with exposure time. We conclude that there is evidence that exposure to lead does affect the circadian clock of *A. thaliana*, and this change manifests itself after approximately three days of free-running conditions.

As discussed in Section 3.1, the 'circadian clock' allows plants to synchronise their internal processes with the external environment (Oakenfull et al., 2018). In particular, it allows the anticipation of daily changes and, therefore, future environmental stresses, such as mid–day drought and midnight coldness (Sanchez et al., 2011). Therefore, a 'circadian clock' provides fitness by anticipating predictable environmental stresses and coordinating appropriate physiological responses (Sanchez et al., 2011). Consequently, if clock function is impaired, for example by changes in the plant's chemical environment, then there could be major consequences for growth (Oakenfull et al., 2018). Our results suggest that exposure to lead does affect the circadian clock of *A. thaliana* which, therefore, may negatively impact growth efficiency. However, the change in the circadian clock manifests itself after approximately three days of free-running conditions. Therefore, transient changes in lead exposure would be less detrimental to the plant. In conclusion, this study suggests that long–term exposure to lead (at this particular concentration) may negatively impact the fitness of *A. thaliana* and hence would not be recommended. This result could be used to inform the appropriate concentration of lead (nitrate) in soil (see Chapter 4 for further details).

3.5.2 Ultradian Dataset

Section 3.1.1.2 introduced this experiment and highlighted the need to detect whether any differences appear in the circadian and ultradian components of the 'Control' and 'Mutant' groups. Hence we are interested in the results of the HFT, specifically developed to identify the scales, rather than the times, at which potential differences arise. Table 15 shows the results for the HFT, including both the Bonferroni correction and FDR. The results indicate rejections of

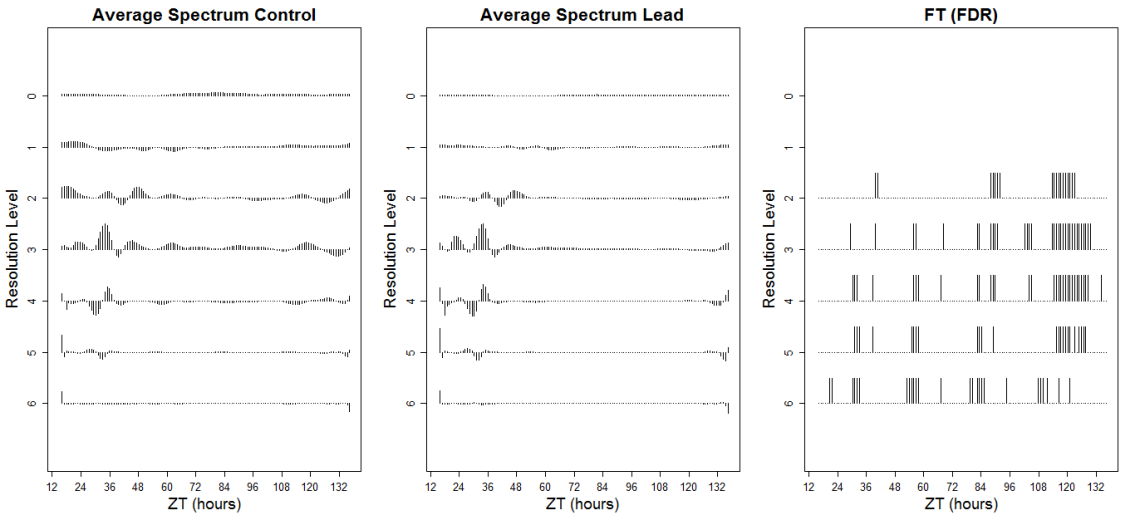

Figure 36: **Lead dataset.** Left: Average estimated spectrum of the 'Control' group; Centre: Average estimated spectrum of the 'Lead' group; Right: 'Barcode' plot for FT (with FDR).

the null hypothesis of spectral equality between the control and mutant plants across a range of scales. The group average estimated spectra and 'barcode' plot for the HFT (with FDR) can be found in Figure 37. Note that the differences between the two spectra lie in the coarsest resolution levels 1–4, associated with circadian rhythms, and higher-frequency levels 6 and 7, corresponding to an ultradian rhythm. We conclude that there is evidence that the mutant plants have altered circadian and ultradian rhythms within *A. thaliana*.

Circadian clocks depend on species–specific clock genes and proteins that interact in complex feedback loops to rhythmically control gene expression (Sanchez et al., 2011; Dusik et al., 2014; Millar et al., 2015). As outlined in Section 3.1.1.2, one approach towards determining and understanding the clock mechanism, is to mutate a gene and examine the resulting behaviour in response to a variety of stimuli. If a mutation affects the circadian rhythms of an organism, this could indicate that this gene is under circadian control within this species. The results in this section indicate that this genetic mutation has altered the circadian rhythm and induced high-frequency behaviour (known as 'ultradian rhythms') in the laboratory model plant *A. thaliana*. These results could reveal new aspects and interactions in the clock mechanism of *A. thaliana*.

3.5.3 Nematode Dataset

The experiment in Section 3.1.1.3 aimed to elucidate the effect of a pharmacological treatment on the *C. elegans* clock. The average estimated spectra of the 'Control' and 'Treatment' groups in Figure 38 share a common profile but with differences in magnitude, indicating that the HT would be appropriate in this context. Table 15 shows that the HT found no significant difference between the shapes of the two spectra, but when tested for equality, the FT (with FDR) found multiple rejections of the null hypothesis of spectral equality between the 'Control' and 'Treatment' groups (refer to the 'barcode' plot in Figure 38). This provides evidence that the two spectra have the same profile within each scale up to an additive non-zero constant. We thus conclude that there is evidence that the treatment significantly affects the intensity of the spectral behaviour, but not its pattern. The spectral differences are present at the highest fre-

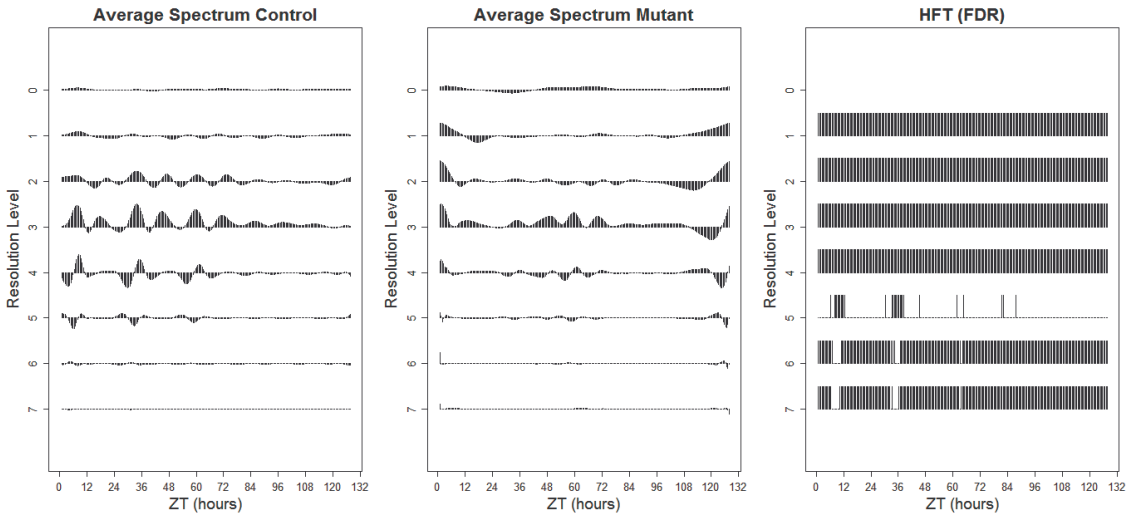

Figure 37: **Ultradian dataset.** Left: Average estimated spectrum of the 'Control' group; Centre: Average estimated spectrum of the 'Mutant' group; Right: 'Barcode' plot for HFT (with FDR).

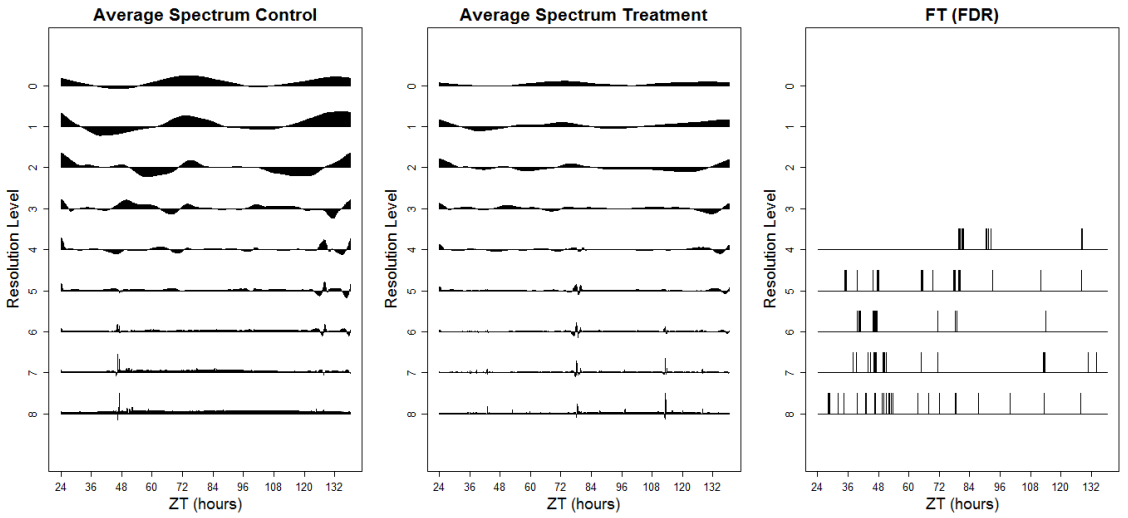

Figure 38: **Nematode dataset.** Left: Average estimated spectrum of the 'Control' group; Centre: Average estimated spectrum of the 'Treatment' group; Right: 'Barcode' plot for FT (with FDR).

quencies (resolution levels 6–8) as an early response to the onset of treatment (prior to time $T = 48$), see Figure 38.

These results indicate that this pharmacological treatment (which has been shown to cause aberrant circadian rhythms in other established mammalian and insect circadian models (Kon et al., 2015; Dusik et al., 2014)) is having an impact on the expression of a gene that has been demonstrated as being under circadian control in the *C. elegans* (Goya et al., 2016). These results support the findings of Goya et al. (2016), that this nematode expresses circadian rhythms. Interestingly, this pharmacological treatment increased the period of the circadian rhythms within other established mammalian and insect circadian models (Kon et al., 2015; Dusik et al., 2014), whereas these results suggest that it affected the intensity of the spectral behaviour, but not its pattern within the *C. elegans*. Nevertheless, these results could aid the precise determination of the elusive circadian clock of *C. elegans*. However, the biological details are beyond the scope of this thesis.

3.6 Conclusions and Further Work

This work was stimulated by a variety of challenging applications faced by the circadian–biology community, which is becoming increasingly aware of the nonstationary characteristics present in much of their data (Hargreaves et al., 2018; Zielinski et al., 2014; Leise et al., 2013). Our methodology fills the gap in the current literature by developing and testing a much needed tool for the formal spectral comparison of nonstationary data. Our methods are developed as testing procedures, analogous to the period analysis techniques currently adopted within circadian community. Motivated by three complementary applications in circadian biology, our new methodology allows the identification of three specific types of spectral difference. Table 29 in Appendix 3.12 provides a summary of the hypothesis tests developed in this manuscript detailing their strengths and weaknesses.

The competitive performance of our methods was comparatively assessed in an extensive simulation study (Section 3.4). Additionally, when compared to existing methods currently adopted within the circadian community, our proposed tests were able to discriminate between real data sets (Table 15 and Figure 38) where the current methodology could not (Table 16, Appendix 3.8).

In the applications provided, we illustrated the important implications in further understanding the mechanisms behind the plant and nematode circadian clocks, and the environmental implications associated with soil pollution. However, we note that our methodology can readily be applied to other circadian datasets, as well as to data originating in other fields, as long as the data share the same dyadic length (T). This assumption is easily achievable for most experimental data, but for other setups might necessitate further specific treatments depending on the discrepancy between the number of observations.

In all of our proposed hypothesis tests, we wish to test many hypotheses of the type $H_0 : S_j^{(1)}(k/T) = S_j^{(2)}(k/T)$ for several values of j and k. In this chapter we adopted the Bonferroni correction and, for a less conservative approach, the false discovery rate (FDR) procedure. Our simulations in Section 3.4 showed that both these methods work well. However, the multiple-hypothesis testing methods we use do not account for the dependence of the spectral coefficients. The hypothesis tests developed in Sections 3.3.2 and 3.3.3 alleviate this problem by transforming the data to produce coefficients that are approximately uncorrelated. However, neither method fully decorrelates the data and multiple-hypothesis testing methods that take the dependence of the (transformed) spectral coefficients into account are an interesting avenue of further work.

There may be practical situations when a measure of the degree of difference between the underlying evolutionary wavelet spectra of two groups of time series would be of interest (see for example Section 4.5). In Section 3.3.1.2 we discussed how, for all our proposed tests, practitioners can be informed by the number of rejections of the null hypothesis (as a coarse dissimilarity measure), with larger values potentially indicating a greater difference between the spectral behaviour of the two groups. However, factors such as the dependence of the spectral coefficients (see discussion in Section 3.3.1.1) mean that such numbers should be treated with caution. The hypothesis tests developed in Sections 3.3.2 and 3.3.3 alleviate this problem by transforming the data to produce coefficients that are approximately uncorrelated. However, transforming the coefficients adds an additional level of complexity when utilising the number of rejections as a dissimilarity measure. For example, the HFT can result in many more time–

localised rejections than the actual number of differing coefficients in the original spectra (see Section 3.4), as potential differences tend to spread throughout the scale. An interesting avenue of further work would be the development of a robust method that measures the degree of difference between the underlying evolutionary wavelet spectra of two groups of time series.

3.7 Appendix: Experimental Details

In this section we outline the experimental details that led to the datasets introduced in Section 3.1 and subsequently analysed in Sections 3.5.1, 3.5.2 and 3.5.3.

3.7.1 Experimental Overview: Lead and Ultradian Datasets

Both Davis and Millar labs used a firefly luciferase reporter system. This involves fusing the gene of interest (here, 'cold and circadian regulated and RNA binding 2', *CCR2*) to a bioluminescent enzyme called luciferase (Doyle et al., 2002). When *CCR2* is expressed, the resultant luciferase emits light which is measured using a TopCount NXT scintillation counter (Perkin Elmer), allowing relative gene expression of *CCR2* to be quantified *in vivo* (Southern and Millar, 2005; Perea-García et al., 2016a).

3.7.2 Lead Nitrate Dataset

Arabidopsis thaliana seeds (Ws–*CCR2:LUC* (Doyle et al., 2002)) were surface sterilised and plated onto Hoagland's media containing 1% sucrose, 1.5% phyto agar (Hoagland et al., 1950). The seeds were stratified for 2 days at 4°C and transferred to growth chambers to entrain under 12:12 light/dark cycles at a constant temperature of 20°C. Six-day-old seedlings were transferred to 96 well microtiter plates containing Hoagland's 1% sucrose, 1.5% agar (Hanano et al., 2006) with or without supplemental $Pb(NO_3)_2$ (lead nitrate) at a concentration of 1.4mM. After 24 hours, the plants were then transferred to the TOPCount machine. Measurements were taken at intervals of approximately 45 minutes. Measurement began after the transition to 12 hours of darkness (known as subjective dusk) on the seventh day of the plants' life. Therefore, the plants experience one 'normal' day in the TOPCount machine (known as entrainment). After this, the plants are exposed to constant light (known as an LL free-run) for approximately four days. This dataset consists of 48 plant signals recorded at $T = 128$ time points, with both the 'Control' and 'Lead' groups containing 24 plants.

3.7.3 Ultradian Dataset

(Millar et al., 2015). This dataset was obtained following a similar method as outlined for the Lead dataset above, but compared 'Control' *A. thaliana* plants (Ws–2 with *CCR2:LUC* (Doyle et al., 2002)) with 'Mutant' *A. thaliana* plants (Ws–2 *cca1 lhy*). Plants were grown on MS media Murashige and Skoog (1962) with 3% sucrose and 1.5% phyto-agar. Plants were entrained in 12:12 L:D conditions at 22°C followed by an LL free-run. Measurements were taken at intervals of approximately 30 minutes. This dataset consists of 48 plant signals recorded at $T = 256$ time points, with both the 'Control' and 'Mutant' groups containing 24 plants.

3.7.4 Nematode Dataset

This dataset was obtained using male *Caenorhabditis elegans* strain PE254 (obtained from the CGC), which expresses firefly luciferase under the promoter of the *sur-5* gene (Lagido et al., 2008). Nematodes expressing luciferase driven by the *sur-5* promoter have previously been reported to show circadian rhythms in luminescence (Goya et al., 2016). Single nematodes were placed in wells containing 100μl S buffer (Stiernagle, 1999), supplemented with 5 mg/mL

cholesterol, 1 g/L wet weight pelleted *Escherichia coli* OP50 strain and 100 µM luciferin. Treatment wells also contained 10 μM SB 203580 (a p38 MAPK inhibitor (Sigma S8307)). Entrainment conditions were 12 hours at 20°C followed by 12 hours at 15°C for two days in constant darkness. Free-running was at 20°C in constant darkness. Luciferase measurements were recorded approximately every 13 minutes. Nematodes that died (shown by a sudden loss of luciferase expression) were excluded from data analysis. Therefore, this dataset consists of 62 signals recorded at $T = 512$ time points, with the 'Control' and 'Treatment' groups containing 32 and 30 time series respectively.

3.8 Appendix: Real Data Analysis: Supplementary Material

In this section, for each motivating example dataset, we report: a summary of the output of the analysis of the motivating datasets in BRASS (Table 16) and the results of the Priestley-Subba Rao test of stationarity (for each time series) in Table 17.

Dataset	Mean Period Estimate: Control Group	Mean Period Estimate: Test Group	Difference	p–value
Lead	27.4	26.8	-0.6	0.16
Ultradian	6.5	6.5	0.0	0.98
Nematode	24.8	25.6	+0.8	0.55

Table 16: A summary of the output of the analysis of the motivating example datasets in BRASS: the mean period estimate for the control and test groups in hours (obtained using FFT-NLLS analysis (Plautz et al., 1997)), the difference between the period estimates and the corresponding p–value.

Dataset	Lead	Ultradian	Nematode
Number of nonstationary time series	39 (81%)	41 (85%)	61 (98%)
Total number of time series	48	48	62

Table 17: Results for the Priestley-Subba Rao test of stationarity, implemented in the `fractal` package in R and available from the CRAN package repository. Number of nonstationary plants indicates the number of time series (in each motivating example dataset) with enough evidence to reject the null hypothesis of stationarity at the 5% significance level (as a percentage in brackets).

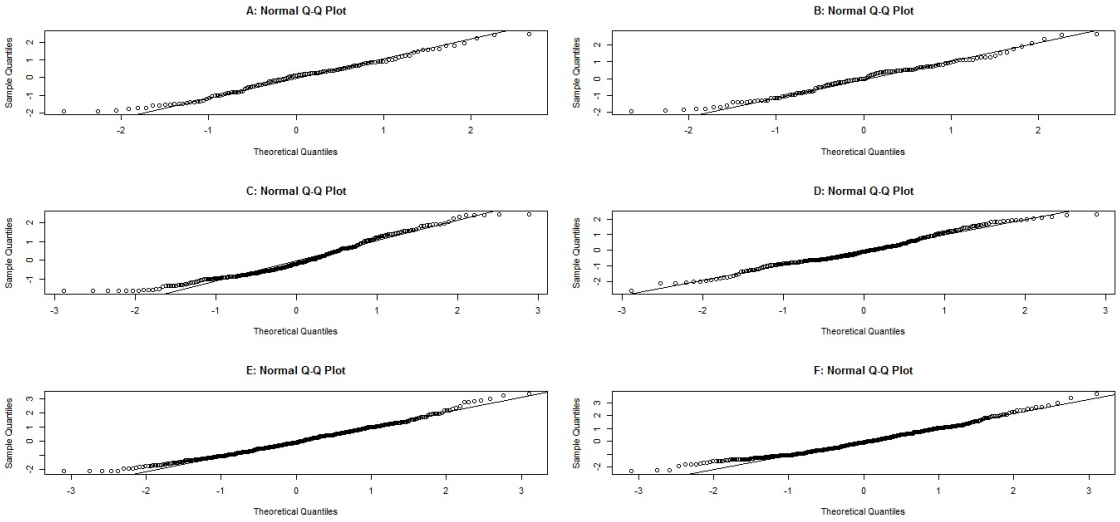

Figure 39: Q–Q plots for a representative series from the control (Plots A, C, E) and test groups (Plots B, D, F) of each of our motivating datasets. Lead Dataset: Plots A and B. Ultradian Dataset: C and D. Nematode Dataset: E and F.

3.9 Appendix: Tenability of the Normality Assumption

In this section we investigate the tenability of the normality assumption for each of our motivating datasets. Following Fryzlewicz (2005), for each series, we standardise the (zero-mean) data using an estimate of the local standard deviation. The estimate was obtained by means of a localised Gaussian kernel with bandwidth chosen using the methods of Fryzlewicz (2005). We then examine the Q–Q plot of the standardised series against the normal quantiles. We report Q–Q plots for a representative series from the control and test groups of each of our motivating datasets in Figure 39. These demonstrate that the normality assumption holds for our motivating data, an assumption also typically undertaken by the circadian community (Perea-García et al., 2016a).

3.10 Appendix: Haar-Fisz Transform

We adapt the definition from Fryzlewicz and Nason (2006), Section 6, which applies the Haar-Fisz transform to the raw wavelet periodogram $I_{j,k;T}$. The algorithm is applied to each scale j of the periodogram separately as follows.

1. Let $c_{J,m} := I_{j,m}$ for $m = 0, \ldots T-1$, where $T = 2^J$.

2. For $l = (J-1), \ldots, 0$, recursively form the vectors

$$d_{l,m} = \frac{c_{l+1,2m} - c_{l+1,2m+1}}{\sqrt{2}} \qquad \text{and} \qquad c_{l,m} = \frac{c_{l+1,2m} + c_{l+1,2m+1}}{\sqrt{2}},$$

   where $m = 1, \ldots, 2^l - 1$, and $d_{l,m}$ and $c_{l,m}$ are the Haar wavelet and scaling coefficient of the raw wavelet periodogram at scale j, respectively.

3. Divide the wavelet coefficients by the scaling coefficients to produce the Haar-Fisz coefficients

$$f_{l,m} = \frac{d_{l,m}}{c_{l,m}} \tag{93}$$

   for $c_{l,m} \neq 0$. For $c_{l,m} = 0$ set $f_{l,m} = 0$.

4. For $l = 0, \ldots, J-1$, recursively modify the vectors c_l:

$$c_{l+1,2m} = c_{l,m} + f_{l,m} \qquad \text{and} \qquad c_{l+1,2m-1} = c_{l,m} - f_{l,m},$$

   where $c_{0,0} = c_{0,0}$ and $m = 1, \ldots, 2^l$.

5. Define $\mathscr{H}_m = c_{J,m}, m = 1, \ldots, 2^J$.

In other words, we have transformed the input vector $\{I_{j,k;T}\}_{k=0}^{T-1}$ into the Haar-Fisz output vector $\{\mathscr{H}_{j,k;T}\}_{k=0}^{T-1}$. This Haar-Fisz processing is then replicated at each scale j of the wavelet periodogram.

3.11 Appendix: Detailed Description of Simulation Studies

In this section we give a more detailed description of the simulation studies outlined in Section 3.4. In Section 3.11.1, we describe the adaptive Neyman test (ANT) of Fan and Lin (1998) (see Section 3.2.2) which provides the benchmark for comparison within the simulation studies outlined in Section 3.4. In Section 3.11.2, we describe the basic structure of each simulated experiment and give a detailed description of each model outlined in Sections 3.4.1 and 3.4.2. In Section 3.11.3 we provide results which support the discussion of the hypothesis tests in Section 3.4.

3.11.1 Detailed Description of Adaptive Neyman Test

In this section, we describe the adaptive Neyman test (ANT) of Fan and Lin (1998) (see Section 3.2.2) which provides the benchmark for comparison with the hypothesis testing methodology we develop in this chapter for the simulation studies outlined in Section 3.4. Firstly, we formulate the motivating applied problem within the framework of Fan and Lin (1998). Secondly, the ANT is based on the adaptive Neyman Test Statistic (Fan, 1996). Therefore, in Section 3.11.1.1 we briefly outline the adaptive Neyman Test Statistic before describing the ANT in Section 3.11.1.2.

For our proposed methodology, we model the observed signals using the LSW framework of Nason et al. (2000) (see Section 1.4.2 for details). We denote each individual profile by $\{X_{t,T}^{(i),r_i}\}_{t=0}^{T-1}$ with $i = 1, 2$ corresponding to one of two groups (e.g. control/ treatment) and potential replicates $r_i = 1, \ldots, N_i$ (i.e. N_i circadian traces in the ith group). We then assume that the signals in each group, $i = 1, 2$, are underpinned by a common wavelet spectrum, denoted $S_j^{(i)}(t/T)$ at scales $j \in \overline{1, J}$ ($J = \log_2 T$) and rescaled times $z = t/T \in (0, 1)$.

In contrast with our proposed methodology, the ANT assumes that the observed signals in each group, $i = 1, 2$, are a random sample from the model

$$X_t^{(i),r_i} = f^{(i)}(t) + \epsilon_t^{(i),r_i},$$

for $t = 1, ..., T$ and potential replicates $r_i = 1, \ldots, N_i$ where the random variables $\epsilon_t^{(i),r_i}$ have mean zero and variance $\left(\sigma_t^{(i)}\right)^2$. Fan and Lin (1998) then test whether there is any statistically significant difference between groups of curves by testing

$$H_0 : f^{(1)}(t) = f^{(2)}(t) \quad \text{vs.} \quad H_A : f^{(1)}(t) \neq f^{(2)}(t),$$

based on the observed signals.

3.11.1.1 Adaptive Neyman Test Statistic

The ANT utilises the **Adaptive Neyman Test Statistic** of Fan (1996). The adaptive Neyman test statistic was developed as a high–dimensional hypothesis testing technique. Formally, let $\mathbf{X}$ be an n-dimensional normal random vector with

$$\mathbf{X} \sim N(\theta, I_n).$$

Fan (1996) wish to test

$$H_0 : \theta = \mathbf{0} \quad \text{vs.} \quad H_A : \theta \neq \mathbf{0}. \tag{94}$$

The maximum likelihood ratio test statistic for problem (94) tests all components of $\mathbf{X}$, but this decreases the power of the test (Fan and Lin, 1998). However, if there is a "vague prior" indicating that most of the large absolute values are located on the first m components of θ, then Fan (1996) propose testing only the first m-dimensional subproblem. Fan (1996) then develop a method of determining the parameter m based on power considerations, which leads to the **adaptive Neyman test statistic** (see Fan (1996) for details). Fan and Lin (1998) note that applying the discrete Fourier transform to the observations, $\mathbf{X}$, before implementing the adaptive Neyman test, obtains the required "vague prior".

3.11.1.2 Adaptive Neyman Test

Fan and Lin (1998) utilise the adaptive Neyman test statistic in the development of the ANT. Denote the **standardised difference**:

$$Z_t = \frac{\bar{X}_t^{(1)} - \bar{X}_t^{(2)}}{\left((\hat{\sigma}_t^{(1)})^2 / N_1 + (\hat{\sigma}_t^{(2)})^2 / N_2\right)^{1/2}}, \tag{95}$$

where:

$$\bar{X}_t^{(i)} = \frac{1}{N_i} \sum_{r_i=1}^{N_i} X_t^{(i),r_i} \tag{96}$$

and

$$\left(\hat{\sigma}_t^{(i)}\right)^2 = \frac{1}{N_i - 1} \sum_{r_i=1}^{N_i} \left(X_t^{(i),r_i} - \bar{X}_t^{(i)}\right)^2, \tag{97}$$

for $i = 1, 2$. Fan and Lin (1998) then define the **standardised difference vector** as follows:

$$\mathbf{Z} = (Z_1, \ldots, Z_T)^T, \tag{98}$$

where $\mathbf{v}^T$ denotes the transpose of the vector $\mathbf{v}$.

Fan and Lin (1998) further assume that the random variables, $\epsilon_t^{(i),r_i}, i = 1, 2$ are normally distributed

$$\epsilon_t^{(i),r_i} \sim N\left(0, \left(\sigma_t^{(i)}\right)^2\right)$$

and are independent for all r_i and t. Then, when N_1 and N_2 are "reasonably large", the standardised difference, Z_t has an approximate normal distribution with mean

$$d_t = \frac{f^{(1)}(t) - f^{(2)}(t)}{\left(\left(\sigma_t^{(1)}\right)^2 / N_1 + \left(\sigma_t^{(2)}\right)^2 / N_2\right)^{1/2}}$$

and variance 1. As for our proposed methodology, Fan and Lin (1998) note that when $\left(\sigma_t^{(1)}\right)^2 = \left(\sigma_t^{(2)}\right)^2$, we can use the pooled variance estimates (see Section 3.3.1.1) in equation 95.

In order to obtain the required "vague prior" of the adaptive Neyman test statistic, Fan and Lin (1998) apply the Fourier transform to the standardized difference vector, $\mathbf{Z}$, and denote the resulting vector $\mathbf{Z}^*$. The adaptive Neyman test statistic (Fan, 1996) is then applied to the vector

$\mathbf{Z}^*$ to obtain a p value for the test (as outlined in Fan and Lin (1998)).

3.11.2 Basic Structure of Hypothesis Tests and Model Details

3.11.2.1 Basic Structure

The basic structure of each simulated experiment can be described as follows. In each case, we assumed that the signal was a realisation of length $T = 256$ from one of $i = 1,2$ possible groups, each having (possibly) different spectral structure. A set of $N_1 = N_2 = 1,10,25,50$ signal realisations for each group was generated either from variously defined: spectra (models **P1**–**P5** and **M1** and **M2**); AR processes (models **P6**, **P7**, **M3** and **M4**) or 'function plus noise' time series (models **P8**–**P12** and **M5**).

For the models defined by group spectra (models **P1**–**P5** and **M1** and **M2**), signal realisations were generated using the `locits` package in R (available from the CRAN package repository) and the representation in equation (77) with the Haar wavelet and a Gaussian orthonormal increment sequence with mean zero and unit variance. (Note that the `wavethresh` package in R preceded the `locits` package and can also be used to generate LSW processes. For more information on how to generate LSW processes from a particular spectrum see Nason (2010).)

3.11.2.2 Model Details

In this section we give a detailed description of each model outlined in Sections 3.4.1 and 3.4.2.

1. **P1: Fixed Spectra.** We follow Krzemieniewska et al. (2014) Section 4.1.1- Fixed spectra where the spectra of the two groups differ only at the finest level by 100 coefficients. We simulate each replicate r_i-th time series of length $T = 256$ of the i-th group from the wavelet spectrum $\{S_j^{(i)}(z)\}_{j=1}^J$ which we define for each of the $i = 1,2$ groups as follows:

$$S_j^{(1)}(z) = \begin{cases} 4\cos^2(2\pi z), & \text{for } j = 3, z \in (0,1) \\ 1, & \text{for } j = 7, z \in (1/256, 56/256) \\ 0, & \text{otherwise;} \end{cases} \tag{99}$$

and

$$S_j^{(2)}(z) = \begin{cases} 4\cos^2(2\pi z), & \text{for } j = 3, z \in (0,1) \\ 1, & \text{for } j = 7, z \in (1/256, 156/256) \\ 0, & \text{otherwise.} \end{cases} \tag{100}$$

Figure 40 provides a visualisation of the wavelet spectra (top row) and an example of a signal realisation from each of the two groups (bottom row).

2. **P2: Fixed Spectra-Fine Difference.** For our next study, we modify the setting above such that the spectra of the two groups differ by 6 coefficients (in resolution level 7). Therefore, $\{S_j^{(1)}(z)\}_{j=1}^J$ is as defined in equation (99) above but we specify the evolutionary wavelet

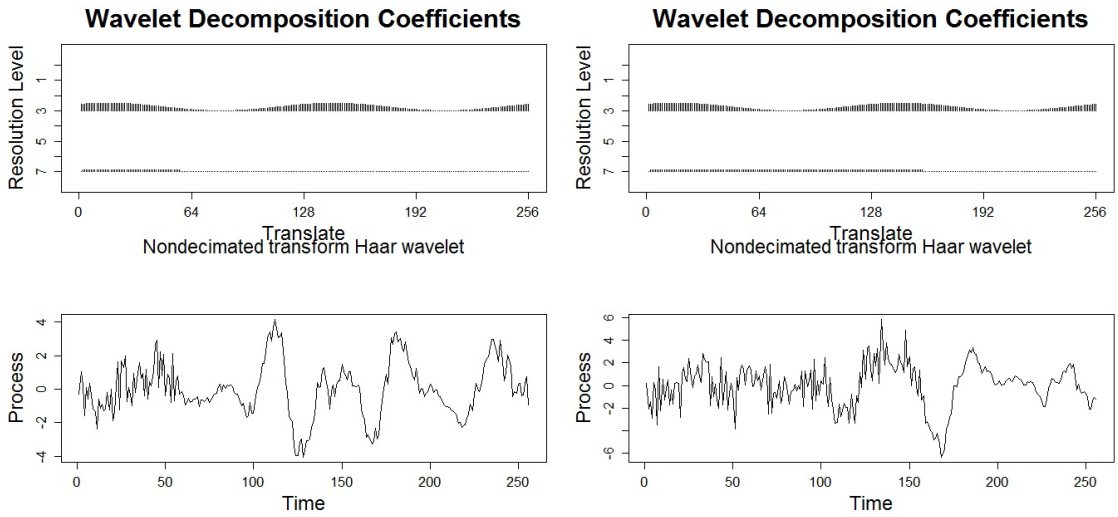

Figure 40: **P1:Fixed Spectra.** Top left: Group 1 wavelet spectrum; Top right: Group 2 wavelet spectrum; Bottom left: Group 1 realisation; Bottom right: Group 2 realisation.

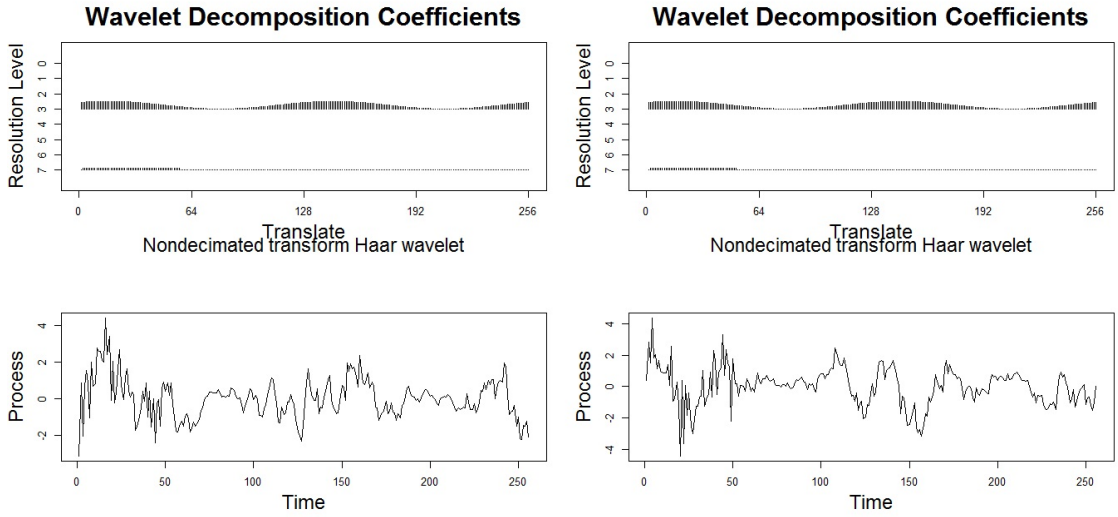

Figure 41: **P2:Fixed Spectra-Fine Difference.** Top left: Group 1 wavelet spectrum; Top right: Group 2 wavelet spectrum; Bottom left: Group 1 realisation; Bottom right: Group 2 realisation.

spectrum $\{S_j^{(2)}(z)\}_{j=1}^J$ as follows:

$$S_j^{(2)}(z) = \begin{cases} 4\cos^2(2\pi z), & \text{for } j = 3, z \in (0,1) \\ 1, & \text{for } j = 7, z \in (1/256, 50/256) \\ 0, & \text{otherwise.} \end{cases} \tag{101}$$

Figure 41 provides a visualisation of the wavelet spectra (top row) and an example of a signal realisation from each of the two groups (bottom row).

3. **P3: Fixed Spectra-Plus Constant.** We now define fixed spectra such that the spectra of the two groups differ by a constant at the finest resolution level. Therefore, $\{S_j^{(1)}(z)\}_{j=1}^J$ is as defined in equation (99) above but we specify the evolutionary wavelet spectrum

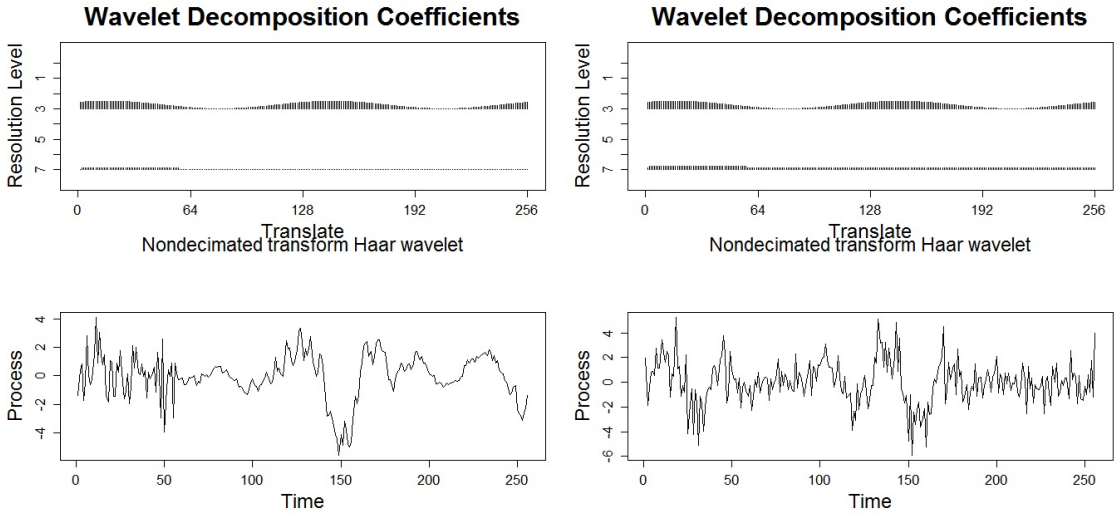

Figure 42: **P3:Fixed Spectra-Plus Constant.** Top left: Group 1 wavelet spectrum; Top right: Group 2 wavelet spectrum; Bottom left: Group 1 realisation; Bottom right: Group 2 realisation.

$\{S_j^{(2)}(z)\}_{j=1}^J$ as follows:

$$S_j^{(2)}(z) = \begin{cases} 4\cos^2(2\pi z), & \text{for } j = 3, z \in (0,1) \\ 2, & \text{for } j = 7, z \in (1/256, 56/256) \\ 1, & \text{for } j = 7, z \in (57/256, 256/256) \\ 0, & \text{otherwise.} \end{cases} \tag{102}$$

Figure 42 provides a visualisation of the wavelet spectra (top row) and an example of a signal realisation from each of the two groups.

4. **P4/P5: Gradual Period Change.** With this simulation study aiming to replicate a typical circadian experiment with changes beyond the stationarity assumption, we define time series as realisations from one of 3 possible groups, each with different spectral characteristics. In particular, each group represents a time series that gradually changes period from 24 to: 25 (Group 1), 26 (Group 2) and 27 (Group 3) over (approximately) two days, before continuing with the relevant period for a further two days. We choose $T = 256$ which is equivalent to a free-running period of 4 days with equally spaced observations every 22.5 minutes. Figure 43 shows the wavelet spectra which display the gradually changing periods that define each of the 3 groups. (Note that the increased period is shown by the movement up through the resolution levels and the gradual increase in period of the wavelet coefficients.) To determine which changes can be discriminated by the methods, we perform two studies within this setting: **P4**: simulations from Group 1 and Group 2 and **P5**: simulations from Group 1 and Group 3.

5. **P6/P7: AR Processes with Time-Varying Coefficients.** The signals in models **P1–P5** are generated from a defined group spectrum, satisfying the underlying LSW modelling assumptions of our proposed tests. The purpose of this study is to asses the performance of our tests when these assumptions are not met. Therefore, we simulate from an important class of nonstationary processes– AR processes with time-varying coefficients.

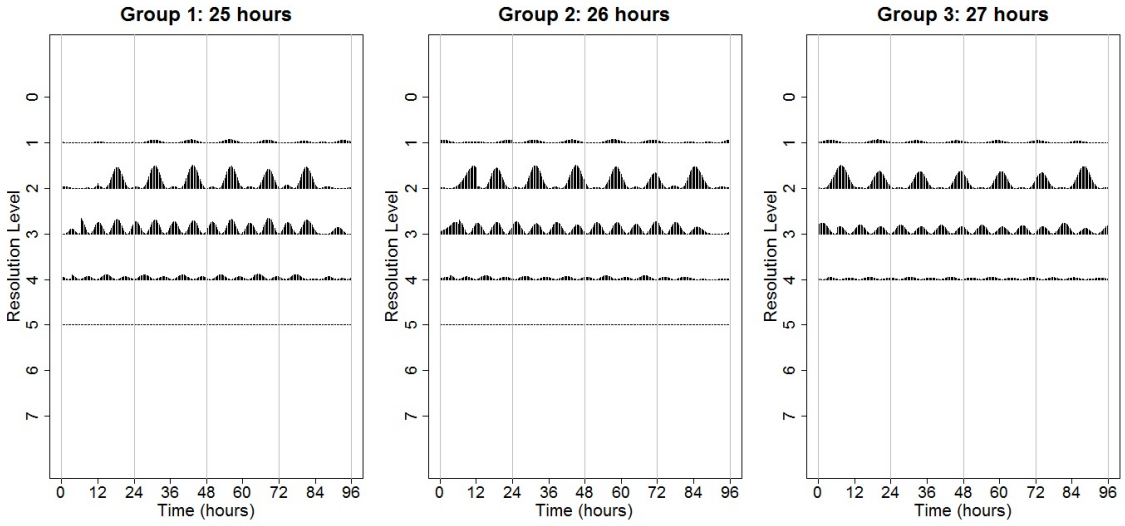

Figure 43: **P4/P5: Gradual Period Change.** Left: Group 1 wavelet spectrum (gradual period change from 24 to 25 hours); Centre: Group 2 wavelet spectrum (gradual period change from 24 to 26 hours); Right: Group 3 wavelet spectrum (gradual period change from 24 to 27 hours).

Time-varying parameters	Time Index	Group $i=1$	Group $i=2$
$\phi_1^{(i)}(t)$	$t=1,\ldots,53$	0.8	0.8
	$t=54,\ldots,128$	-0.9	-0.3
	$t=129,\ldots,256$	0.8	0.8
$\phi_2^{(i)}(t)$	$t=1,\ldots,256$	-0.81	-0.81

Table 18: **P6: AR Processes with Abruptly Changing Parameters.** The abruptly changing parameters of two nonstationary autoregressive processes.

We propose a simulation study in a setting as described in Fryzlewicz and Ombao (2009) Section 4.1 Cases 1 and 2.

P6: AR Processes with Abruptly Changing Parameters. The r_i-th time series from group $i=1,2$, denoted $X_{n,t}^{(i),r_i}$ is generated from the process defined by:

$$X_t^{(i),r_i} = \phi_1^{(i)}(t) X_{t-1}^{(i),r_i} + \phi_2^{(i)}(t) X_{t-2}^{(i),r_i} + \epsilon_t^{(i),r_i}, \tag{103}$$

where the innovations $\epsilon_t^{(i),r_i}$ are independent and identically distributed (iid) Gaussian with zero mean and unit variance. In this study, the squared difference between the group spectra is relatively small and the abruptly changing parameters for the two groups are shown in Table 18. Representative time series plots from each group and the estimated spectra are shown in Figure 44.

P7: AR Processes With Slowly Changing Parameters. The r_i-th time series from group $i=1,2$, denoted $X_t^{(i),r_i}$ is generated from the process defined by:

$$X_t^{(i),r_i} = \phi_1^{(i)}(t) X_{t-1}^{(i),r_i} + \phi_2^{(i)}(t) X_{t-2}^{(i),r_i} + \epsilon_t^{(i),r_i}, \tag{104}$$

where the innovations $\epsilon_t^{(i),r_i}$ are iid Gaussian with zero mean and unit variance. In this study, the group wavelet spectra are highly similar and hence the squared difference between group spectra is relatively small. The slowly changing parameters for groups

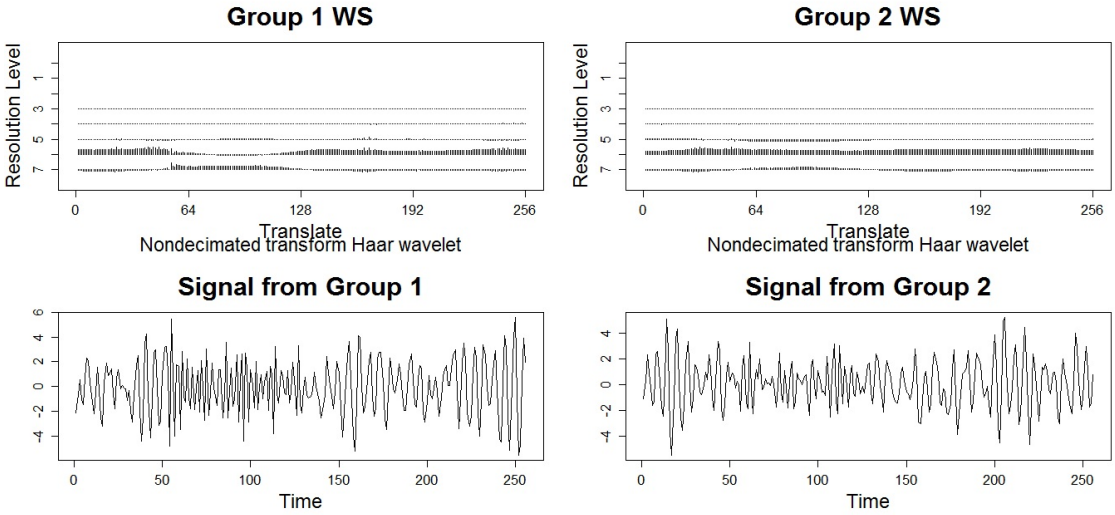

Figure 44: **P6: AR Processes with Abruptly Changing Parameters.** Nonstationary autoregressive processes. Top left: Estimated wavelet spectrum of Group 1; Top right: Estimated wavelet spectrum of Group 2; Bottom left: Group 1 realisation; Bottom right: Group 2 realisation.

Time-varying parameters	Group $i = 1$	Group $i = 2$
$\phi_1^{(i)}(t)$	$-0.8[1-0.7\cos(\pi t/T)]$	$-0.8[1-0.1\cos(\pi t/T)]$
$\phi_2^{(i)}(t)$	-0.81	-0.81

Table 19: **P7: AR Processes With Slowly Changing Parameters.** The slowly changing parameters of two nonstationary autoregressive processes.

$i = 1,2$ are shown in Table 19. Representative time series plots from each group and the estimated spectra are shown in Figure 45.

6. **P8–P12: 'Function Plus Noise' Time Series (Constant Period).** This study follows Zielinski et al. (2014) and generates each time series using an underlying cosine curve with additive noise, which also coincides with the theoretical assumptions of the ANT. As in Models **P4** and **P5**, we choose $T = 256$, which is equivalent to a free-running period of 4 days with equally spaced observations every 22.5 minutes. The r_i-th time series from group $i = 1,2$, denoted $X_t^{(i),r_i}$ is generated from the process defined by:

$$X_t^{(i),r_i} = f^{(i)}(t) + \epsilon_t^{(i),r_i}, \tag{105}$$

where the random variables $\epsilon_t^{(i),r_i}$ are iid Gaussian with zero mean and unit variance and the functions $f^{(i)}(t)$ are defined below. We define time series as realisations from one of 6 possible groups, each with a different (constant) period. The function $f^{(i)}(t)$ is set as a cosine curve with an amplitude of 2 and a period of: 24 hours (Group 1), 21 hours (Group 2), 22 hours (Group 3), 23 hours (Group 4), 23.5 hours (Group 5) and 23.75 hours (Group 6). Representative time series plots and the estimated spectra for Groups 1 and 4 are shown in Figure 46. To determine which period changes can be discriminated by the methods, we perform five studies within this setting: simulations from Group 1 and Groups 2–6 (models **P8–P12** respectively).

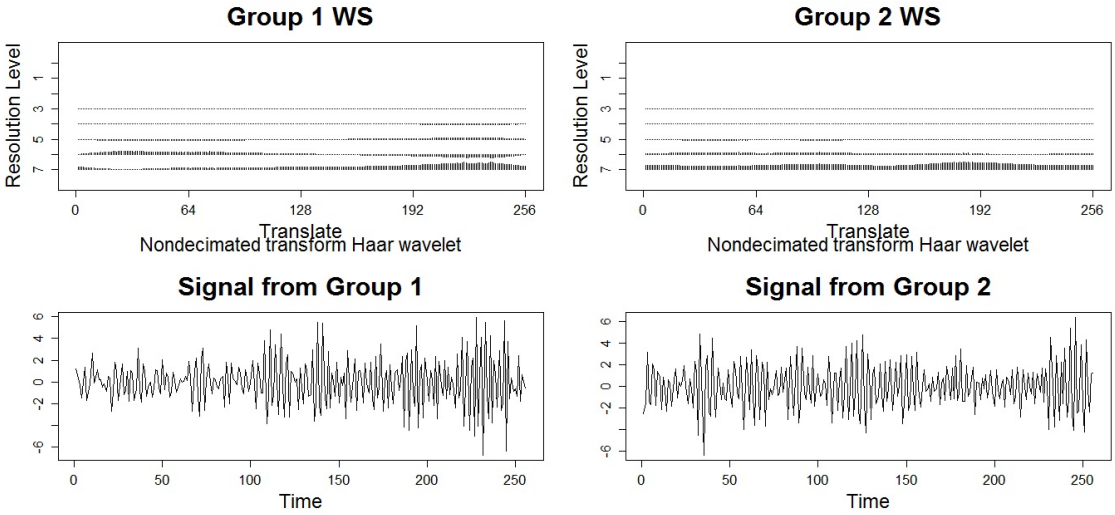

Figure 45: **P7: AR Processes with Slowly Changing Parameters.** Top left: Estimated wavelet spectrum of Group 1; Top right: Estimated wavelet spectrum of Group 2; Bottom left: Group 1 realisation; Bottom right: Group 2 realisation.

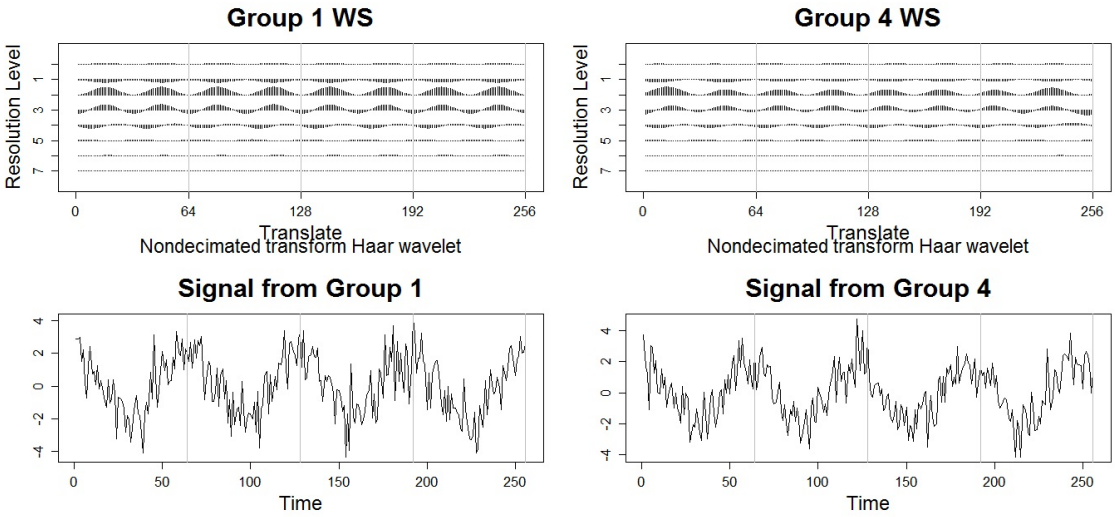

Figure 46: **P10: 'Function Plus Noise' Time Series with Constant Period.** Top left: Estimated wavelet spectrum of Group 1 (24 hour period); Top right: Estimated wavelet spectrum of Group 4 (23 hour period); Bottom left: Group 1 realisation; Bottom right: Group 4 realisation. Grey lines indicate a 24 hour period.

Model	P1	P2	P3	P4	P5	P6	P7	M1	M2	M3	M4
HFT (Bon.)	69.4	3.8	72.6	4.1	51.3	2.5	21.8	2.8	4.1	0.8	1.5
HFT (FDR)	77.7	4.9	79.0	5.4	57.9	15.2	35.9	3.2	4.8	1.7	2.1

Table 20: Simulated power and size estimates (%) for the HFT for models P1-P7 and M1-M4 with nominal size of 5% and $N_1 = N_2 = 1$ realisation from each group.

N	Model	WST (Bon.)	WST (FDR)	FT (Bon.)	FT (FDR)	HFT (Bon.)	HFT (FDR)	HT (Bon.)	HT (FDR)
10	**P1**	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
10	**P2**	3.5	4.6	**51.9**	**54.3**	4.1	6.5	16.9	17.4
10	**P3**	100.0	100.0	100.0	100.0	100.0	100.0	4.2	4.3
10	**P4**	0.5	0.6	8.4	10.8	4.8	7.0	**50.4**	**55.4**
10	**P5**	0.4	1.1	22.6	31.0	73.4	80.2	**95.8**	**98.4**
10	**P6**	**92.2**	**99.7**	14.7	16.4	3.4	30.7	11.6	12.2
10	**P7**	**99.2**	**100.0**	11.5	12.1	30.0	54.7	75.6	77.4
50	**P1**	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
50	**P2**	94.8	97.2	**100.0**	**100.0**	87.1	88.5	**100.0**	**100.0**
50	**P3**	100.0	100.0	100.0	100.0	100.0	100.0	5.3	5.3
50	**P4**	11.8	28.0	96.0	99.0	92.0	94.8	**100.0**	**100.0**
50	**P5**	60.2	86.6	100.0	100.0	100.0	100.0	100.0	100.0
50	**P6**	100.0	100.0	100.0	100.0	96.7	100.0	99.3	99.8
50	**P7**	100.0	100.0	99.0	100.0	100.0	100.0	100.0	100.0

Table 21: Simulated power estimates (%) for models P1-P7 with nominal size of 5%. $N = N_1 = N_2$ is the number of realisations in each group. Highest empirical power estimates are highlighted in bold.

3.11.3 Supplementary Tables

In this section we provide results which support the discussion of the hypothesis tests in Section 2.4. We report the simulated power and size estimates for $N_1 = N_2 = 1, 10, 50$ for the simulation studies outlined in Sections 3.4.1 and 3.4.2 in tables 20 – 23. Additionally, we report the number of rejections for the FT for model **M4** with $N_1 = N_2 = 10$ and 25 and both multiple-hypothesis testing methods in Table 24.

We also report the simulated power and size estimates for $N_1 = N_2 = 25$ for the simulation studies outlined in: Section 3.4.3.1 in Tables 25 and 26 and Section 3.4.3.2 in Tables 27 and 28.

N	Model	Test Group Period	WST (FDR)	FT (FDR)	HFT (FDR)	HT (FDR)	ANT
10	**P8**	21	100.0	100.0	100.0	100.0	100.0
10	**P9**	22	100.0	100.0	93.3	100.0	100.0
10	**P10**	23	100.0	100.0	31.9	100.0	100.0
10	**P11**	23.5	100	96.1	9.5	99.4	100.0
10	**P12**	23.75	81.2	14.6	5.6	32.4	100.0
10	**M5**	24	2.0	2.1	3.1	4.1	7.9
25	**M5**	24	3.0	2.7	2.7	3.5	4.8

Table 22: Simulated size and power estimates (%) for models P8-P12 and M5 with nominal size of 5% and using the false discovery rate procedure (FDR). $N = N_1 = N_2$ is the number of realisations in each group. Note: Control group period is 24 hours in each model.

N	Model	WST (Bon.)	WST (FDR)	FT (Bon.)	FT (FDR)	HFT (Bon.)	HFT (FDR)	HT (Bon.)	HT (FDR)
10	**M1**	0.3	0.5	2.6	3.3	1.0	2.6	2.5	2.7
10	**M2**	0.0	0.2	2.4	3.6	2.0	5.0	3.3	3.3
10	**M3**	0.3	1.2	4.1	4.4	0.2	1.4	1.9	2.1
10	**M4**	0.4	1.6	**5.1**	**5.6**	0.9	1.8	2.1	2.2
50	**M1**	0.4	1.1	2.4	3.9	0.3	2.4	3.1	3.3
50	**M2**	0.3	0.6	3.1	3.8	1.4	3.1	2.5	2.6
50	**M3**	0.5	1.2	4.4	4.8	0.2	2.2	3.9	4.2
50	**M4**	0.2	1.1	4.4	4.8	1.3	2.6	2.8	2.9

Table 23: Simulated size estimates (%) for models M1-M4 with nominal size of 5%. $N = N_1 = N_2$ is the number of realisations in each group. Empirical size estimates over the nominal size of 5% are highlighted in bold.

N	Multiple-hypothesis Testing Method	1 Rej.	2 Rej.	3 Rej.	4 Rej.	>5 Rej.	Modified Empirical Size Estimate
10	Bon.	44	5	2	0	0	0.7
10	FDR	40	12	3	0	1	1.6
25	Bon.	38	8	0	0	0	0.8
25	FDR	31	16	3	2	0	2.1
50	Bon.	39	5	0	0	0	0.5
50	FDR	32	10	3	0	3	1.6

Table 24: **M4: AR Process with Slowly Changing Parameters.** Numbers of rejections in empirical size estimates for the **Raw Periodogram F-Test** (FT), with Bonferroni Correction (Bon.) and false discovery rate (FDR) and with nominal size of 5%. "Modified Empirical Size Estimate" is calculated by examining only cases with more than one significant coefficient.

Model	Test	N(0,1)	t_5	t_3
P1	WST	100.0	100.0	100.0
P1	FT	100.0	100.0	100.0
P1	HFT	100.0	99.9	88.4
P1	HT	100.0	100.0	93.3
P2	WST	48.0	30.5	11.6
P2	FT	100.0	100.0	100.0
P2	HFT	31.8	5.6	1.6
P2	HT	86.4	53.5	17.9
P3	WST	100.0	100.0	100.0
P3	FT	100.0	100.0	100.0
P3	HFT	100.0	100.0	97.0
P3	HT	4.4	2.4	1.9
P4	WST	2.7	1.2	0.6
P4	FT	54.5	49.1	35.7
P4	HFT	36.5	4.7	4.0
P4	HT	100.0	79.7	40.0
P5	WST	14.6	1.1	0.3
P5	FT	99.9	76.6	32.9
P5	HFT	100.0	30.1	11.3
P5	HT	100.0	81.6	38.4
M1	WST	1.3	0.7	0.1
M1	FT	3.1	4.1	**14.5**
M1	HFT	2.0	3.2	1.8
M1	HT	2.7	2.7	0.6
M2	WST	0.6	0.1	0.4
M2	FT	3.9	4.5	4.5
M2	HFT	3.3	2.9	2.1
M2	HT	2.7	0.9	0.9

Table 25: **Potential Non-Gaussian Innovations:** Simulated size and power estimates (%) for models P1-P5 and M1, M2 with nominal size of 5% and $N_1 = N_2 = 25$ realisations from each group. Innovations are distributed as: standard normal (denoted N(0,1)) or t-distribution with 5 or 3 degrees of freedom (denoted t_5, t_3 respectively). For the FT, the modified size and power estimates are recorded (i.e. only consider cases when more than 5 rejections are reported– see Section 3.4.2). Empirical size estimates over the nominal size of 5% are highlighted in bold.

Model	Test	N(0,1)	t_5	t_3
P8	WST	100.0	100.0	100.0
P8	FT	100.0	100.0	100.0
P8	HFT	100.0	100.0	99.6
P8	HT	100.0	100.0	100.0
P9	WST	100.0	100.0	100.0
P9	FT	100.0	100.0	100.0
P9	HFT	100.0	99.9	79.4
P9	HT	100.0	100.0	100.0
P10	WST	100.0	100.0	100.0
P10	FT	100.0	100.0	100.0
P10	HFT	92.0	59.5	25.6
P10	HT	100.0	100.0	100.0
P11	WST	100.0	100.0	100.0
P11	FT	100.0	100.0	100.0
P11	HFT	31.8	15.1	8.1
P11	HT	100.0	100.0	99.5
P12	WST	100.0	98.5	52.4
P12	FT	97.9	83.6	80.0
P12	HFT	9.1	5.9	3.6
P12	HT	98.3	77.2	31.6
M5	WST	3.0	1.0	1.5
M5	FT	2.7	1.7	**10.4**
M5	HFT	2.7	2.0	0.9
M5	HT	3.5	4.2	1.5

Table 26: **Potential Non-Gaussian Errors:** Simulated size and power estimates (%) for models P8-P12 and M5 with nominal size of 5% and $N_1 = N_2 = 25$ realisations from each group. The noise term in equation (105) is distributed as: standard normal (denoted N(0,1)) or t-distribution with 5 or 3 degrees of freedom (denoted t_5, t_3 respectively). For the FT, the modified size and power estimates are recorded (i.e. only consider cases when more than 5 rejections are reported– see Section 3.4.2). Empirical size estimates over the nominal size of 5% are highlighted in bold.

Model	Test	Haar wavelet (1 V.M.)	Daubechies' least-asymmetric (4 V.M.)	Daubechies' extremal phase (10 V.M.)
P1	WST	100.0	100.0	100.0
P1	FT	100.0	100.0	100.0
P1	HFT	100.0	100.0	100.0
P1	HT	100.0	100.0	100.0
P2	WST	48.0	55.7	44.6
P2	FT	100.0	100.0	100.0
P2	HFT	31.8	78.2	73.9
P2	HT	86.4	99.9	99.6
P3	WST	100.0	100.0	100.0
P3	FT	100.0	100.0	100.0
P3	HFT	100.0	100.0	100.0
P3	HT	4.4	4.2	6.0
P4	WST	2.7	23.5	25.1
P4	FT	54.5	91.9	89.4
P4	HFT	36.5	96.5	78.0
P4	HT	100.0	50.8	12.3
P5	WST	14.6	55.3	68.0
P5	FT	99.9	98.6	100.0
P5	HFT	100.0	74.7	99.8
P5	HT	100.0	36.5	52.6
M1	WST	1.3	0.3	0.2
M1	FT	3.1	2.5	2.9
M1	HFT	2.0	2.0	1.6
M1	HT	2.7	1.3	1.8
M2	WST	0.6	0.0	0.2
M2	FT	3.9	1.8	2.8
M2	HFT	3.3	2.8	3.0
M2	HT	2.7	2.6	2.0

Table 27: **Sensitivity to Generation and Estimation Wavelet Mismatch:** Simulated size and power estimates (%) for models P1-P5 and M1, M2 with nominal size of 5% and $N_1 = N_2 = 25$ realisations from each group. In all settings, the Haar wavelet is used for spectral estimation, but the following wavelets are used to generate the true spectra: Haar wavelets, Daubechies' least-asymmetric wavelets with 4 vanishing moments (V.M.) and Daubechies' extremal phase wavelets with 10 vanishing moments, respectively.

Model	Test	Haar wavelet (1 V.M.)	Daubechies' least-asymmetric (4 V.M.)	Daubechies' extremal phase (10 V.M.)
P6	WST	100.0	100.0	100.0
P6	FT	100.0	89.2	100.0
P6	HFT	100.0	89.5	87.6
P6	HT	100.0	68.7	66.3
P7	WST	100.0	100.0	100.0
P7	FT	100.0	92.0	93.0
P7	HFT	100.0	100.0	100.0
P7	HT	100.0	100.0	100.0
P8	WST	100.0	100.0	100.0
P8	FT	100.0	100.0	100.0
P8	HFT	100.0	100.0	100.0
P8	HT	100.0	100.0	100.0
P9	WST	100.0	100.0	100.0
P9	FT	100.0	100.0	100.0
P9	HFT	100.0	100.0	100.0
P9	HT	100.0	100.0	100.0
P10	WST	100.0	100.0	100.0
P10	FT	100.0	100.0	100.0
P10	HFT	92.0	92.5	92.0
P10	HT	100.0	100.0	100.0
P11	WST	100.0	100.0	100.0
P11	FT	100.0	100.0	100.0
P11	HFT	31.8	28.8	32.6
P11	HT	100.0	100.0	100.0
P12	WST	100.0	100.0	100.0
P12	FT	97.9	99.4	98.1
P12	HFT	9.1	7.1	7.9
P12	HT	98.3	98.8	99.1

Table 28: **Sensitivity to the Change of Modelling Wavelet:** Simulated power estimates (%) for models P6-P12 with nominal size of 5% and $N_1 = N_2 = 25$ realisations from each group. Different wavelets are used for the wavelet spectral estimation: Haar wavelets, Daubechies' least-asymmetric wavelets with 4 vanishing moments (V.M.) and Daubechies' extremal phase wavelets with 10 vanishing moments, respectively.

Name (Acronym)	Designed to ...	Strengths	Weaknesses
Wavelet Spectrum Test (WST)	Detect whether two groups display significant differences in the evolution of their spectral structures, and if so, the particular scales and times at which such differences occur.	Utilises CLT-type idea, therefore not sensitive to normality assumption when number of replicates is large.	Power heavily dependent on sample size.
Raw periodogram F-Test (FT)	Detect whether two groups display significant differences in the evolution of their spectral structures, and if so, the particular scales and times at which such differences occur.	Designed for (Gaussian) LSW processes, therefore can identify fine differences between spectra.	Sensitive to normality assumption.
Haar-Fisz Test (HFT)	Detect differences when the total power within a scale differs between groups.	Can identify differences when the total power within a scale differs between groups.	Reduced performance if there is similar overall power within each scale.
Haar Test (HT)	Detect whether groups evolve according to spectra that have the same shape (up to an additive constant) at each scale.	Can identify small differences between spectra.	It needs to be used in conjunction with WST or FT. The plot indicating where significant differences are located in the series is less easy to interpret than the 'barcode' plots of the other tests.

Table 29: A summary of the hypothesis tests developed in this chapter.

3.12 Appendix: Summary Table

Table 29 provides a summary of the hypothesis tests developed in this chapter detailing the test name, its acronym, strengths and weaknesses for each of the proposed tests.

4 Investigating the Effect of Soil Pollution on the Plant Circadian Clock

The methodology developed throughout this thesis was motivated by a specific application in the field of circadian biology– the effect of industrial and agricultural pollutants on the plant circadian clock (Foley et al., 2005; Senesil et al., 1998; Hargreaves et al., 2018; Nicholson et al., 2003). The 'Cerium dataset' that motivated the work in Chapter 2 and the 'Lead dataset' that motivated the development of the raw periodogram F–test in Chapter 3 were taken from a broad investigation of the effect of various salt stresses on plants (Oakenfull et al., 2018). Therefore, in this chapter, we apply the wavelet spectral testing and clustering methodologies to the dataset in Oakenfull et al. (2018), to organize and understand the impact on plant circadian rhythms of a comprehensive range of environmentally relevant pollutants. A key strength of the new methodologies developed in this thesis is that, compared to existing Fourier-based methods, they allow a much more comprehensive investigation of the large datasets encountered in important practical problems, such as the dataset analysed in this chapter.

Thus, the aims of this chapter are to facilitate understanding of the environmental ramifications associated with soil pollution, thereby demonstrating the utility and additional insight our wavelet spectral testing and clustering methodology can provide.

4.1 Introduction and Motivation

Soil pollution is defined as an alteration in the natural soil environment. Some of the most common causes are: industrial activity, application of agricultural chemicals (such as fertilisers and pesticides) and improper disposal of waste. As a result, the growth conditions of many plants are changing in various ways, such as exposure to essential nutrients at toxic levels, or exposure to non—essential elements never before encountered by species in their natural environment (Foley et al., 2005).

As discussed in Section 3.5.1, the circadian clock enhances survival by directing anticipatory changes in physiology, synchronised with environmental fluctuations (Hanano et al., 2006). Therefore, it is vitally important to understand the effects that soil contaminants have on the plant circadian clock (Nicholson et al., 2003). For example, soil contamination of agricultural land typically alters plant metabolism, often causing a reduction in crop yields. Soil contaminants can also have significant consequences for ecosystems. In particular, changes in soil chemistry which effect the numbers and fitness levels of plants will in turn have major consequences for consumer species (and the rest of the food chain) as they respond to changes in the food supply (Foley et al., 2005).

Part 2A of the Environmental Protection Act (1990) developed a procedure for the identification (and treatment) of 'contaminated land' (where contaminated land was defined 'according to whether it poses a significant risk to human health and/or the environment'). The Department for Environment, Food and Rural Affairs (DEFRA) then developed 'Soil Guideline Values' (SGVs) that can be used to determine appropriate concentrations of certain chemicals in soil. Oakenfull et al. (2018) investigated the impact of exposure to the chemicals at the concentrations outlined in this report on the plant circadian clock (see Table 30). However, the SGVs do not comprise an exhaustive list of the potential chemicals that plants can be exposed to in the modern world. In particular, advances in technology utilising the latest developments in material science rely on a growing range of previously unused chemicals (Nicholson et al.,

2003). Therefore, Oakenfull et al. (2018) also investigated the effects of an extensive list of chemicals on the circadian clock of *A. thaliana* (see Tables 35 and 36).

This chapter is organised as follows. Section 4.2 outlines the experimental details that led to the datasets analysed in this chapter. Section 4.3 reports the results of the analysis a circadian biologist would typically use. In Sections 4.4 and 4.5, we apply the wavelet spectral testing developed in Chapter 3 to the motivating circadian datasets before applying the clustering methodology in Section 4.6. Section 4.7 concludes with a brief discussion and suggests topics for further investigation.

4.2 Experimental Details

A comprehensive description of the biological experimental details (carried out in the Davis Lab, University of York) can be found in Oakenfull et al. (2018). Briefly, each dataset was obtained following the method outlined for the Lead dataset in Chapter 3 (Appendix 3.7). However, we generalise this method to include other salt stresses as follows: six-day-old seedlings were transferred to 96 well microtiter plates containing Hoagland's media (Hoagland et al., 1950) with or without a supplemental chemical at a specific concentration. Therefore, each microtiter plate comprises a control group and 3 chemical treatment groups (each containing 24 plants). A full list of the exact chemicals used and their concentrations can be found in Tables 30, 35 and 36.

For the elements described in the DEFRA guidelines (henceforth referred to as the 'DEFRA chemicals'), the maximum permissible concentration was tested (denoted 'Max') as well as half of the maximum concentration (denoted 'Half') for the Ph of the media used (5.5<6.0). Note that the 'Lead dataset' from Chapter 3 corresponds to the Lead (Max) group from this investigation. For the remaining elements, multiple concentrations were tested: the final concentration for each chemical (appearing in Tables 35 and 36) was the maximum concentration possible before becoming toxic to the plant. For each element, more than one compound was tested (where possible), with the intention of helping to establish whether the effects on the clock were due to the anion or cation of each compound.

A control group was included on each microtiter plate for a number of reasons. Firstly, since we are investigating the effect of exposure to a particular chemical, we should compare it to a control group that was not exposed to the chemical, but otherwise experienced identical growth conditions. In particular, in the control groups in Chapter 2, we noted individual–level variability in plant response to stimuli, despite their sharing identical genetic characteristics (Doyle et al., 2002) which we concluded may be due to the individual plants in some instances showing a stress response, perhaps induced by the experimental method itself (Hargreaves et al., 2018). This demonstrates that it is of the utmost importance that the control and treatment groups should experience identical growth conditions, and therefore the same stresses, as otherwise the specific experimental stress response may be confounded with the effect of the chemical treatment. Furthermore, the machine that measures the luminescence (a TopCount NXT scintillation counter (Perkin Elmer)– see Chapters 2 and 3) iterates through multiple microtiter plates within an experiment. Therefore, the exact time of a given observation is not identical for each plate (see for example the slightly different timings in Figures 47 and 48). Since we will use the coefficients of the evolutionary wavelet spectrum (which are indexed by time) to compare the control and treatment groups, the timings of the observations

should also be identical for both groups.

4.3 Traditional Fourier Analysis

As discussed in previous chapters, period estimation has traditionally been central to the analysis of circadian data (see for example Perea-García et al. (2016a), Costa et al. (2011)). Oakenfull et al. (2018) used the Microsoft Excel macro BRASS (see Section 1.3.2) to produce period estimates for the control and treatment groups respectively (using FFT–NLLS analysis (see Section 1.3.2.2) over a window of ZT36 to 120, considering only period estimates between 15 and 40 hours). For each concentration of DEFRA chemical, Table 30 shows: the mean period estimate; the difference in the mean period estimates for the (appropriate) control and treatment group; the number of plants that were analysed and the mean relative amplitude error (RAE). (Recall: RAE is a value between 0 and 1 and gives information about the goodness of fit of the model with a value of 0 indicating that the estimated cosine curve perfectly fits the data–see Chapters 1 and 2 for details.)

Hypothesis testing (a two-tailed t-test at the 5% significance level) was then used to compare the control and treatment period estimates (see for example Perea-García et al. (2016a)) and the results for the DEFRA chemicals can also be found in Table 30. This analysis found significant differences in period for 6 out of the 24 treatment groups: Zinc (Max), Selenium (both concentrations), Molybdenum (both) and Lead (Half). Figure 47 displays the individual time series for these chemicals.

4.3.1 Discussion of Findings

The results of the BRASS analysis in Table 30 suggest that Zinc (Max) and Selenium (both) increase period whereas Molybdenum (both) and Lead (Half) decrease period. To an extent, this is supported by the individual time series in Figure 47, as the average time series for the Zinc (Max) and Selenium (both) treatment groups appear to display an increased period and the average time series for Molybdenum (Half) seems to have a shorter period. However, the rhythmic behaviour of the control and treatment groups does not appear to be accurately described by a single cosine curve with a constant period (the period and amplitude of all time series appear to gradually change throughout the experiment).

The results in Table 30 indicate that Molybdenum (Max) causes a significant decrease in period (of approximately 3 hours) with an RAE of 0.65. In the circadian community, standard practice dictates that results with an RAE value above the threshold of 0.4 are discarded (Doyle et al., 2002). Therefore, this finding would not be considered as statistically reliable using existing Fourier–based methods. The decision to discard this result is validated upon examining Figure 50– the Molybdenum (Max) time series appear to have a shorter period before becoming what is known in the circadian community as 'arrhythmic' (after approximately 48 hours). This non–sinusoidal behaviour could explain the high RAE and confirms that these time series should not be modelled by a single cosine curve with a constant period. This highlights the urgent need for more statistically advanced approaches to analyse these types of data.

In Chapters 2 and 3, we have repeatedly seen that some changes are not detected by BRASS, even though qualitative differences can be noted by eye. Therefore, Figure 48 displays the individual time series for a selection of the DEFRA chemicals which were not identified as causing a significant change in period. Of the time series displayed in Figure 48, only Cadmium (Half)

Treatment	Chemical	Concentration	Period Estimate (hours)	Period Difference	RAE	Number Analysed
Fluorine	NaF	26mM (Max)	29.33	3.27	**0.56**†	3
Fluorine	NaF	13mM (Half)	28.45	0.39	0.18	22
Chromium	$KCr(SO_4)_2$	7mM (Max)	34.04	NA	NA	1
Chromium	$KCr(SO_4)_2$	3.5mM (Half)	25.89	-1.18	**0.60**†	5
Nickel	$NiCl_2$	10mM (Max)	29.31	0.96	**0.51**†	4
Nickel	$NiCl_2$	500μM (Half)	29.23	1.41	**0.53**†	5
Copper	$CuSO_4$	1.6mM (Max)	30.82	2.82	**0.92**†	4
Copper	$CuSO_4$	800μM (Half)	24.98	-2.66	**0.88**†	3
Zinc	$ZnSO_4$	3mM (Max)	27.97	**0.56**∗	0.17	24
Zinc	$ZnSO_4$	1.5mM (Half)	27.74	0.15	0.14	22
Arsenic	$KAsO_4$	670μM (Max)	29.13	1.94	**0.42**†	15
Arsenic	$KAsO_4$	335μM (Half)	28.78	1.59	0.31	24
Selenium	Na_2SeO_4	40μM (Max)	31.63	**3.83**∗	0.21	19
Selenium	Na_2SeO_4	20μM (Half)	29.59	**2.48**∗	0.20	22
Molybdenum	Na_2MoO_4	4mM (Max)	24.86	**-3.18**∗	**0.65**†	11
Molybdenum	Na_2MoO_4	2mM (Half)	23.89	**-3.99**∗	0.32	21
Cadmium	$CdCl_2$	26μM (Max)	27.19	0.17	0.22	23
Cadmium	$CdCl_2$	13μM (Half)	27.46	0.38	0.22	24
Cadmium	$CdSO_4$	26μM (Max)	26.96	-0.32	0.20	24
Cadmium	$CdSO_4$	13μM (Half)	27.19	0.28	0.21	24
Mercury	$HgCl_2$	5μM (Max)	26.94	-0.06	0.15	23
Mercury	$HgCl_2$	2.5μM (Half)	27.43	0.13	0.18	23
Lead‡	$Pb(NO_3)_2$	1.4mM (Max)	26.82	-0.62	0.32	21
Lead	$Pb(NO_3)_2$	700μM (Half)	26.74	**-0.70**∗	0.20	23

Table 30: **BRASS Results– DEFRA Chemicals**. Summary of the output of the analysis of the DEFRA chemicals in BRASS. "Treatment" represents the element under investigation within the chemical compound. ∗ indicates a significant change in period from the respective control group. † denotes an RAE value above the 0.4 threshold. "Number Analysed" is the number of time series for which BRASS was able to return a period estimate. There are 24 plants in each treatment group. ‡ Note that the Lead (Max) treatment group coincides with the 'Lead dataset' from Chapter 3.

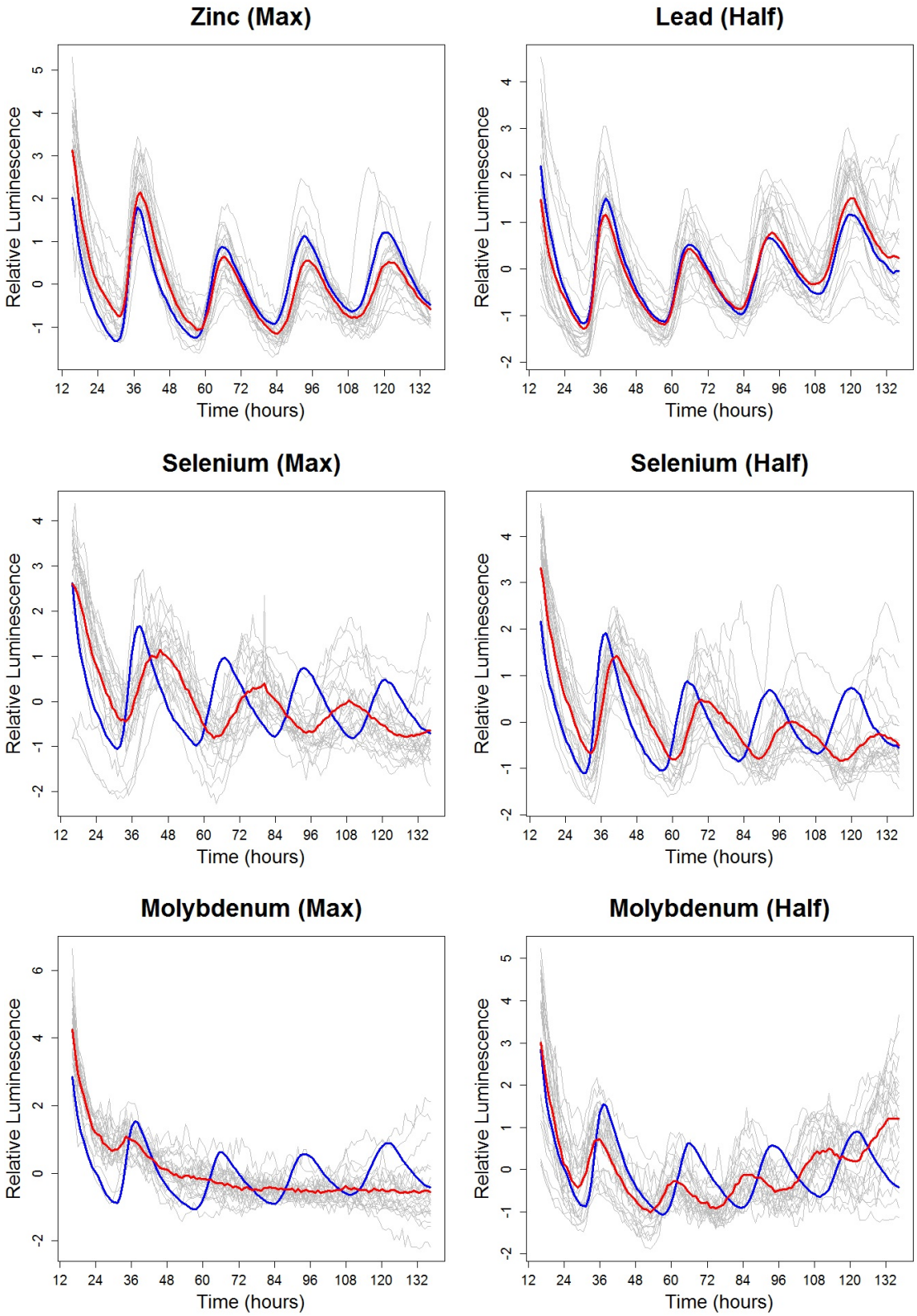

Figure 47: **DEFRA Chemicals:** Luminescence profiles over time for *A. thaliana* plants exposed to a selection of the DEFRA chemicals. Each Panel: Individuals in the chemical treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero.

and Mercury (Max) appear to have no significant effect. This again demonstrates that more statistically advanced approaches to analyse these types of data.

The results of the Fourier analysis in Table 30 suggest that a higher concentration of lead has no effect on the circadian clock of *A. thaliana*, whereas the lower concentration does. The individual time series in Figures 47 and 48 do not support this conclusion, in fact, quite the opposite. The Lead (Half) treatment group (Figure 47) does not visually appear to be significantly different from the control but the Lead (Max) treatment group (Figure 48) does. (Recall the 'Lead dataset' and associated discussion from Chapter 3 and note that the Lead (Max) treatment group is equivalent to this dataset.)

Three of the remaining treatment groups in Figure 48 appear to display similar behaviour to certain series in Figure 47, yet the FFT–NLLS analysis found no significant difference in period. For example, Arsenic (Max) (Figure 48) seems to display similar behaviour to Molybdenum (Max) (Figure 47)– the average time series appears to have a similar period to the control (albeit with a different amplitude), followed by a slight increase in period, before becoming arrhythmic (after approximately 60 hours). Furthermore, only 15 (out of 24) time series in the Arsenic (Max) group were analysed by BRASS giving a mean RAE of 0.42 (which is above the threshold of 0.4, indicating a poor fit).

These examples illustrate that the time series arising from this circadian experiment display nonstationary behaviour (Price et al., 2008; Hargreaves et al., 2018) such as changes in both period and amplitude. Therefore, traditional methods that assume a rhythm of fixed period and amplitude and determine period length from experimental datasets are not appropriate (see Leise et al. (2013) and Chapter 1) and can lead to inaccurate results and misleading conclusions (Harang et al., 2012; Hargreaves et al., 2018).

Visual inspection of Figure 48 shows that Copper (Max) caused the clock to become arrhythmic yet this was not detected by the BRASS analysis. The reported difference in period estimates (Table 30) instead indicates that Copper (Max) increased period (which does not seem credible) with an RAE of 0.92. This could be due to the constraints imposed on the FFT–NLLS procedure (to only consider period estimates between 15 and 40 hours) which clearly are not appropriate for this dataset. Therefore, this example highlights another flaw with this methodology– high–frequency behaviour cannot be captured using BRASS (also see the Ultradian dataset in Chapter 3 and associated discussion).

4.3.2 Testing for Stationarity

As discussed in Chapter 3, one limitation of the traditional Fourier analysis is that the employed methodology does not typically evaluate the crucial underpinning assumption of data stationarity. In Chapter 3, we noted that the Lead (Max) dataset displayed a number of nonstationary features, so we investigated whether the individual time series in the Lead (Max) dataset were (second–order) stationary via hypothesis testing. We found that over 80% of the time series provided evidence to reject the null hypothesis of stationarity. Similarly, in Section 4.3.1 above we discussed the nonstationary features of a number of time series from the DEFRA chemical dataset. For example, the period and amplitude of the mean time series of the (Arsenic) control and Arsenic (Half) groups appeared to change throughout the experiment. Therefore, we investigated whether the time series in the DEFRA chemical dataset are (second-order) stationary. We employed the Priestley-Subba Rao test (Priestley and Rao, 1969) and a selection of the

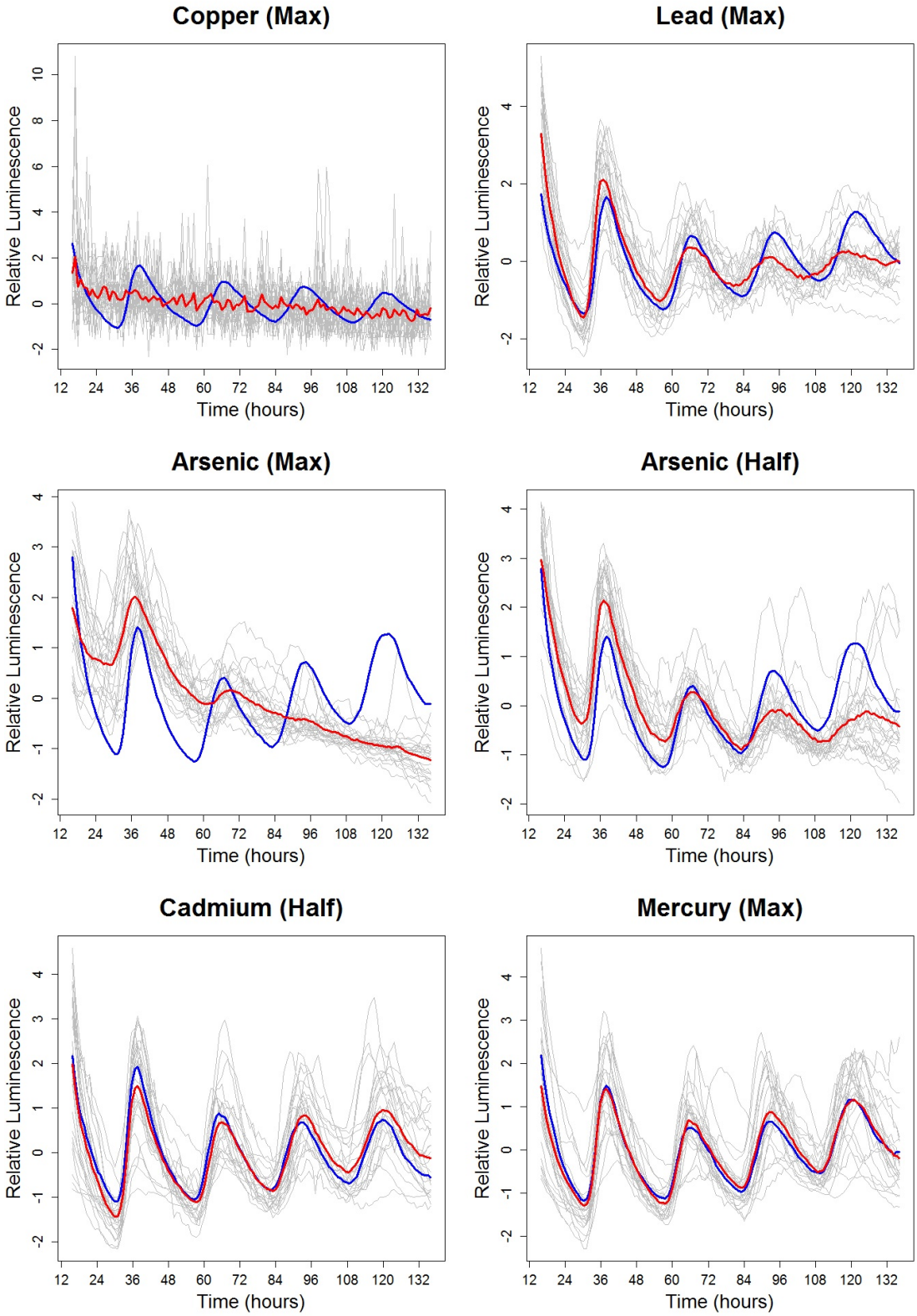

Figure 48: **DEFRA Chemicals:** Luminescence profiles over time for *A. thaliana* plants exposed to a selection of the DEFRA chemicals. Each Panel: Individuals in the chemical treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero.

Treatment Group	Control (Arsenic)	Arsenic (Max)	Arsenic (Half)	Copper (Max)
Number of nonstationary time series	19 (79%)	24 (100%)	23 (96%)	6 (25%)

Table 31: Results for the Priestley-Subba Rao test of stationarity, implemented in the `fractal` package in R and available from the CRAN package repository. Number of nonstationary time series indicates the number of time series (in each treatment group) with enough evidence to reject the null hypothesis of stationarity at the 5% significance level (as a percentage in brackets).

results can be found in Table 31.

Table 31 confirms our assertion that both arsenic treatment groups display nonstationary behaviour (Section 4.3.1). We also note that 79% of the (Arsenic) control group provided enough evidence to reject the null hypothesis of stationarity. These results support the argument in Section 4.3.1, that the rhythmic behaviours of the time series arising from this experiment do not appear to be accurately described by a single cosine curve with a constant period and amplitude. Therefore, the application of the current Fourier–based methodology (which assumes data stationarity) would be inappropriate for these time series. This highlights the urgent need for more statistically advanced approaches for formal spectral comparison.

4.4 Wavelet Spectral Testing Using the Methodology Developed in Chapter 3

FFT-NLLS analysis with software packages such as BRASS or BioDare assumes that time series are stationary and can be represented by sinusoidal waveforms. However, we have demonstrated throughout this chapter that many of the time series in the DEFRA chemicals dataset displayed broadly periodic behaviour, but with time-varying period and amplitude, conducive to a time–evolving period. Therefore, we now use the methodology developed in Chapter 3 for the formal spectral comparison of nonstationary time series to analyse the effects of the DEFRA chemicals.

In this investigation we want to determine whether each DEFRA chemical affects the *Arabidopsis thaliana* circadian clock and, if so, when and how? Hence, we choose to use the raw periodogram F-Test ('FT'), which was developed in Chapter 3 to detect whether the two groups display significant differences in the evolution of their spectral structures, and if so, to identify the scales and times at which such differences occur.

For wavelet representations, the data is often required to be of dyadic length, $T = 2^J$. Therefore, as in Chapter 3, our approach is to analyse a (dyadic length) segment of the data, with the truncation decided after consultation with the experimental scientists. We then model each circadian time series as an LSW process, estimate its corresponding group-average raw wavelet periodogram and then construct the test statistic proposed in equation (87). For each DEFRA chemical and concentration, the corresponding number of rejections of spectral equality between the treatment and control groups can be found in Table 32, with a selection of the corresponding representative 'barcode' plots in Figures 50 and 51.

4.4.1 Discussion of Findings

In this section, we present the results of the wavelet spectral testing methodology proposed in Chapter 3 and compare them with the results in Section 4.3 which represent the traditional

Treatment	Chemical	Concentration	Number of Rejections FT (FDR)	Period Difference
Fluorine (F)	NaF	26mM (Max)	501 (56%)	3.27
Fluorine (F)	NaF	13mM (Half)	15 (2%)	0.39
Chromium (Cr)	$KCr(SO_4)_2$	7mM (Max)	594 (66%)	NA
Chromium (Cr)	$KCr(SO_4)_2$	3.5mM (Half)	544 (61%)	-1.18
Nickel (Ni)	$NiCl_2$	10mM (Max)	534 (60%)	0.96
Nickel (Ni)	$NiCl_2$	500μM (Half)	498 (56%)	1.41
Copper (Cu)	$CuSO_4$	1.6mM (Max)	475 (53%)	2.82
Copper (Cu)	$CuSO_4$	800μM (Half)	442 (49%)	-2.66
Zinc (Zn)	$ZnSO_4$	3mM (Max)	90 (10%)	**0.56**$*$
Zinc (Zn)	$ZnSO_4$	1.5mM (Half)	**3 (0%)**†	0.15
Arsenic (As)	$KAsO_4$	670μM (Max)	458 (51%)	1.94
Arsenic (As)	$KAsO_4$	335μM (Half)	123 (14%)	1.59
Selenium (Se)	Na_2SeO_4	40μM (Max)	196 (22%)	**3.83**$*$
Selenium (Se)	Na_2SeO_4	20μM (Half)	198 (22%)	**2.48**$*$
Molybdenum (Mo)	Na_2MoO_4	4mM (Max)	346 (39%)	**-3.18**$*$
Molybdenum (Mo)	Na_2MoO_4	2mM (Half)	284 (32%)	**-3.99**$*$
Cadmium (Cd)	$CdCl_2$	26μM (Max)	**3 (0%)**†	0.17
Cadmium (Cd)	$CdCl_2$	13μM (Half)	**1 (0%)**†	0.38
Cadmium (Cd)	$CdSO_4$	26μM (Max)	**1 (0%)**†	-0.32
Cadmium (Cd)	$CdSO_4$	13μM (Half)	**1 (0%)** †	0.28
Mercury (Hg)	$HgCl_2$	5μM (Max)	**1 (0%)**†	-0.06
Mercury (Hg)	$HgCl_2$	2.5μM (Half)	**1 (0%)**†	0.13
Lead (Pb)	$Pb(NO_3)_2$	1.4mM (Max)	133 (15%)	-0.62
Lead (Pb)	$Pb(NO_3)_2$	700μM (Half)	**1 (0%)**†	**-0.70**$*$

Table 32: **FT (FDR) results– DEFRA Chemicals**. The number of rejections (as a percentage in brackets) for the FT with FDR (at the 5% significance level) for the DEFRA Chemicals with † denoting 0% rejections. "Treatment" represents the element under investigation within the chemical compound. The estimated mean difference in period (using FFT–NLLS) is also shown for reference with $*$ indicating a significant change in period from the respective control group.

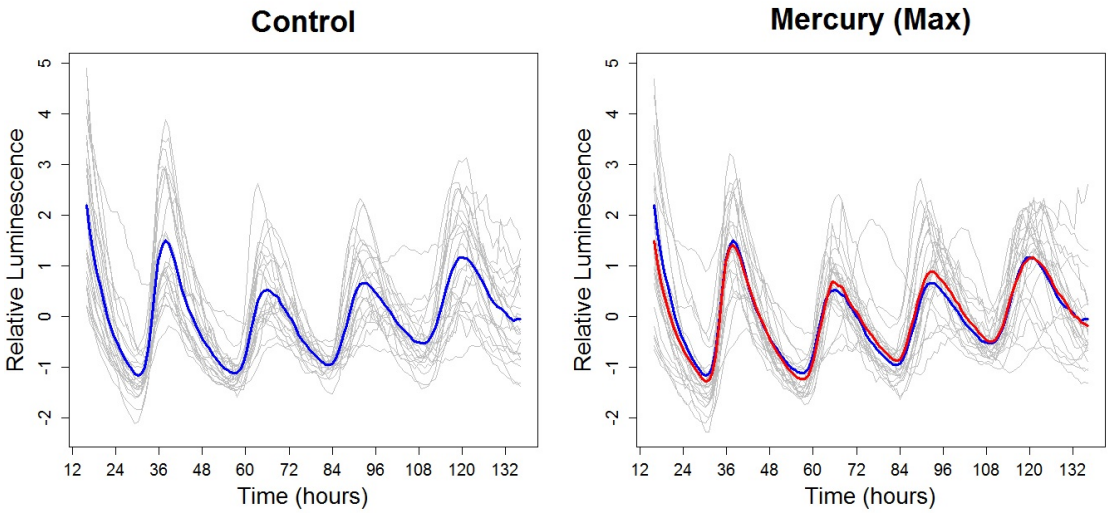

Figure 49: **Mercury (Max):** Luminescence profiles over time for untreated *A. thaliana* plants (denoted 'Control') and those exposed to mercuric chloride ($HgCl_2$) at a concentration of 5μM (denoted 'Mercury (Max)'). Left: Individuals in the control group (in grey) along with the group average (blue). Right: Individuals in the Mercury (Max) treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero.

Fourier–based analysis a circadian biologist would typically perform. We begin with presenting examples of datasets where the wavelet spectral testing supports the results of the classical BRASS analysis. This also allows us to demonstrate the additional insight our proposed methodology can provide. We then discuss examples where the wavelet spectral testing does not support the BRASS analysis but confirms results that were visually apparent. This highlights another advantage of our proposed methodology over traditional methods– it can discriminate between real data sets where the current methodology cannot.

4.4.1.1 Examples of the FT Supporting the Classical Analysis

The results in Table 32 indicate that the FT found very few rejections of the null hypothesis of spectral equality for 8 DEFRA chemicals. As discussed in Chapter 3, circadian scientists often choose to disregard situations where very few coefficients are significantly different and this is our approach here. Throughout this chapter, we will not infer that chemicals cause a significant change to the spectral behaviour when the percentage of rejections is 0 (indicated by † in Table 32). This result is supported upon visual examination of the raw time series (see, for example, Figure 48). Figure 49 displays the raw time series for both the control group (untreated *A. thaliana* plants), as well as for those in the Mercury (Max) group. The raw time series in Figure 49 show very small differences between the control and treatment groups that do not appear to be significant. Therefore, this example illustrates that the FT supports the visually–apparent result that these 8 chemicals have no effect on the circadian clock of *A. thaliana*. This result is also supported by the BRASS analysis as, excluding Lead (Half) (see discussion below), all of the remaining 7 chemicals (which the FT found had no significant effect) also corresponded to a small change in period (using FFT–NLLS) that was not statistically significant.

Table 30 shows that the Fourier analysis (using FFT–NLLS implemented in BRASS) found significant differences in period for 6 out of the 24 treatment groups: Zinc (Max), Selenium

(both concentrations), Molybdenum (both) and Lead (Half). To an extent, these findings are reinforced by the FT (FDR) as, for all these chemicals (other than Lead (Half)), the FT found a number of significant differences (over 0%), which indicates that these chemicals have an effect on the circadian clock of *A. thaliana*.

As discussed in Section 3.3.1, practitioners can also be (cautiously) informed by the number of rejections of the null hypothesis of spectral equality, with larger values potentially indicating a greater departure from the null hypothesis. In this investigation, this could suggest that a chemical has a greater effect on the circadian clock of *A. thaliana*. For example, the FFT-NLLS analysis (Table 30) and time series (Figure 47) indicated that Zinc (Max) caused a small (yet statistically significant) increase in period whereas selenium (both concentrations) caused a larger (statistically significant) increase in period. These results were reinforced by the FT (FDR), which found relatively few significant differences (10%) between the control and treatment group spectra for Zinc (Max) and found a large number of significant differences (22%) between the treatment and control group spectra for both concentrations of selenium (Table 32). However, as discussed in Section 3.3.1, there are a number of factors which could influence the number of rejections, therefore, these values should be treated with caution.

In contrast with the traditional Fourier–based analysis which is limited to identifying a fixed change in period of the circadian component of a signal, the FT can provide additional insight by identifying the time point at which the control and treatment groups start to have different circadian rhythms. For example, the FFT-NLLS analysis (Table 30) found that Molybdenum (Half) caused a significant decrease in period. This result was supported by the FT (FDR) which found a number of significant differences between the spectra (32% (Table 32), which is greater than the 0% threshold, see discussion above). Figure 47 visually indicated that this difference in period manifested itself after 24 hours and the barcode plot in Figure 50 supports this assertion as the rejections of spectral equality occur after ZT24.

4.4.1.2 Examples of the FT Not Supporting the Classical Analysis

There are also a number of instances where the wavelet spectral testing does not coincide with the Fourier–based analysis. For example, the BRASS analysis (Table 30) of Lead (Half) reported a small but significant decrease in period, though this was not visually apparent in the raw time series (Figure 47). However, the FT (FDR) found 0% rejections of the null hypothesis of spectral equality. In Section 4.4.1.1, we stated that we will not infer that a chemical causes a significant change to the spectral behaviour when the percentage of rejections is 0. Therefore, the FT supports the result that was visually apparent– Lead (Half) has no effect on the circadian clock of *A. thaliana*.

We now analyse the chemicals in Figure 48 which BRASS reported as causing no significant effect on the circadian clock of *A. thaliana*. Recall: of the time series displayed in Figure 48, only Cadmium (Half) and Mercury (Max) appeared to have no significant effect. Again, the FT supports this (intuitive) result. Table 32 indicates that Cadmium (Half) and Mercury (Max) have no effect on the circadian clock of *A. thaliana* (with 0% rejections) whereas the remaining chemicals in Figure 48 all had a number of rejections of the null hypothesis of spectral equality.

Figure 48 visually indicated that Copper (Max) caused the clock to become arrhythmic (indicated by high–frequency behaviour throughout the experiment). The FT (FDR) found a number of significant differences between the spectra (53% (Table 32), which is greater than the 0%

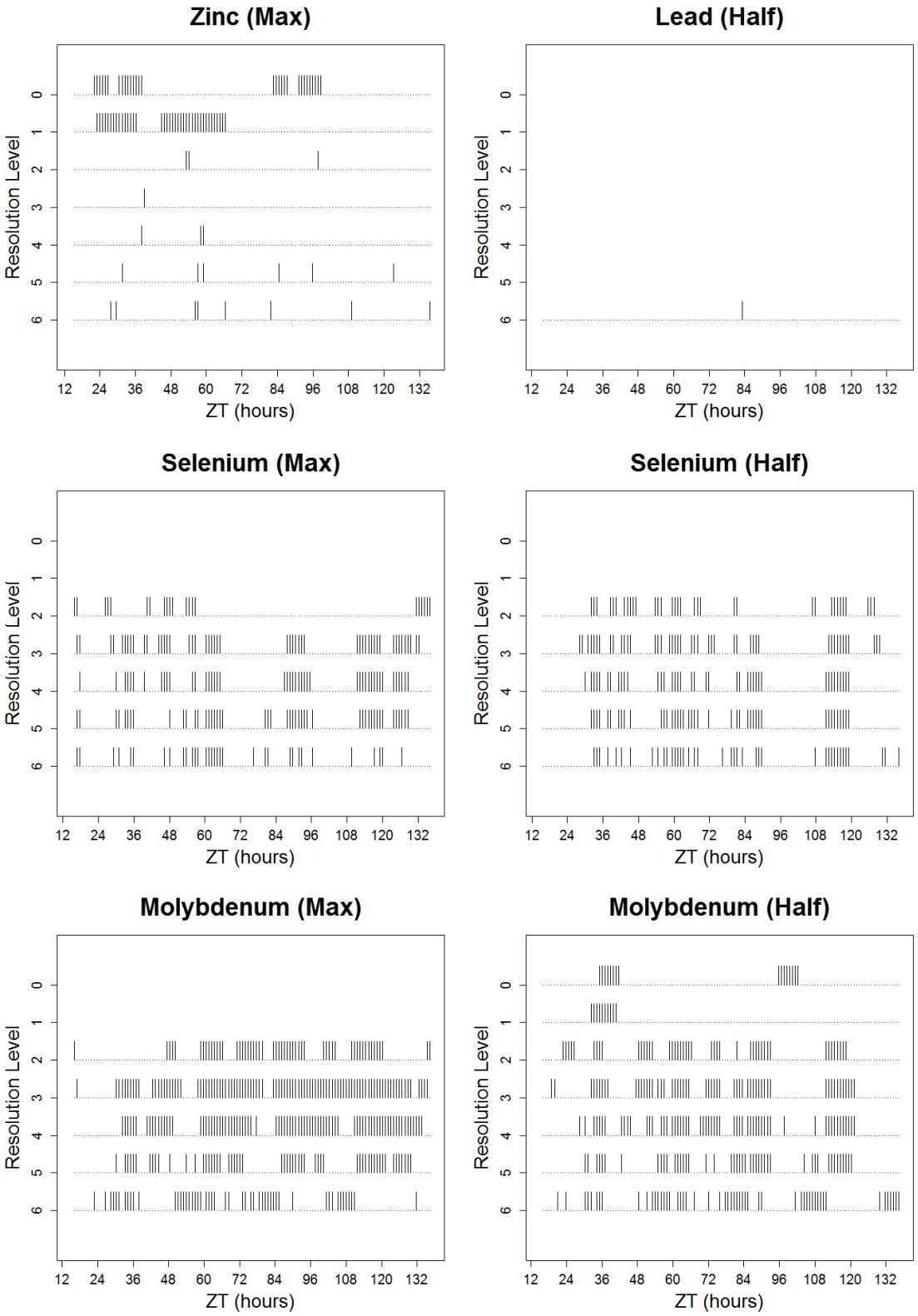

Figure 50: **'Barcode' plots** for FT (with FDR) for the time series shown in Figure 47.

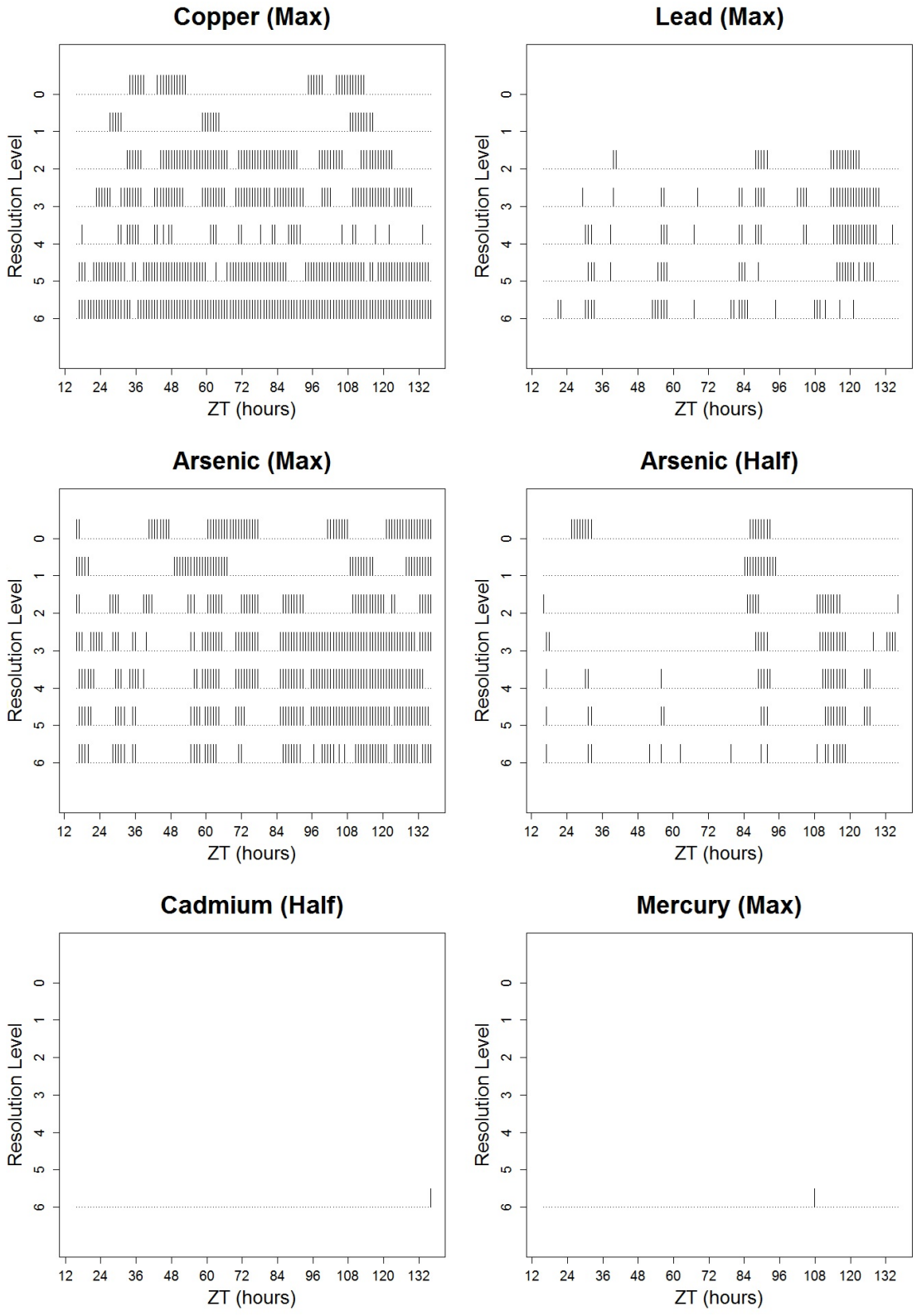

Figure 51: **'Barcode' plots** for FT (with FDR) for the time series shown in Figure 48.

threshold, see Section 4.4.1.1). The barcode plot in Figure 51 indicates that these differences are located in the coarsest resolution levels 1–4, associated with circadian rhythms, and higher-frequency levels 6 and 7, corresponding to an ultradian rhythm (see Chapter 3). We conclude that there is evidence that the Copper (Max) alters the circadian and ultradian rhythms within *A. thaliana*.

Arsenic (Half) appeared to cause a period lengthening effect (Figure 48) and the FFT-NLLS analysis supported this (though the result was not statistically significant). The FT (FDR) found enough evidence to reject the null hypothesis of spectral equality at the 5% level (14% of coefficients tested found to be significantly different (Table 32), which is greater than the 0% threshold, see Section 4.4.1.1). The time series (Figure 48) also indicated that the change in periodicity only occurred after ZT84 and this is reflected in the barcode plot in Figure 51 where we note that most significant differences appear after ZT84. We conclude that there is evidence that Arsenic (Half) does affect the circadian clock of *A. thaliana*, and this change manifests itself after approximately three days of free-running conditions.

Arsenic (Max) appeared to spike at ZT36 before increasing in period while decreasing in amplitude at ZT72 and becoming arrhythmic (see Figure 48). The FFT–NLLS analysis found no significant difference in period between the two groups whereas the FT (FDR) found a number of significant differences between the spectra (51% (Table 32), which is greater than the 0% threshold, see Section 4.4.1.1). Furthermore, the barcode plot reflects the visual differences noted above, with rejections of the null hypothesis of spectral equality (between the control and treatment group) located (at all scales) at ZT36, ZT72 and after ZT84 in Figure 51. We conclude that there is evidence that Arsenic (Max) alters circadian and ultradian rhythms within *A. thaliana*.

4.4.2 Conclusions

In Section 4.1, we introduced the 'Soil Guideline Values' (SGVs) that can be used to determine appropriate concentrations of certain chemicals in soil. We also recall that 'contaminated land' was defined 'according to whether it poses a significant risk to human health and/or the environment' (Environmental Protection Act, 1990). In Sections 3.5.1 and 4.1 we noted that altered plant circadian clocks (e.g. due to changes in the plant's chemical environment) could have a large impact on the numbers and fitness levels of plants, which would in turn have major consequences for consumer species (and thus entire ecosystems) as they respond to a reduction in the food supply (Foley et al., 2005). Thus, if a certain chemical at a particular concentration affects the plant circadian clock, it would indeed pose a 'significant risk to... the environment' and hence would satisfy the definition of 'contaminated land' as outlined in the Environmental Protection Act (1990). Therefore, if the DEFRA guidelines are appropriate, all treatments should have no effect on the circadian clock of *A. thaliana*, as the chemicals were tested at or below the maximum permitted concentrations. However, the wavelet spectral testing in Section 4.4 (Table 32) reveals that many of the DEFRA chemicals do have an effect. This suggests that the DEFRA guidelines may need to be revised for all chemicals in Table 30 excluding cadmium and mercury. In particular, the results in Table 32 indicate that for most of these chemicals (fluorine, chromium, nickel, copper, arsenic, selenium and molybdenum) half of the recommended maximum permitted concentration significantly affects the circadian clock of *A. thaliana*. These results suggest that the SGVs of these seven chemicals should be below

half the current value.

The results of this section are of particular importance as they suggest that currently acceptable levels of a large number of chemicals could be having a detrimental effect on many ecosystems, leading to the potential extinction of certain species.

4.5 Extension to Other Chemicals

The DEFRA chemicals do not encompass all elements in the periodic table. This could be due to the fact that, when the guidelines were written in 1990, certain chemicals were not anticipated to be found in UK soils. However, advances in technology mean that, in the modern world, plants are being exposed to a wider range of chemicals than ever before (Foley et al., 2005). Alternatively, the exclusion of certain chemicals from the guidelines could also imply that they are permitted at any concentration. Therefore, Oakenfull et al. (2018) also tested a comprehensive range of environmentally relevant pollutants to ascertain whether these chemicals have an effect on the plant circadian clock and hence determine if the SGVs should be extended to include other chemicals. A full list of chemicals and concentrations tested can be found in Tables 35 and 36 (Appendix 4.8).

As in Section 4.3, FFT-NLLS analysis was implemented to establish whether each chemical induced a change in periodicity and the results can be found in Tables 35 and 36 (Appendix 4.8). As discussed in Section 4.3.1, the results of the FFT–NLLS analysis can be used to group the tested chemicals by effect: 'No Change', 'Period Lengthening' or 'Period Shortening'. Oakenfull et al. (2018) visualised these groupings in a colour–coded periodic table (Figure 52A). Figure 52A can then be used by circadian biologists to ascertain if certain groups of elements (such as the rare earth metals) are having a similar effect on the circadian clock of *A. thaliana*. Such results could offer biological insight into the mechanistic basis for the plant circadian clock. However, the biological details are beyond the scope of this thesis.

The FT (FDR) was also implemented and the results can be found in Tables 35 and 36 (Appendix 4.8). Figures 53 and 54 display the individual time series and corresponding representative 'barcode' plots for a selection of the extension chemicals. As discussed in Section 3.3.1.2, there may be practical situations where practitioners can also be (cautiously) informed by the number of rejections of the null hypothesis of spectral equality, with larger values potentially indicating a greater departure from the null hypothesis. In this application, larger numbers of rejections could suggest a greater difference between the spectral behaviour in the control and chemical treatment groups and hence could indicate that a chemical has a greater effect on the circadian clock of *A. thaliana*. Therefore, Oakenfull et al. (2018) also used the percentage of rejections for the FT with FDR (at the 5% significance level) as a (coarse) dissimilarity measure, to produce a colour–coded periodic table (Figure 52B). Figure 52B can then be used by practitioners as a quick reference guide to deduce which chemicals are potentially the most hazardous to the environment.

4.5.1 Discussion of Findings

4.5.1.1 Examples of the FT Supporting the Classical Analysis

The results in Tables 35 and 36 indicate that the FT found 0% rejections of the null hypothesis of spectral equality for 11 chemicals. Therefore, (as discussed in Section 4.4) we will not assume

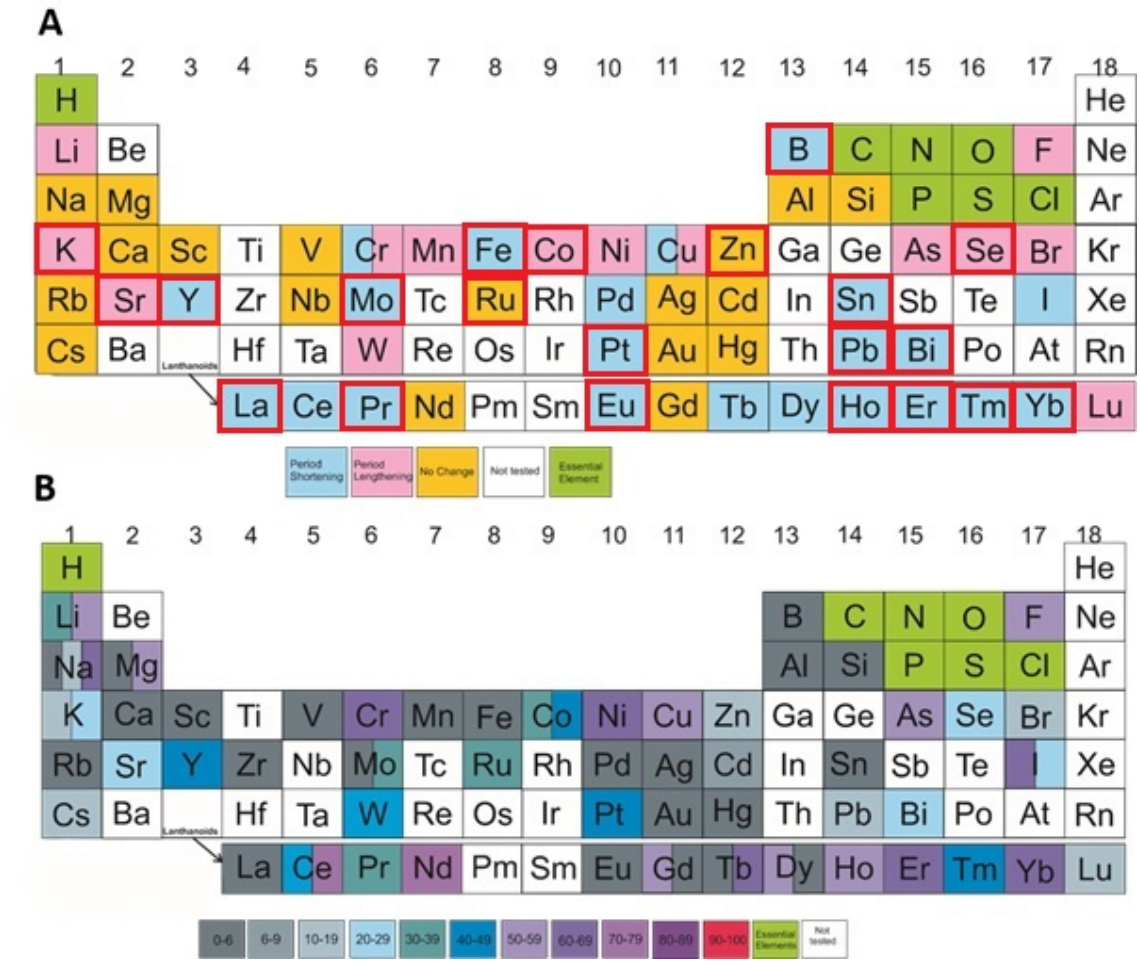

Figure 52: **Periodic tables**, coloured by effect on the circadian clock of *A. thaliana* (Oakenfull et al., 2018). **A**: Coloured by FFT-NLLS period estimates (red outlines indicate a statistically significant change in period for all compounds tested). **B**: Coloured by percentage change from control using FT (FDR) analysis. **A** and **B**: Green elements are essential to life and were not tested individually; White elements were not tested due to safety or solubility. The actinoids and group 7 elements have been omitted as they were not tested.

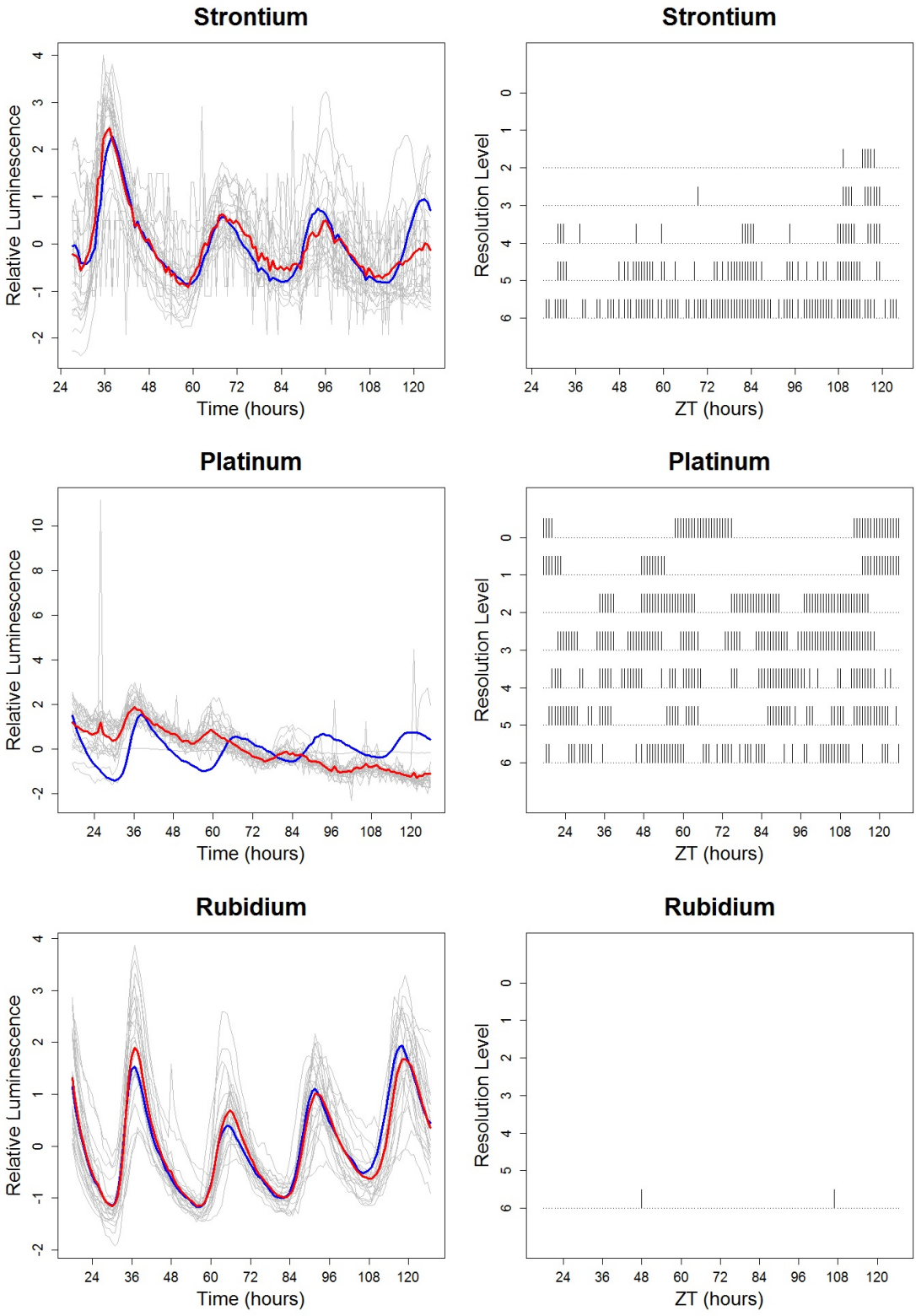

Figure 53: **Time series and Barcode plots for Strontium, Platinum and Rubidium.** Time series (left panels): Blue lines indicate the control average for each chemical; grey lines indicate individual time series within each chemical treatment group and red lines indicate the average time series for the chemical treatment group. Barcode plots (right panels): Barcode plots for FT (with FDR) at the 5% significance level.

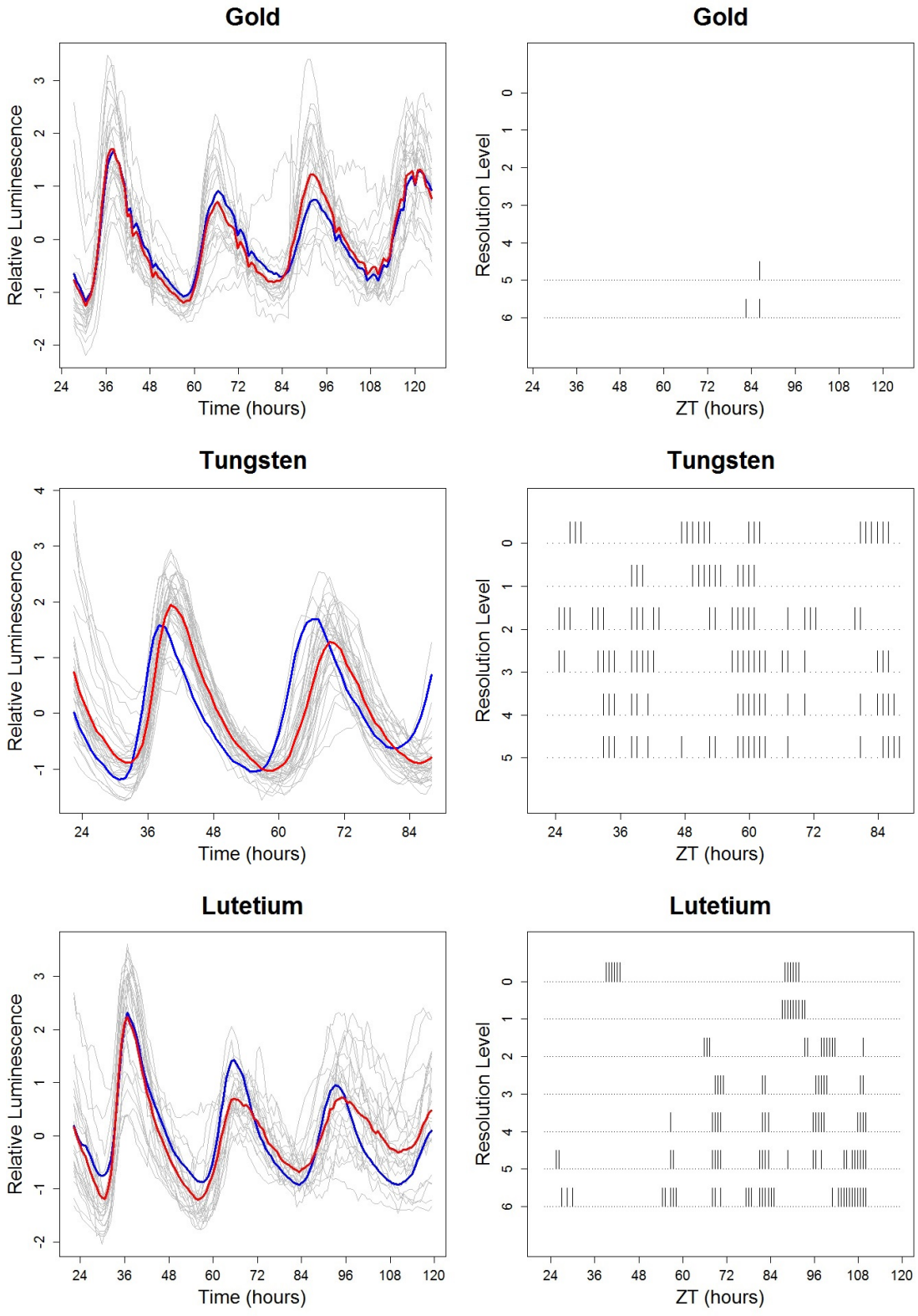

Figure 54: **Time series and Barcode plots for Gold, Tungsten and Lutetium.** Time series (left panels): Blue lines indicate the control group average for each chemical; grey lines indicate individual time series within each chemical treatment group and red lines indicate the average time series for the chemical treatment group. Barcode plots (right panels): Barcode plots for FT (with FDR) at the 5% significance level.

that these chemicals have a significant effect on the circadian clock of *A. thaliana*. These results are also supported by the BRASS analysis as, excluding Ruthenium (see discussion below), all of the remaining chemicals (which the FT found had no significant effect) also corresponded to a small change in period (using FFT–NLLS) that was not statistically significant. Furthermore, this conclusion is also supported by visual examination of the raw time series of each treatment group and its respective control. For example, Figures 53 and 54 display the raw time series and corresponding barcode plots for the Rubidium and Gold treatment groups, respectively. The raw time series in Figures 53 and 54 suggest that there may be very small differences between the average time series of control and treatment groups; however, these differences do not appear to be significant.

Tables 35 and 36 indicate that the Fourier analysis (using FFT–NLLS implemented in BRASS) found significant differences in period for 28 treatment groups. To an extent, these findings are reinforced by the FT (FDR) as for all these chemicals (other than Ruthenium– see discussion below) the FT found a number of significant differences (over the 0% threshold, see Section 4.4.1.1), indicating that these chemicals have an effect on the circadian clock of *A. thaliana*. For example, on examining Table 36, note that the BRASS analysis found that Platinum caused a significant decrease in period and the FT also found a number of rejections of spectral equality (46%, which is greater than the 0% threshold). Furthermore, the barcode plot (Figure 53) shows that the differences between the treatment group and control lie in resolution levels 2–4 (directly corresponding to a circadian rhythm). We conclude that there is evidence that Platinum does affect the circadian clock of *A. thaliana*.

Combining the results of the FFT-NLLS analysis (where appropriate) and the FT can also provide more detail regarding the change in period. For example, the BRASS analysis found that Strontium and Platinum both cause a significant change in period (see Table 36) but the barcode plots (Figure 53) show that Platinum is faster–acting than Strontium since the differences between the two spectra in resolution levels 2–4 are present throughout the experiment for Platinum but only appear after ZT106 for Strontium. This conclusion is also visually supported by the time series in Figure 53.

The FT can also provide additional insight that cannot be captured through a single period estimate, such as changes in spectral behaviour at multiple scales. For example, Strontium also induces high–frequency behaviour throughout the experiment (see Figure 53). This is reflected in the large number of significant differences in the finest resolution level (level 6) of the barcode plot in Figure 53 throughout the experiment. We conclude that there is evidence that Strontium alters the circadian and ultradian rhythms within *A. thaliana*. Furthermore, Strontium induces ultradian rhythms from the start of the experiment, but the changes to the circadian rhythms occur after ZT106.

4.5.1.2 Examples of the FT Not Supporting the Classical Analysis

There are also a number of instances where the wavelet spectral testing does not coincide with the Fourier–based analysis. For example, the BRASS analysis of Ruthenium reported a small but significant increase in period (see Table 36). Conversely, the FT (FDR) found 0% rejections of the null hypothesis of spectral equality (also see Table 36). As discussed in Section 4.4.1.1, circadian scientists often choose to disregard situations where very few coefficients are significantly different. Therefore, throughout this chapter, we have applied a 0% threshold for in-

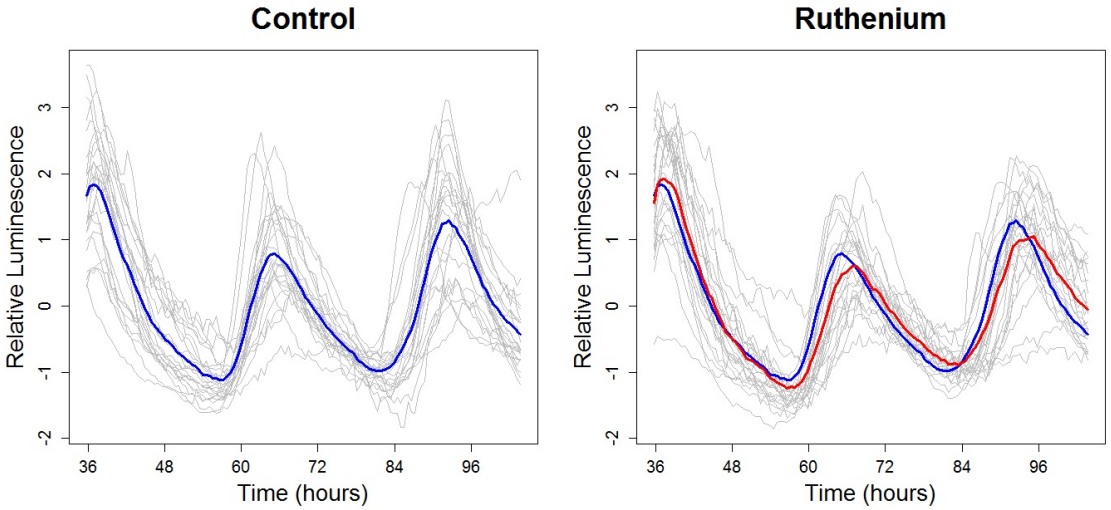

Figure 55: **Ruthenium:** Luminescence profiles over time for untreated *A. thaliana* plants (denoted 'Control') and those exposed to ruthenium chloride ($RuCl_3$) at a concentration of 2mM (denoted 'Ruthenium'). Left: Individuals in the control group (in grey) along with the group average (blue). Right: Individuals in the Ruthenium treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero.

ferring that chemicals cause a significant change to the spectral behaviour. Hence, the results of the FT suggest that there is not enough evidence that Ruthenium affects the circadian clock of *A. thaliana*. Figure 55 displays the raw time series for both the control group (untreated *A. thaliana* plants), as well as for those in the Ruthenium treatment group. The average time series in Figure 55 indicate a very small difference. However, given the variation in the raw time series, it could also be expected that this difference between the control and treatment groups would not be found to be statistically significant. Therefore, there is an argument for both conclusions. This uncertainty may be due to the resolution of the data. As an avenue of further work, we would recommend repeating this experiment for this treatment group but increasing the length of the free-run and taking observations at shorter intervals, which would improve the resolution of both methods.

Finally, there were also a large number of instances when the FT was able to detect a significant change in behaviour but the BRASS analysis could not (see Figure 52 and Tables 35 and 36). For example, the raw time series of Tungsten and Lutetium (Figure 54) indicate that both chemicals increased period. The BRASS analysis reported an increase in period for both chemicals, however, it was not found to be statistically significant (see Table 36). Conversely, on examining Table 36, we note that the FT found that Tungsten and Lutetium both displayed enough evidence to reject the null hypothesis of spectral equality between the treatment groups and their respective control (over the 0% threshold, see Section 4.4.1.1). Hence, based on the FT, we can conclude that there is evidence that Tungsten and Lutetium affect the circadian clock of *A. thaliana*. Additionally, the barcode plots for both chemicals (Figure 54) show that Tungsten is faster–acting than Lutetium, since the differences between the treatment and control group spectra are present throughout the experiment for Tungsten, but only appear after ZT60 for Lutetium. This conclusion is also visually supported by the time series in Figure 54.

4.5.2 Conclusions

In Section 4.4.2, we argued that, if a certain chemical at a particular concentration affects the plant circadian clock, this could have major consequences for entire ecosystems. Hence, by the definition of 'contaminated land' in the Environmental Protection Act (1990), this particular concentration of this chemical should not be permitted in soils and, consequently, the SGV for this chemical should be below this particular value. The SGVs in the DEFRA guidelines do not encompass all elements in the periodic table. If this is appropriate, all treatments tested in Section 4.5 should have no effect on the circadian clock of *A. thaliana*, as their omission means that they are permitted at any concentration. However, the wavelet spectral testing in Section 4.5 (Tables 35 and 36) found that many of these chemicals do have an effect. This suggests that the SGVs should be extended to include other chemicals.

This result is particularly relevant as advances in technology mean that, in the modern world, plants are being exposed to a wider range of chemicals than ever before (Foley et al., 2005). Within this context, we have demonstrated that a large number of potentially harmful chemicals have been historically overlooked by the procedures (such as Part 2A of the Environmental Protection Act (1990)) which were designed to identify (and subsequently treat) 'contaminated land'. These results are of particular importance as they suggest that a large number of chemicals could pose a significant risk to the environment, yet are going undetected by current assessment methods.

4.6 Cluster Analysis Using the Methodology Developed in Chapter 2

In Chapter 2, we developed a procedure for clustering inherently nonstationary rhythmic data by modelling them as locally stationary wavelet processes and exploiting their local time-scale spectral properties by means of a functional principal component analysis. We demonstrated the method's suitability in organising and understanding multiple nonstationary time series, such as the gene expression levels in this dataset. In this section we apply the clustering methodology of Chapter 2 to a selection of the DEFRA chemicals. This will facilitate answering the question, 'Which elements in the periodic table (and at which concentrations) produce similar kinds of reactions in plants?'

To answer this question, we analysed a number of different subsets of the DEFRA chemical dataset and the results are detailed below. The basic structure of each study is described as follows: as the LSW model is underpinned by wavelets and requires the data to be of dyadic length ($T = 2^J$), in our analysis we chose a segment of length $T = 128$ out of the original dataset, as in Section 4.4. For each plant we estimated the (Haar) wavelet spectrum by means of the corrected wavelet periodogram estimate (using the `locits` R package). Each periodogram was level smoothed by log transform, followed by translation invariant global universal thresholding and then the inverse transform was applied. For each scale of the wavelet periodogram, only levels 3 and finer were thresholded. Using the estimated spectral information, we obtained a dissimilarity matrix. As in Chapter 2, we determined the number of principal components to retain based on a screeplot. The resulting dissimilarity matrix was the input of a PAM algorithm (performed in the `cluster` R package) which clustered the data into a user–defined number of groups. We used the methods outlined in Section 2.3.4.3 (Chapter 2) to determine the optimal number of clusters.

4.6.1 Clustering DEFRA Chemicals

We began by applying our proposed LSW-PCA clustering method to analyse the 12 chemicals (and their respective controls) displayed in Figures 47 and 48. On examining the screeplot and for ease of interpretation, we retained two principal components to cluster the data. The methods outlined in Section 2.3.4.3 were used to determine the optimal number of clusters and all methods indicated that we should cluster the data into 2 groups. This was supported by experimental scientists who confirmed that, as a preliminary analysis, it would be useful to cluster the data into 2 groups: 'No Change' and any distinct departures from this group, thus indicating which chemicals have an effect on the circadian clock of *A. thaliana* and which do not. The LSW-PCA clustering method yielded the results detailed in Table 37 (Appendix 4.8).

4.6.2 Discussion of Findings

On examining Table 37, we can see that the LSW-PCA clustering method has clustered the behaviour of the data into the following two groups: Cluster 2 identifies similar behaviour of plants in the control groups and the Lead (Half), Mercury (Max) and Cadmium (Half) treatment groups and Cluster 1 contains all 24 plants in the remaining treatment groups. These results are in agreement with Figures 47 and 48 which provided visual evidence that the plants in the Lead (Half), Mercury (Max) and Cadmium (Half) treatment groups seemed to display similar behaviour to the control groups, indicating that these chemicals had no effect on the circadian clock of *A. thaliana*. This conclusion was also supported by the wavelet spectral testing (Section 4.4) which found 0% rejections of the null hypothesis of spectral equality for these chemicals. Therefore, Cluster 2 can be conceptualised as essentially 'No Change' and Cluster 1 as 'Change'.

Table 37 shows that, for nine chemical treatments, all 24 plants are in Cluster 1 ('Change'). (Note: these correspond with chemical treatments that the FT indicated had a statistically significant effect on the circadian clock (Section 4.4).) However, there are no chemical treatment or control groups where all 24 plants are in Cluster 2 ('No Change'). That is, a number of plants from the control and chemical treatment groups that were identified as having no significant effect on the circadian clock, can be found in Cluster 1 ('Change'). The presence of these 'No Change' plants in the 'Change' cluster highlights individual-level variability in plant response to stimuli, despite their sharing identical genetic characteristics (Doyle et al., 2002). This result may be due to the individual plants in some instances showing a stress response, particularly those individuals from the chemical treatment groups in Cluster 1 (which have more plants in Cluster 1 than the control groups). Alternatively, this may be due to stress induced by the experimental method itself. This result supports the discussion in Section 4.5.1, that although the average time series for some chemicals could indicate a very small difference (see for example Ruthenium in Figure 55), the variation in the raw time series, even within the control groups, means that such small differences between the control and treatment groups may not be found to be statistically significant.

4.6.3 Example: Clustering Within Individual Microtiter Plates

We now attempt to answer the questions 'Does exposure to different elements in the periodic table produce a generic type of reaction in plants?' and, if not, 'Which elements induce similar

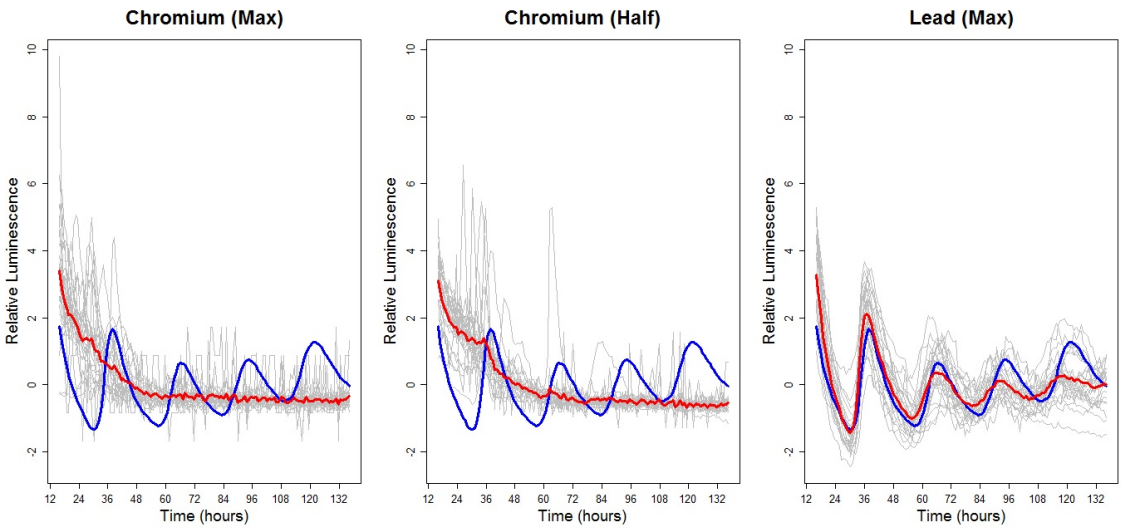

Figure 56: **DEFRA Chemicals (plate 0953):** Luminescence profiles over time for *A. thaliana* plants exposed to a selection of the DEFRA chemicals. Each Panel: Individuals in the chemical treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero.

Number of plants	Chromium (Max)	Chromium (Half)	Lead (Max)	Total
Cluster 1	**24**	**24**	9	57
Cluster 2	0	0	**15**	15
Total	24	24	24	72

Table 33: Results of clustering plate 0953 into two clusters using the proposed LSW-PCA method. The modal cluster for each treatment group is highlighted in bold.

kinds of reactions in plants?' by using the LSW–PCA clustering methodology. In Section 4.6.1 it was useful (as a preliminary analysis) to cluster data arising from different mictrotiter plates. However, as highlighted in Section 4.2, it is preferable to perform data analysis on time series from the same plate. Therefore, we applied the LSW–PCA clustering methodology to the individual microtiter plates within the DEFRA chemical dataset. A representative selection of the results are presented in this section and in Appendix 4.9.

In Section 4.6.1, we demonstrated that our LSW-PCA clustering method can effectively discriminate between the control and treatment groups. Hence, we began by applying our proposed LSW-PCA clustering method to analyse the 3 chemicals (not their respective control) on plate 0953. This plate constituted: a control group, Chromium (both concentrations) and Lead (Max). Figure 56 displays the individual time series for the 3 DEFRA chemicals on plate 0953. On examining the screeplot and for ease of interpretation, we retained two principal components to cluster this data. The methods outlined in Section 2.3.4.3 were used to determine the optimal number of clusters and all methods indicated that we should cluster the data into 2 groups. The LSW-PCA clustering method yielded the results detailed in Table 33.

4.6.4 Discussion of Findings

On examining Table 33, we can see that the LSW-PCA clustering method has clustered the behaviour of the data into the following two groups: Cluster 1 identifies similar behaviour of plants in both Chromium treatment groups (conceptualised as essentially 'Chromium') and

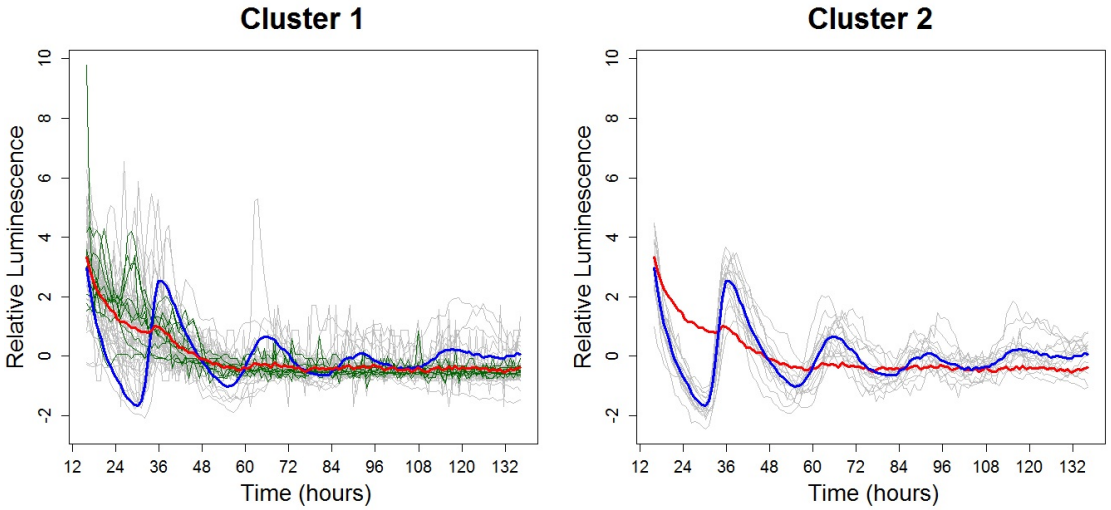

Figure 57: The results of clustering the DEFRA Chemicals (plate 0953) into 2 groups using the LSW-PCA method. The individual signals (grey) along with the cluster average in: red for Cluster 1 and blue for Cluster 2. The individual signals of the Lead (Max) treatment group in Cluster 1 are plotted in green.

Cluster 2 is the modal cluster of the Lead (Max) treatment group (conceptualised as 'Lead (Max)'). These results are in agreement with Figure 56 which provided visual evidence that the plants in both Chromium treatment groups seemed to display similar behaviour, while the Lead (Max) group seems to display average behaviour which is distinct from the Chromium groups and from the control group.

Combining the results of the cluster analysis and the FT, we can conclude that, although all three chemicals have an effect on the circadian clock of *A. thaliana* (Table 32), they do not induce the same effect. This suggests that the chemicals may not simply induce a generic chemical stress response, but may actually induce a chemical–specific response. This could be due to specific chemical reactions within the circadian oscillator. For example, exposure to certain chemicals has been shown to inhibit the uptake of other (essential) ions (see for example Silver et al. (1981)). Therefore, this result is of particular biological interest as it provides insight into the different chemical input mechanisms of the circadian oscillator (Oakenfull et al., 2018). For example, Perea-García et al. (2016a) examined the effect of copper on the circadian clock of *A. thaliana* and the results of this investigation provided insight (on a chemical level) into the structure and composition of a model proposed for the *A. thaliana* central oscillator (Bujdoso and Davis, 2013). Similarly, these results could also offer biological insight into the mechanistic basis for the plant circadian clock. However, the biological details are beyond the scope of this thesis.

Our proposed method also allows us to characterise the behaviour associated with each cluster. The signals within each cluster are shown (in grey) along with the cluster averages (in bold) in Figure 57. Figure 57 also shows the individual time series from the Lead (Max) treatment group that were assigned to Cluster 1 (plotted in green).

Note in Figure 57 that Cluster 1 is characterised by a marked amplitude dampening with time, resulting in a rhythmicity loss at approximately ZT48. In particular, the individual time series seem to display a burst of relative luminescence prior to ZT36, but at different points in time. Also, note that the plants from the Lead (Max) treatment group in Cluster 1, also share

Chemical Treatment Group	Chemical Treatment Group	Number of Rejections FT (FDR)
Chromium (Max)	Chromium (Half)	264 (29%)
Chromium (Max)	Lead (Max)	553 (62%)
Chromium (Half)	Lead (Max)	533 (59%)
Chromium (both)	Lead (Max)	576 (64%)

Table 34: **FT (FDR) results– DEFRA Chemicals (plate 0953)**. The number of rejections (as a percentage in brackets) for the FT with FDR (at the 5% significance level) for the DEFRA Chemicals (plate 0953).

this behaviour. This illustrates the point in Section 4.2, that although plants in each treatment group share identical genetic characteristics and have been treated in identical conditions, they can respond differently.

In contrast to Cluster 1, Cluster 2 displays broadly periodic behaviour but with a gradual decrease in period throughout the experiment and amplitude dampening with time. Furthermore, the variation between the individual time series seems to increase throughout the experiment: the individual signals display very similar behaviour prior to time ZT48 (with a trough at ZT30 and peak just after ZT36) and broadly similar behaviour thereafter.

The results of the LSW-PCA clustering method (Tables 37 and 33) indicate that all 3 chemicals have an effect on the circadian clock of *A. thaliana* but both concentrations of chromium seem to have a similar effect which is distinct to the effect of Lead (Max). Since the data originates from the same microtiter plate, it is possible to apply the wavelet spectral testing of Chapter 3 to test this hypothesis. Therefore, following the methods outlined in Section 4.4, we applied the FT to plate 0953 and the results can be found in Table 34.

As discussed in Section 4.5, practitioners can be (cautiously) informed by the number of rejections of the null hypothesis of spectral equality (as a dissimilarity measure), with larger values indicating a greater departure from the null hypothesis. In our application, larger numbers of rejections could indicate a greater difference between the spectral behaviour of the two chemical treatment groups. Therefore, the results in Table 34 confirm the conclusions of the LSW–PCA clustering– the chromium treatment groups display similar behaviour, whereas the Lead (Max) group displays distinct behaviour from the chromium treatment groups. This is reflected in the greater number of percentage rejections when comparing the Lead (Max) group with the chromium groups (approximately 60%), than the chromium groups with each other (29%). However, there are still a large number of rejections of spectral equality between the two chromium treatment groups. This suggests that though the chromium treatment groups are more similar than the Lead (Max) group, they are still significantly different. This result supports the discussion in Chapter 2, that recent research has shown that certain compounds can produce different effects on plant growth at low and high doses (Yang et al., 2016). Furthermore, it also demonstrates the complementary utility of the methodology developed in Chapters 2 and 3– wavelet spectral testing can identify relatively small differences in spectral behaviour between two groups of nonstationary time series, whereas the LSW–PCA clustering methodology can identify broadly similar spectral behaviour.

4.7 Conclusions and Further Work

In this chapter, we applied our proposed wavelet spectral testing and clustering methodologies (of Chapters 3 and 2, respectively) to the dataset that motivated the work in this thesis. This allowed us to organise and understand the impact of a comprehensive range of environmentally relevant pollutants on plant circadian rhythms. Our proposed methodology was able to discriminate between treatment groups (Table 32) when the current methodology could not (Table 30). This facilitated the understanding of the environmental ramifications associated with soil pollution and demonstrated the additional insight our wavelet spectral testing and clustering methodology can provide.

We also applied a period analysis technique currently adopted within the circadian community and contrasted it with one of the hypothesis tests developed in Chapter 3. The application of the FT alongside the industry–standard BRASS analysis demonstrates that the hypothesis tests developed in Chapter 3 fill the gap in the current literature by developing a much needed tool for the formal spectral comparison of nonstationary data, analogous to the techniques currently adopted within the circadian community.

The FT was also used to characterise the different types of behaviour present in the data as the barcode plots (Figures 50, 51, 53 and 54) are able to identify the times and scales at which spectral differences occur. In Section 4.5, we illustrated the additional insight our methodology can provide as the FT is therefore able to identify how fast a chemical effects the plant circadian clock and also identify spectral differences at multiple scales. Additionally, we demonstrated that practitioners can also be informed by the number of rejections of the null hypothesis of spectral equality (see Figure 52), with larger values (potentially) indicating that a particular chemical has a greater effect on the circadian clock of *A. thaliana*.

We then applied the clustering methodology of Chapter 2 to a selection of the DEFRA chemicals. The results in Section 4.6 demonstrated the ability of the LSW–PCA clustering method to determine the different types of reactions present in the DEFRA dataset and subsequently identify which elements in the periodic table (and at which concentrations) produce similar kinds of reactions in plants. The complementary examples in Section 4.6 and Appendix 4.9 demonstrate the method's suitability in organizing and understanding multiple nonstationary time series, such as the gene expression levels in the DEFRA chemical dataset.

This chapter also showcases the complementary utility of the methodology developed in Chapters 2 and 3. In particular, while wavelet spectral testing can identify relatively small differences in spectral behaviour between two groups of nonstationary time series, the LSW–PCA clustering methodology can identify broadly similar spectral behaviour.

In Section 4.6.4 we combined the results of the cluster analysis and the FT, this enabled us to conclude that, although three chemicals had an effect on the circadian clock of *A. thaliana*, they did not induce the same effect. By extension, although a large number of chemicals tested in Sections 4.4 and 4.5 had an effect on the circadian clock, they may not simply display a generic chemical stress response but may induce a chemical–specific response or a similar response to a selection of other chemicals. In Section 4.6.4 we discussed how the results of this chapter could also offer biological insight into the mechanistic basis for the plant circadian clock. These studies could also enable deeper understanding of the circadian clock mechanisms and its adaptations to change (Perea-García et al., 2016a). However, the biological details are beyond the scope of this thesis.

In Section 4.4.2 we argued that the DEFRA guidelines should be revised for all chemicals in Table 30 excluding cadmium and mercury since we would expect that all the treatments have no effect as the chemicals were tested at (or below) the recommended maximum permitted concentrations according to the DEFRA guidelines. We also demonstrated that, for a large number of the DEFRA chemicals, the recommended maximum permitted concentration should be below half the current value. An interesting area of further work would be to determine the threshold at which these chemicals have an effect and hence produce new recommendations for the SGVs. This could be achieved by varying the supplementary concentrations of the 18 DEFRA chemicals that had a significant effect (as discussed in Section 4.2 for the extension chemicals), and then applying our hypothesis testing methodology to test for statistically significant differences. Upper and lower bounds for the concentrations to be investigated would be guided by the results of Section 4.4. For example, since Chromium (Half) had a significant effect (Table 32), only concentrations below this value should be investigated. Alternatively, since Lead (Max) had a significant effect but Lead (Half) did not, concentrations between these values should be investigated.

The dataset used throughout this chapter was specifically designed for the period analysis techniques and spectral testing methodology discussed in Section 4.3 and 4.4, respectively. Thus, the optimal configuration of the microtiter plates was 4 groups of 24 plants, as approximately this number of realisations is necessary for good performance of the wavelet spectral testing procedures (see the results of the simulation studies in Chapter 3). On the other hand, this format restricted the application of the clustering methodology in Section 4.6: since it is preferable to perform cluster analysis on time series from the same plate (see Section 4.2 for details), we were only able to cluster the behaviour of three chemicals. Nevertheless, Section 4.6 still demonstrated the additional insight the LSW–PCA method can provide for this application. In particular, identifying the different types of reactions present in a particular dataset and which elements in the periodic table (and at which concentrations) produce similar kinds of reactions in plants. An area of further work would be to repeat these experiments with smaller treatment groups so that more chemicals could be compared on a single microtiter plate.

4.8 Appendix: Supplementary Tables

In this section we provide supplementary tables that support the discussion throughout this chapter. Tables 35 and 36 provide a full list of the exact chemicals and concentrations used in the salt stress experiment. Table 37 reports the results of clustering the 12 DEFRA Chemicals in Figures 47 and 48 and their respective controls into 2 groups using the LSW-PCA method.

AN	Treatment	Chemical	Concentration	Rejections FT (FDR)	Period Difference
3	Lithium (Li)	$LiCl_2$	20mM	280 (31%)	**4.54**∗
3	Lithium (Li)	$LiSO_4$	15mM	455 (51%)	6.76
5	Boron (B)	$Na_2B_4O_7$	3mM	34 (4%)	**-1.68**∗
11	Sodium (Na)	NaCl	2mM	1 (0%)	-0.21
11	Sodium (Na)	NaBr	100mM	114 (13%)	**1.33**∗
11	Sodium (Na)	NaI	100mM	545 (61%)	0.32
12	Magnesium (Mg)	$MgCl_2$	5mM	38 (4%)	0.01
12	Magnesium (Mg)	$C_4H_6O_4Mg$	5mM	512 (57%)	**2.00**∗
12	Magnesium (Mg)	$Mg(NO_3)_2$	5mM	2 (0%)	0.05
13	Aluminium (Al)	$AlCl_3$	300μM	6 (1%)	-0.45
14	Silicon (Si)	Na_2SiO_3	25mM	7 (1%)	0.56
19	Potassium (K)	KCl	100mM	146 (16%)	**1.55**∗
19	Potassium (K)	KBr	100mM	95 (11%)	**1.60**∗
19	Potassium (K)	KI	100mM	252 (28%)	**-1.42**
20	Calcium (Ca)	$CaCl_2$	50mM	9 (1%)	**1.77**∗
20	Calcium (Ca)	$Ca(NO_3)_2$	1mM	2 (0%)	0.08
21	Scandium (Sc)	$Sc(SO_3CF_3)_3$	100μM	1 (0%)	0.20
21	Scandium (Sc)	ScF_3	300μM	1 (0%)	-0.58
23	Vanadium (V)	H_3NO_3V	25μM	4 (1%)	-0.41
25	Manganese (Mn)	$MnCl_2$	1mM	19 (2%)	**0.87**∗
25	Manganese (Mn)	$MnSO_4$	200μM	1 (0%)	0.48
26	Iron (Fe)	$FeCl_3$	300μM	16 (2%)	**-1.27**∗
27	Cobalt (Co)	$CoCl_2$	250μM	133 (35%)	**1.70**∗
27	Cobalt (Co)	$CoSO_4$	250μM	158 (41%)	**1.82**∗

Table 35: **Extension chemicals Part 1** (atomic numbers 3–27). The chemicals and concentrations used in the salt stress experiment (Section 4.5), where "Treatment" represents the element under investigation within the chemical compound (corresponding to the periodic table representation used in Figure 52) and "AN" represents the associated atomic number. For each chemical, the number of rejections (as a percentage in brackets) for the FT with FDR (at the 5% significance level) and the estimated mean difference in period (using FFT–NLLS), with ∗ indicating a significant change in period from the respective control group. ‡ indicates time series and a barcode plot for the chemical are shown in Figures 53 or 54.

AN	Treatment	Chemical	Concentration	Rejections FT (FDR)	Period Difference
37	**Rubidium (Rb)**‡	RbCl	200μM	2 (0%)	0.38
38	**Strontium (Sr)**‡	$SrCl_2$	30mM	189 (21%)	**1.42**∗
39	Yttrium (Y)	YCl_3	3mM	418 (47%)	**-3.18**∗
41	Niobium (Nb)	$NbCl_5$	500μM	2 (1%)	-0.39
44	Ruthenium (Ru)	$RuCl_3$	2mM	1 (0%)	**0.64**∗
47	Silver (Ag)	$AgNO_3$	200μM	50 (6%)	-0.46
50	Tin (Sn)	$SnCl_2$	1.5mM	43 (11%)	**-1.81**∗
55	Caesium (Cs)	CsCl	200μM	4 (0%)	0.27
57	Lanthanum (La)	$LaCl_3$	5mM	420 (47%)	**-3.33**∗
58	Cerium (Ce)	$CeCl_3$	3mM	630 (70%)	**-2.83**∗
58	Cerium (Ce)	$(NH_4)_2Ce(NO_3)_6$	150μM	281 (31%)	-1.40
59	Praseodymium (Pr)	$PrCl_3$	2mM	625 (70%)	**-2.53**∗
60	Neodymium (Nd)	$NdCl_3$	1.5mM	40 (4%)	0.62
63	Europium (Eu)	$EuCl_3$	5mM	490 (55%)	**-2.02**∗
64	Gadolinium (Gd)	$(CF_3SO_3)_3Gd$	500μM	27 (3%)	0.57
64	Gadolinium (Gd)	$GdCl_3$	600μM	1 (0%)	0.03
65	Terbium (Tb)	$TbCl_3$	1.5mM	541 (60%)	**-2.60**∗
66	Dysprosium (Dy)	$DyCl_3$	3mM	501 (56%)	**-1.56**∗
66	Dysprosium (Dy)	DyF_3	100μM	2 (1%)	0.66
67	Holmium (Ho)	$HoCl_3$	1mM	447 (50%)	**-2.51**∗
68	Erbium (Er)	$ErCl_3$	1mM	617 (69%)	**-1.92**∗
69	Thulium (Tm)	$TmCl_3$	1mM	412 (46%)	**-2.48**∗
70	Ytterbium (Yb)	$YbCl_3$	1mM	592 (66%)	**-2.64**∗
71	**Lutetium (Lu)**‡	$LuCl_3$	1mM	119 (13%)	0.92
74	**Tungsten (W)**‡	Na_2WO_4	20g/L	119 (31%)	1.61
78	**Platinum (Pt)**‡	K_2PtCl_4	200μM	409 (46%)	**-3.62**∗
79	**Gold (Au)**‡	$KAuCl_4$	50μM	3 (0%)	0.10
83	Bismuth (Bi)	$BiCl_3$	2mM	179 (20%)	**-1.10**∗

Table 36: **Extension chemicals Part 2** (atomic numbers 37–83). The chemicals and concentrations used in the salt stress experiment (Section 4.5), where "Treatment" represents the element under investigation within the chemical compound (corresponding to the periodic table representation used in Figure 52) and "AN" represents the associated atomic number. For each chemical, the number of rejections (as a percentage in brackets) for the FT with FDR (at the 5% significance level) and the estimated mean difference in period (using FFT–NLLS), with ∗ indicating a significant change in period from the respective control group. ‡ indicates time series and a barcode plot for the chemical are shown in Figures 53 or 54.

Treatment Group	Cluster 1 (Number of Plants)	Cluster 2 (Number of Plants)
Control 1	17	7
Copper (Max)∗	24	0
Selenium (Max)∗	24	0
Control 2	14	10
Lead (Half)	19	5
Mercury (Max)	20	4
Control 3	17	7
Lead (Max)∗	24	0
Control 4	15	9
Selenium (Half)∗	24	0
Cadmium (Half)	20	4
Control 5	10	14
Zinc (Max)∗	24	0
Control 6	16	8
Molybdenum (Max)∗	24	0
Molybdenum (Half)∗	24	0
Control 7	16	8
Arsenic (Max)∗	24	0
Arsenic (Half)∗	24	0

Table 37: **Results of clustering the 12 DEFRA Chemicals** in Figures 47 and 48 and their respective controls into 2 groups using the LSW-PCA method. There are 24 plants in each treatment group. ∗ indicates a treatment with 0 plants in cluster 2.

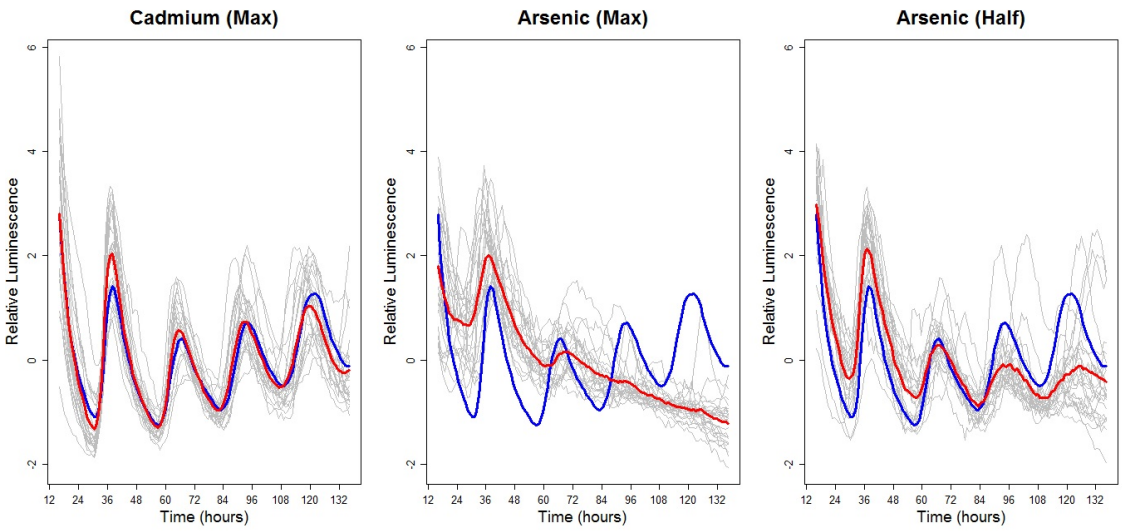

Figure 58: **DEFRA Chemicals (plate 0952):** Luminescence profiles over time for *A. thaliana* plants exposed to a selection of the DEFRA chemicals. Each Panel: Individuals in the chemical treatment group (in grey) along with the treatment group average (red) and the control group average (blue). Each time series has been standardised to have mean zero.

Number of plants	Cadmium (Max)	Arsenic (Max)	Arsenic (Half)	Total
Cluster 1	**14**	1	3	18
Cluster 2	2	**21**	6	29
Cluster 3	8	2	**15**	25
Total	24	24	24	72

Table 38: Results of clustering plate 0952 into three clusters using the proposed LSW-PCA method. The modal cluster for each treatment group is highlighted in bold.

4.9 Appendix: Additional Clustering Example

In this section, we apply our proposed LSW-PCA clustering method to analyse the three chemicals on plate 0952: Arsenic (both concentrations) and Cadmium (Max). The individual time series for these chemicals are displayed in Figure 58. On examining the screeplot and for ease of interpretation, we retained two principal components to cluster this data. The methods outlined in Section 2.3.4.3 were used to determine the optimal number of clusters and all methods indicated that we should cluster the data into 3 groups. The LSW-PCA clustering method yielded the results detailed in Table 38.

4.9.1 Discussion of Findings

On examining Table 38, we can see that the LSW-PCA clustering method has clustered the behaviour of the data into the following three groups: Cluster 1 is the modal cluster of the Cadmium (Max) treatment group (conceptualised as 'Cadmium (Max)'); Cluster 2 is the modal cluster of the Arsenic (Max) treatment group (conceptualised as 'Arsenic (Max)') and Cluster 3 is the modal cluster of the Arsenic (Half) treatment group (conceptualised as 'Arsenic (Half)'). These results are in agreement with Figure 58 which provided visual evidence that the plants in each treatment group display distinct behaviour (i.e. no two treatments provide a similar effect). This conclusion was also supported by the wavelet spectral testing (Section 4.4) which found very different numbers of rejections of the null hypothesis of spectral equality for each

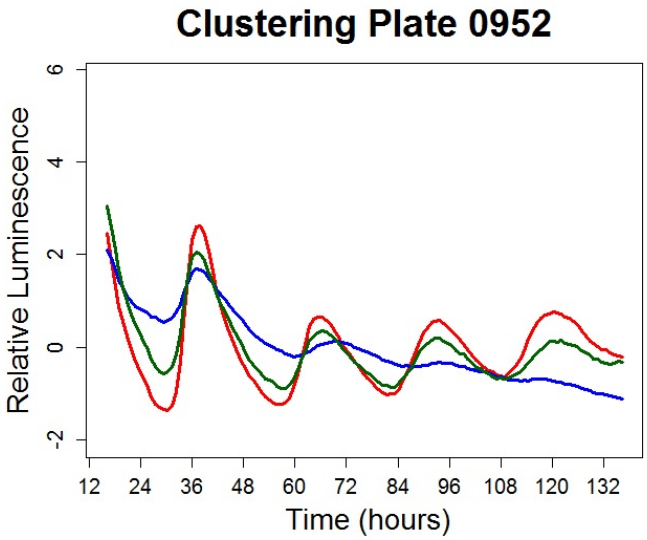

Figure 59: The results of clustering the DEFRA Chemicals (plate 0952) into 3 groups using the LSW-PCA method. The cluster average time series in: red for Cluster 1 (conceptualised as 'Cadmium (Max)'); blue for Cluster 2 (conceptualised as 'Arsenic (Max)') and green for Cluster 3 (conceptualised as 'Arsenic (Half)').

treatment group (see Table 32) indicating that these chemicals do not have a similar effect on the circadian clock of *A. thaliana*. However, the fact that each chemical treatment group appears in each cluster again highlights individual–level variability in plant response to stimuli which may result in individual plants displaying a similar response to different treatments. In Chapter 2, we proposed that this may be due to the individual plants in some instances showing a more general stress response, perhaps induced by the experimental method itself (Hargreaves et al., 2018).

The LSW–PCA clustering method also allows us to characterise the behaviour associated with each cluster. The average time series for each cluster are shown in Figure 59. The signals within each cluster are shown (in grey) along with the cluster averages (in bold) in Figure 60. For each cluster, Figure 60 also shows the individual signals in the non–modal treatment group (plotted in: red for Cadmium (Max); blue for Arsenic (Max) and green for Arsenic (Half)).

Note in Figures 59 and 60 that all three clusters appear to have a similar period prior to ZT48 (albeit with a different amplitude). However, Clusters 1 and 3 display broadly periodic behaviour but with a time–varying period and amplitude dampening with time. The main difference between the clusters seems to be in the amplitude, with Cluster 1 having a higher amplitude. However, more subtle differences in period can also be noted. Furthermore, there appears to be more variation between the individual time series in Cluster 3 and this individual–level variability seems to increase throughout the experiment. This is also confirmed by Table 38 as this cluster contains the largest proportion of plants from the non–modal treatment groups. In contrast to Clusters 1 and 3, Cluster 2 is characterised by an increase in period with a marked amplitude dampening with time and a decreasing mean (decreasing linear trend).

The results of the LSW-PCA clustering method (Table 38) indicate that Arsenic (Max) seems to display the most distinct behaviour (with 87.5% of the plants in Cluster 2). Cadmium (Max) appears to most closely resemble Arsenic (Half) as 33% of its plants are in Cluster 3 (conceptualised as 'Arsenic (Half)') but only 8% are in Cluster 2 (conceptualised as 'Arsenic (Max)').

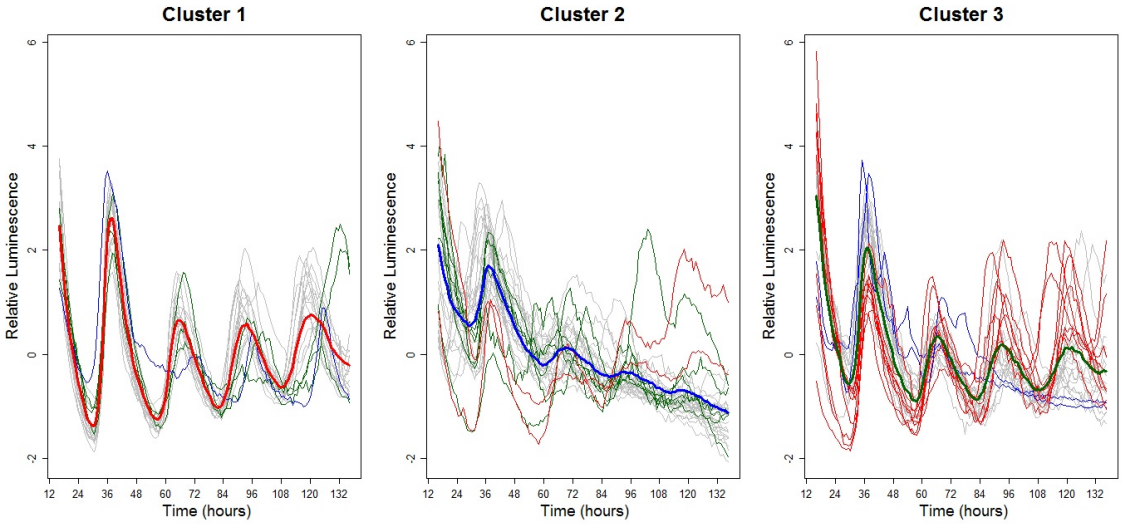

Figure 60: The results of clustering the DEFRA Chemicals (plate 0952) into 3 groups using the LSW-PCA method. The individual signals of the modal treatment group (grey) along with the cluster average in: red for Cluster 1; blue for Cluster 2 and green for Cluster 3. For each cluster, the individual signals in the non–modal treatment group are plotted in: red for Cadmium (Max); blue for Arsenic (Max) and green for Arsenic (Half).

Chemical Treatment Group	**Chemical Treatment Group**	**Number of Rejections FT (FDR)**
Cadmium (Max)	Arsenic (Max)	525 (59%)
Cadmium (Max)	Arsenic (Half)	154 (17%)
Arsenic (Max)	Arsenic (Half)	325 (36%)

Table 39: **FT (FDR) results– Comparing DEFRA Chemicals (plate 0952)**. The number of rejections (as a percentage in brackets) for the FT with FDR (at the 5% significance level) for the DEFRA Chemicals (plate 0952).

Interestingly, Arsenic (Half) is more similar to Arsenic (Max) as 25% of its plants are in Cluster 2 (conceptualised as 'Arsenic (Max)') but only 13% are in Cluster 1 (conceptualised as 'Cadmium (Max)'). Following the methods outlined in Section 4.4, we applied the FT to plate 0952 to test these hypotheses and the results are reported in Table 39.

The results in Table 39 confirm the conclusions of the LSW–PCA clustering– the Cadmium (Max) and Arsenic (Half) treatment groups display some similar behavioural traits, whereas the Arsenic (Max) group displays distinct behaviour from the other treatment groups. These results are in agreement with Figure 58 which provided visual evidence that the plants in each treatment group display distinct behaviour but the Cadmium (Max) and Arsenic (Half) treatment groups display more similar behaviour.

5 Conclusions and Further Work

This thesis has developed wavelet–based methodology, motivated by selected applied problems arising in nonstationary time series analysis of circadian signals. In particular, the work in this thesis was motivated by the limitations of current time series analysis in the field of circadian biology. In this chapter, we briefly summarise the main contributions made in Chapters 2–4 before discussing possible directions for future research. Further discussions of each element of work are also provided by the individual summaries at the end of each chapter.

Chapter 2

In Chapter 2 we discussed the phenomenon of individual-level variability in plant response to stimuli, despite their sharing identical genetic characteristics Doyle et al. (2002). The presence of multiple behaviours within the same treatment group within our motivating dataset motivated our development of a clustering procedure that can detect these different characteristics and analyse them separately. In Chapter 2 we also investigated the lack of stationarity of the circadian plant rhythms that motivated the work of this thesis. This result, along with several others recorded in the literature (Price et al., 2008; Leise et al., 2013; Harang et al., 2012) motivated the development of clustering techniques that can account for nonstationarity. The clustering method combines locally stationary wavelet time series modelling with functional principal components analysis and thus extracts the time-scale patterns arising in a range of rhythmic data. We demonstrated the advantages of our methodology over alternative approaches by means of a comprehensive simulation study and real data applications. Although the data analysed throughout this chapter is from the field of circadian biology, the methodology is general and can be readily applied to data originating in a range of fields (e.g. finance, climatology, seismic problems).

Chapter 3

In Chapter 3 we addressed the problem of comparing circadian oscillation behaviour between two groups of observations. The work in this chapter was motivated by three circadian datasets, each posing a different research question. As a response, we developed a new methodology for comparing nonstationary time series in the wavelet spectral domain through hypothesis testing, embedding replicate information when available, analogous in spirit to the techniques currently adopted within the circadian community, but accounting for the nonstationarity in the data. Under the LSW modelling framework of Nason et al. (2000), we developed four different hypothesis tests which detect three types of spectral differences between two groups. Our methodology was applied to the motivating problems from circadian biology, illustrating the practical use of the proposed techniques and the additional insight they provide.

Chapter 4

The methodology developed throughout this thesis was motivated by a specific application in the field of circadian biology– the effect of industrial and agricultural pollutants on the plant circadian clock (Foley et al., 2005; Senesil et al., 1998; Hargreaves et al., 2018; Nicholson et al., 2003). The 'Cerium dataset' that motivated the work in Chapter 2 and the 'Lead dataset' that motivated the development of the raw periodogram F–test in Chapter 3 were taken from a broad investigation of the effect of various salt stresses on plants (Oakenfull et al., 2018). Specifically, the Department for Environment, Food and Rural Affairs (DEFRA) developed 'Soil Guideline Values' (SGVs) that can be used to determine appropriate concentrations of certain

chemicals in soil. Therefore, in Chapter 4, we applied the clustering methodology and wavelet spectral testing (of Chapters 2 and 3, respectively) to investigate the impact of exposure to the chemicals at the concentrations outlined in the DEFRA report, as well as to chemicals not included in the report, on the plant circadian clock. Our findings provided novel evidence that many of the tested chemicals do indeed have an effect on the plant circadian clock. Therefore, the results of Chapter 4 could be used to inform a revision of the SGVs. Critically, for certain chemicals, our findings suggest that the recommended maximum permitted concentration should be below half the current value.

The results of Chapter 4 also demonstrated the additional insight our methodology can provide. In particular, we identified how fast a chemical effects the plant circadian clock and the spectral differences at multiple scales. Additionally, the analysis in Chapter 4 illustrated the utility of our proposed methodologies. We showed that practitioners can be informed by the number of rejections of the null hypothesis of spectral equality, with larger values indicating that a particular chemical has a greater effect on the circadian clock of *A. thaliana*. The results in Chapter 4 also demonstrated the ability of the LSW–PCA clustering method to determine which elements in the periodic table (and at which concentrations) produce similar kinds of reactions in plants and to identify the different types of reactions present in a particular dataset. This application of the clustering methodology also highlighted the method's suitability in organizing and understanding multiple nonstationary time series and revealed that individual–level plant variability is more prevalent under certain treatments. Therefore, the methodologies developed in Chapters 2 and 3 have complementary utility: wavelet spectral testing can identify relatively small differences in spectral behaviour between two groups of nonstationary time series, whereas the LSW–PCA clustering methodology can identify broadly similar spectral behaviour.

Directions for future research

In Chapters 2–4 we discussed specific areas of further work within the conclusions at the end of each chapter. We describe more general potential avenues of future research next.

High dimensional multivariate time series often exhibit multi–collinearities. This suggests that such signals can be decomposed into uncorrelated principal components with possibly lower dimension than that of the original signal. In Chapter 2 we developed a clustering method that combines locally stationary wavelet time series modelling with functional principal components analysis. An interesting area of future work would be to develop a time-localised frequency domain principal components analysis method for signals that exhibit locally stationary (wavelet) behaviour. In addition, it would also be of interest to develop formal statistical procedures for testing the significance of the time–varying weights (components of an eigenvector) at a particular channel, and whether they do indeed change over time.

Consider the situation where we have training data containing a number of (nonstationary) time series having known group membership and test data with unknown group membership and we wish to classify the test data into the least dissimilar group. This is reminiscent of the soil pollutant investigation that motivated the work in this thesis. For this particular application, classification methodology could be used to determine if a new soil pollutant has no effect or a similar effect to a particular known chemical. Hence, an interesting avenue of future research would be to utilise the theoretical basis of the wavelet spectral testing in Chapter 3 to develop dissimilarity measures that take account of the distribution of the spectral co-

efficients which could be embedded within a classification procedure for nonstationary time series. Furthermore, the results of Chapter 3, suggest that it may be beneficial to employ a transform that brings the raw periodogram ordinates closer to Gaussianity and decorrelates within each scale, for example the Haar or Haar–Fisz transform. Thus, the transformed evolutionary wavelet spectrum could also be used as an alternative classification signature.

The Department for Environment, Food and Rural Affairs (DEFRA) developed 'Soil Guideline Values' (SGVs) that can be used to determine appropriate concentrations of certain chemical elements in soil. In Chapter 4, we investigated the impact of exposure to these elements at the concentrations outlined in the DEFRA report on the circadian clock of *A. thaliana*. However, it is impossible to add only one element to the growth media of the plants. Therefore, to investigate the impact of a specific element, a compound containing that element has to be added to the growth media. This makes it difficult to establish whether any effects on the clock were due to the anion or cation of each compound. An area of further work would be to derive a procedure for determining the individual effects of each element within a tested compound.

References

Abramovich, F., Bailey, T. C. and Sapatinas, T. (2000). Wavelet analysis and its statistical applications. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(1):1–29.

Andrés-Colás, N., Perea-García, A., Puig, S. and Peñarrubia, L. (2010). Deregulated copper transport affects Arabidopsis development especially in the absence of environmental cycles. *Plant Physiology*, 153(1):170–184.

Antoniadis, A., Brossat, X., Cugliari, J. and Poggi, J.-M. (2013). Clustering functional data using wavelets. *International Journal of Wavelets, Multiresolution and Information Processing*, 11(01):1350003.

Atkinson, A. D., Hill, R. R., Pignatiello Jr, J. J., Vining, G. G., White, E. D. and Chicken, E. (2017). Wavelet ANOVA approach to model validation. *Simulation Modelling Practice and Theory*, 78:18–27.

Bell-Pedersen, D., Cassone, V. M., Earnest, D. J., Golden, S. S., Hardin, P. E., Thomas, T. L. and Zoran, M. J. (2005). Circadian rhythms from multiple oscillators: lessons from diverse organisms. *Nature Reviews Genetics*, 6(7):544–556.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57:289–300.

Brillinger, D. R. (2001). *Time Series: Data Analysis and Theory*. SIAM.

Bujdoso, N. and Davis, S. J. (2013). Mathematical modeling of an oscillating gene circuit to unravel the circadian clock network of Arabidopsis thaliana. *Frontiers in Plant Science*, 4:3.

Burg, J. P. (1972). The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics, Society of Exploration Geophysicists*, 37(2):375–376.

Chiann, C. and Morettin, P. A. (1998). A wavelet analysis for time series. *Journal of Nonparametric Statistics*, 10(1):1–46.

Cho, H. (2016). A test for second-order stationarity of time series based on unsystematic subsamples. *Stat*, 5(1): 262–277.

Cho, H., Goude, Y., Brossat, X. and Yao, Q. (2013). Modeling and forecasting daily electricity load curves: a hybrid approach. *Journal of the American Statistical Association*, 108(501):7–21.

Costa, M. J., Finkenstädt, B. R., Roche, V., Lévi, F., Gould, P. D., Foreman, J., Halliday, K., Hall, A. and Rand, D. A. (2011). Estimating periodicity of oscillatory time series through resampling techniques. *Biostatistics*, 14(4):792–806.

Costa, M. J., Finkenstädt, B., Gould, P. D., Foreman, J., Halliday, K. J., Hall, A. J. W. and Rand, D. A. (2013). Inference on periodicity of circadian time series. *University of Warwick. Centre for Research in Statistical Methodology*.

Cressie, N. and Wikle, C. K. (2015). *Statistics for Spatio-temporal Data*. John Wiley & Sons.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics*, 25(1):1–37.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7):909–996.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM.

Das, S. and Nason, G. P. (2016). Measuring the degree of non-stationarity of a time series. *Stat*, 5(1):295–305.

Dodd, A. N., Salathia, N., Hall, A., Kévei, E., Tóth, R., Nagy, F., Hibberd, J. M., Millar, A. J. and Webb, A. R. (2005). Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science, American Association for the Advancement of Science*, 309(5734): 630–633.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Doyle, M. R., Davis, S. J., Bastow, R. M., McWatters, H. G., Kozma-Bognár, L., Nagy, F., Millar, A. J. and Amasino, R. M. (2002). The ELF4 gene controls circadian rhythms and flowering time in Arabidopsis thaliana. *Nature*, 419(6902): 74–77.

Dusik, V., Senthilan, P. R., Mentzel, B., Hartlieb, H., Wülbeck, C., Yoshii, T., Raabe, T. and Helfrich-Förster, C. (2014). The MAP kinase p38 is part of Drosophila melanogaster's circadian clock. *PLoS Genetics*, 10(8):e1004565.

Edwards, K. D., Akman, O. E., Knox, K., Lumsden, P. J., Thomson, A. W., Brown, P. E., Pokhilko, A., Kozma-Bognar, L., Nagy, F., Rand, D. A. and Millar, A. J. (2010). Quantitative analysis of regulatory flexibility under changing environmental conditions. *Molecular Systems Biology*, 6(1):424.

Enright, J. T. (1965). The search for rhythmicity in biological time-series. *Journal of Theoretical Biology*, 8(3):426–468.

Environmental Protection Act Part IIA Contaminated Land. *DETR Circular*(2):1-2000.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association*, 91(434):674–688.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman & Hall.

Fan, J. and Lin, S.-K. (1998). Test of significance when data are curves. *Journal of the American Statistical Association*, 93(443):1007–1021.

Fiecas, M. and Ombao, H. (2016). Modeling the evolution of dynamic brain processes during an associative learning experiment. *Journal of the American Statistical Association*, 111:1440–1453.

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., Chapin, F. S., Coe, M. T., Daily, G. C., Gibbs, H. K. and others (2005). Global consequences of land use. *Science, American Association for the Advancement of Science*, 309(5734):570–574.

Fryzlewicz, P. (2005). Modelling and forecasting financial log-returns as locally stationary wavelet processes. *Journal of Applied Statistics*, 32(5):503–528.

Fryzlewicz, P. and Nason, G. P. (2006). Haar–fisz estimation of evolutionary wavelet spectra. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(4):611–634.

Fryzlewicz, P. and Ombao, H. (2009). Consistent classification of nonstationary time series using stochastic wavelet representations. *Journal of the American Statistical Association*, 104:299–312.

Goya, M. E., Romanowski, A., Caldart, C. S., Bénard, C. Y. and Golombek, D. A. (2016). Circadian rhythms identified in Caenorhabditis elegans by in vivo long-term monitoring of a bioluminescent reporter. *Proceedings of the National Academy of Sciences*, 201605769.

Guo, W., Dai, M., Ombao, H. C. and von Sachs, R. (2003). Smoothing spline ANOVA for time-dependent spectral analysis. *Journal of American Statistical Association*, 98(463):643–652.

Hanano, S., Domagalska, M. A., Nagy, F. and Davis, S. J. (2006). Multiple phytohormones influence distinct parameters of the plant circadian clock. *Genes to Cells*, 11(12):1381–1392.

Harang, R., Bonnet, G. and Petzold, L. R. (2012). WAVOS: a MATLAB toolkit for wavelet analysis and visualization of oscillatory systems. *BMC Research Notes, BioMed Central*, 5(1):163.

Hargreaves, J. K., Knight, M. I., Pitchford, J. W., Oakenfull, R. and Davis, S. J. (2018). Clustering nonstationary circadian plant rhythms using locally stationary wavelet representations. *SIAM Multiscale Modeling and Simulation*, 16(1):184–214.

Hargreaves, J. K., Knight, M. I., Pitchford, J. W., Oakenfull, R., Chawla, S., Munns, J. and Davis, S. J. (2018). Wavelet spectral testing: application to nonstationary circadian rhythms. *arXiv preprint*, arXiv:1803.09507.

Hoagland, D. R. and Arnon, D. I. (1950). The water-culture method for growing plants without soil. *California Agricultural Experiment Station, Circular*, 347.

Holan, S. H., Wikle, C. K., Sullivan-Beckers, L. E. and Cocroft, R. B. (2010). Modeling complex phenotypes: generalized linear models using spectrogram predictors of animal communication signals. *Biometrics*, 66(3):914–924.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

Keogh, E. J. and Pazzani, M. J. (1998). An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. *Proc. of the* 4^{th} *International Conference of Knowledge Discovery and Data Mining, AAAI Press* , 98:239–243.

Kon, N., Sugiyama, Y., Yoshitane, H., Kameshita, I. and Fukada, Y. (2015). Cell-based inhibitor screening identifies multiple protein kinases important for circadian clock oscillations. *Communicative & Integrative Biology*, 8(4):e982405.

Krzemieniewska, K., Eckley, I. A. and Fearnhead, P. (2014). Classification of non-stationary time series. *Stat*, 3(1):144–157.

Lagido, C., Pettitt, J., Flett, A. and Glover, L. A. (2008). Bridging the phenotypic gap: real-time assessment of mitochondrial function and metabolism of the nematode Caenorhabditis elegans. *BMC physiology*, 8(1):7.

Leise, T. L., Indic, P., Paul, M. J. and Schwartz, W. J. (2013). Wavelet meets actogram. *Journal of Biological Rhythms*, 28(1):62–68.

Lomb, N. R. (1976). Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39(2):447–462.

Mallat, S. G. (1989a). Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Transactions of the American Mathematical Society*, 315(1):69–87.

Mallat, S. G. (1989b). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693.

Martinez, J. G., Bohn, K. M., Carroll, R. J. and Morris, J. S. (2013). A study of Mexican free-tailed bat chirp syllables: Bayesian functional mixed modeling for nonstationary acoustic time series. *Journal of American Statistical Association*, 108(502):514–526.

McClung, C. R. (2006). Plant circadian rhythms. *The Plant Cell*, 18(4):792–803.

McKay, J. L., Welch, T. D. J., Vidakovic, B. and Ting, L. H. (2012). Statistically significant contrasts between EMG waveforms revealed using wavelet-based functional ANOVA. *Journal of Neurophysiology*, 109(2):591–602.

Michael, T. P., Salomé, P. A., Yu, H. J., Spencer, T. R., Sharp, E. L., McPeek, M. A., Alonso, J. M., Ecker, J. R. and McClung, C. R. (2003). Enhanced fitness conferred by naturally occurring variation in the circadian clock. *Science, American Association for the Advancement of Science*, 302(5647):1049–1053.

Minors, D. S. and Waterhouse, J. M. (2013). *Circadian Rhythms and the Human*. Butterworth-Heinemann.

Millar, A. J., Carrington, J. T., Tee, W. V. and Hodge, S. K. (2015). Changing planetary rotation rescues the biological clock mutant lhy cca1 of Arabidopsis thaliana. *bioRxiv at Cold Spring Harbor Laboratory*.

Millar, A. J. and Kay, S. A. (1991). Circadian control of cab gene transcription and mRNA accumulation in Arabidopsis. *The Plant Cell*, 3(5):541–550.

Moore, D. S. (2007). *The Basic Practice of Statistics (Vol. 2)*. WH Freeman, New York.

Moore, A., Zielinski, T. and Millar, A. J. (2014). Online period estimation and determination of rhythmicity in circadian data, using the BioDare data infrastructure. *Methods in Molecular Biology*, 1158:13–44.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and its Application*, 2:321–359.

Morris, J. S., Baladandayuthapani, V., Herrick, R. C., Sanna, P. and Gutstein, H. (2011). Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *Annals of Applied Statistics*, 5(2A):894–923.

Murashige, T. and Skoog, F. (1962). A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiologia Plantarum*, 15(3):473–497.

Nason, G. (2010). *Wavelet Methods in Statistics with R (use R)*. Springer Science & Business Media.

Nason, G. (2013). A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):879–904.

Nason, G. P. and Savchev, D. (2014). White noise testing using wavelets. *Stat*, 3(1):351–362.

Nason, G. P. and Silverman, B. W. (1995). The stationary wavelet transform and some statistical applications. *Wavelets and Statistics*, 281–299.

Nason, G. P., von Sachs, R. and Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):271–292.

Nason, G. P. and Stevens, K. (2015). Bayesian Wavelet Shrinkage of the Haar-Fisz Transformed Wavelet Periodogram. *PloS one, Public Library of Science*, 10(9):e0137662.

Nicholson, F. A., Smith, S. R., Alloway, B. J., Carlton-Smith, C. and Chambers, B. J. (2003). An inventory of heavy metals inputs to agricultural soils in England and Wales. *Science of the Total Environment*, 311(1):205–219.

Oakenfull, R., Hargreaves, J. K., Knight, M. I., Pitchford, J. W. and Davis, S. J. (2018). Out of the sewage rises new maths... *In Preparation*.

Ogden, T. R. (1997). On preconditioning the data for the wavelet transform when the sample size is not a power of two. *Communications in Statistics-Simulation and Computation*, 26(2):467–486.

Oh, H.-S., Ammann, C. M., Naveau, P., Nychka, D. and Otto-Bliesner, B. L. (2003). Multiresolution time series analysis applied to solar irradiance and climate reconstructions. *Journal of Atmospheric and Solar-terrestrial Physics*, 65(2):191–201.

Percival, D. B. and Walden, A. T. (2006). *Wavelet Methods for Time Series Analysis (Vol. 4)*. Cambridge University Press, Cambridge.

Perea-García, A., Andrés-Bordería, A., de Andrés, S. M., Sanz, A., Davis, A. M., Davis, S. J., Huijser, P. and Peñarrubia, L. (2016a). Modulation of copper deficiency responses by diurnal and circadian rhythms in Arabidopsis thaliana. *Journal of Experimental Botany*, 67(1):391–403.

Perea-García, A., Sanz, A., Moreno, J., Andrés-Bordería, A., de Andrés, S. M., Davis, A. M., Huijser, P., Davis, S. J. and Peñarrubia, L. (2016b). Daily rhythmicity of high affinity copper transport. *Plant Signaling & Behavior*, 11(3):e1140291.

Plautz, J. D., Straume, M., Stanewsky, R., Jamison, C. F., Brandes, C., Dowse, H. B., Hall, J. C. and Kay, S. A. (1997). Quantitative analysis of Drosophila period gene transcription in living animals. *Journal of Biological Rhythms*, 12(3): 204–217.

Price, T. S., Baggs, J. E., Curtis, A. M., FitzGerald, G. A. and Hogenesch, J. B. (2008). WAVECLOCK: wavelet analysis of circadian oscillation. *Bioinformatics*, 24(23): 2794–2795.

Priestley, M. B. (1965). Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society, Series B (Methodological)*, 27:204–237.

Priestley, M. and Rao, T. S. (1969). A test for non-stationarity of time-series. *Journal of the Royal Statistical Society, Series B (Methodological)*, 31:140–149.

Priestley, M. B. (1982). *Spectral Analysis and Time Series*. Academic Press.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer.

Rouyer, T., Fromentin, J.-M., Stenseth, N. C. and Cazelles, B. (2008). Analysing multiple time series and extending significance testing in wavelet analysis. *Marine Ecology Progress Series*, 359:11–23.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

von Sachs, R. and Neumann, M. H. (2000). A wavelet-based test for stationarity. *Journal of Time Series Analysis*, 21(5):597–613.

Sanchez, A., Shin, J. and Davis, S. J. (2011). Abiotic stress and the plant circadian clock. *Plant Signalling & Behaviour*, 6(2):223–231.

Senesil, G. S., Baldassarre, G., Senesi, N. and Radina, B. (1998). Trace element inputs into soils by anthropogenic activities and implications for human health. *Chemosphere*, 39(2):343–377.

Shumway, R. H. (1988). Applied statistical time series analysis. *Statistics & Probability Letters*, 63(3):307–314.

Shumway, R. H. (2003). Time-frequency clustering and discriminant analysis. *Statistics & Probability Letters*, 63(3):307–314.

Shumway, R. H. and Stoffer, D. S. (2000). *Time Series Analysis and its Applications*. Springer Science & Business Media.

Silver, S., Budd, K., Leahy, K. M., Shaw, W. V., Hammond, D., Novick, R. P., Willsky, G. R., Malamy, M. H. and Rosenberg, H. (1981). Inducible plasmid–determined resistance to arsenate, arsenite, and antimony (III) in Escherichia coli and Staphylococcus aureus. *Journal of Bacteriology*, 146(3):983–996.

Southern, M. M. and Millar, A. J. (2005). Circadian genetics in the model higher plant, Arabidopsis thaliana. *Methods in Enzymology*, 393:23–35.

Stiernagle, T. (1999). Maintenance of C. elegans. *Wormbook*, 1-11.

Straume, M., Frasier-Cadoret, S. G. and Johnson, M. L. (2002). Least-squares analysis of fluorescence data. *Topics in Fluorescence Spectroscopy*, 177–240.

Tavakoli, S. and Panaretos, V. M. (2016). Detecting and localizing differences in functional time series dynamics: a case study in molecular biophysics. *Journal of the American Statistical Association*, 111(515):1020–1035.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Torrence, C. and Compo, G. P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78.

Van Bellegem, S. and von Sachs, R. (2008). Locally adaptive estimation of evolutionary wavelet spectra. *Annals of Statistics*, 36(4):1879–1924.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. John Wiley & Sons.

Vidakovic, B. (2001). Wavelet-based functional data analysis: theory, applications and ramifications. *Proceedings of the 3rd Pacific Symposium on Flow Visualization and Image Processing*, Maui, HI.

Vitaterna, M. H., Takahashi, J. S. and Turek, F. W. (2001). Overview of circadian rhythms. *Alcohol Research and Health*, 25(2):85–93.

Yang, X., Pan, H., Wang, P. and Zhao, F. (2016). Particle-specific toxicity and bioavailability of cerium oxide (CeO2) nanoparticles to Arabidopsis thaliana. *Journal of Hazardous Materials*, 322:292–300.

Zielinski, T., Moore, A. M., Troup, E., Halliday, K. J. and Millar, A. J. (2014). Strengths and limitations of period estimation methods for circadian data. *PloS one, Public Library of Science*, 9(5):96462.