Exploring Peaks and Trends in Fungi Nuclei Division Through SiZer Analysis

Client: Grace Mclaughlin
Investigator: Geonhyeok Jeong

2024-04-11

1. Abstract.

The client, Grace, seeks to quantify the burstiness of fungi nuclei. To address this inquiry, we employ SiZer (Significant Zero Crossing), a non-parametric method used to approximate the response function and its derivatives. SiZer examines how these functions change across the range of the explanatory variable. We apply SiZer to two types of data: non-transformed data and binned data. SiZer with non-transformed data aims to identify statistically significant peaks, while SiZer with binned data aims to uncover trends in the division of nuclei.

Results from SiZer with non-transformed actual data reveal no statistically significant findings due to data sparseness. However, analysis of simulated data shows significant peaks of division of nuclei based on the larger number of data.

For SiZer with binned actual data, 2 out of the 5 binned actual data show the decreasing trends. The assumption that it would show the increasing trend because of an idea that the more number of nuclei leads to the more division. Also, the Kolmogorov-Smirnov test, which is statistically significant, supports the result, saying the distribution of counts of binned data does not follow uniform distribution. Conversely, some SiZer with binned simulated data indicate the increasing trends. This is because the cycle of division of nuclei is manually decided when Client Grace build a model. It is supported by the KS test as well.

2. Data

The dataset provided by the client comprises 10 variables, consisting of 5 actual data and 5 simulated data. Each variable has a different number of observations, ranging from a minimum of 1.7 to a maximum of 495. The data represent the time points at which nuclei division occurs. For instance, if a nucleus divides at time 10, the corresponding observation is recorded as 10.

3. Analysis.

In this section, we employ the SiZer method to find out patterns in the timing of nuclei division. Before getting into the results, let's first outline the SiZer method. Initially, we construct kernel density estimation plots using various window widths—a common technique in statistical smoothing methods. Subsequently, we generate derivative plots that highlight significant increases or decreases in density across the range of the explanatory variable. Finally, we extract insights regarding any noticeable patterns or trends present in the dataset.[1]

[1]Hannig and J. S. Marron (2006): Advanced Distribution Theory for SiZer, Journal of the American. Statistical Association

3.1 SiZer with non-transformed data

3.1.1 SIZer with actual data 1

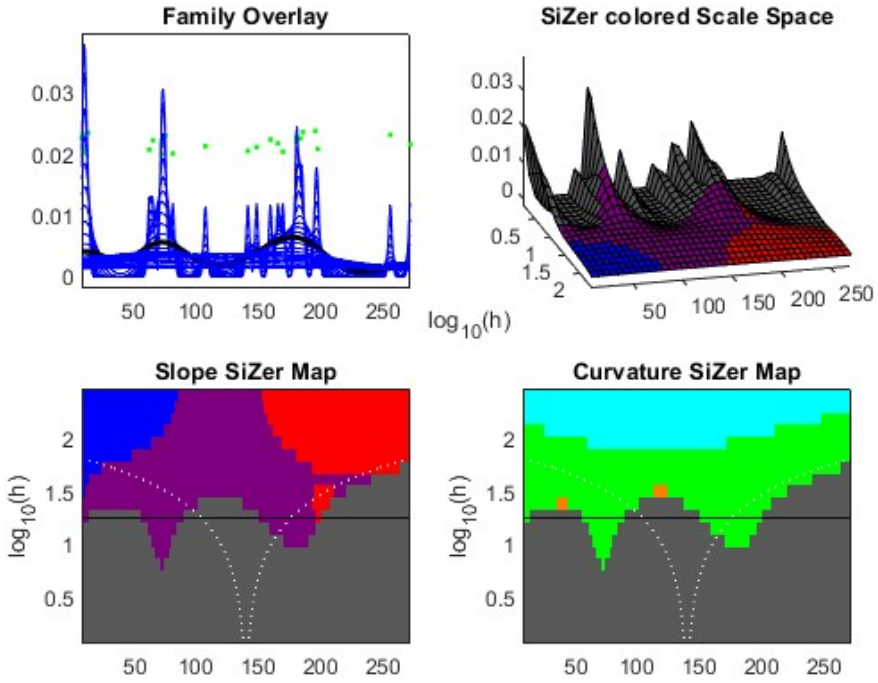

Figure 1: SiZer analysis with actual non-transformed data 1. The figure comprises four plots: The upper left plot displays the kernel density estimation with all window widths. The upper right plot depicts a 3D visualization of the density estimation and SiZer analysis. The lower left plot represents the SiZer map, indicating significant slopes with color-coded regions. The lower right plot shows the curvature SiZer map, highlighting significant concave and convex features.

This section presents the SiZer analysis results obtained using Marron's SiZer function in MATLAB for one of the actual datasets. The analysis produces four plots, each providing unique insights into the dataset. Located in the upper left quadrant, this plot illustrates the kernel density estimation with varying window widths, assuming a normal distribution of the data. The density estimation plot reveals the distribution of the data across different window sizes. Positioned beneath the density estimation plot, the SiZer map displays the significance of derivatives across different window sizes, which is y-axis of the map. The color indicates the direction and significance of slopes: blue denotes a significant upward slope, red indicates a significant downward slope, purple signifies an insignificant slope, and gray represents regions with insufficient data for inference. Notably, areas of overly smoothed data, where log10(h) = 2, exhibit a significant increase on the left side and a significant decrease on the right side. Next to the SiZer map, the Curvature SiZer Map highlights significant concave or convex regions within the dataset. Cyan indicates significant concavity, while orange denotes significant convexity. Similar to the SiZer map, green areas indicate insignificance and gray areas indicate regions with insufficient data for inference. Located in the upper right quadrant, the 3D plot depicts the relationship between density estimation and SiZer analysis in a three-dimensional space. Despite this analysis, the dataset's limited size (only 39 observations) hinders the identification of features. Additional SiZer results for actual datasets 2, 3, 4, and 5 are presented in Section 5.1, revealing similar non-meaningful outcomes.

Click : Go to Section 5.1

3.1.2 SIZer with Simulated non-transformed data 1

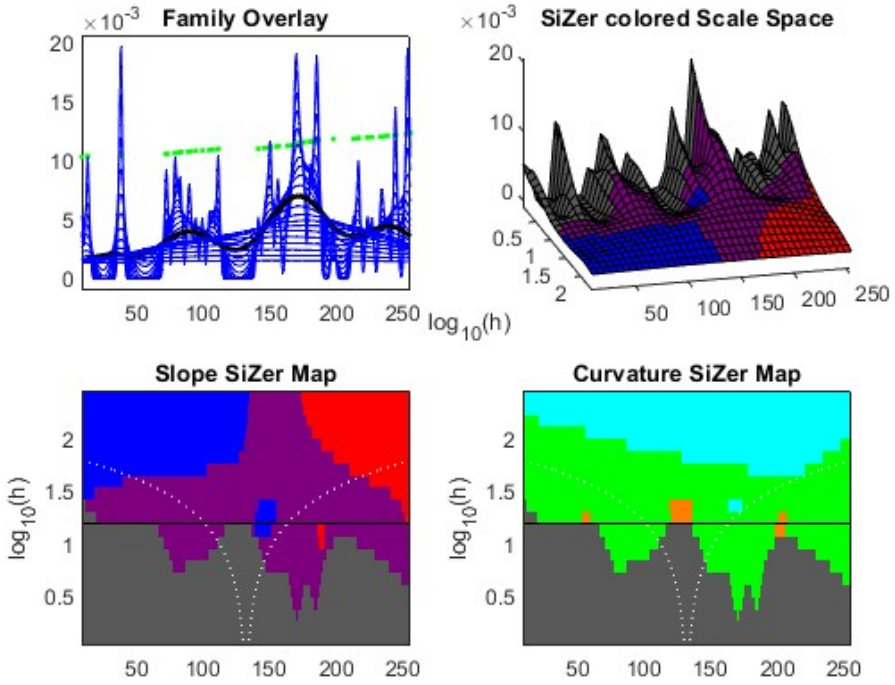

Figure 2: SiZer analysis with simulated non-transformed data 1.

In this section, we present the SiZer analysis results for simulated data 1. The SiZer map exhibits a more diverse color distribution compared to the actual dataset, attributed to the larger sample size of 62 observations. Notably, the region where the window width, log10(h), equals 2 displays a pattern similar to that observed in the actual dataset analysis. However, noteworthy is the observation near the black line in the SiZer map, indicating a significant increase in the number of nuclei divisions around time 150, followed by a decrease near time 200. This peak can be seen in the curvature SiZer map. These findings suggest that in the simulated model, a significant number of nuclei divisions occur around time 150. Further analysis of the curvature SiZer map confirms this observation. Additionally, four more simulated datasets are presented in Section 5.2.

Click : Go to Section 5.2

3.2 SiZer with binned data

3.2.1 SiZer with actual binned data 2

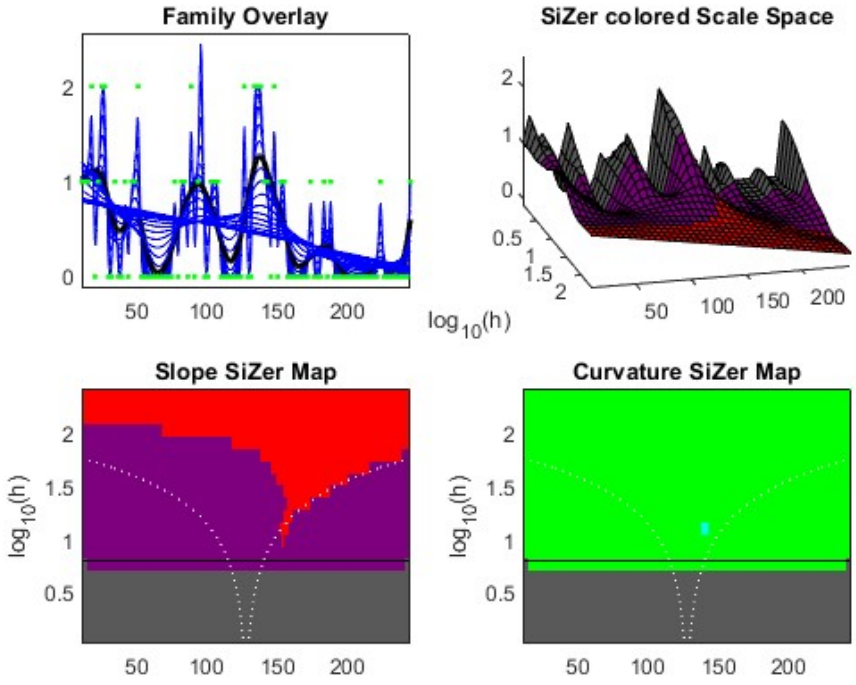

Figure 3: SiZer analysis with actual binned data 2

To overcome the limitations encountered with SiZer in detecting trends in the division of nuclei, we employ SiZer with binned data. Binned data refers to the aggregation of counts within predefined bins. In our analysis, time intervals are divided into 100 bins, and the count of nuclei divisions within each bin is used. The plot of the binned data indicates the trend of division of nuclei. For example, if the number of divisions of nuclei increases as time progresses, it would indicate a significant increasing slope; if it decreases, it would show a significant decreasing slope, or a non-significant slope if neither increasing nor decreasing. Before analysis, we assumes it would show an increasing slope. This assumption is based on the understanding that the number of nuclei increases as it divides, resulting in a higher occurrence of division events. However, the results show a different trend compared to the actual binned data 2. It indicates a significantly decreasing trend, as evidenced by the red area where log10(h) equals 2 in the Slope SiZer map. Since there is no convexity and curvature, all areas are green. (The result with actual binned data 1 shows non-significance, so we decide to present the results with actual binned data 2).

Table 1: The result of KS test

KS.test.Uniform	Statistics	p.value
	0.26347	0.004009

To support this result, we conduct a Kolmogorov-Smirnov (KS) test. This test compares the empirical cumulative distribution function of our data to a specified theoretical distribution. The test result (shown in Table 1 below) rejects the null hypothesis, indicating that the data does not follow a uniform distribution (p-value = 0.004009). It supports the significant trend of the result above as it suggests that data is skewed

in certain area. Typically, a p-value less than 0.05 is considered statistically significant. Additional SiZer results for actual binned datasets 1, 3, 4, and 5 are presented in Section 5.3.

Click : Go to Section 5.3

3.2.2 SiZer with simulated binned data 2

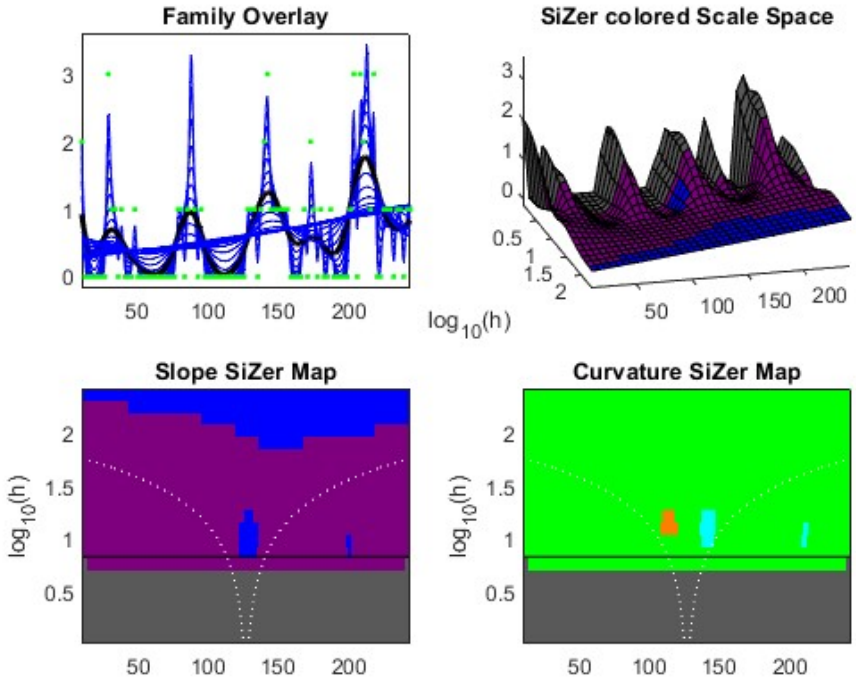

Figure 4: SiZer with simulated binned data 2

This is the result for simulated data 2. Unlike the result of the actual data 2, it shows the increase ternd as it has blue area in Slope SiZer map. We believe it is because the cycle of division of nuclei is set manually before builing a model. As a result, the number of division of nuclei should increase as the number of nuclei increaes by time.

Table 2: The result of KS test

KS.test.Uniform	Statistics	p.value
	0.20112	0.01127

Also, to see if these occurrences are uniformly distributed, I conduct another KS test. The resulting p-value of 0.01127 indicates rejection of the null hypothesis, suggesting that the data does not follow a uniform distribution. Additional SiZer results for simulated binned datasets 1, 3, 4, and 5 are presented in Section 5.4.(Like 3.2.1, simulated data 1 is not used because it does not show significance trend.)

Click : Go to Section 5.4

4. ChatGPT

4.1 Check grammar and contents

The client, Grace, would like to find out how we can quantify the level of burstiness of fungi nuclei. To solve this question, we use SiZer(significant Zero Crossing), which is a non-paramatric method to approximate the response function and its derivatives and then exaaamines how thos functions change across the rage of the explanatory variable. As a result of the analysis, due to lack of the data, the results of the functions do not show valuable information. The other analysis would be needed for further investigation.

4.2 Check if my title goes along with the contents. if not, recommend options.

Using Sizer analysis, finding out trends of fungi nuclei division.

recommended : Exploring Trends in Fungi Nuclei Division Through SiZer Analysis

5. Appendix

5.1 SiZer with actual non-transformed data 2,3,4,5

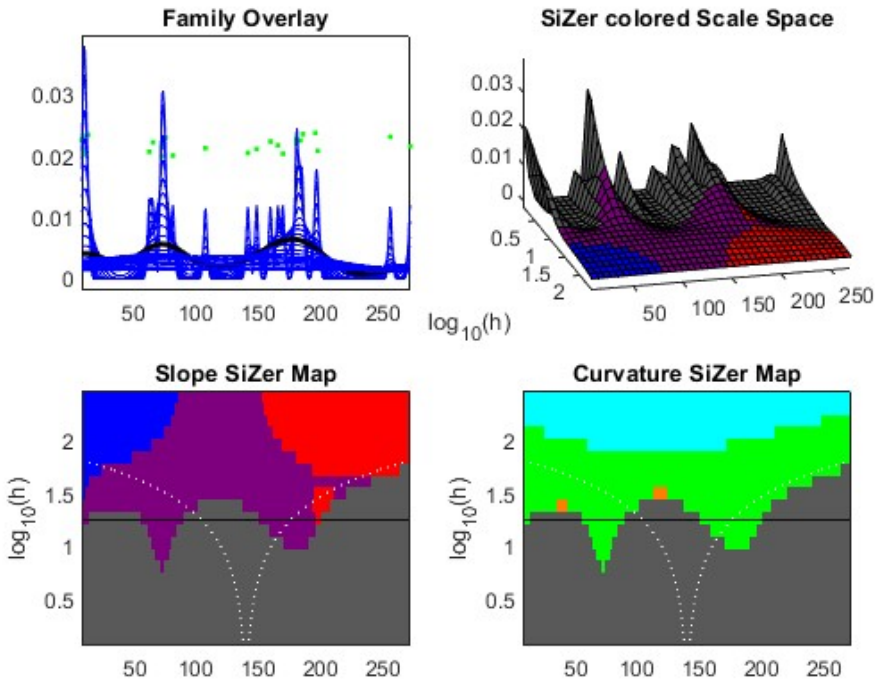

Figure 5: SiZer with actual non-transformed data 2

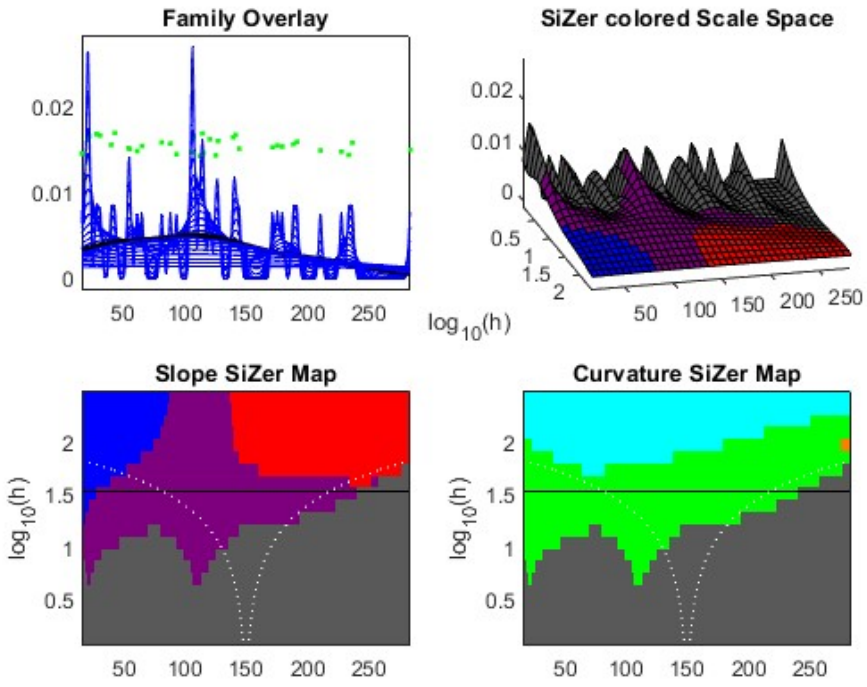

Figure 6: SiZer with actual non-transformed data 3

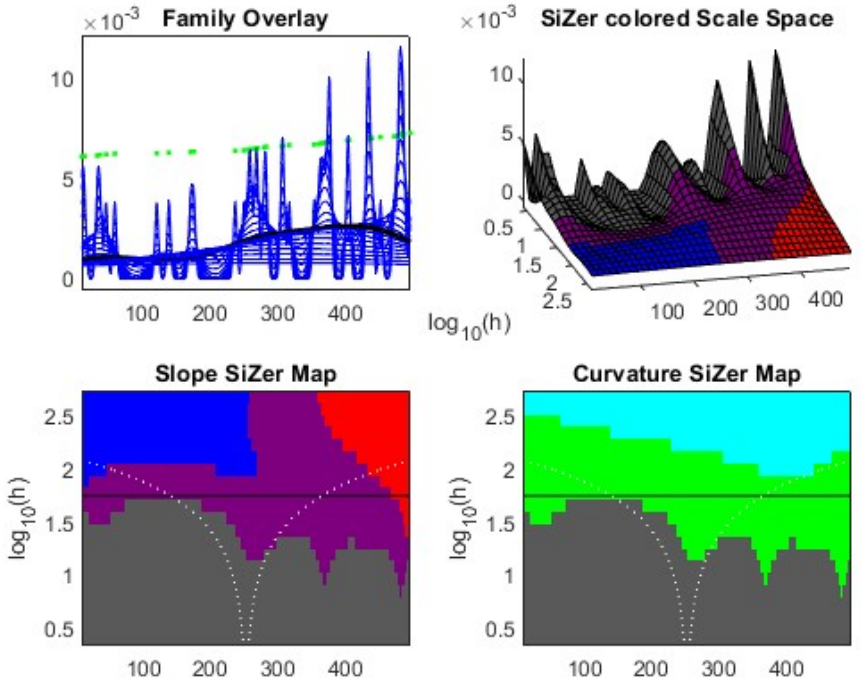

Figure 7: SiZer with actual non-transformed data 4

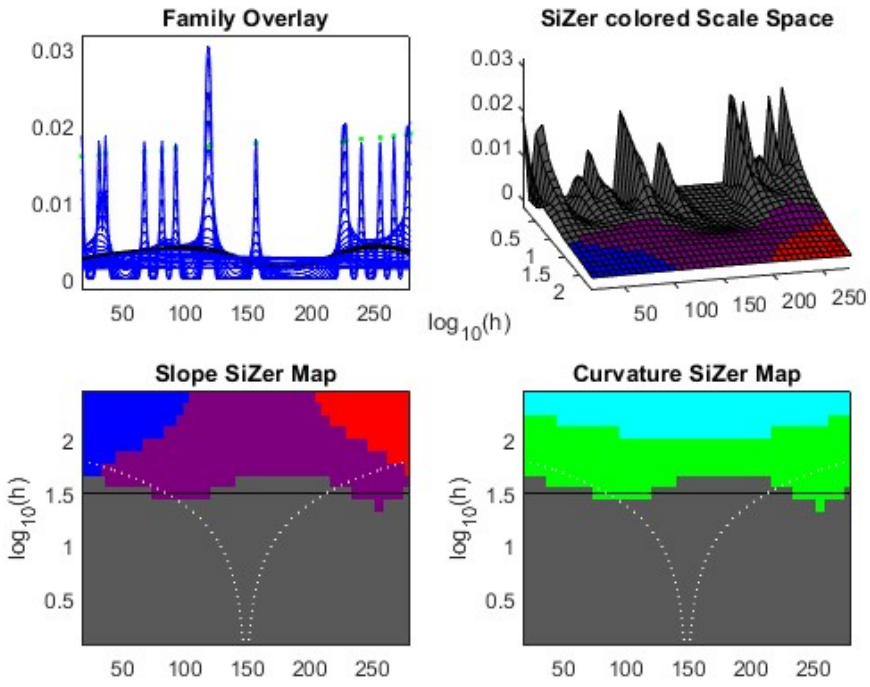

Figure 8: SiZer with actual non-transformed data 5

5.2 SiZer with simulated non-transformed data 2,3,4,5

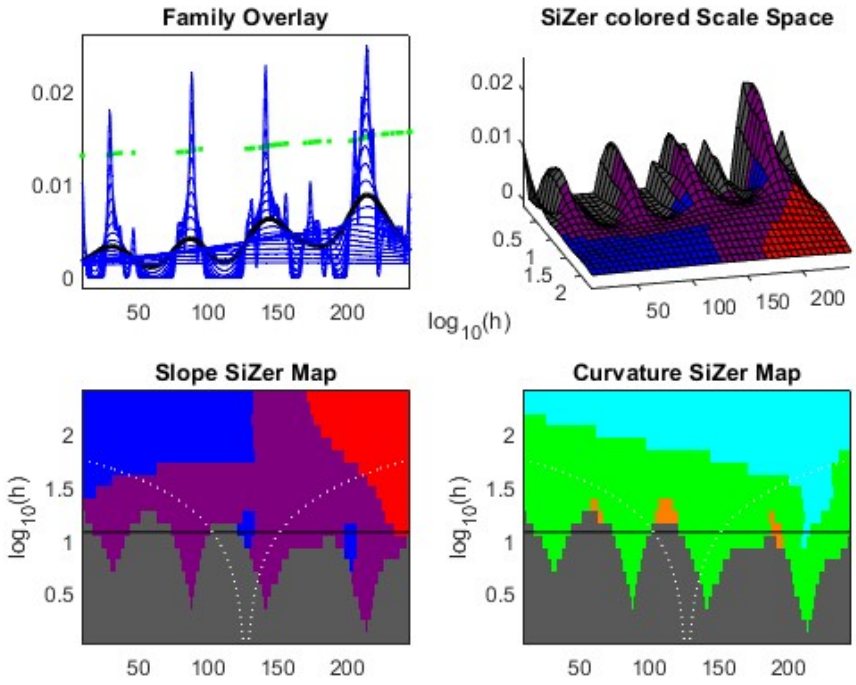

Figure 9: SiZer with simulated non-transformed data 2

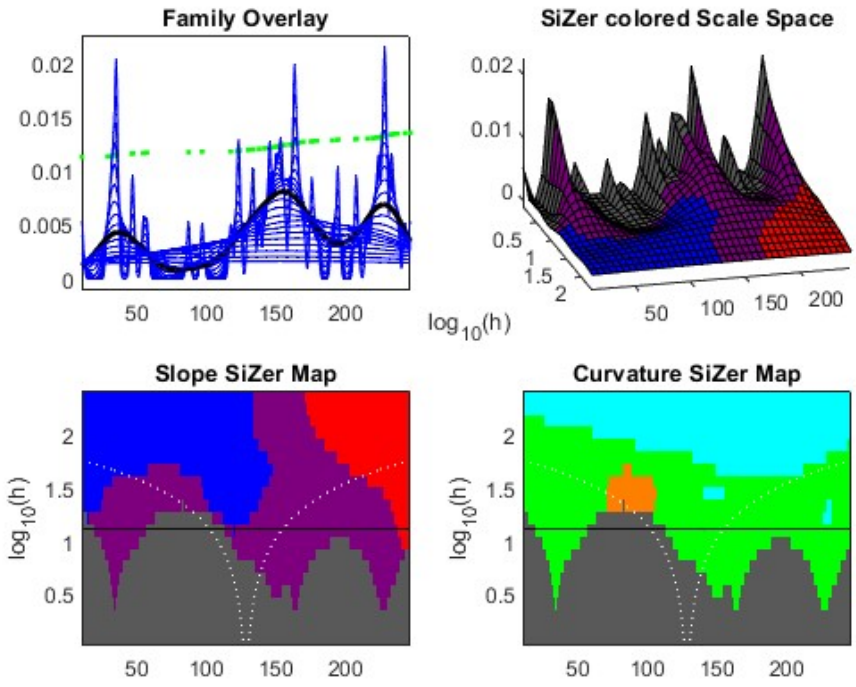

Figure 10: SiZer with simulated non-transformed data 3

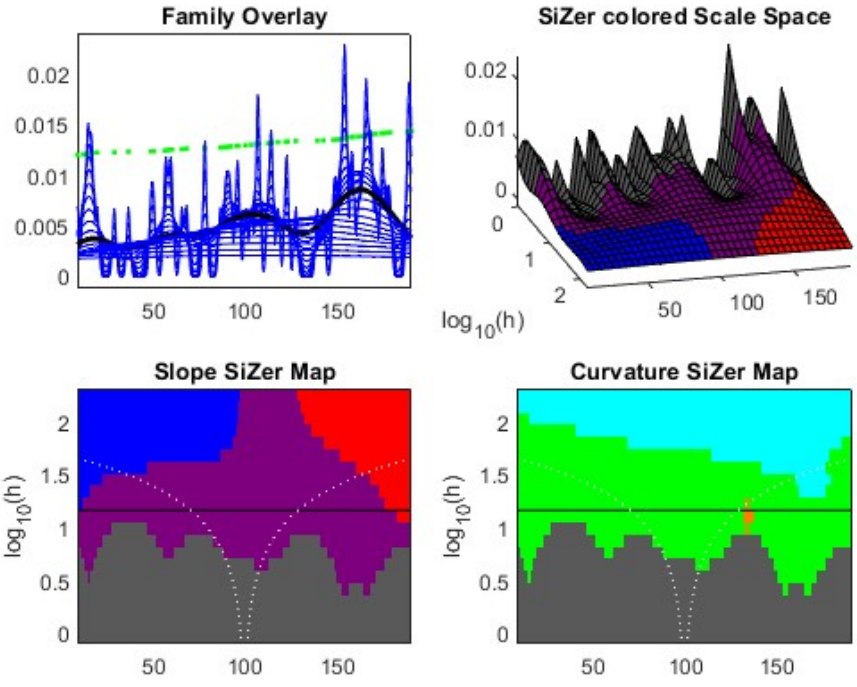

Figure 11: SiZer with simulated non-transformed data 4

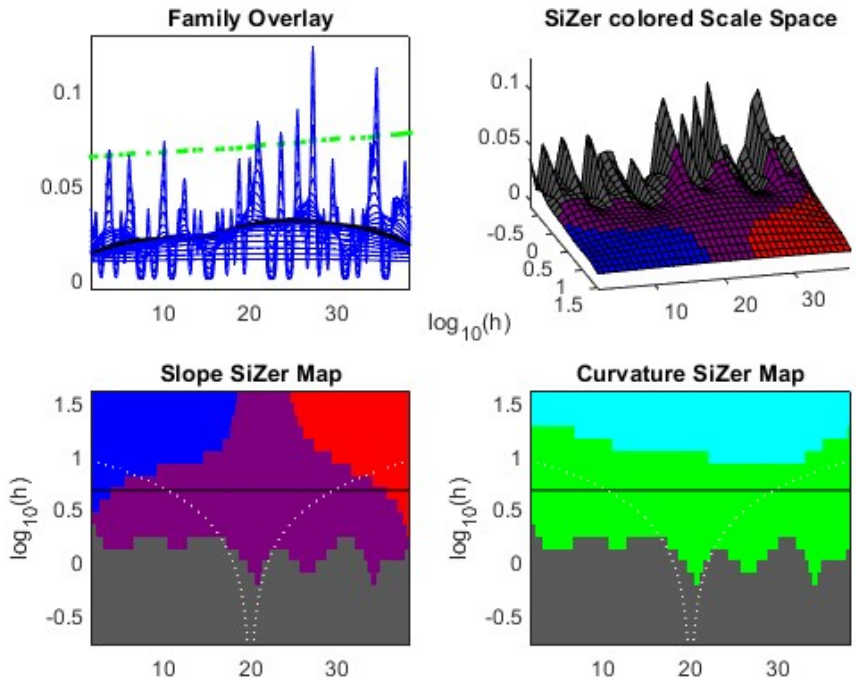

Figure 12: SiZer with simulated non-transformed data 5

5.3 SiZer with actual binned data 1,3,4,5

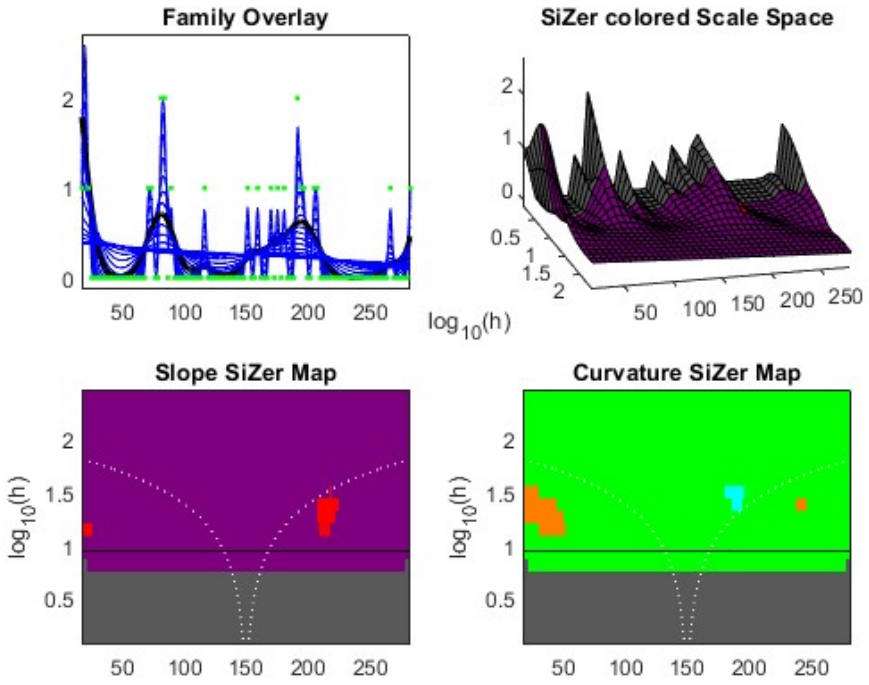

Figure 13: SiZer with actual binned data 1

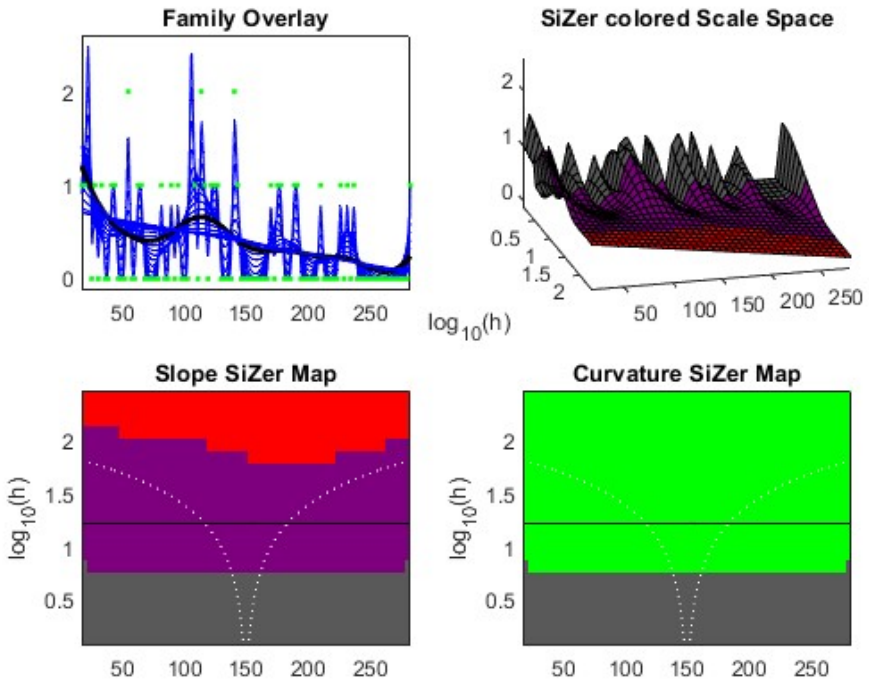

Figure 14: SiZer with actual binned data 3

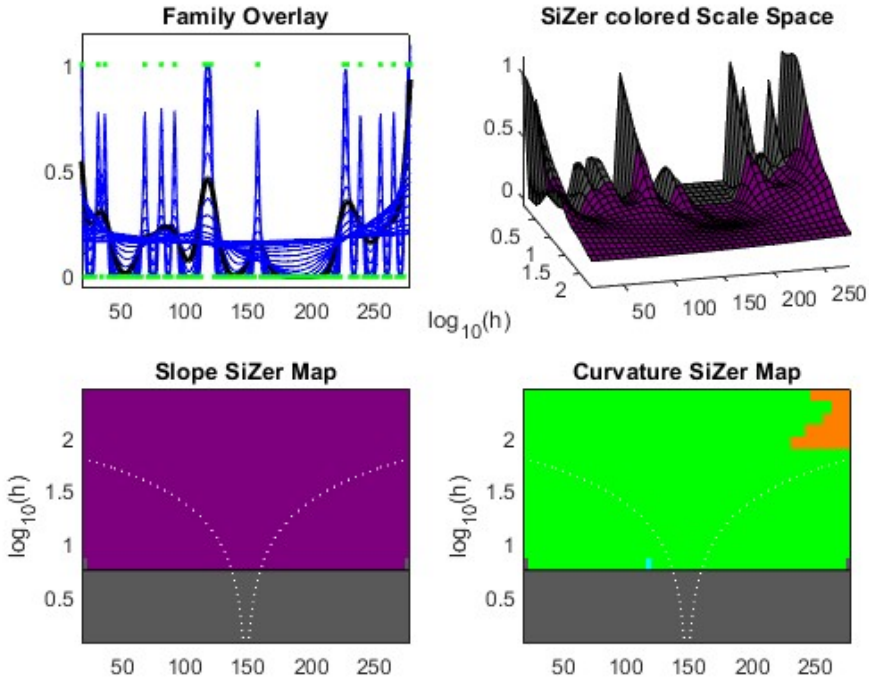

Figure 15: SiZer with actual binned data 4

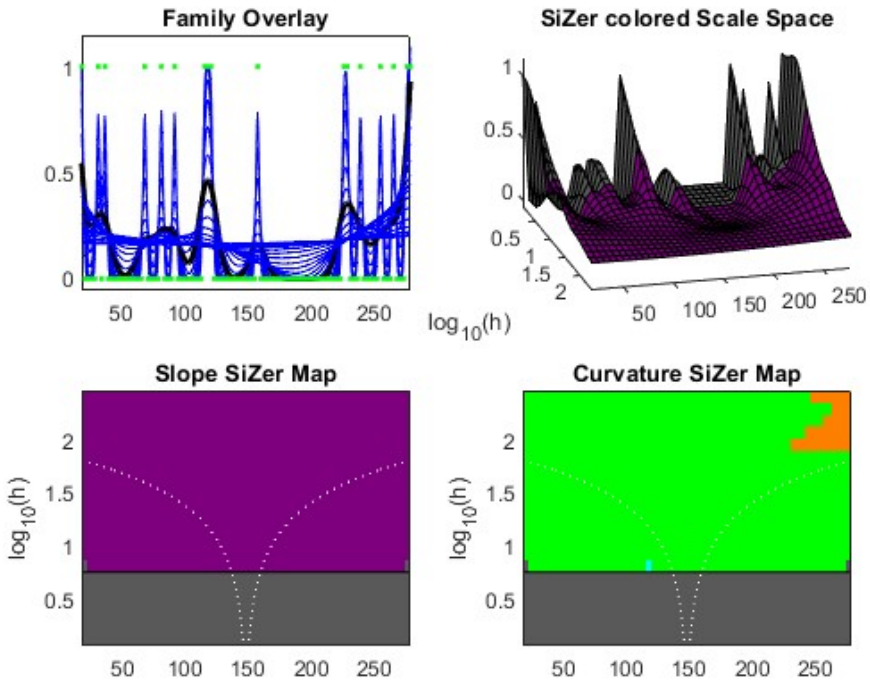

Figure 16: SiZer with actual binned data 5

5.4 SiZer with simulated binned data 1,3,4,5

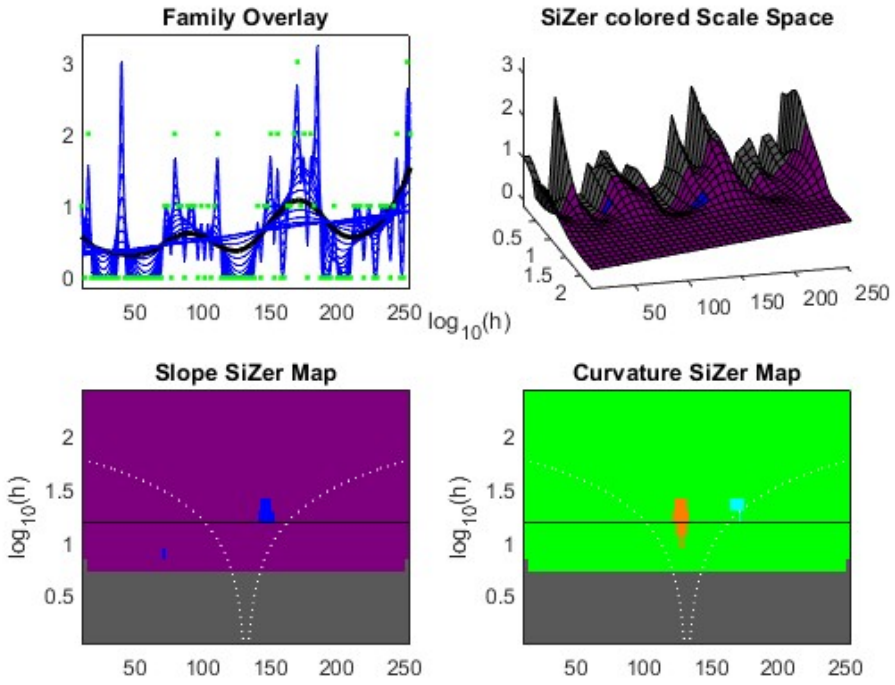

Figure 17: SiZer with simulated binned data 1

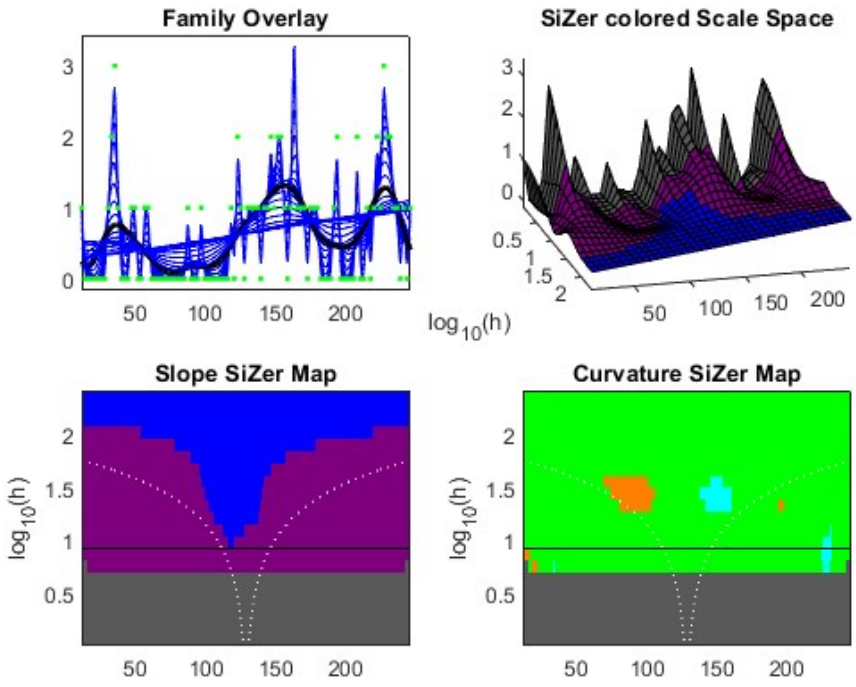

Figure 18: SiZer with simulated binned data 3

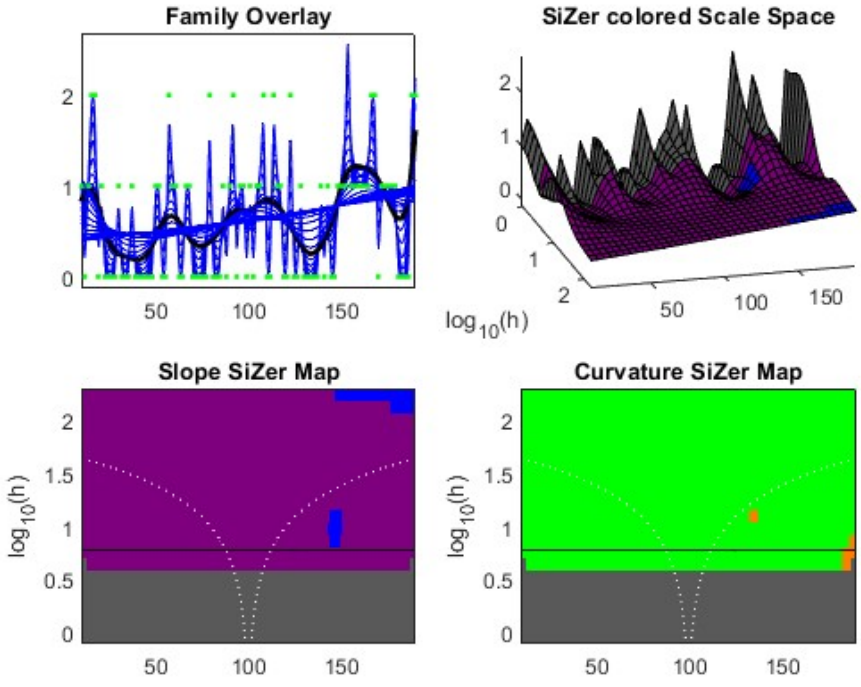

Figure 19: SiZer with simulated binned data 4

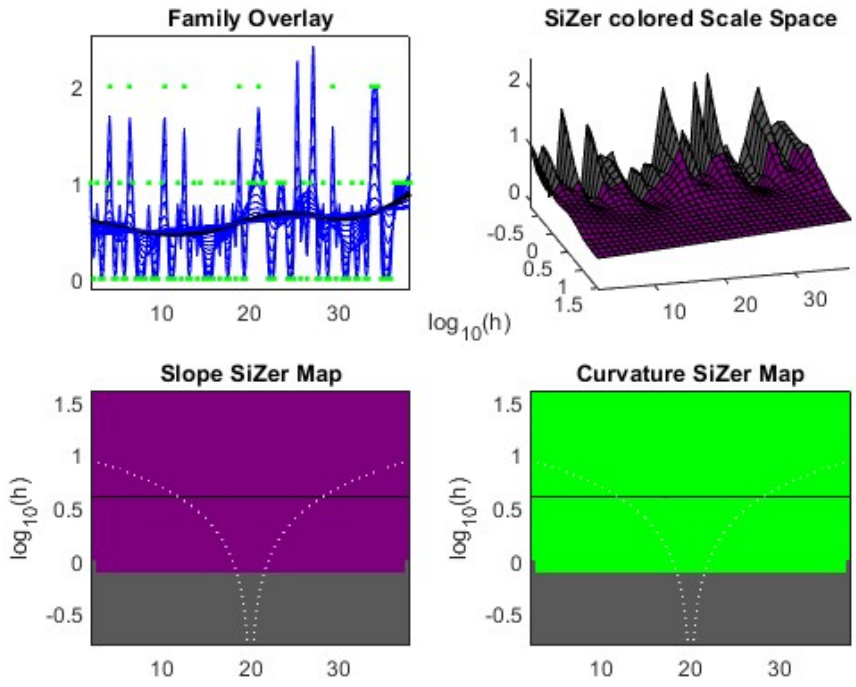

Figure 20: SiZer with simulated binned data 5