Nonparametric Estimation of Microbial Temporal Dynamics

——Eliminate the Impact of Missing Data

Submitted in partial fulfillment of the requirements

for the degree of Bachelor of Statistics, Honours

Ziye Tian

Dalhousie University

Supervisor: Dr. Toby Kenney

April 23, 2024

Introduction

The microbiome is the collection of all microbes, such as bacteria, fungi, viruses, and their genes, that naturally live on our bodies and inside us. In 1988, Whipps et al. proposed the definition of the microbiome, which is a combination of the words "micro" and "biome". The specifically defined microbiome is "a characteristic microbial community occupying a reasonably well-defined habitat which has distinct physio-chemical properties. The term thus not only refers to the microorganisms involved but also encompasses their theatre of activity." (Berg G., 2020) The definition shows the microbiome is closely related to ecology and that microbiota with different functions interact with the ecological environment. After that, more and more researchers have become conscious of the close connection between the microbiome and the human body, and the microbiome plays an important role in the human body. There are many different microbiota in the human body which is called the human microbiome. The human microbiome consists of 10-100 trillion symbiotic microbial cells in each human body. (Luke K.U., 2012) For example, Escherichia coli which is in the human intestine, and Veillonella which is in the oral saliva are all beneficial microbiota of humans. Similarly, viruses and harmful bacteria in the human body can cause illness and infection, thereby leading to death. Therefore, the microbiome which is in the human body can maintain the internal balance of the human body. The human and the microbiome which is in the human are co-evolved. These microorganisms ensure the orderliness of the human body's ecosystem. The microbiome inside the human body is constantly changing over time. However, many researchers treat microbial communities as static and the time sequence of the human microbiome is ignored.

To study the dynamic changes in the human microbiome, researchers collect genes from the human microbiome and use gene sequencing and other methods to research it. Then they can observe the changes in the human microbiome genes, and analyze the human microbiome genes to acquaint themselves and the role that the microbiome plays in the ecological balance of the human body. These dynamic gene data are the time series with sequential characteristics. By researching the sequential characteristics of the dynamic gene data, we can find the changes of the gene following the time so that we can know the cycle and predict with the no rule of change. We also can use the sequential characteristics of the dynamic gene data to know the interaction between different genes. Then we can grasp the relationship with two genes.

Time series data can effectively explain the change of variables over time which can research the relation between the sample on one day, and the sample on the following day. How to find the relation between the sample on one day, and the sample on the following day from time series data has become a difficult problem. Therefore, non-parametrically estimating the relationship between the sample on one day and the sample on the second question has become a good way to solve this problem.

Nonparametric estimation is a statistical method that allows the functional form of a fit to data to be obtained in the absence of any guidance or constraints from theory. As a result, the procedures of nonparametric estimation have no meaningful associated parameters. Using nonparametric estimation does not need to have the distribution of the data which makes the research more convenient. However, research on the human microbiome genes needs to have long-term data, and most of the data is collected manually. Subjects cannot record the data every day such as being unable to attend the lab or provide the sample on that day, thus cause missing data. Due to the influence of the missing data, using nonparametric estimation methods to find the relationship between the sample on one day, and the sample on the following day has the larger error. In this case, how to reduce the impact of missing data and ensure the accuracy of nonparametric estimation has become a difficult question in researching the human microbial genome. This report uses the moving picture dataset published by Caporaso, J.G. in 2011, which has the missing data as the time series dataset to use the nonparametric estimation method to find the relationship between the samples of one day and the samples of the next day.

This data collects DNA and uses the PCR to sequence, so it occurs the sequencing depth issues. Sequencing depth can correct some problems that occur during the sequencing process and improve sequencing accuracy. However, due to the influence of factors unrelated to the sample, the total number of reads recorded by the sequencer changes randomly, so consider proportions to eliminate the multiplicative noise. However, proportions will also be affected by counts of unrelated microbes, causing data instability. Therefore, the log ratio is used to make it more stable.

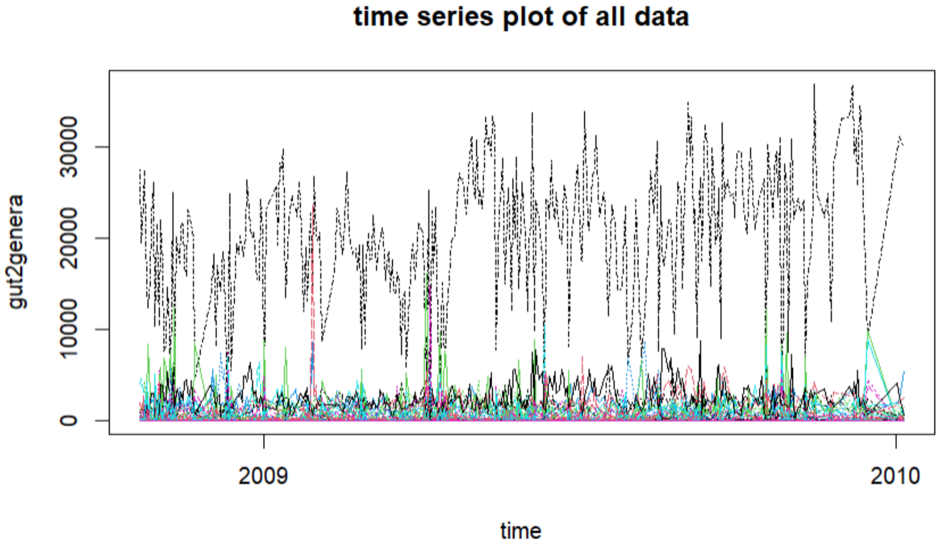
Method

Missing Data Type

There are three types of missing data, which are Missing Completely at Random (MCAR), Missing at Random (MAR) and Not Missing at Random (NMAR). MCAR means the events that lead to any particular data item being missing are independent both of observable variables and unobservable parameters of interest and occur entirely at random. MAR means when the missingness is not random, but where missingness can be fully accounted for by variables where there is complete information. NMAR is data that is neither MAR nor MCAR.

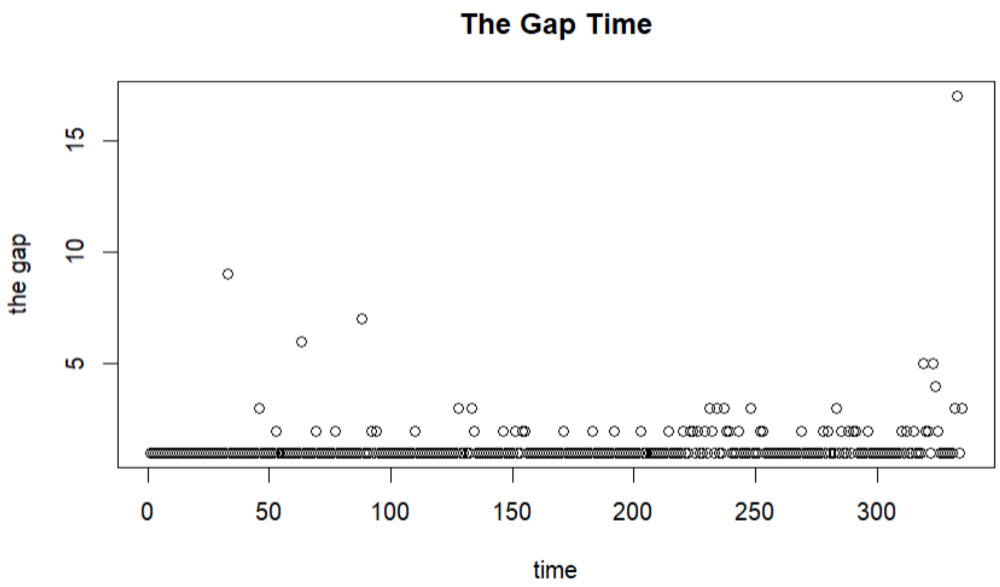
Dataset

This report uses the moving picture dataset published by Caporaso, J.G. in 2011 as the basic data for research. The data is collected daily from three body sites (gut (feces), mouth, and skin (left and right palms)) of two healthy subjects (one male and one female). DNA data are collected, and PCR, sequencing, and read quality filtering are performed. Based on the original paper data, we use the pre-processed dataset. This report focuses on the gut microbiome of person 2, aggregated at genus level. The time series plot of this dataset is:



Because most of the data values are small no obvious results can show during the research process, which is not conducive to nonparametric estimation. Therefore, the (Bacteroides) genus with the largest abundance is selected for the nonparametric estimate.

Because the data date is from 2008-10-21 to 2010-01-06, but there are just 336 days in the data, then calculate the gap time to visualize the missing data. The gap time is the time between two consecutive samples. For example, the first data of gap time is the difference date from 2008-10-21 to 2008-10-22. The gap time plot is:



Research Tool

All research processes in this report use R software for visualization, screening, nonparametric estimation and model fitting of original data.

Research Methods

Nonparametric Estimation

In Bayesian inference, the known prior distribution and conditional probability are used to calculate the posterior distribution. Because the form of the conditional probability is unknown in actual data, nonparametric estimation is useful. Nonparametric estimation refers to methods that directly use prior known categories of learning samples to conduct statistical testing and judgment analysis without considering the original data distribution, making assumptions about parameters or assuming a mathematical model. Use the sample to make inferences about the overall distribution and then make assumptions. Because the nonparametric estimation does not assume a mathematical model, it can avoid large errors caused by inappropriate assumptions about the overall distribution, it has the advantage of better robustness. Nonparametric estimation generally uses training data to directly estimate the probability density. Nonparametric estimation commonly uses Kernel Density Estimation (KDE), Histograms, Empirical Distribution Function (EDF) and so on to estimate the probability density of the data. This report uses Histograms to conduct an estimate of the probability density.

Histograms

Histograms are the simplest and most intuitive method for nonparametric estimation of probability density. The topic of Histograms is to divide the variable X into K bins with equal spacing and count the number falling into each bin, we can get

$$H(i) = \sum_{j=1}^{n} I(X \in B_i) \ \forall i = 1,2,\ldots,m$$

Then use $\hat{p}(i) = \frac{H(i)}{\sum_{j=1}^{m} H(j)}$ to get the approximates the actual p(X).

Log transform

Log transform takes the logarithm of a variable. Log transform can make data normalize. Therefore, log transform is used to keep the data stable, and facilitate the comparison of data between different ranges. It also helps to fit some nonlinear relations.

Linear regression

linear regression is a statistical model which estimates the linear relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). (Wikipedia, 2024) In the data analysis, a model that uses unknown distribution data to linearly predict and establish it is called a linear model. The form of the linear regression is $Y = X\beta + \varepsilon$. The plot of this model is a line.

Restricted cubic spline (RCS) regression

Restricted cubic spline (RCS) regression is a type of nonlinear regression. RCS divides the data into multi-segment polynomials and fits cubic regression on them. RCS requires the data set to be continuous and second order differentiable at each segmentation point, the smoothness of the curve needs to be ensured, and the data at both ends of the data range must be linear. RCS can ensure smoothness so that it can avoid overfitting while capturing nonlinear relationships by imposing constraints on the spline function. The best number of RCS regression nodes is 3-5 and this report selects 3 nodes for fitting regression.

Covariance

Covariance is a measure of the joint variability of two random variables. (Rice J., 2007) The formula of the covariance is

$$cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

$$= E[XY] - 2E[Y]E[X] + E[X]E[Y]$$

$$= E[XY] - E[X]E[Y]$$

Therefore, when two variables influence each other and the changing trend is consistent, the covariance is positive, otherwise, it is negative.

Generalized additive model

A generalized additive model is a generalized linear model in which the transformed response variable depends linearly on unknown smooth functions of some predictor variables, and interest focuses on inference about these smooth functions. GAMs were originally developed by Trevor Hastie and Robert Tibshirani to blend the properties of generalized linear models with additive models. (Hastie, T. J.& Tibshirani, R. J., 1990)

The model is:

$$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + + f_m(x_m)$$

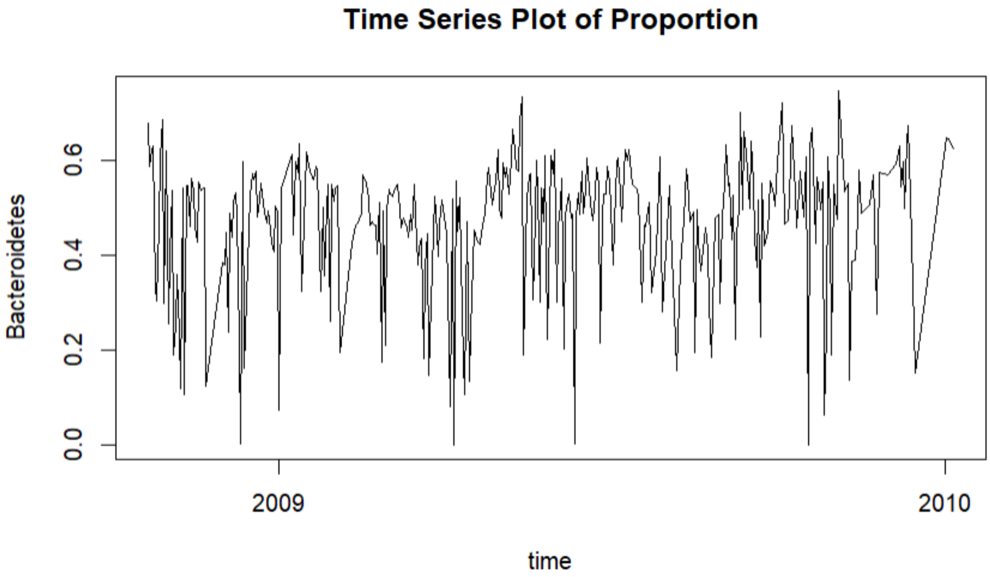
Analysis

Missing Data Type

According to the time gap plot, we can find that most of the missing dates are random. Moreover, the length of the missing dates is also random. Except from days 422 to 437, the gaps are all within 15 days. Therefore, we can think of the missing data type of this dataset to be missing completely at random.

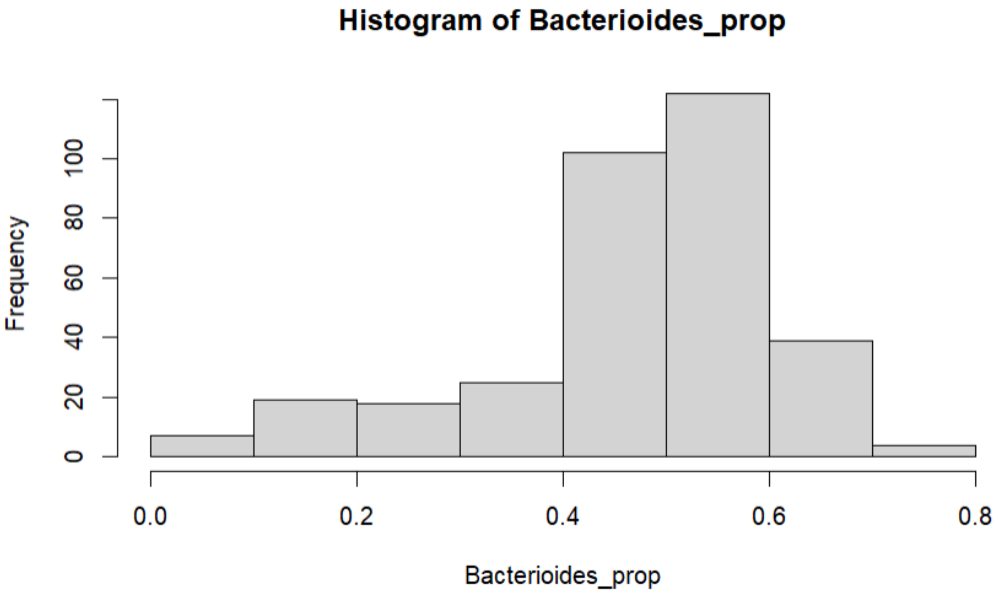
Time Series Plot

To account for sequencing depth, we use the proportion formula $p(x) = \frac{x}{sum(x)}$ of Bacteroides, and plot its time series:



According to this figure, we can find that the Bacteroides data fluctuates greatly, and the proportion is mostly above 0.2. Because the fluctuations are irregular, the time series data does not have obvious periodic signals. Since there is no obvious increasing or decreasing trend, there is not a long-term trend.

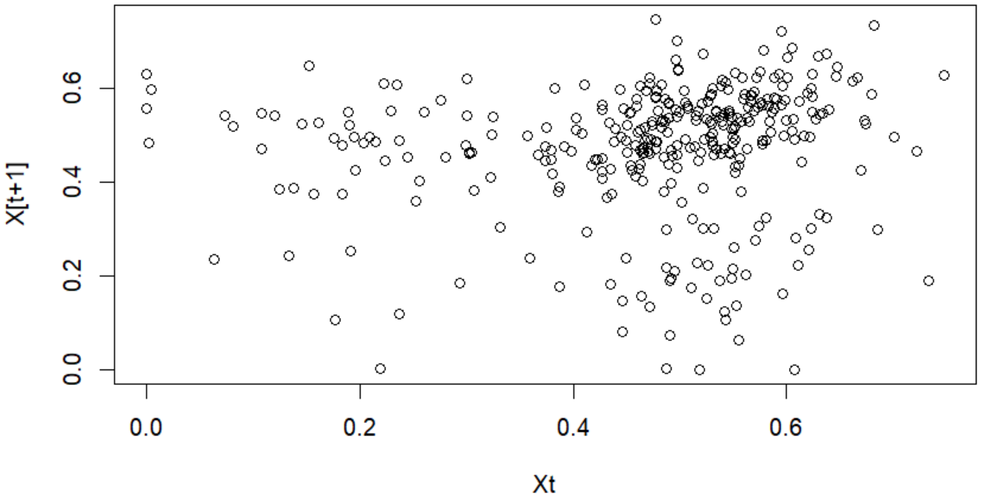
Then make the histograms:



According to the histogram, we can find the probability of this Bacteroides genome in the daily overall data is between 0.4-0.6, and the probability of 0.5-0.6 has the most number. So, the median probability is 0.55. The histogram shows that the probability density is right-skewed. The data does not follow a normal distribution, so the data may have outliers or special trends.

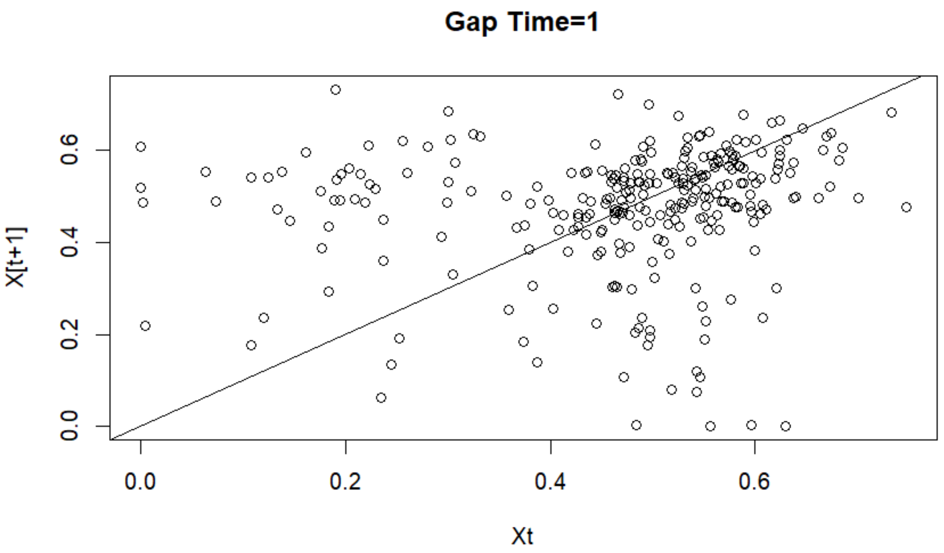
Based on the preliminary calculation and analysis of the data, nonparametric estimation of microbial dynamics is researched on the probabilistic data.

1. First, one day of the probability data is used as the independent variable, and the next day is used as the response variable to analyze and calculate the covariance. Then plot it:



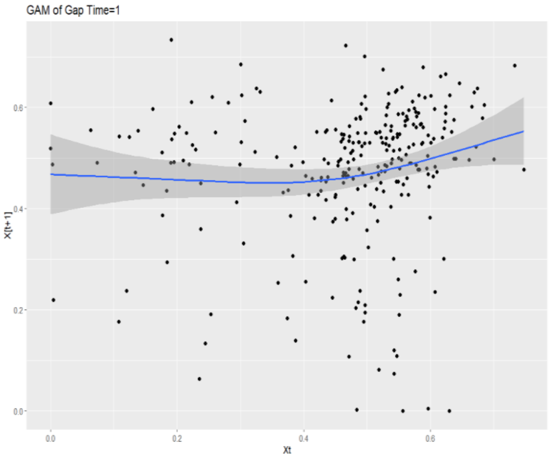
According to the figure, most of the data has a range between 0.4 and 0.6, and the figure shows points in the two areas of the x-axis from 0.4 to 0.6, y-axis from 0 to 0.2 and x-axis from 0 to 0.2, y-axis from 0.4 to 0.6 have the similar distribution, so the data are probably linearly symmetric. The covariance is 0.1019937, so there is some continuity between consecutive points.

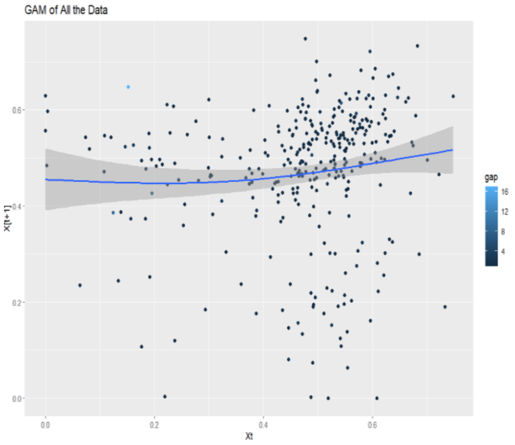
2. Due to there are missing data, in order to reduce the impact of missing data on the analysis, we filtered out consecutive data, which is data with gap time=1. Then repeated the above process, and added y=x lines:



According to this figure, on this dataset, the next day also shows a linear relationship with the previous day, and its covariance is 0.1054199. That means having the same trend, the next day has a positive correlation with the previous day.

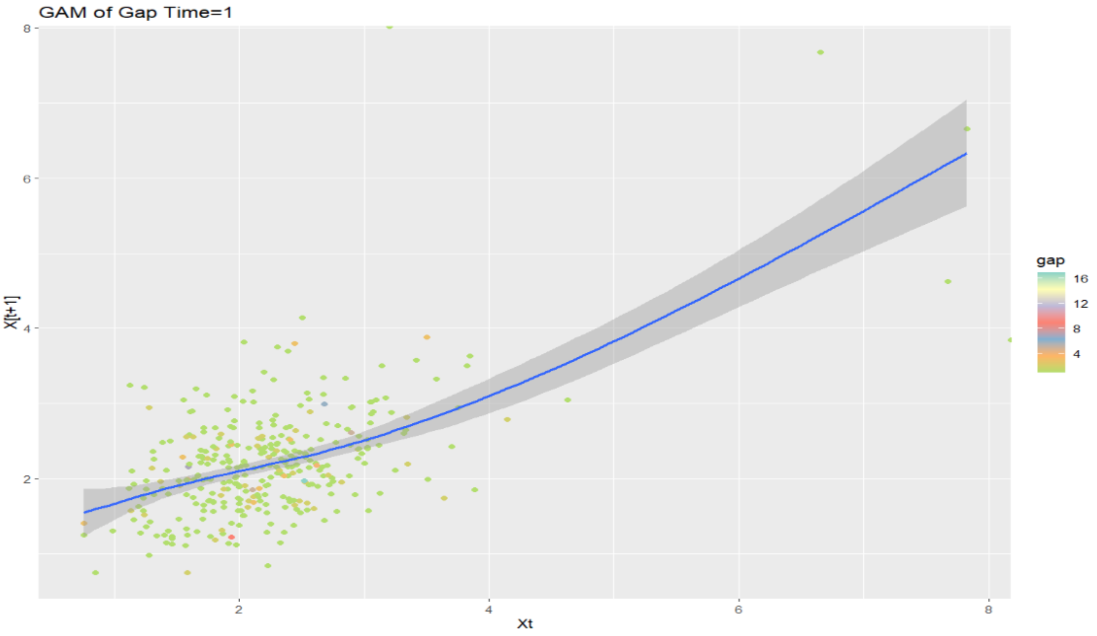
Fitting aa GAM model to the Bacteroides genus proportion data and for all pairs of consecutive samples, and another for days with gap time=1. The results are:



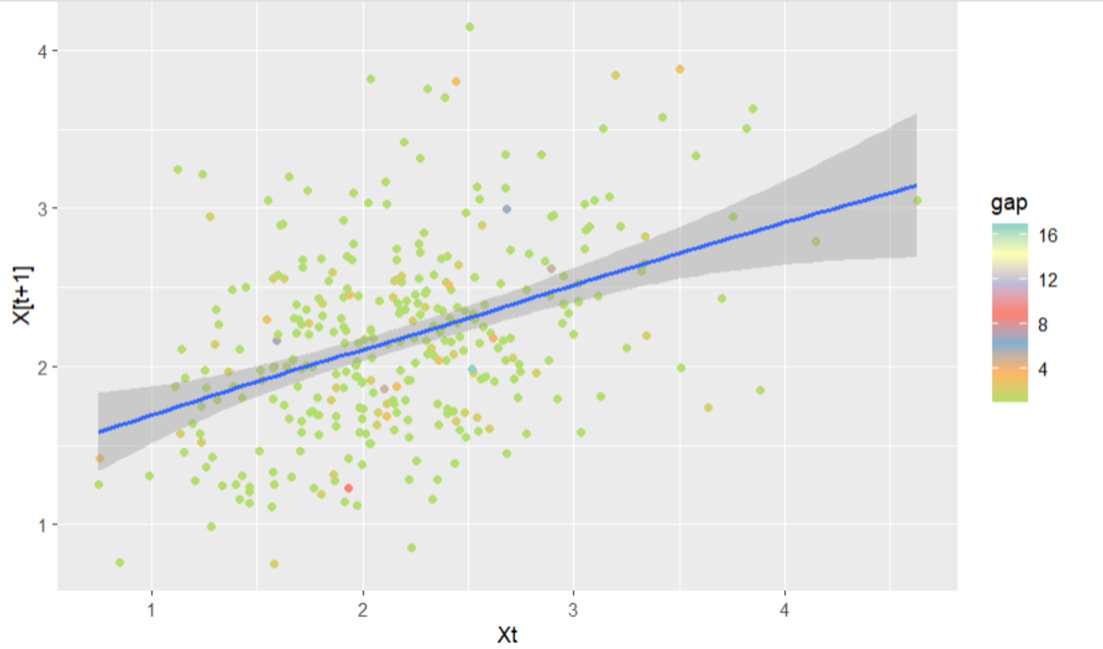


Comparing the two plots, we can find the dataset which has all the data has a more linear correlation between the data on the second day and the first day than the data with gap time=1.

Based on the analysis of the probabilistic data, we found that the data showed a positive linear correlation in the temporal dynamic change, which means the second day's data is linearly related with the first day's data. A better way to correct for sequencing depth is to take the logarithm of the ratio of the abundances of two genera. This measure is not affected by sequencing depth or by the abundance of other microbes, and is therefore a more reliable measurement. We therefore study the temporal dynamics of the log ratio between the two most abundant genera-Bacteroides and Faecalibacterium. We fit the same GAM model to this log ratios as we fitted for the proportion of Bacteroides. Then plot the GAM model:

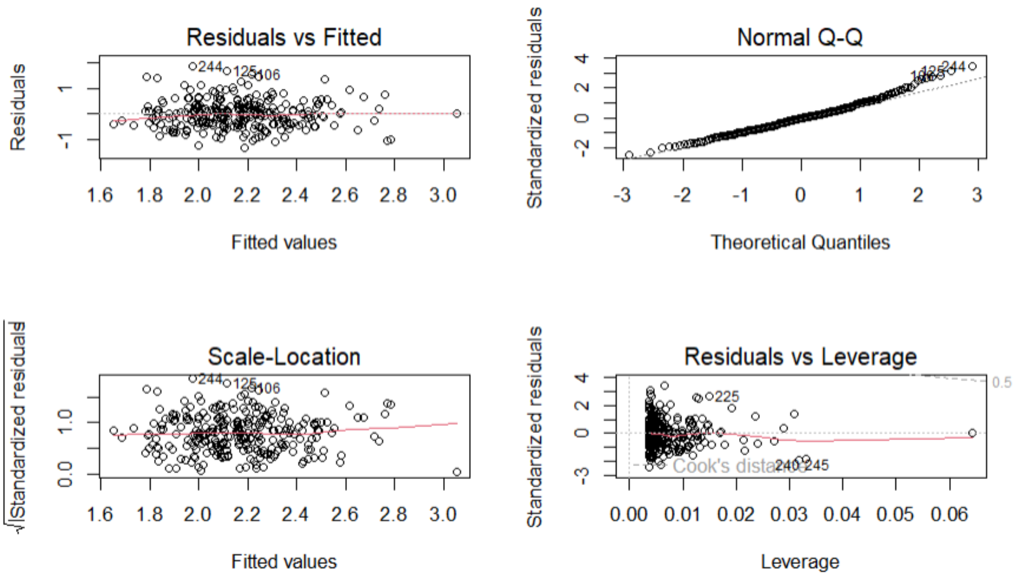


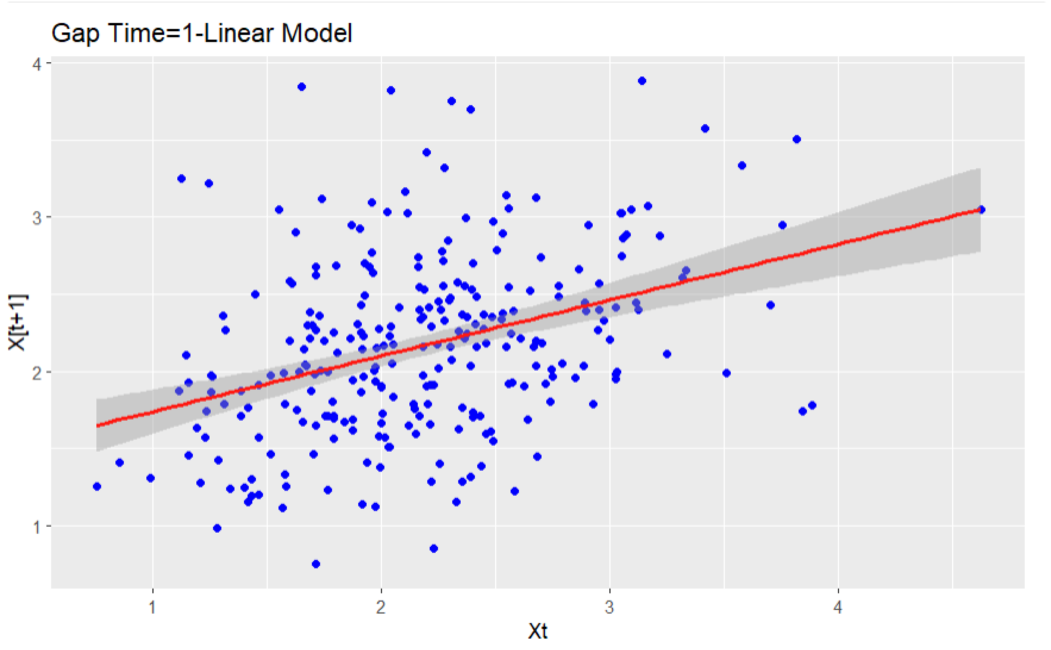
According to this figure, we can find the Bacteroides has some outliers which affect the model fitting, so we will remove the outliers and obtain the following figure:



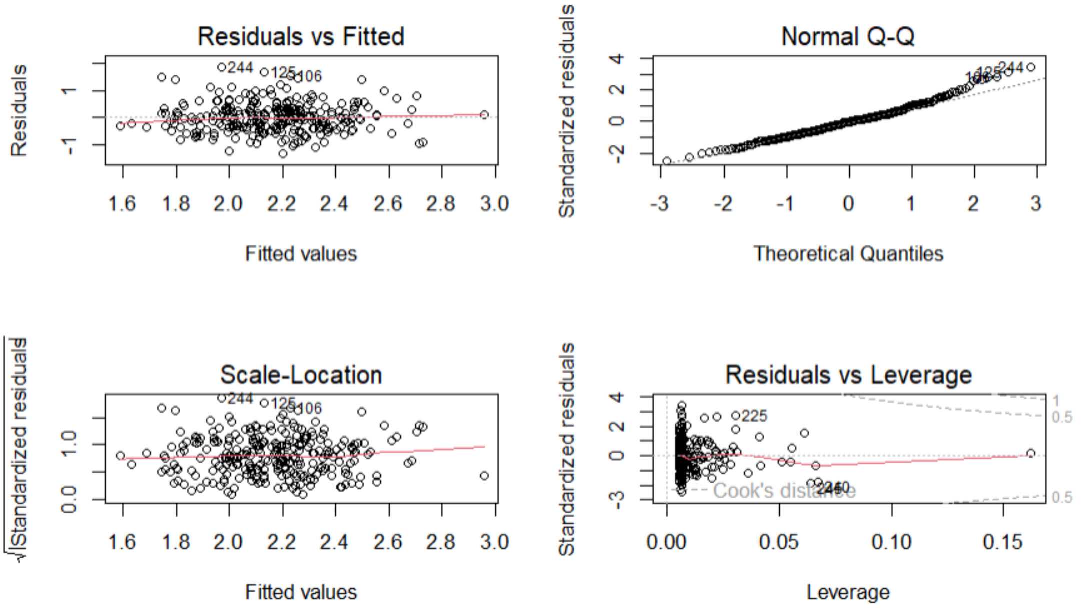
According to the figure with outliers removed, the log ratio shows a linear correlation. To eliminate the impact of missing data on nonparametric estimation, we separately extracted the data with gap time=1. Because the type of the missing data is completely missing at random, to ensure enough amount of data, we chose the data with gap time=4 as another dataset. Then use the two datasets to both fit a linear model and a nonlinear model. The nonlinear model uses the RCS model, and there are the results:

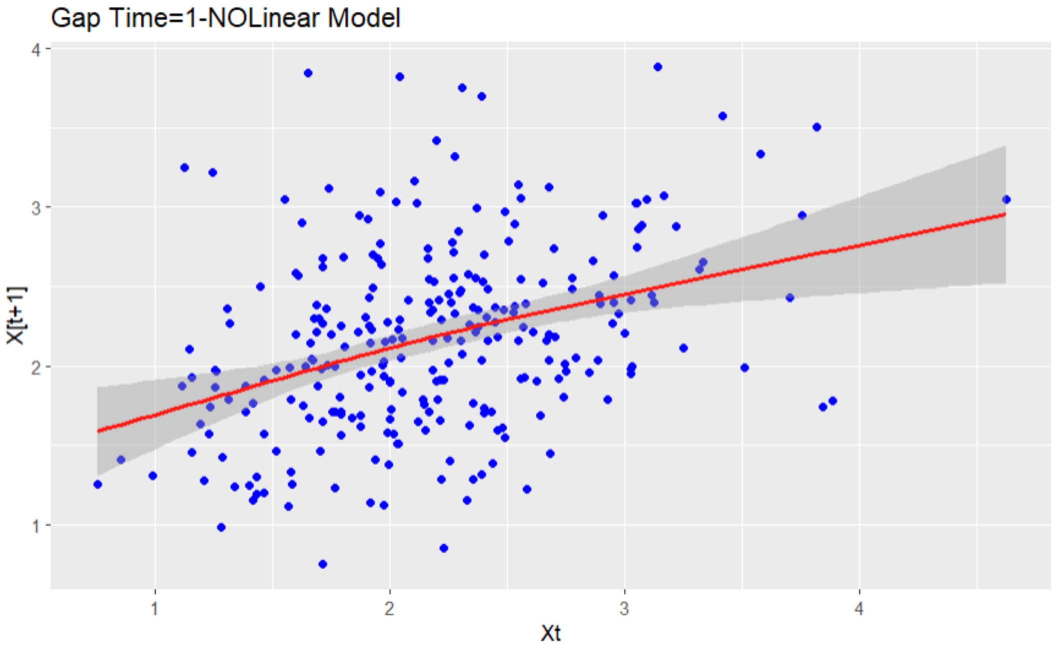
1. Linear model of gap time=1:



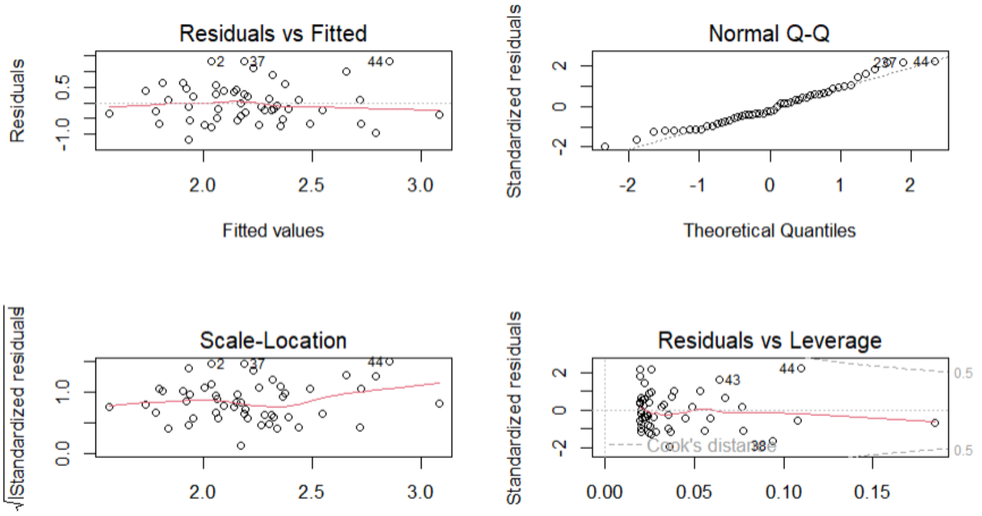


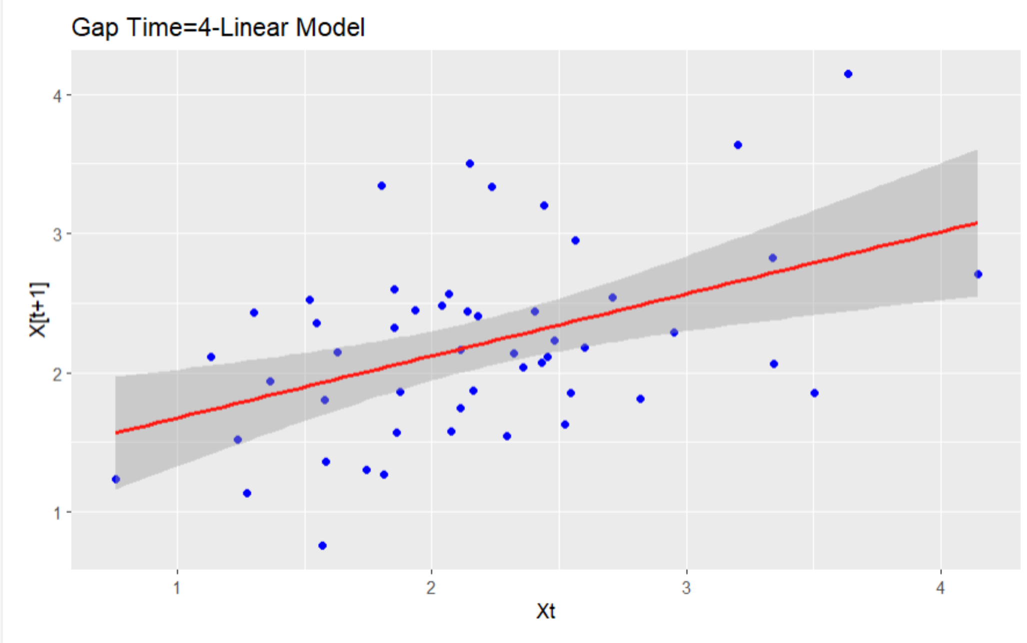
2. Nonlinear model of gap time=1:



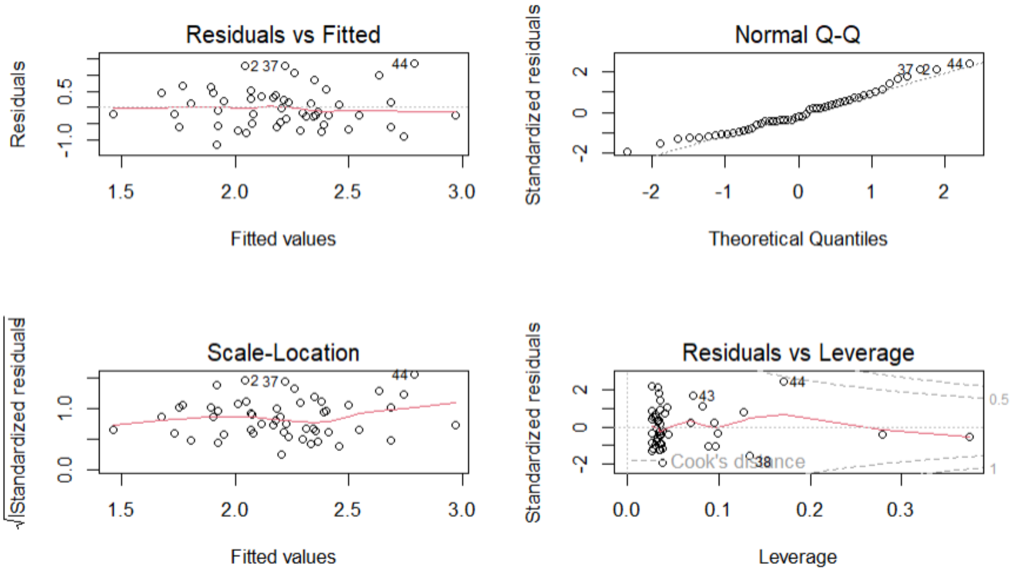


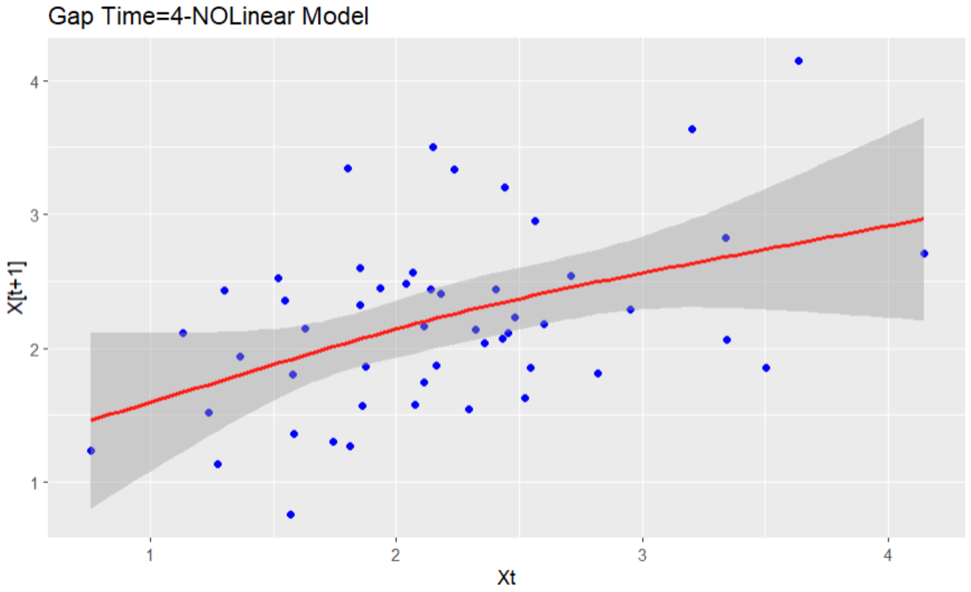
3. Linear model of gap time=4:





4. Nonlinear model of gap time=4:

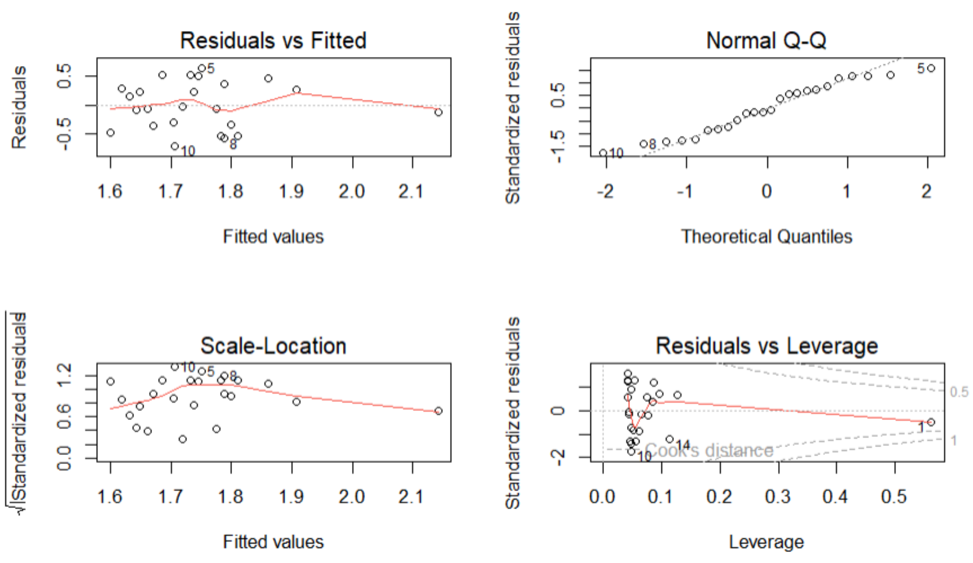


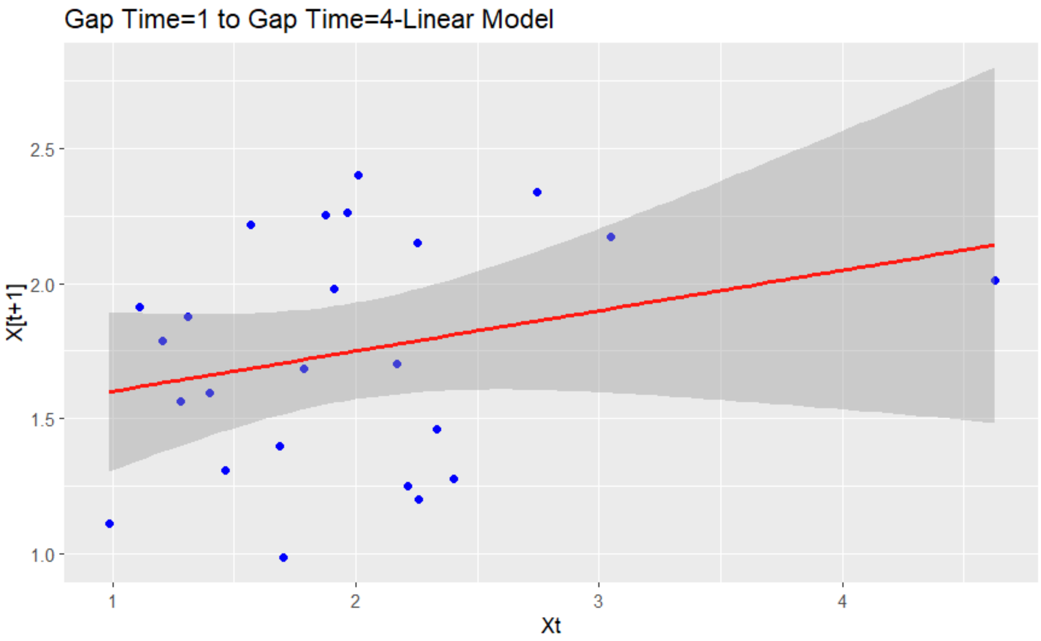


The training mean-square error (MSE) of the four model is:

Model	MSE
Linear model of gap time=1	0.5463
Nonlinear model of gap time=1	0.547
Linear model of gap time=4	0.6112
Nonlinear model of gap time=4	0.6164

According to the above four models, we can find the data of Bacteroides is linear both in the gap time=1 and gap time=4. This is consistent, as a linear relation with gap time=1 should give rise to a linear relation for gap=4, but with a different slope. However, the slopes fitted by the two models are not consistent, suggesting that the truth may be more complicated than a simple AR (1) process. To make sure this is not because our assumption that the data are missing completely at random is incorrect, we restrict to the first 24 time points, where an observation was present every day, for the gap time=4 data, and refit the model. Then plot the result:





The slope of the regression line on these points is much smaller, and is consistent with the slope fitted for time gap=1. Thus, these results are consistent with an AR (1) process where the data are missing not at random. It is also consistent with a non-stationary process where the auto-correlation is higher in the later period, or with process with more complicated dynamics.

Conclusion

Nonparametric estimation is a common method to analyze the unknown data distributions. The application of nonparametric estimation to the temporal dynamic data of the microbiome can help research into the human microbiome. From the dataset of this report, we can conclude from the time series plot that the data has no obvious periodicity and long-term trend. The dataset has a lot of missing data which do not have the relationship with others. It seems plausible that the missing data are missing completely at random. From the histogram, the median and mode proportion of this

Bacteroides microbiome in the overall data is between 0.5-0.6, and the probability density is concentrated at 0.4-0.6. The histogram of this Bacteroides microbiome is right-skewed, so it has some outliers or special trends. This Bacteroides microbiome shows a positive linear correlation trend in the time gap =1 and time gap =4 datasets eliminating the impact of missing data. Looking at all data points, the slopes for the gap=1 and gap=4 regression are not consistent with an AR (1) model. However, when we look at only the first 24 time points, the slopes are consistent. Thus, it could be that the data are not missing at random, or that the auto-correlation varies over time. The relations between consecutive days and between samples 4 days apart are both linear, with no evidence of nonlinear relations.

Reference

1. Luke K Ursell, Jessica L Metcalf, Laura Wegener Parfrey, Rob Knight, Defining the human microbiome, Nutrition Reviews, Volume 70, Issue suppl_1, 1 August 2012, Pages S38–S44, https://doi.org/10.1111/j.1753-4887.2012.00493.x

2. Berg G, Rybakova D, Fischer D, Cernava T, Vergès MC, Charles T, Chen X, Cocolin L, Eversole K, Corral GH, Kazou M, Kinkel L, Lange L, Lima N, Loy A, Macklin JA, Maguin E, Mauchline T, McClure R, Mitter B, Ryan M, Sarand I, Smidt H, Schelkle B, Roume H, Kiran GS, Selvin J, Souza RSC, van Overbeek L, Singh BK, Wagner M, Walsh A, Sessitsch A, Schloter M. Microbiome definition re-visited: old concepts and new challenges. Microbiome. 2020 Jun 30;8(1):103. doi: 10.1186/s40168-020-00875-0. Erratum in: Microbiome. 2020 Aug 20;8(1):119. PMID: 32605663; PMCID: PMC7329523.

3. Wikipedia contributors. (2024, March 10). Time series. In Wikipedia, The Free Encyclopedia. Retrieved 17:19, March 19, 2024, from https://en.wikipedia.org/w/index.php?title=Time_series&oldid=1213036071

4. Caporaso, J.G., Lauber, C.L., Costello, E.K. et al. Moving pictures of the human microbiome. Genome Biol 12, R50 (2011). https://doi.org/10.1186/gb-2011-12-5-r50

5. Wikipedia contributors. (2024, February 13). Linear regression. In Wikipedia, The Free Encyclopedia. Retrieved 12:27, March 20, 2024, from https://en.wikipedia.org/w/index.php?title=Linear_regression&oldid=1206942648

6. Rice, John (2007). Mathematical Statistics and Data Analysis. Brooks/Cole Cengage Learning. p. 138. ISBN 9780534399429.

7. Wikipedia contributors. (2023, December 20). Generalized additive model. In Wikipedia, The Free Encyclopedia. Retrieved 18:41, March 20, 2024, from https://en.wikipedia.org/w/index.php?title=Generalized_additive_model&oldid=1190862328

8. Hastie, T. J.; Tibshirani, R. J. (1990). Generalized Additive Models. Chapman & Hall/CRC. ISBN 978-0-412-34390-2.