A Statistical Approach for the Background CO2 Concentration Measurement at the Korea GAW Center

2014. 10. 21.

Yung-Seop Lee

Department of Statistics
Dongguk University, Seoul, Korea

Contents

I. Estimating the Background CO2 Concentration

1. Data Preprocessing
2. Data Analysis
3. Conclusion

➢ Raw Data (Level 0) : hourly measurement before preprocessing

OBS	day	month	year	time	avg	std	num
1	1	1	2000	0	419.09	4.40	108
2	1	1	2000	1	415.72	0.86	90
3	1	1	2000	2	408.92	2.38	114
4	1	1	2000	3	405.36	0.75	89
5	1	1	2000	4	404.19	1.18	115
6	1	1	2000	5	402.19	1.23	89
7	1	1	2000	6	395.58	4.01	114
8	1	1	2000	7	393.42	1.29	90

Provided by Korea GAW Center (Anmyeondo)

- 01/01/2000~ 03/31/2014 Hourly data
- 112,351 CO2 concentration measurements
- **Variables**: Day, Month, Year, Time(Hour), Average CO2 concentration (Avg), Standard Deviation (std), Number of measurement(num)

➢ Raw Data (Level 0) → Preprocessed Hourly data (Level 1)

Data Preprocessing Steps (Cho *et al.,* 2007):

- Step 1: Discard averaged hourly raw values when the number of measurements to be averaged are less than 60 which is the half of the measurement counts during the given hourly period.

- Step 2: Exclude averaged hourly raw values when the standard deviation of the measurement during the given hourly period is above 1.8ppm.

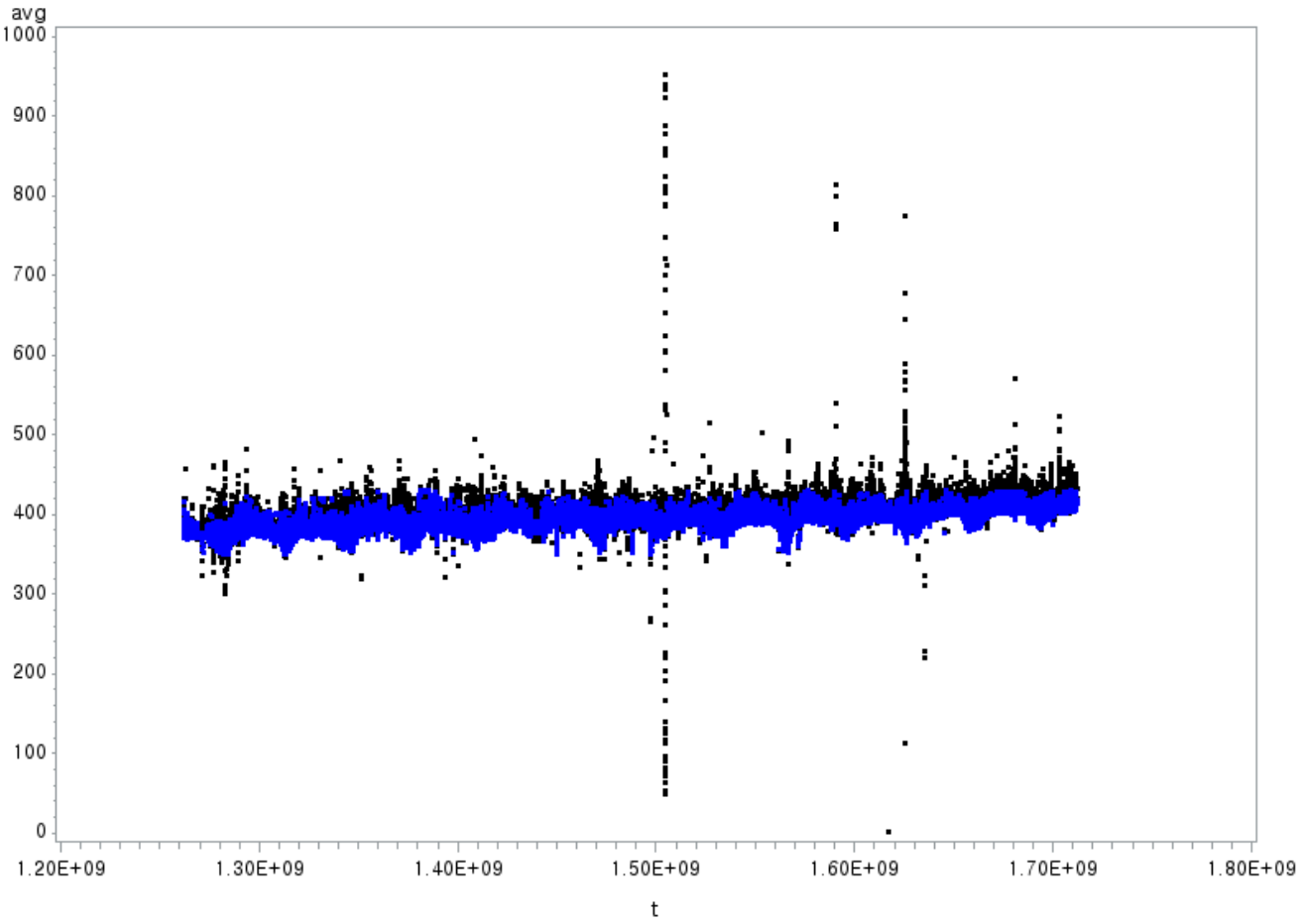
=> After Step 1 and 2, the averaged measurement of specific hour is going to be the preprocessed Hourly data (level 1).

➢ Preprocessed Hourly Data (Level 1) → Daily Data (Level 2)

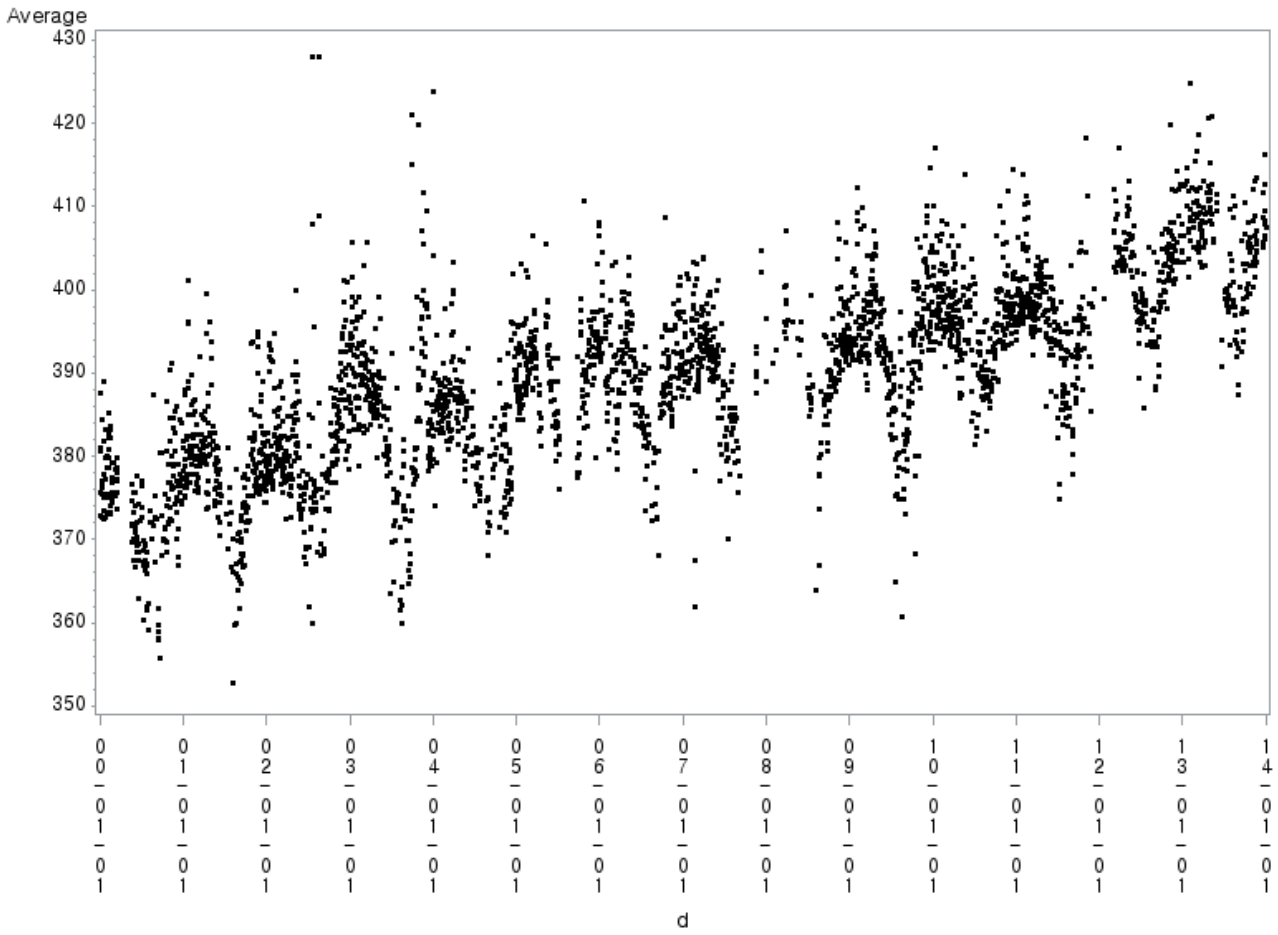
Data Preprocessing Steps (Cho *et al.,* 2007; JMA, 2007):

- Step 1: Exclude the averaged hourly concentration data which are below 350ppm or above 430ppm.
- Step 2: Exclude the averaged hourly data if the difference between consecutive hourly average concentration is larger than 1.8ppm.
- Step 3: Exclude the averaged hourly data when the number of hourly measurement are less than 12 which is the half of the daily hours(24).

=> the averaged measurement of hourly values of specific day is going to be Daily Data(Level 2).



Raw Data (Level 0) (Black Dot) / Hourly Data (Level 1) (Blue Dot)



Preprocessed Daily Data(Level 2) (Missing values may exist)

I. Estimating the Background CO2 Concentration

1. Data Preprocessing
2. Data Analysis
3. Conclusion

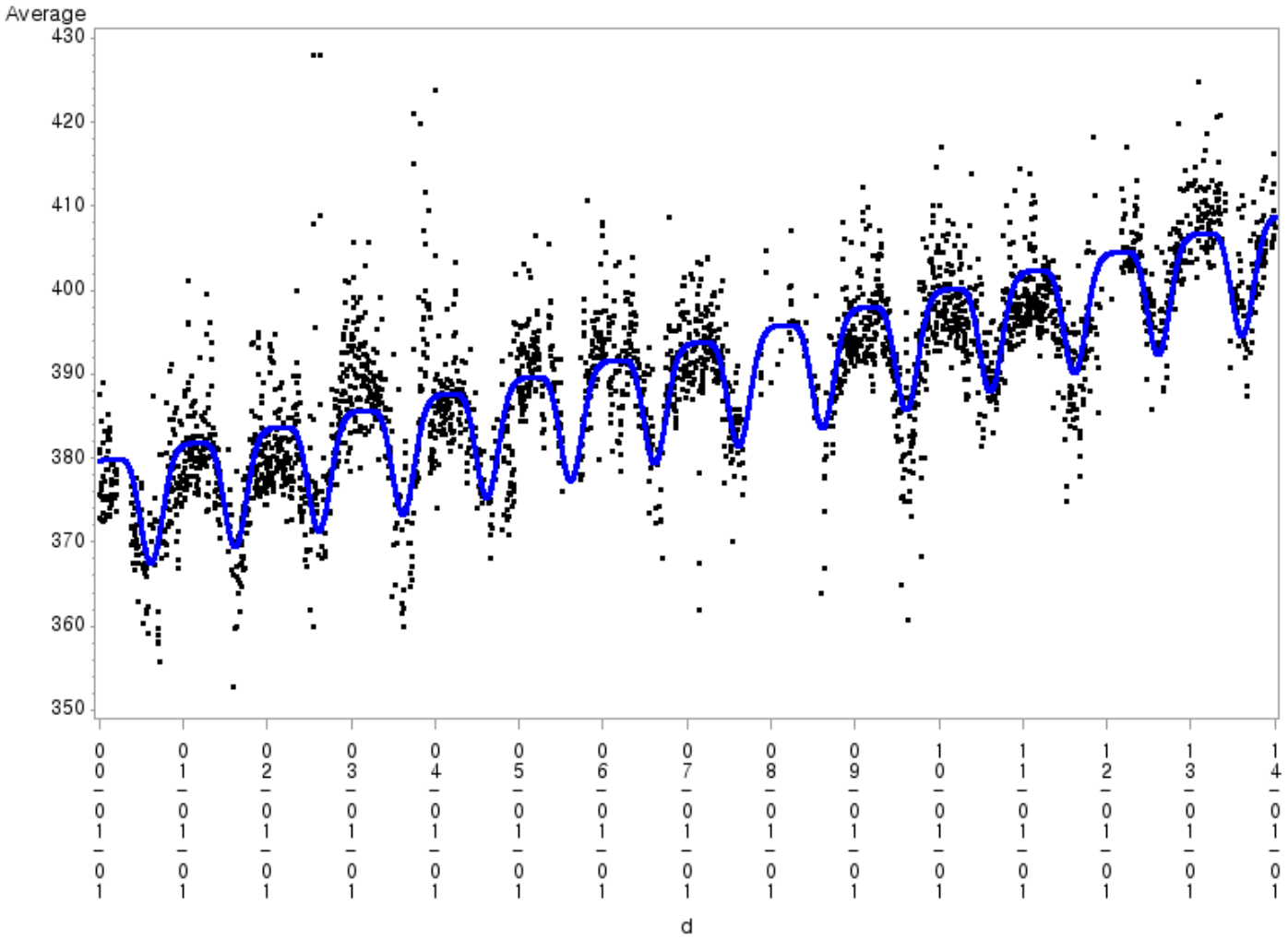
- Estimate the time-series model include inter-annual trend and annual cycle: Curve fitting method (Masarie and Tans, 1995).

$$g(t) = a_0 + a_1 t + a_2 t^2 + \sum_{k=1}^{3} [b_{2k-1} \sin(w_k t) + b_{2k} \cos(w_k t)] \qquad (1)$$

$a_0 + a_1 t + a_2 t^2$: second-order polynomial representing one of the long-term trend
$\sum_{k=1}^{3} [b_{2k-1} \sin(w_k t) + b_{2k} \cos(w_k t)]$: A series of two harmonics representing the average seasonal cycles

a_i: parameters of the inter-annual trend to be determined
b_i: parameters of the annual cycle
t: time in days
w_k: period of kth trigonometry term ($w_k = \frac{2\pi k}{\mathrm{T}}$, $\mathrm{T} = 365.25\ days$)

Estimating a Curve function g(t) of equation(1) using Level 2 data.



Selected Daily (Level 2) Data (Black Dot) / Estimated g(t) for Trend and Cycle (Blue Curve)

- A “residual”, a difference(or deviation) between the level 2 measured data(Black) and the fitted curve (Blue) is generated. It might also be the difference between the daily time series(Black) and the inter-annual trend and annual cycle (Blue) from the figure of the previous slide.

- Apply a residual interpolation method to the missing values for the spectral analysis since it should not have missing value to perform.
  => used spline interpolation

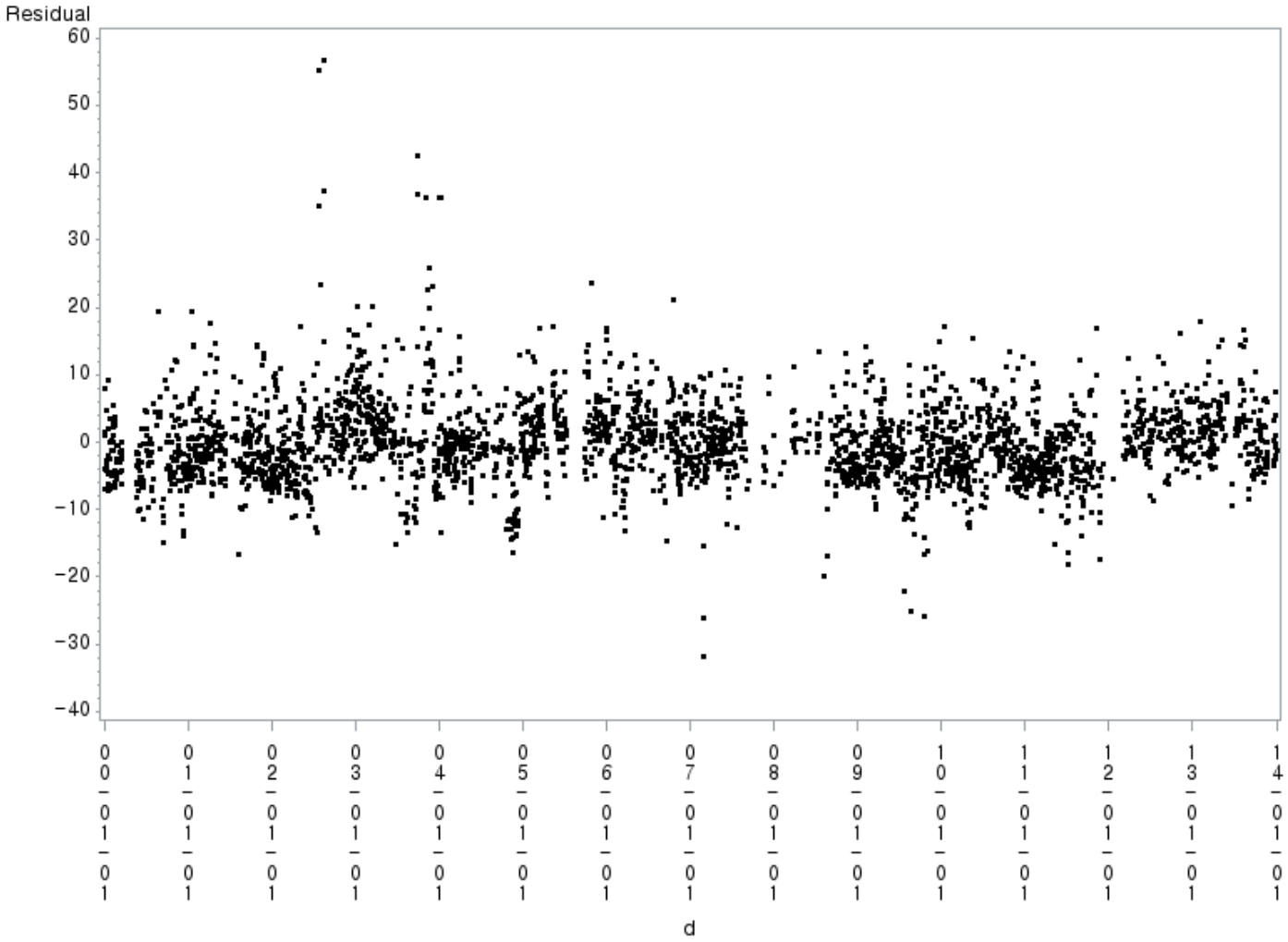
- Why spectral analysis?

⇒ To find the hidden frequencies of the residuals for the background concentration determination not affected the short term period.

⇒ In general, uniform mixing background concentration in the northern hemisphere occurs after 2 or 3 months (60~90 days).

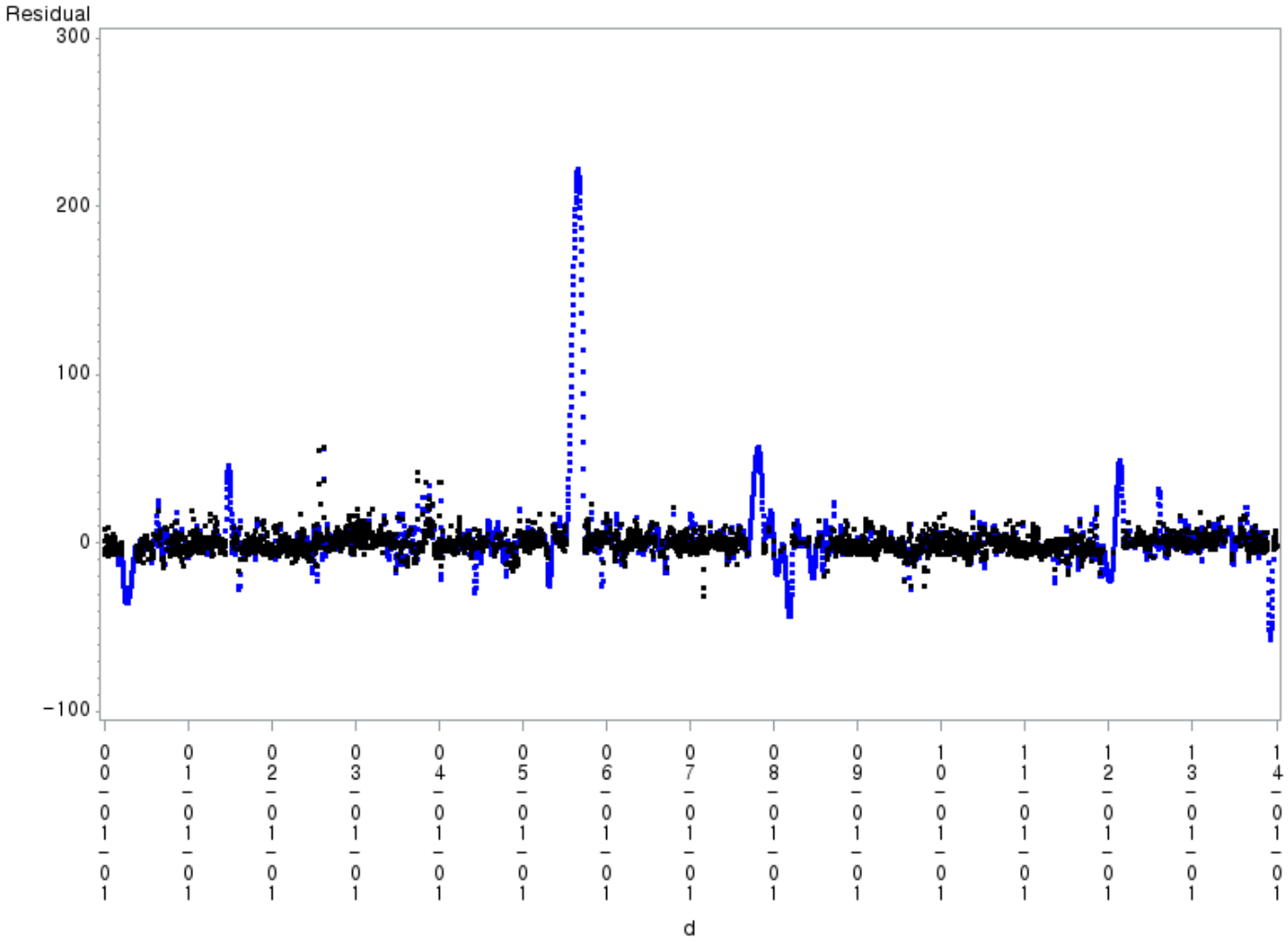
- Spline Interpolation method is the one where the interpolated points are connected by a polynomial curve.

Residual Plot



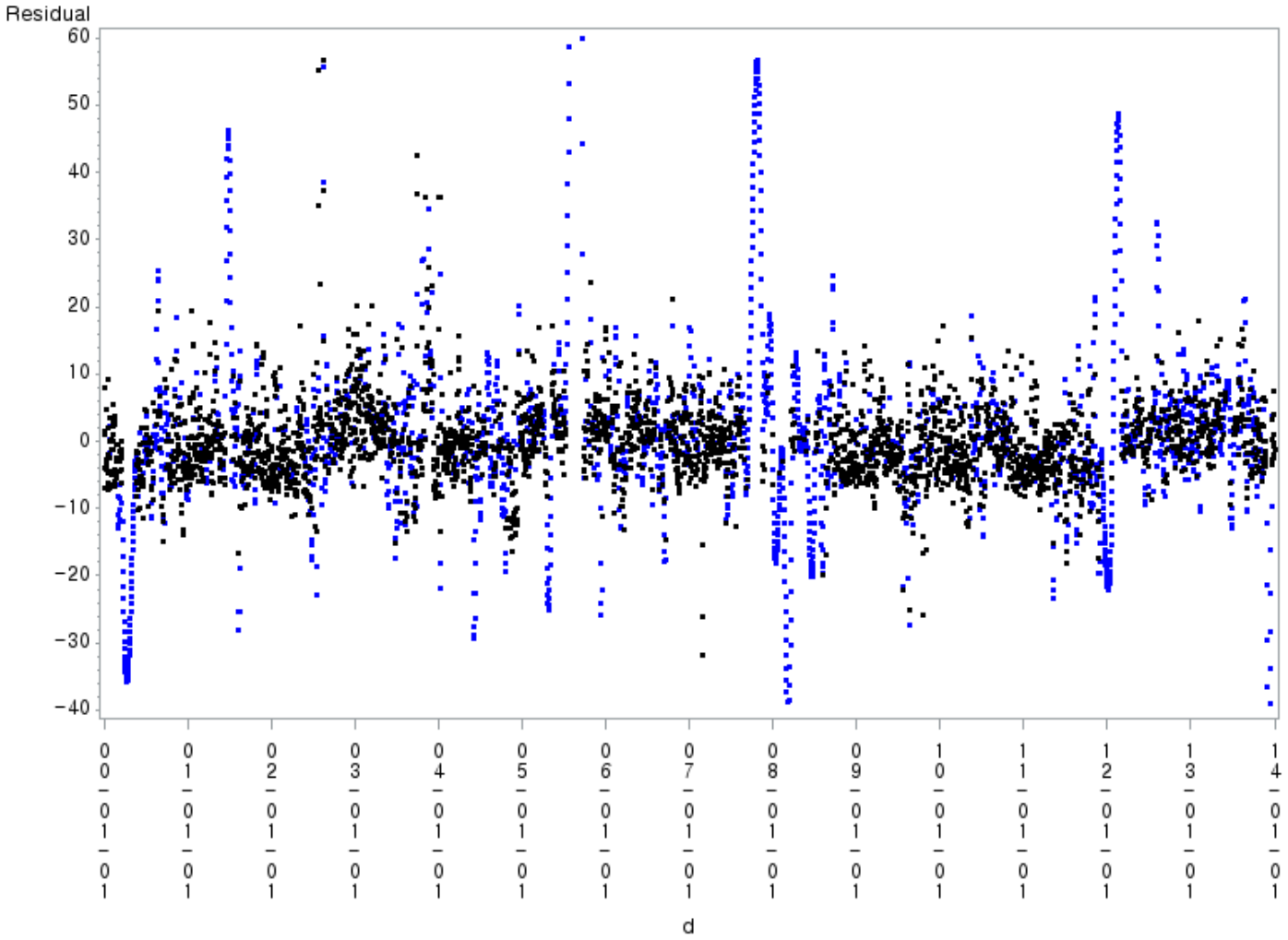
Scatter Plot of Residuals by daily averaged measurements (Level2) for 1/1/2000 ~3/31/2014 (Missing values may exist).

Residual Plot after applying spline interpolation method : Original Scale



Scatter Plot of Residuals with Interpolant (Blue Dot)

Residual Plot after applying spline interpolation method : Zoomed



Scatter Plot of Residuals with Interpolant (Blue Dot)
=> there are several unusual peaks after interpolated

- Calculate periodogram based on the frequency(period) domain using interpolated residuals through FFT(Fast Fourier Transformation) algorithm

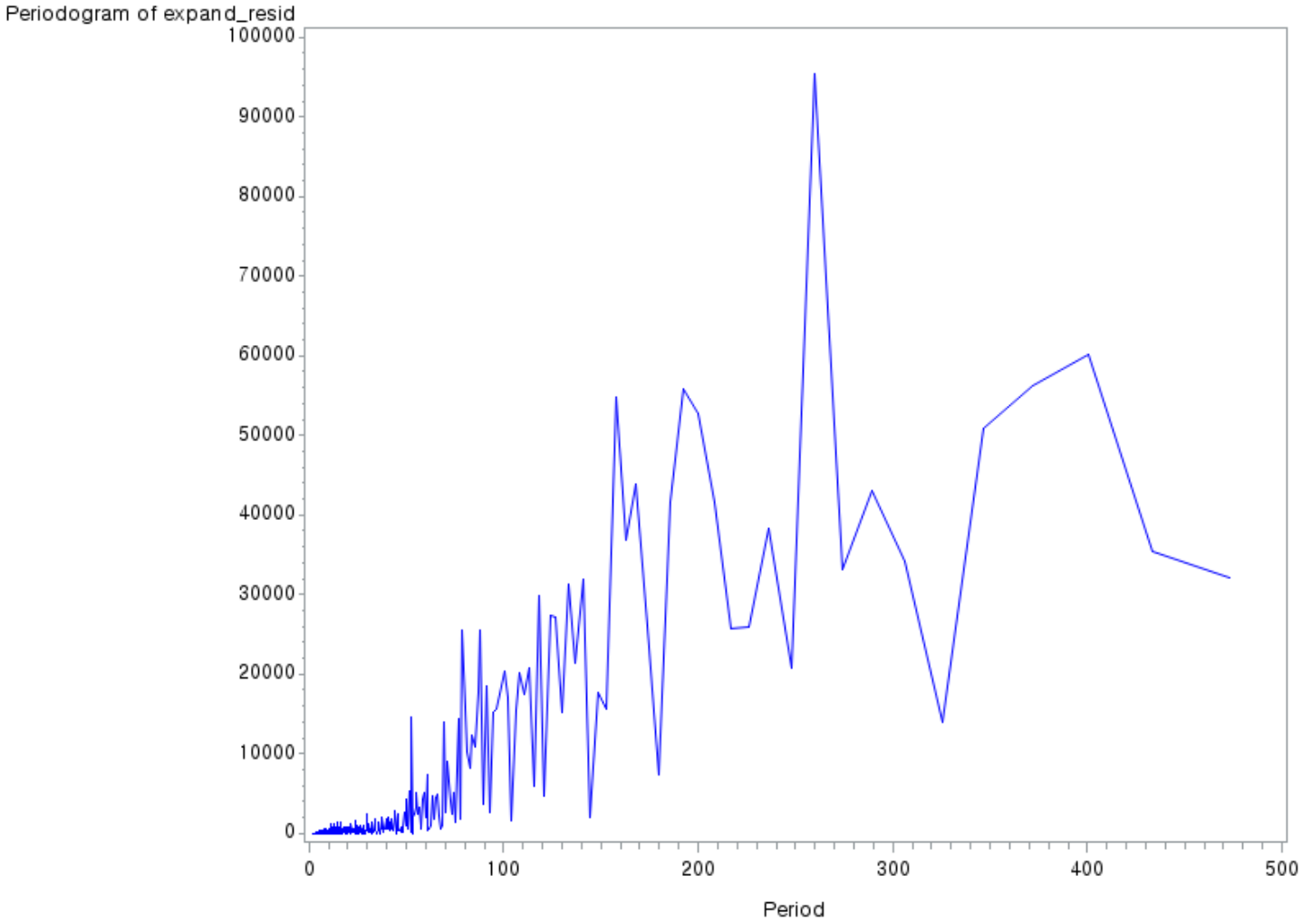
=> Refer to the Figure silde.

- Applying low pass filter to eliminate any frequencies higher than 7.3 cycle/yr (that is, any periods lower than 50 days) (Thoning, 1989)

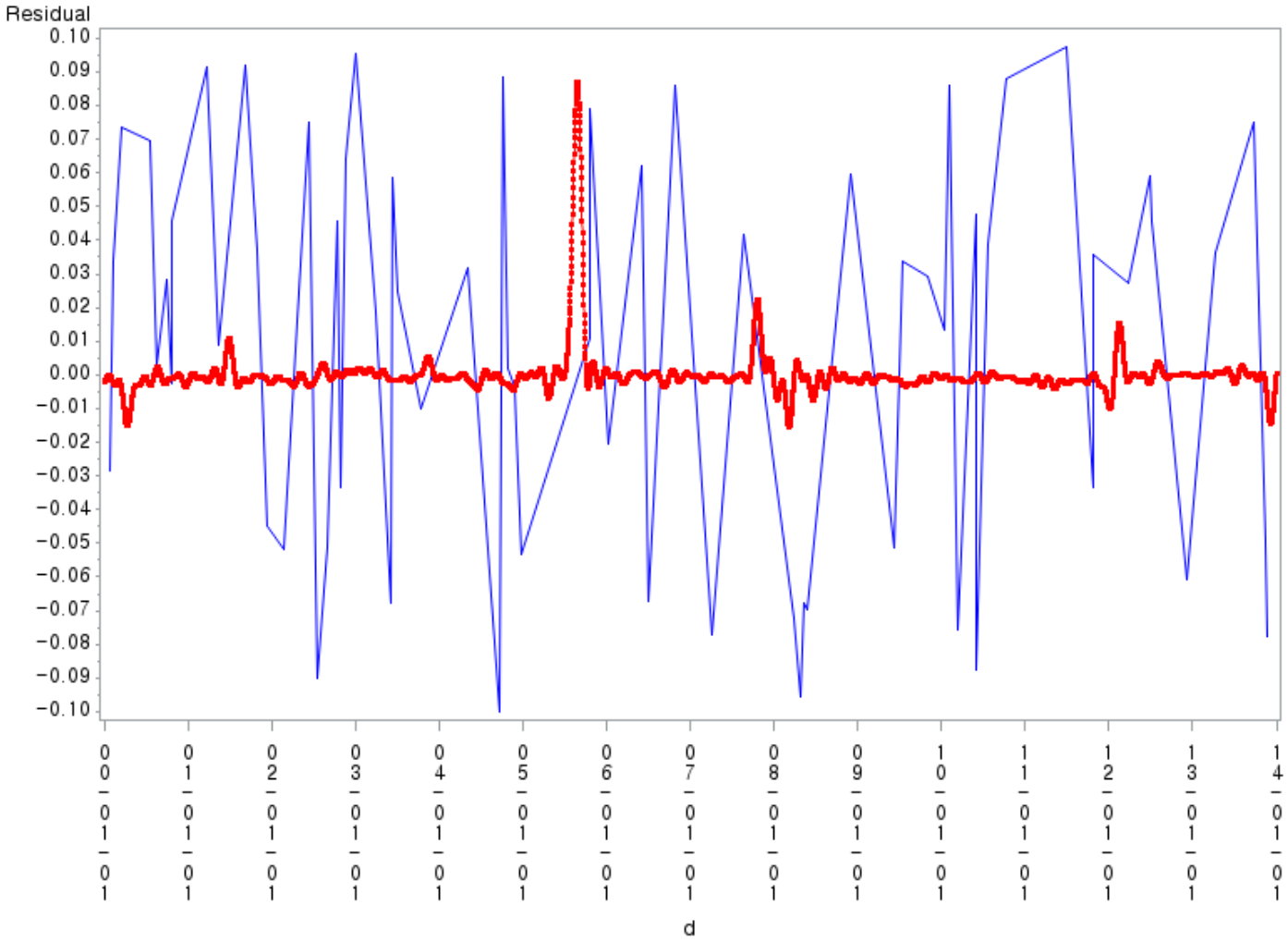
- Convert the remaining lower frequencies after filtered to the time domain model of residual using inverse FFT algorithm.

=> Estimated time domain residual model after converted is called $\{r(t)\}_{50d}$.

=> Refer to the residual and $\{r(t)\}_{50d}$ plots



Plot of Periodogram against Period



Plot Residuals(Blue Line) and $\{r(t)\}_{50d}$ (Red Line) in the time domain : Zoomed

Data quality control using estimated model $S(t)$.

- $\{r(t)\}_{50d}$ after low pass filtering is added to the fitted curve g(t)

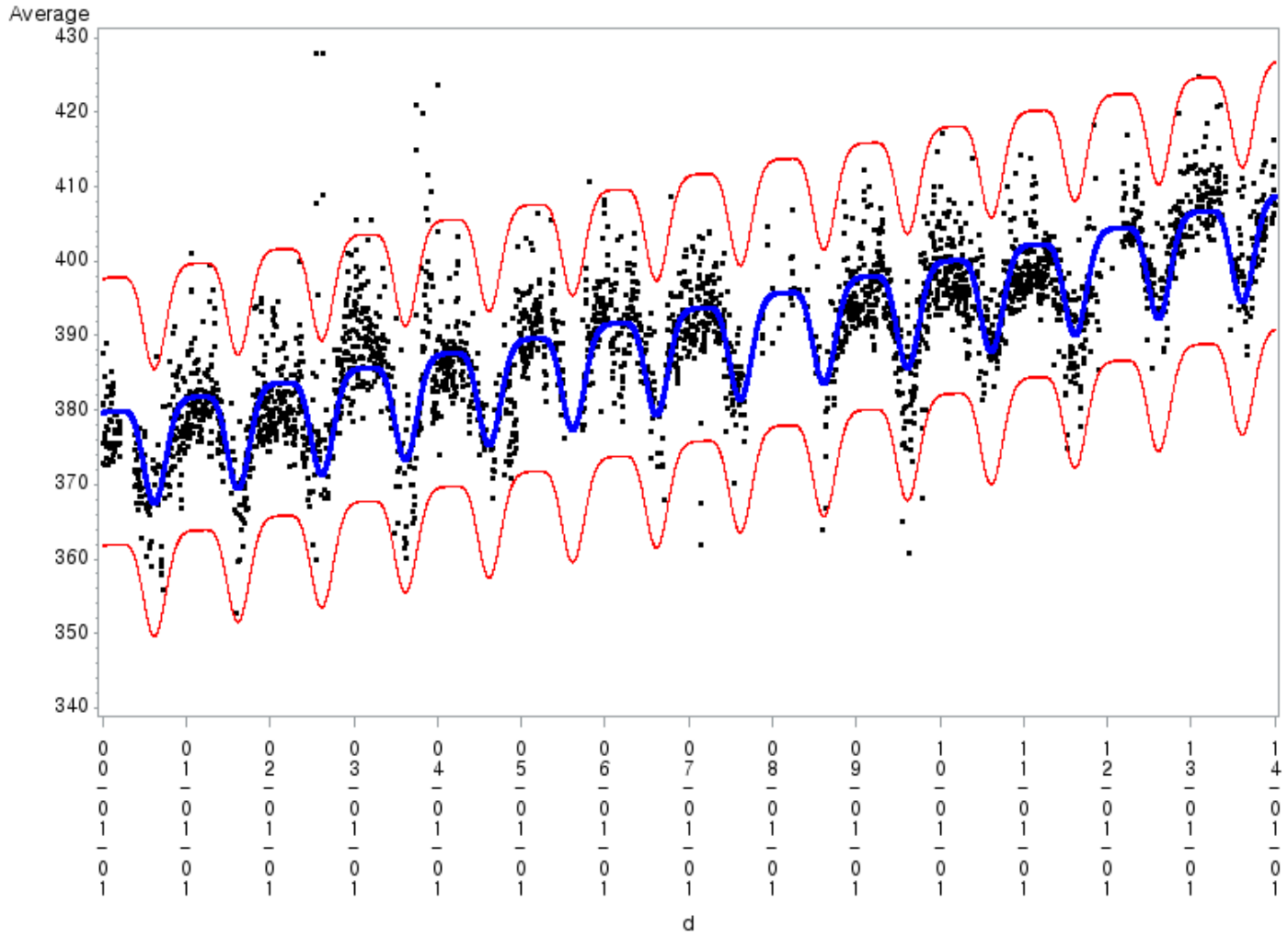
  => $S(t) = g(t) + \{r(t)\}_{50d}$. Click for the fitted $S(t)$.

- Remove observations outside $\pm 3\sigma$ of the fitted model $S(t)$
  => Click for the plot.

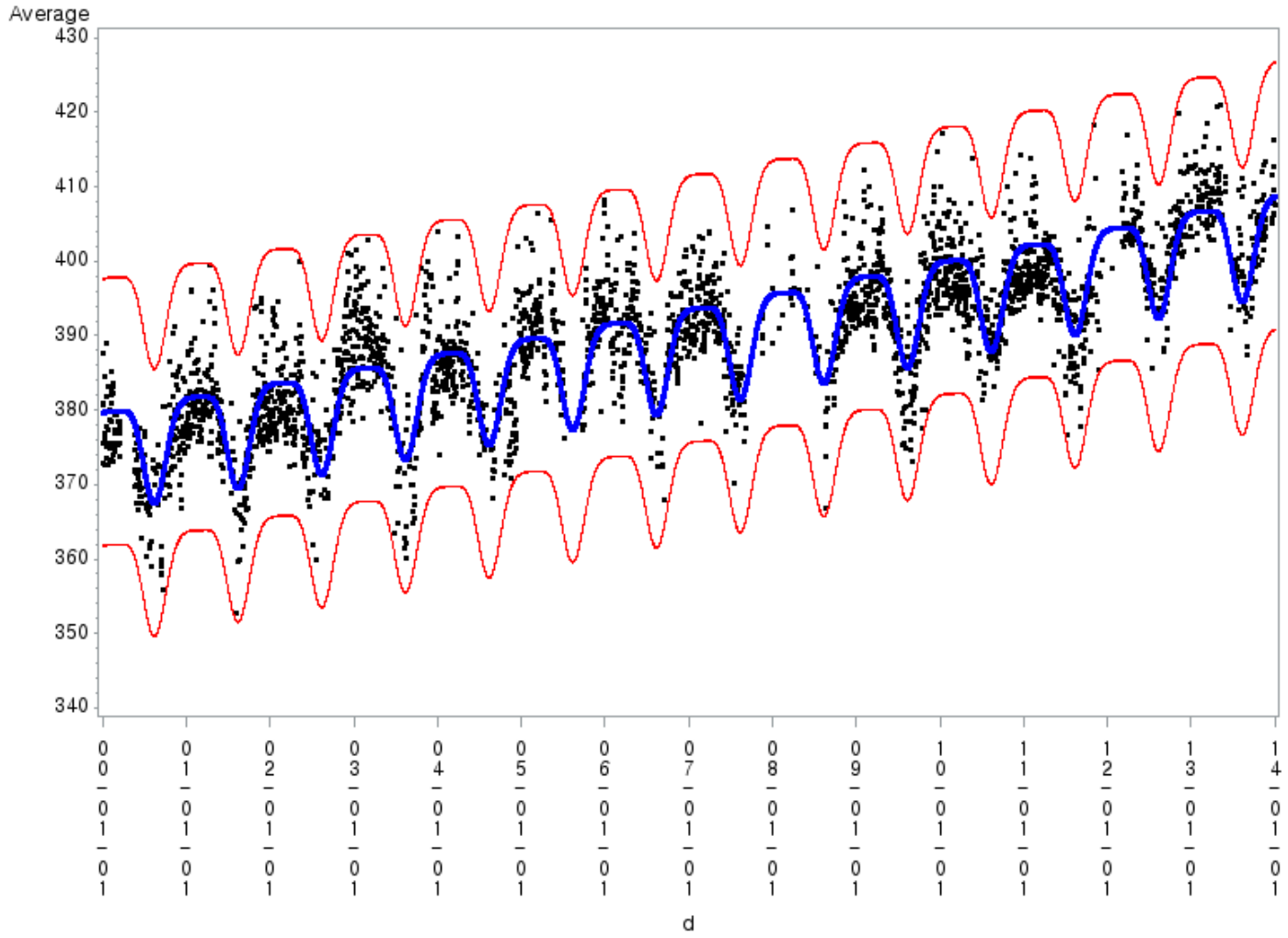
- Iterate these process until all of the remained data are within $\pm$ 3σ of the re-fitted model $S(t)$.

That is, re-estimate $g(t)$ from the remained data -> residuals -> interpolation -> convert residuals to frequency domain -> low pass filtering -> re-convert to time domain -> get $\{r(t)\}_{50d}$ -> re-fit $S(t)$ -> remove obs. outside $\pm 3\sigma$ of $S(t)$.
=> Repeat

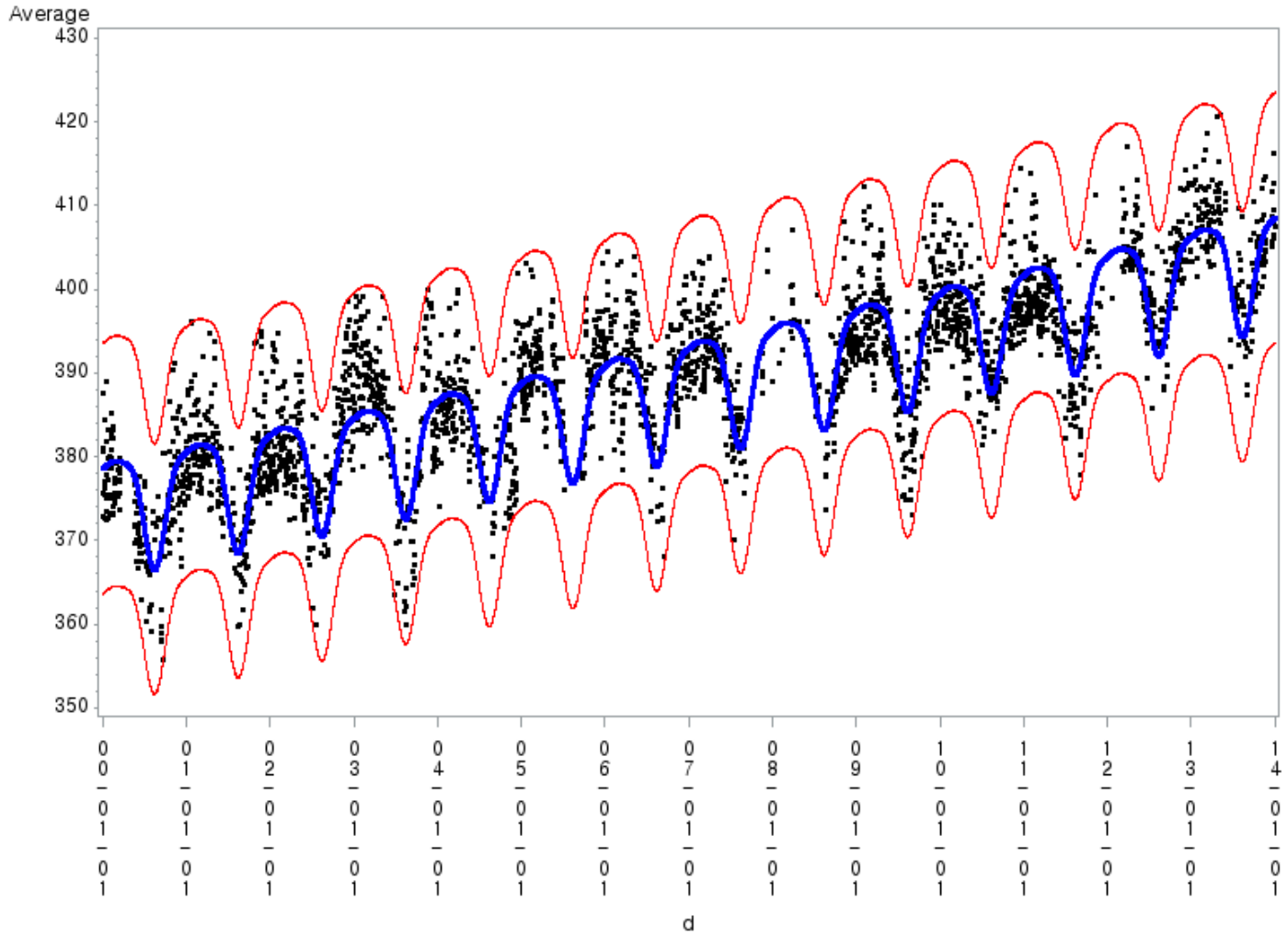
- In this analysis, we repeated 4 times.
  => Click for the plot.



Plot of $S(t)$(Blue Line) and 3σ Range (Red Line) against Date



After removing the data outside of 3σ range



Iterated Output (4 Times)

- To estimate background CO2 concentration, we use only remained data after QC process in the previous slides.
  => Refer to the remained and removed data plots

- To account for inter-annual variability in long-term trends, the residuals can be applied the low pass filter to eliminate frequency higher than 0.55 cycle/yr (period lower than 667 days) (Cho *et al.,* 2007).

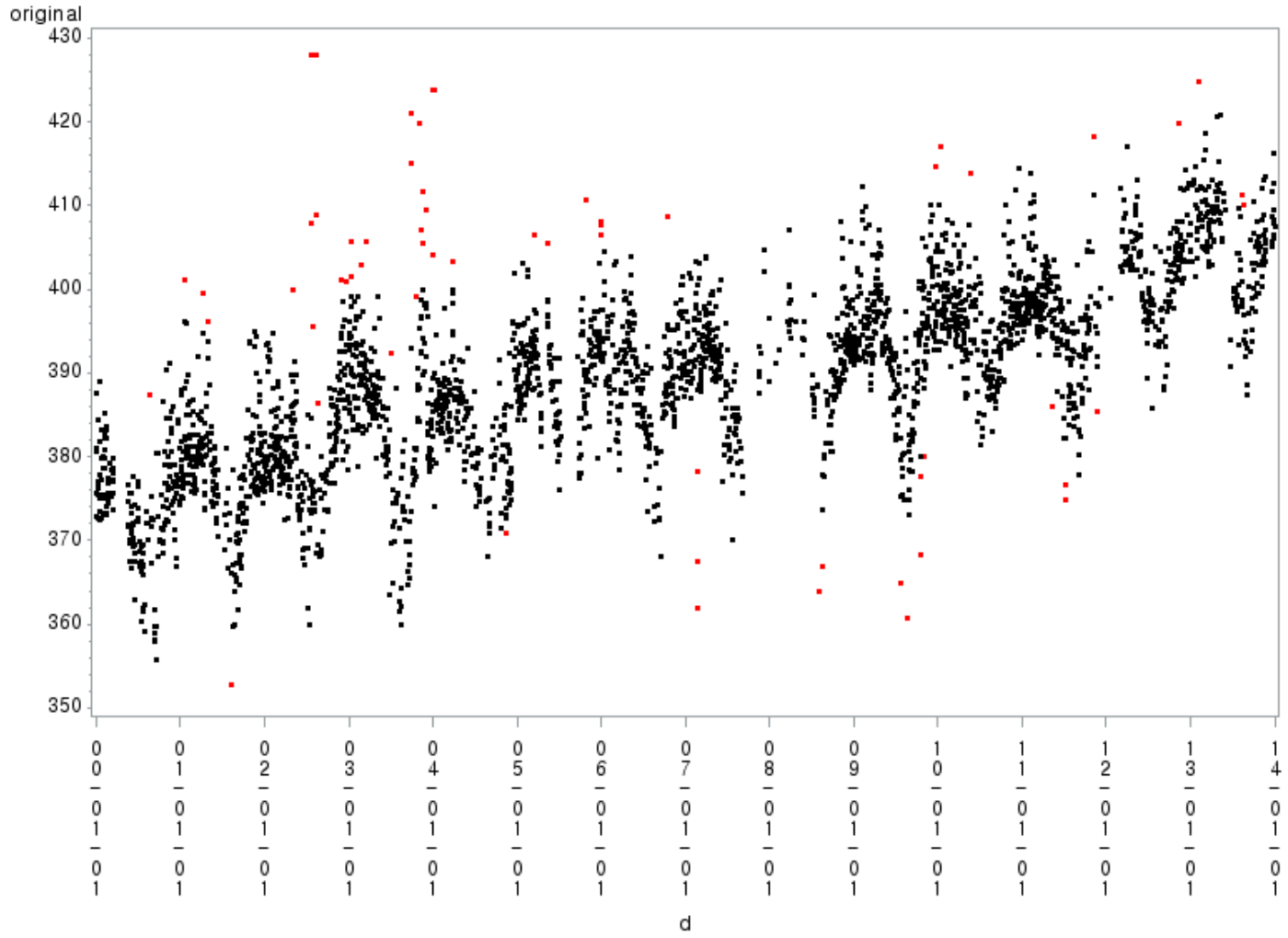
- Convert the remaining lower frequencies (any periods higher than 667 days) after filtering to the time domain residual model using inverse FFT algorithm.

=> Estimated time domain residual model after converting is called $\{r(t)\}_{667d}$.

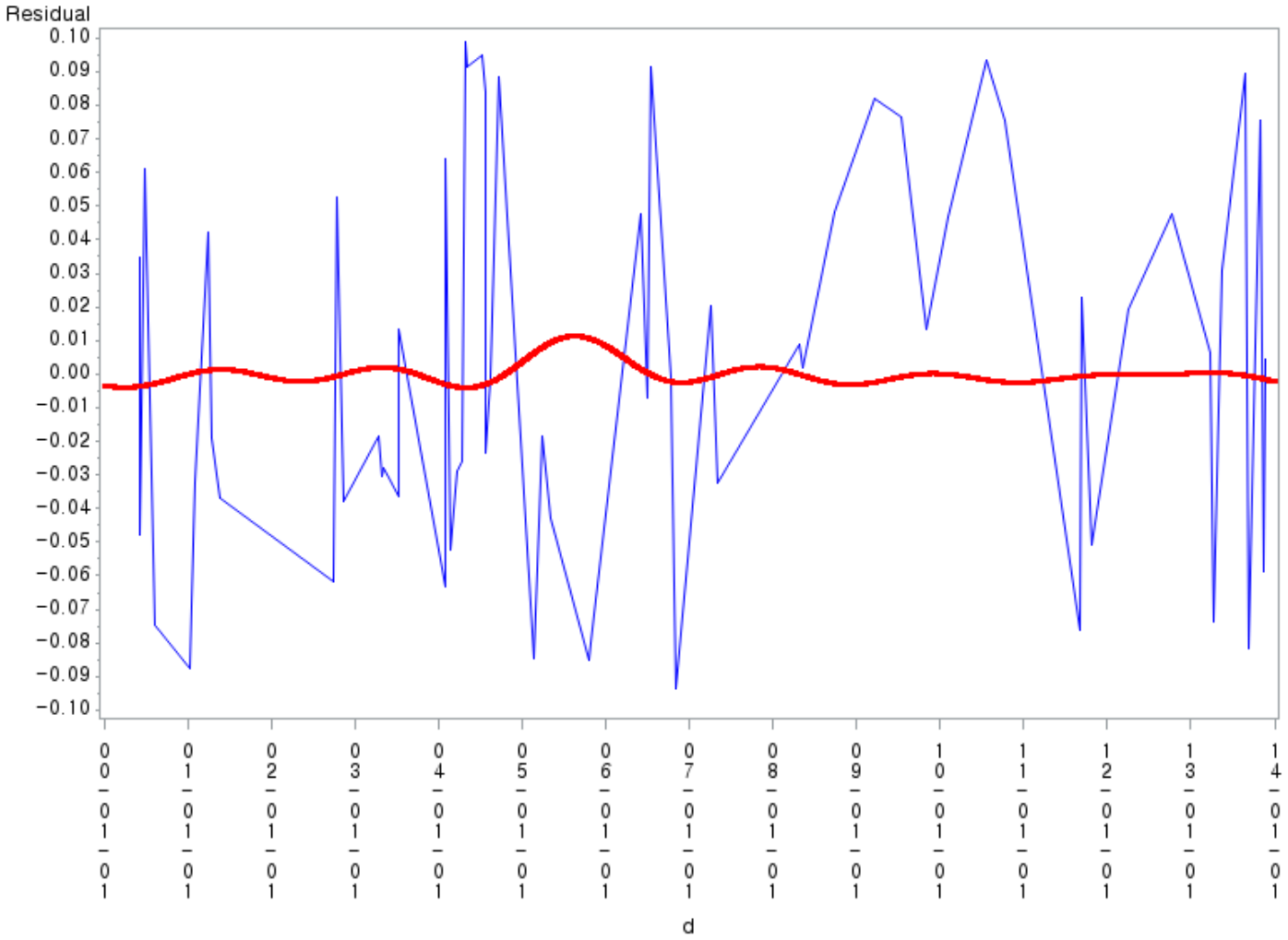
=> Refer to the residual and $\{r(t)\}_{667d}$ plots

- $\{r(t)\}_{667d}$ after low pass filtering is added to the $a_0 + a_1 t + a_2 t^2$, the second-order polynomial. That is, $T(t) = a_0 + a_1 t + a_2 t^2 + \{r(t)\}_{667d}$
  => Fit $T(t)$ based on the remaining observations in the plot(black dots).

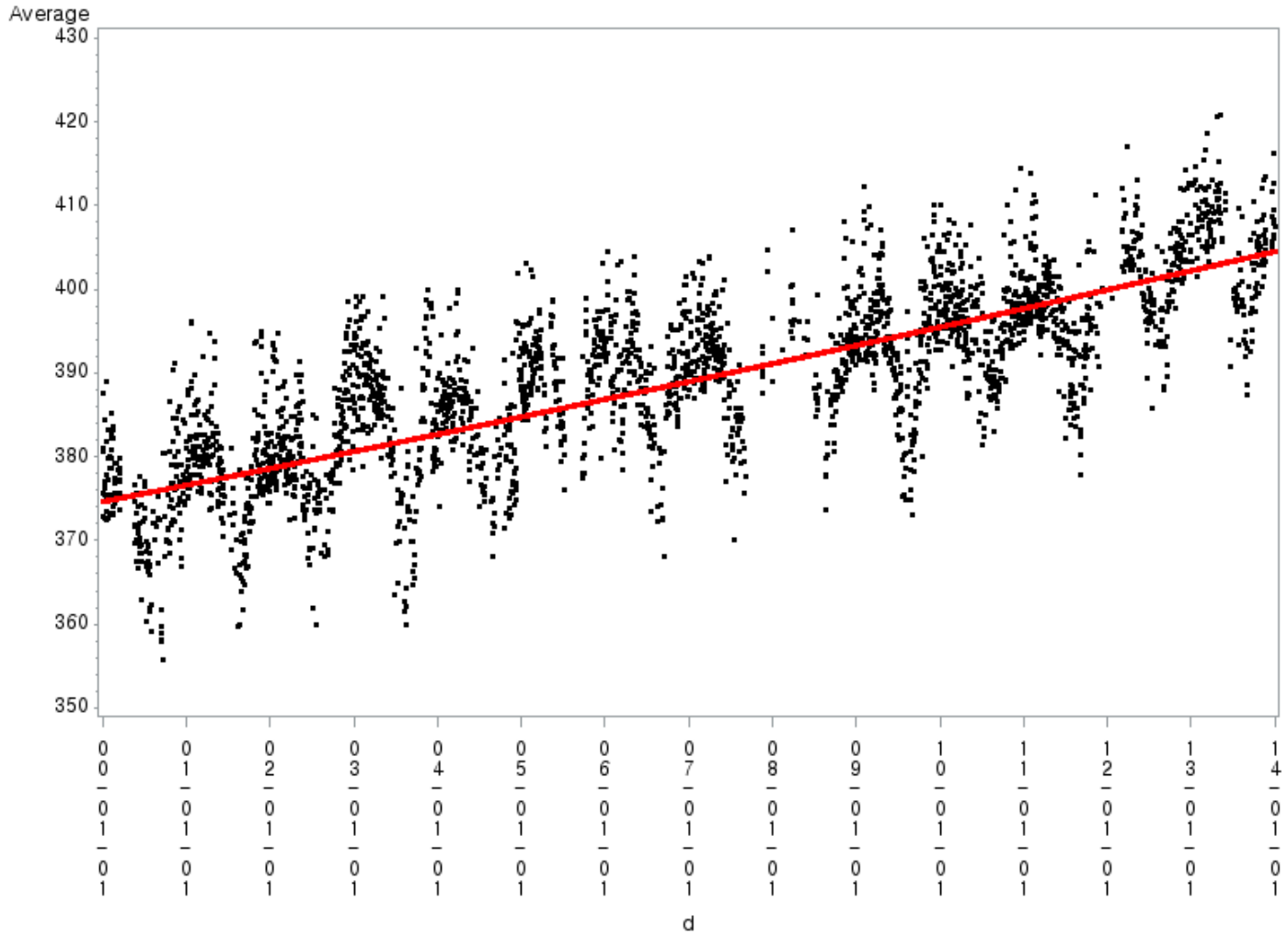
=> The fitted $T(t)$ is the final estimated line for the background concentration determination. Click for the fitted $T(t)$.



Remained data after QC process (Black Dot) and Removed Observations (Red Dot)



Plot Residuals(Blue Line) and Periodic Model of Residuals $\{r(t)\}_{667d}$ (Red Line) :Zoomed



Quality Controlled Daily Average CO2 Concentration and Inter-Annual Trend

I. Estimating the Background CO2 Concentration

1. Data Preprocessing
2. Data Analysis
3. Conclusion

- From the cleaned daily averaged data (Figure), estimate the yearly average background CO2 concentration based on the red line of the previous figure from 2000 to 2013

Year	2000	2001	2002	2003	2004	2005	2006
Estimated Average Background CO2 Concentration (ppm)	375.607	377.598	379.605	381.640	383.698	385.794	387.892
	2007	2008	2009	2010	2011	2012	2013
	390.022	392.178	394.361	396.564	398.793	401.049	403.329

Annual Average Increase: 2.1348ppm

In case of 2014, there are insufficient data to calculate annual average.
But we can estimate it **405.4638ppm** (2013 average concentration + annual average)

$$\text{Annual Average Increase} = \frac{[\text{Max}(T(t)) - Min(T(t))]}{Total\ \ Days}$$

$T(t)$: Inter-annual trend (red line of the figure)

II. Future Challenges

- Discussions in this study

- **In the Preprocessing Step:**

  - Why are the threshold values are 350ppm and 430ppm ?

  - Is that meaningful to remove the data which the number of measurements are less than half in spite of using averaged value ?

  - Why standard deviation should be below 1.8ppm ?
    Where are the value comes from ?

- **In the Interpolation Step:**

  - There are several methods to interpolate the missing observations.
    Other techniques? (for example, join or forecasted values from the model?)

- Discussions in this study (continued)

- **In the Low Pass Filtering Step:**

  - There are basic assumption that it takes 2 or 3 months (60~90 days) to mixing the background atmospheric concentration in the northern hemisphere.

  - Any other low pass filtering threshold rather than below 50 days or below 667 days ?

  - Or Instead, can we consider to select some powerful and leading frequencies because they can explain almost all of the patterns of residuals ?

Any Other Comments or issues?

Thank You

Q & A

Yung-Seop Lee

yung@dongguk.edu

III. Appendix

1. Threshold of Preprocessing
2. Interpolation Method for Missing Value
3. Threshold of Low Pass Filtering
4. Using Specific Period for Filtering

➢ Raw Data (Level 0)

OBS	day	month	year	time	avg	std	num
1	1	1	2000	0	419.09	4.40	108
2	1	1	2000	1	415.72	0.86	90
3	1	1	2000	2	408.92	2.38	114
4	1	1	2000	3	405.36	0.75	89
5	1	1	2000	4	404.19	1.18	115
6	1	1	2000	5	402.19	1.23	89
7	1	1	2000	6	395.58	4.01	114
8	1	1	2000	7	393.42	1.29	90

- From January 01, 2000 to March 31, 2014. Time Interval: 1 Hour.
- 112,351 Observations
- **Variables**: Date, Month, Year, Time(Hour), Average Raw Data of Specific Hour Standard Deviation of Average Raw Data, Number of Raw Data

✓ Preprocessing for change the form of data from time scale to date scale

✓ There are 5 conditions for this preprocessing

1. To remove average raw values when the number of raw data is less than 60 which is half of the hourly collected
2. To exclude average raw values when the standard deviation of the average raw value is above 1.8ppm
3. To remove hourly data over 430ppm or under 350ppm
4. To reject hourly data if the difference between consecutive hourly average concentration is larger than 1.8ppm
5. To remove average hourly values when the number of hourly data is less than 12 which is half of the daily collected

Question >

Then, why that value is 350ppm and 430ppm?
Is that values are statistically significant?

1. To remove average raw values when the number of raw data is less than 60 which is half of the hourly collected

2. To exclude average raw values when the standard deviation of the average raw value is above 1.8ppm

3. To remove hourly data over 430ppm or under 350ppm

4. To reject hourly data if the difference between consecutive hourly average concentration is larger than 1.8ppm

5. To remove average hourly values when the number of hourly data is less than 12 which is half of the daily collected

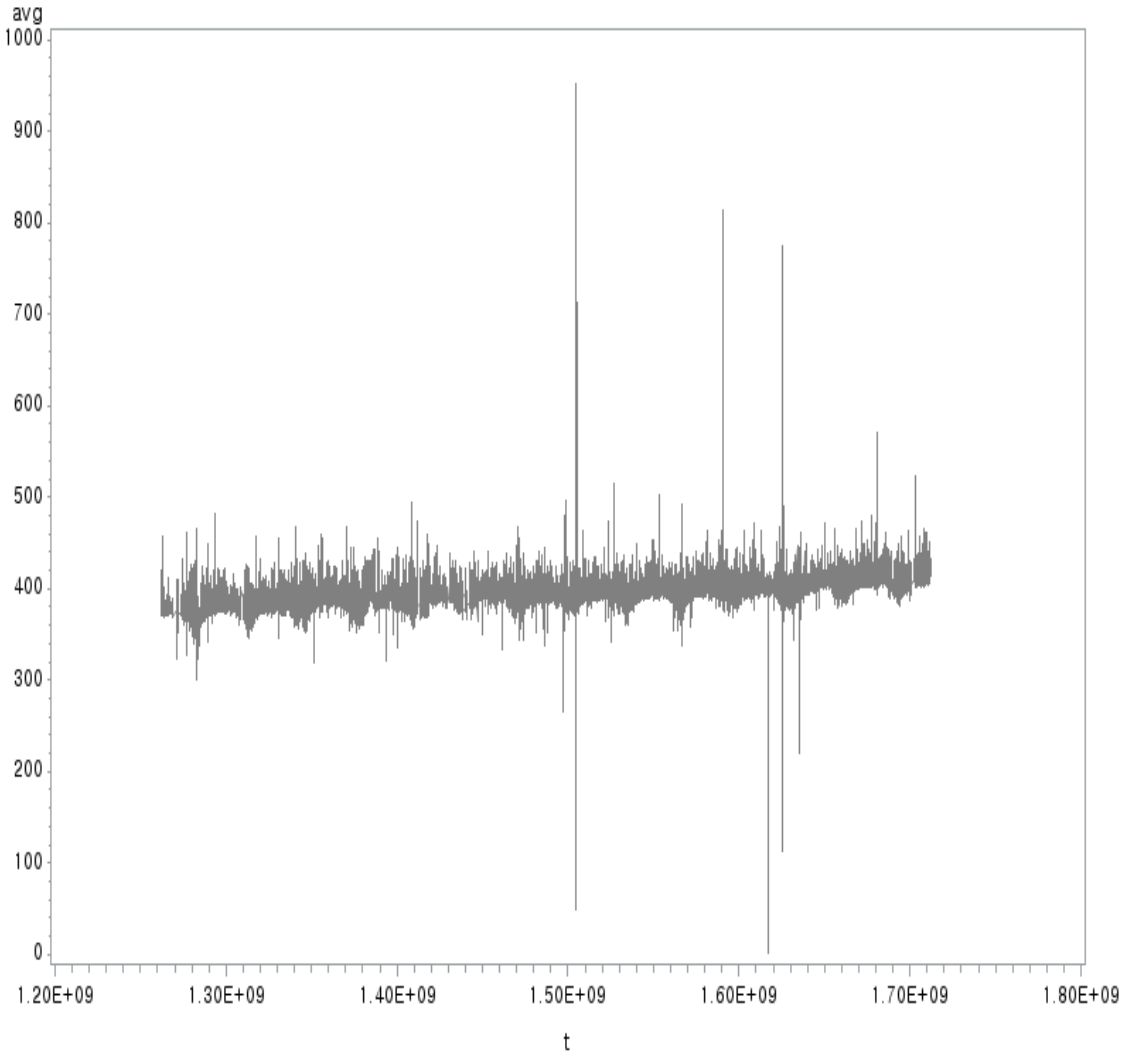
✓ **We think 3 different method to reject hourly data and compare it**

A. To remove hourly data over 430ppm or under 350ppm

B. To estimate time-series model for hourly data. And then remove hourly data over maximum or under minimum value of 95% confidence interval of the model

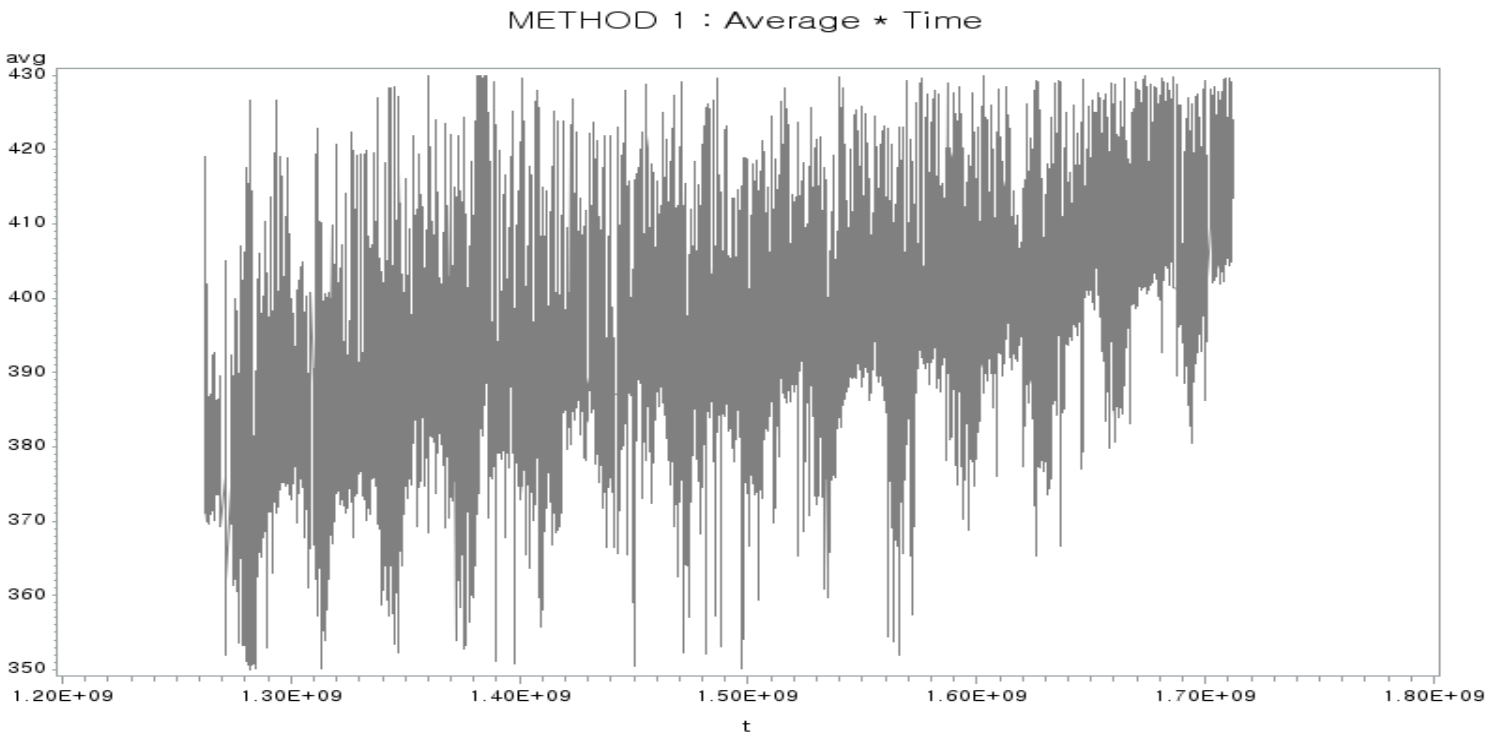
C. To estimate time-series model for hourly data. And then reject all of the data out of 95% confidence interval of the model

Other conditions are all same. Just change this method for comparing



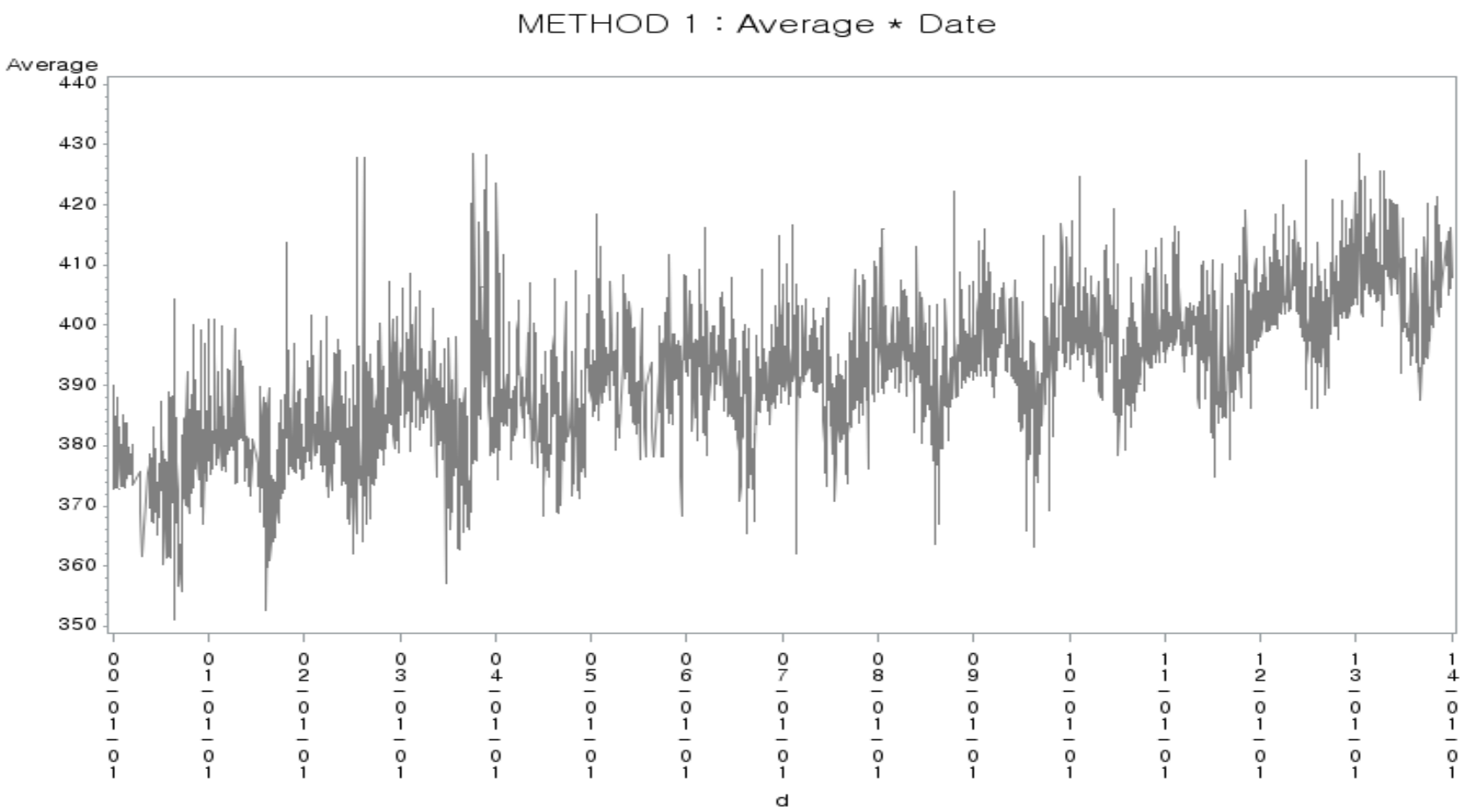
Plot of Raw Data against Time

A. To remove hourly data over 430ppm or under 350ppm



Filtered data under given condition (time scale)

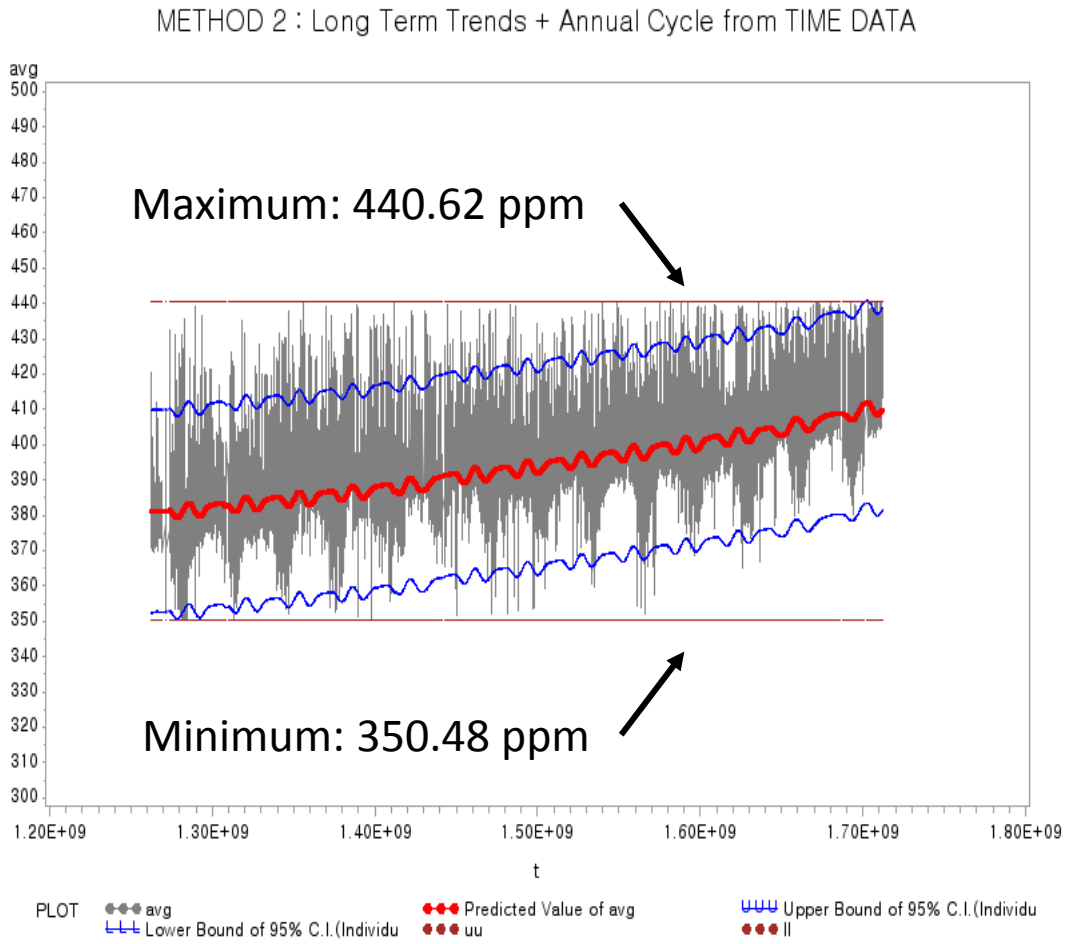
- **1819** data rejected from 112351 data
- Data rejected which have over 430ppm or under 350ppm concentration



Filtered data under given condition (date scale)

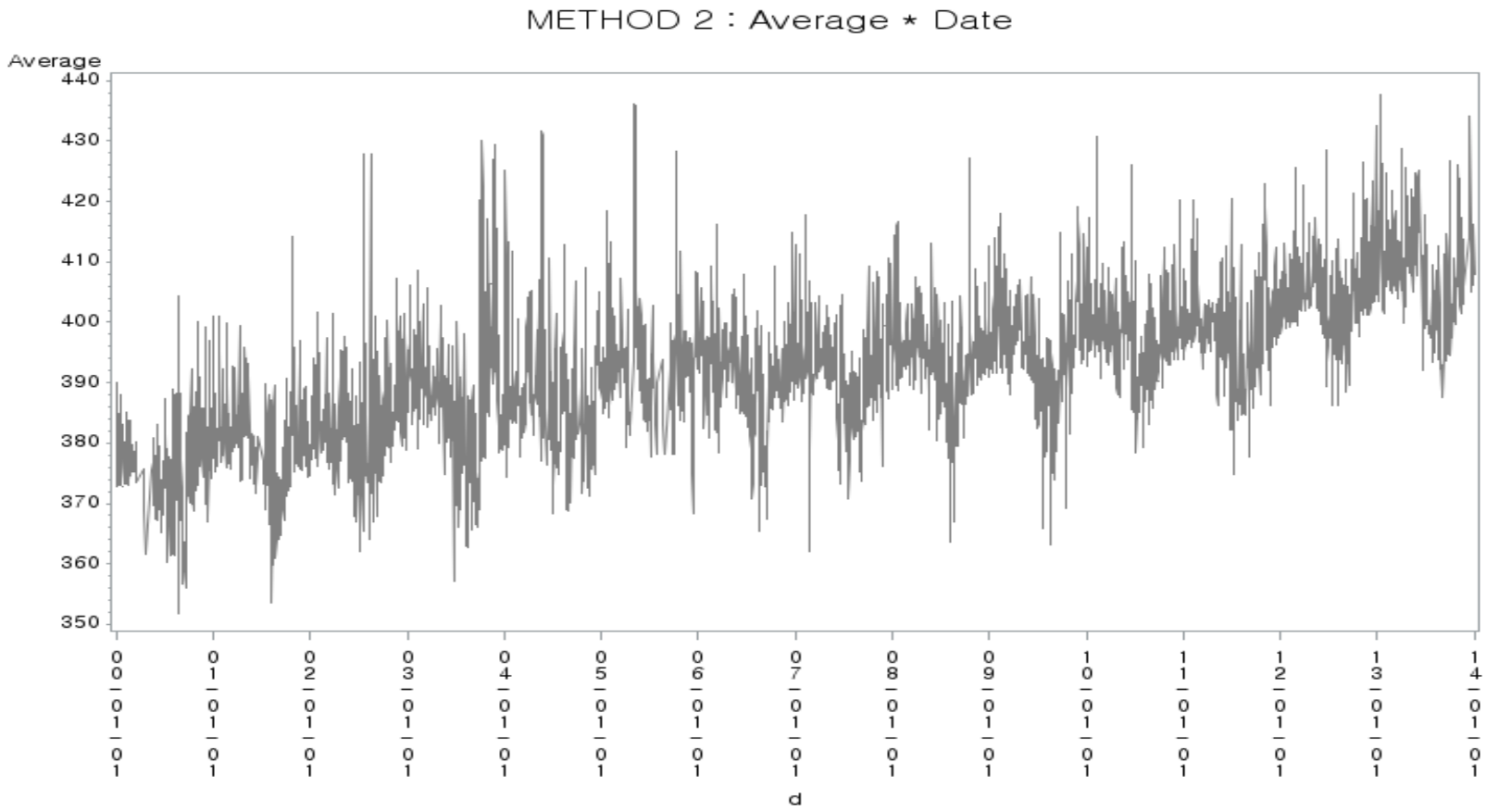
- There are **266** missing values in this daily data

B. To estimate time-series model for hourly data. And then remove hourly data over maximum or under minimum value of 95% confidence interval of the model



Filtered data under given condition (time scale)

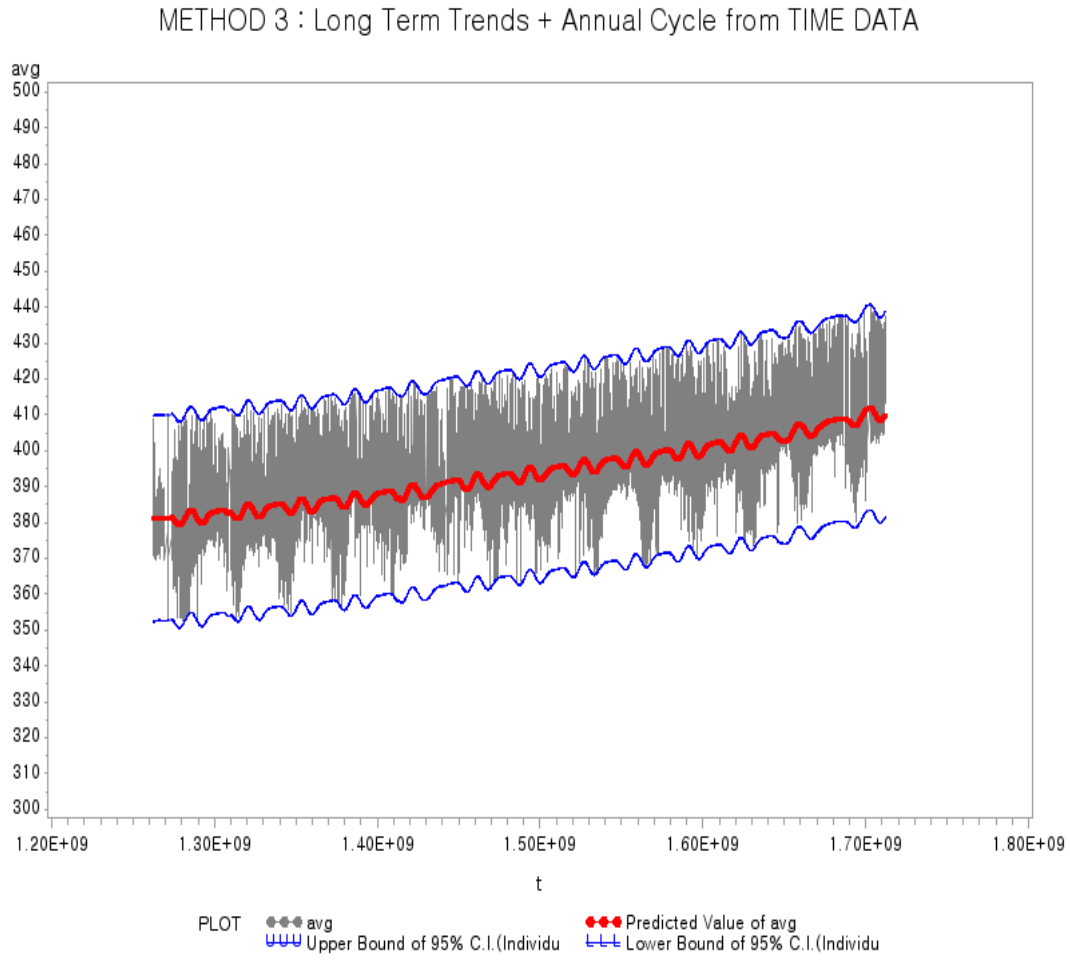
- **702** data rejected from 112351 data
- Estimate time-series model (red) and calculate maximum and minimum value of 95% confidence interval (Burgundy)
- Reject over maximum or under minimum values



Filtered data under given condition (date scale)

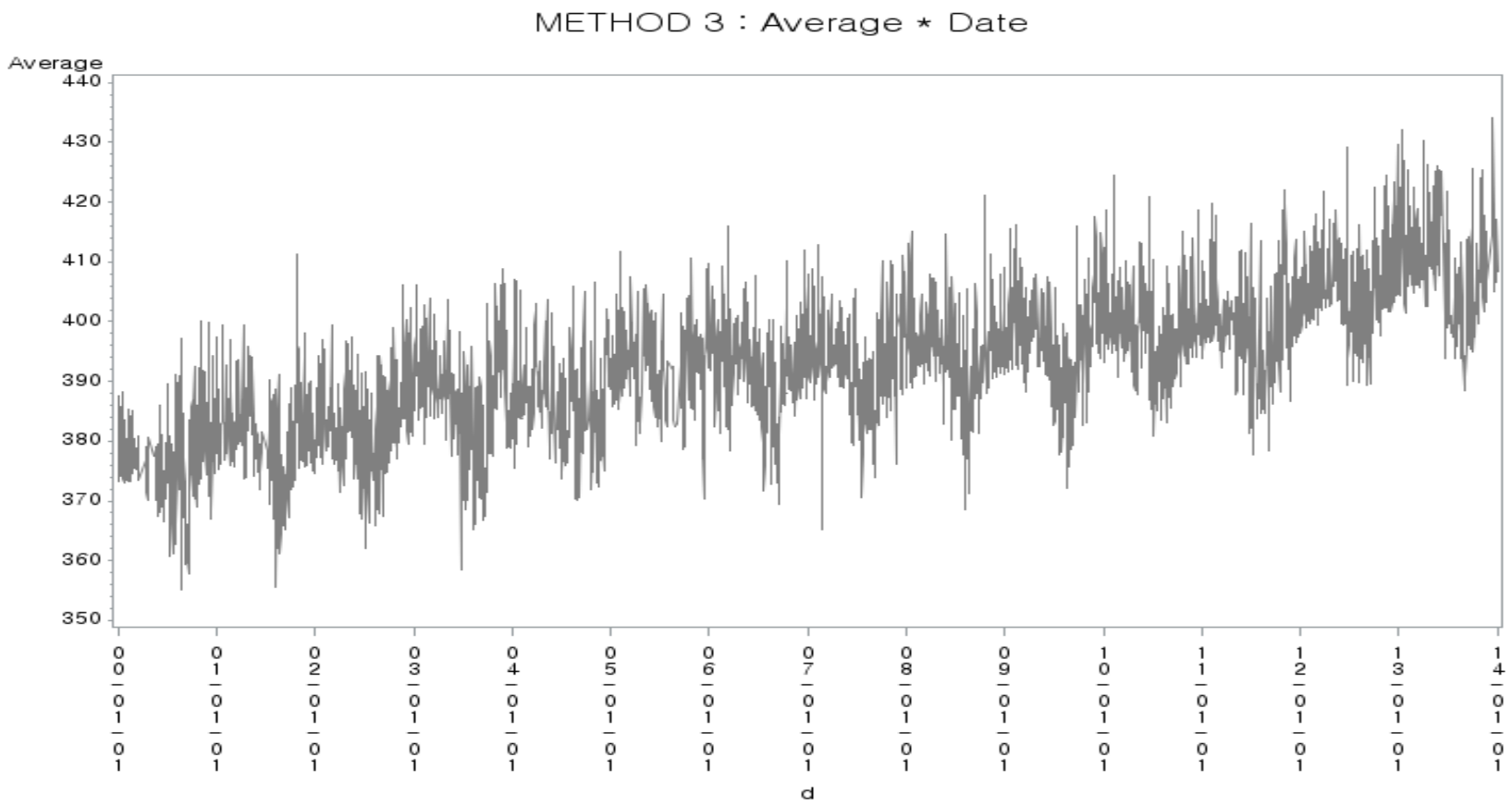
- There are **257** missing values in this daily data

C. To estimate time-series model for hourly data. And then reject all of the data out of 95% confidence interval of the model



Filtered data under given condition (time scale)

- **3080** data rejected from 112351 data
- Estimate time-series model (red) and calculate 95% confidence interval (blue)
- Reject all data out of 95% confidence interval



Filtered data under given condition (date scale)

- There are **273** missing values in this daily data

- If the thresholds are changed, the rejected time data and daily data could be changed.
- The numbers of data are almost same, but which one is statistically confidence for preprocessing?
- It might be interesting issue in respect of accuracy of model

Another Question >

Is that meaningful to reject whole data because the number of data is less than half in spite of using average daily value?

1. **To remove average raw values when the number of raw data is less than 60 which is half of the hourly collected**

2. To exclude average raw values when the standard deviation of the average raw value is above 1.8ppm

3. To remove hourly data over 430ppm or under 350ppm

4. To reject hourly data if the difference between consecutive hourly average concentration is larger than 1.8ppm

5. **To remove average hourly values when the number of hourly data is less than 12 which is half of the daily collected**

Another Question >

Why standard deviation should be 1.8ppm. Where are the value comes from?

1. To remove average raw values when the number of raw data is less than 60 which is half of the hourly collected

2. **To exclude average raw values when the standard deviation of the average raw value is above 1.8ppm**

3. To remove hourly data over 430ppm or under 350ppm

4. To reject hourly data if the difference between consecutive hourly average concentration is larger than 1.8ppm

5. To remove average hourly values when the number of hourly data is less than 12 which is half of the daily collected

Another Question >

Where are the value comes from? How they select this value of 1.8ppm?

1. To remove average raw values when the number of raw data is less than 60 which is half of the hourly collected
2. To exclude average raw values when the standard deviation of the average raw value is above 1.8ppm
3. To remove hourly data over 430ppm or under 350ppm
4. **To reject hourly data if the difference between consecutive hourly average concentration is larger than 1.8ppm**
5. To remove average hourly values when the number of hourly data is less than 12 which is half of the daily collected

➔ **There are many interesting topic in this preprocessing step.**

➔ **And we may enhance the process to additional study**

The distribution of standard deviation of raw data

Coefficient of variance ($\frac{Standard\ Deviation}{Average}$)

The distribution of difference average of time data

And so forth…

III. Appendix

1. Threshold of Preprocessing
2. Interpolation Method for Missing Value
3. Threshold of Low Pass Filtering
4. Using Specific Period for Filtering

✓ INTERPOLATION

- We find hidden **frequency** of data using spectral analysis and FFT
- If the data has missing value in the middle of consequence data, it might be a problem to find frequency of the data
- Therefore, we have to INTERPOLATE the missing value in the middle of consequence data

SPLINE

Interpolate using second order polynomial curves

JOIN

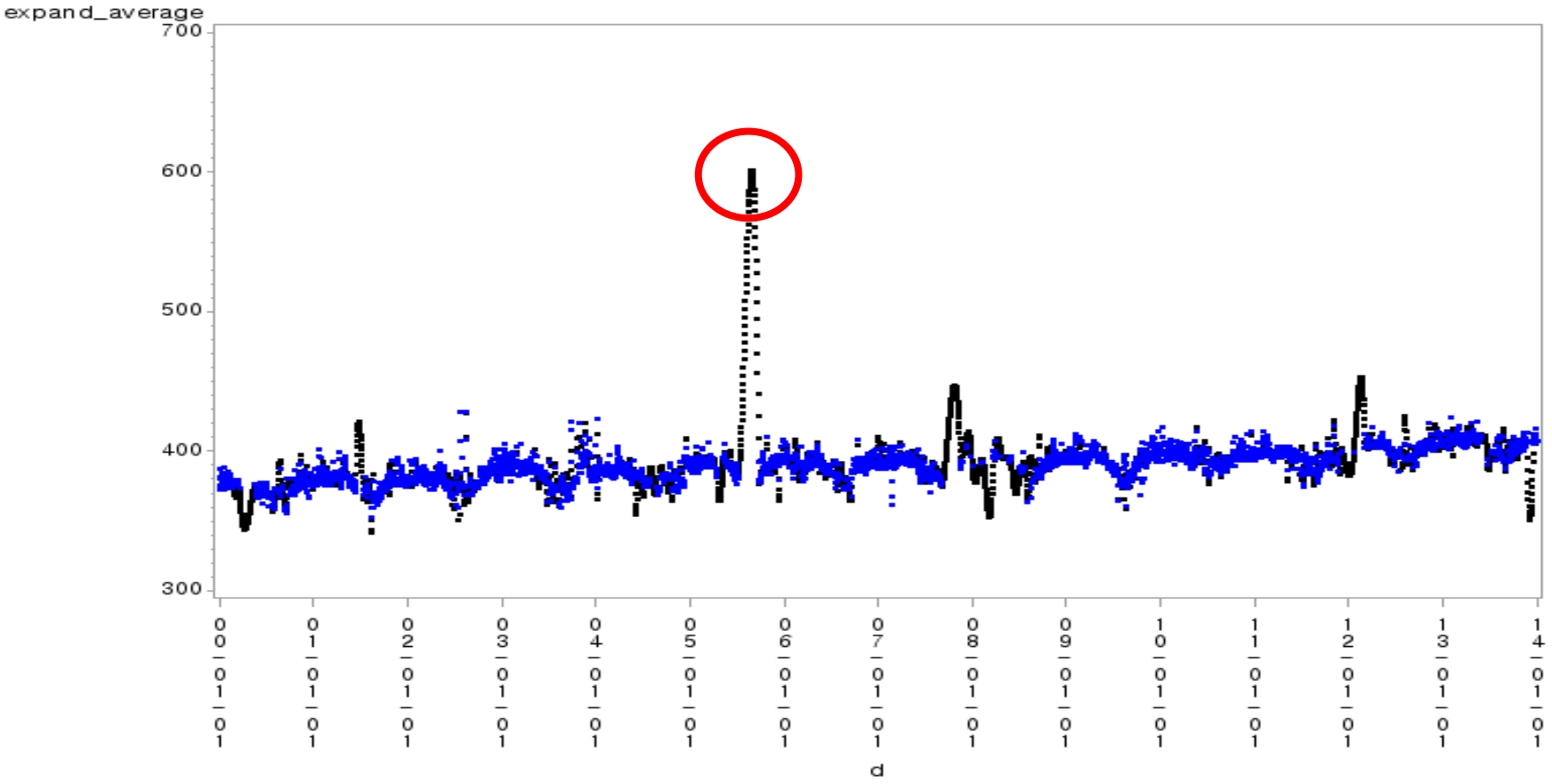
Interpolate using first order polynomial lines

TIME-SERIES PREDICTED VALUE

Estimate time series model using given data. Then fill the missing value with predicted time series value

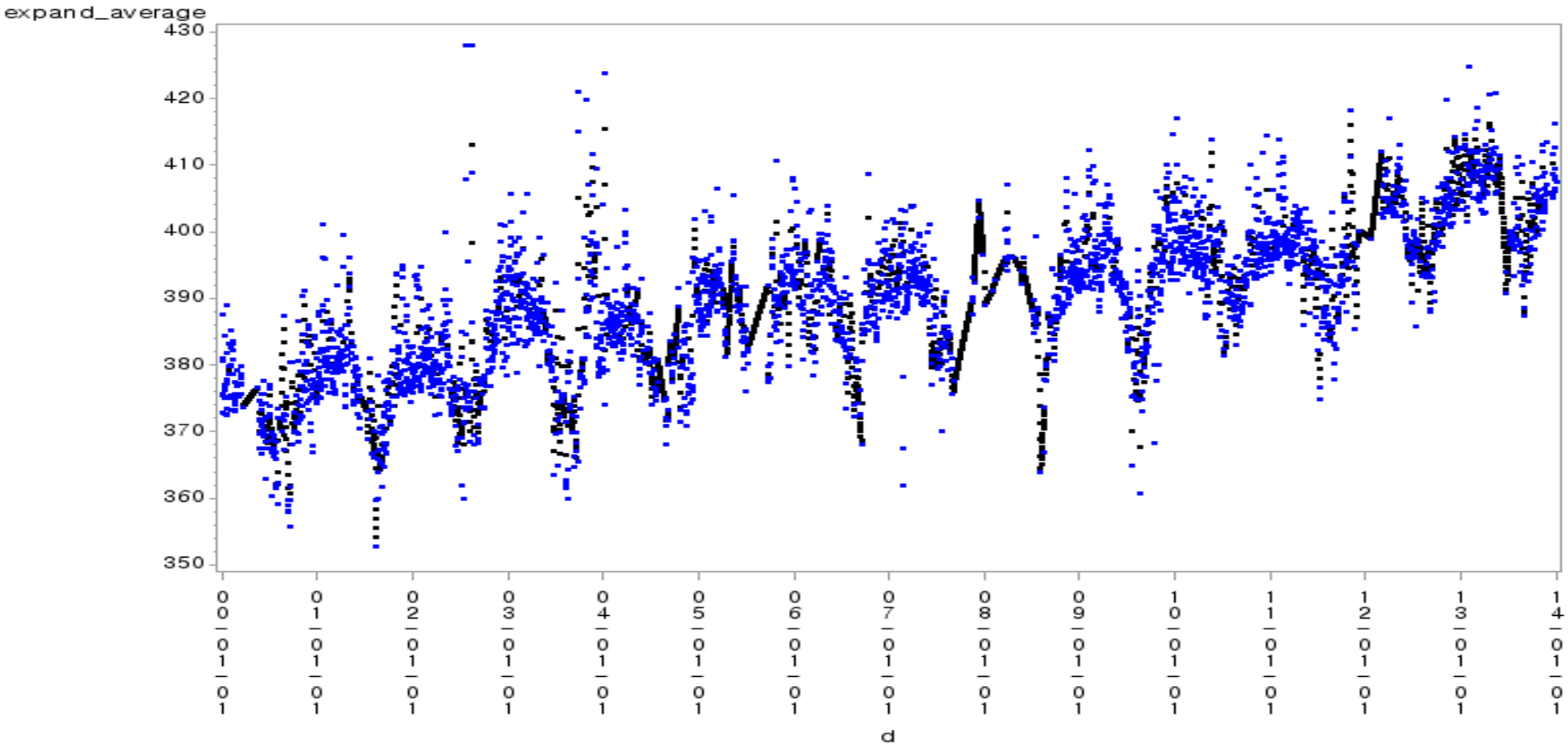
There are diverse method to fill up the data and it's important issue for statisticians
➔ **Which interpolation method can minimize the loss of information?**

A. SPLINE



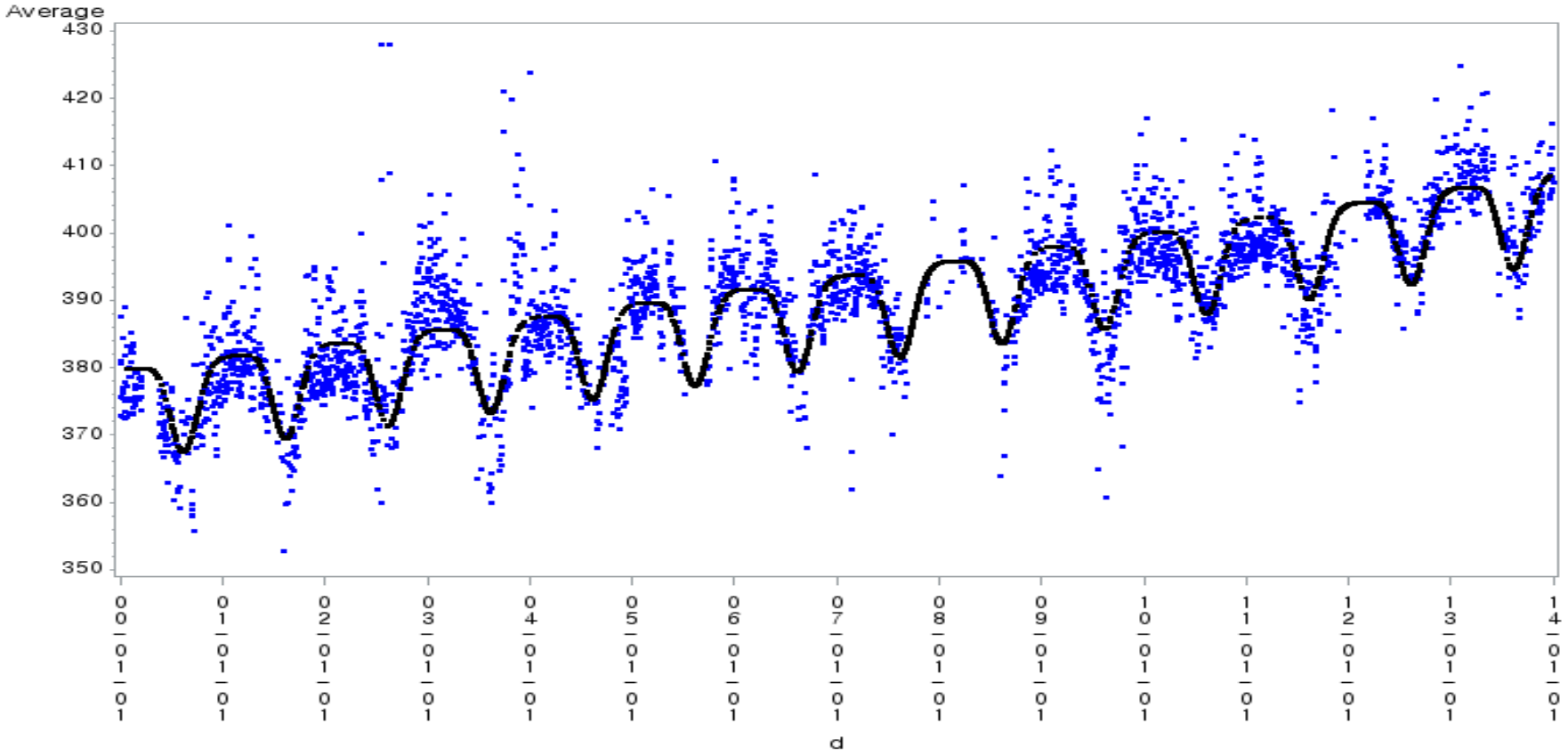
- If data has long space between two point, it causes serious distortion of data

B. JOIN



- This method simply connect two separate point using line.
- So it's inadequate method to apply smooth curve of the data

C. TIME-SERIES PREDICTED VALUE



- This method requires to adjust at the end of the consequence line because predicted value can differ from original daily data at the specific time point.

- Each method has strengths and weaknesses for each cases

- Additional studies are required to find optimal interpolation method which can minimize loss of information of the data

III. Appendix

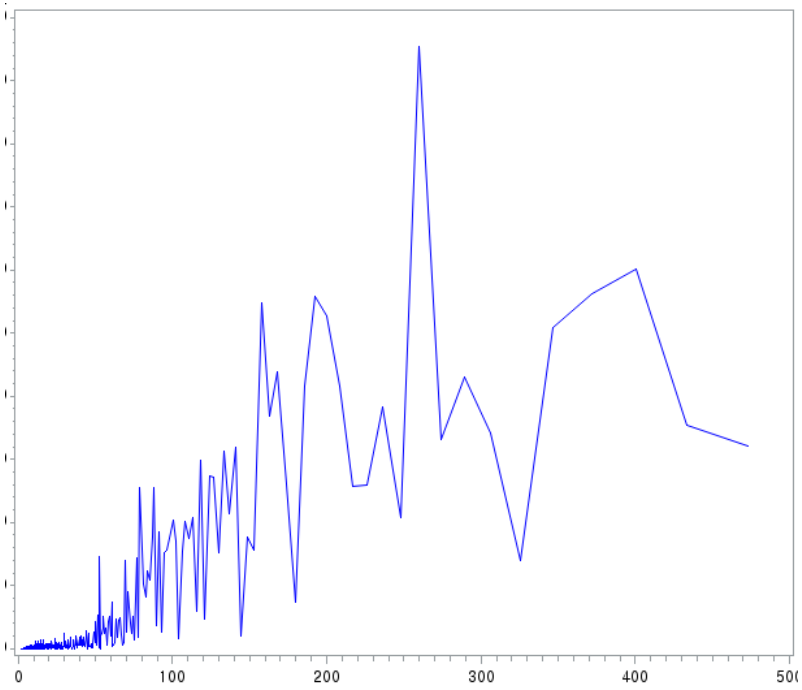
1. Threshold of Preprocessing
2. Interpolation Method for Missing Value
3. Threshold of Low Pass Filtering
4. Using Specific Period for Filtering

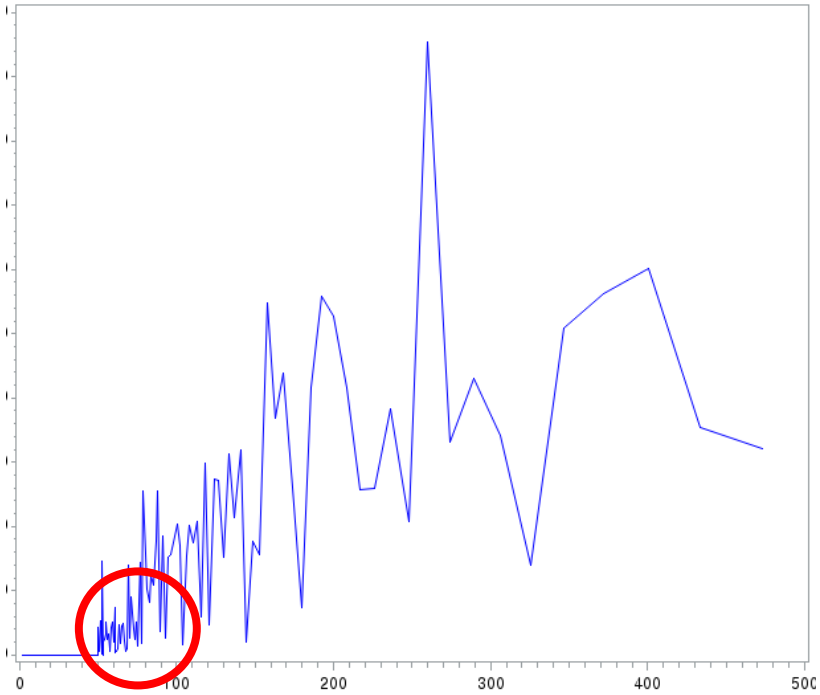
- The thresholds of low pass filtering in the data analysis part are in below..

  - Applying low pass filter to eliminate frequency higher than **7.3 cycle/yr (period lower than 50 days)**

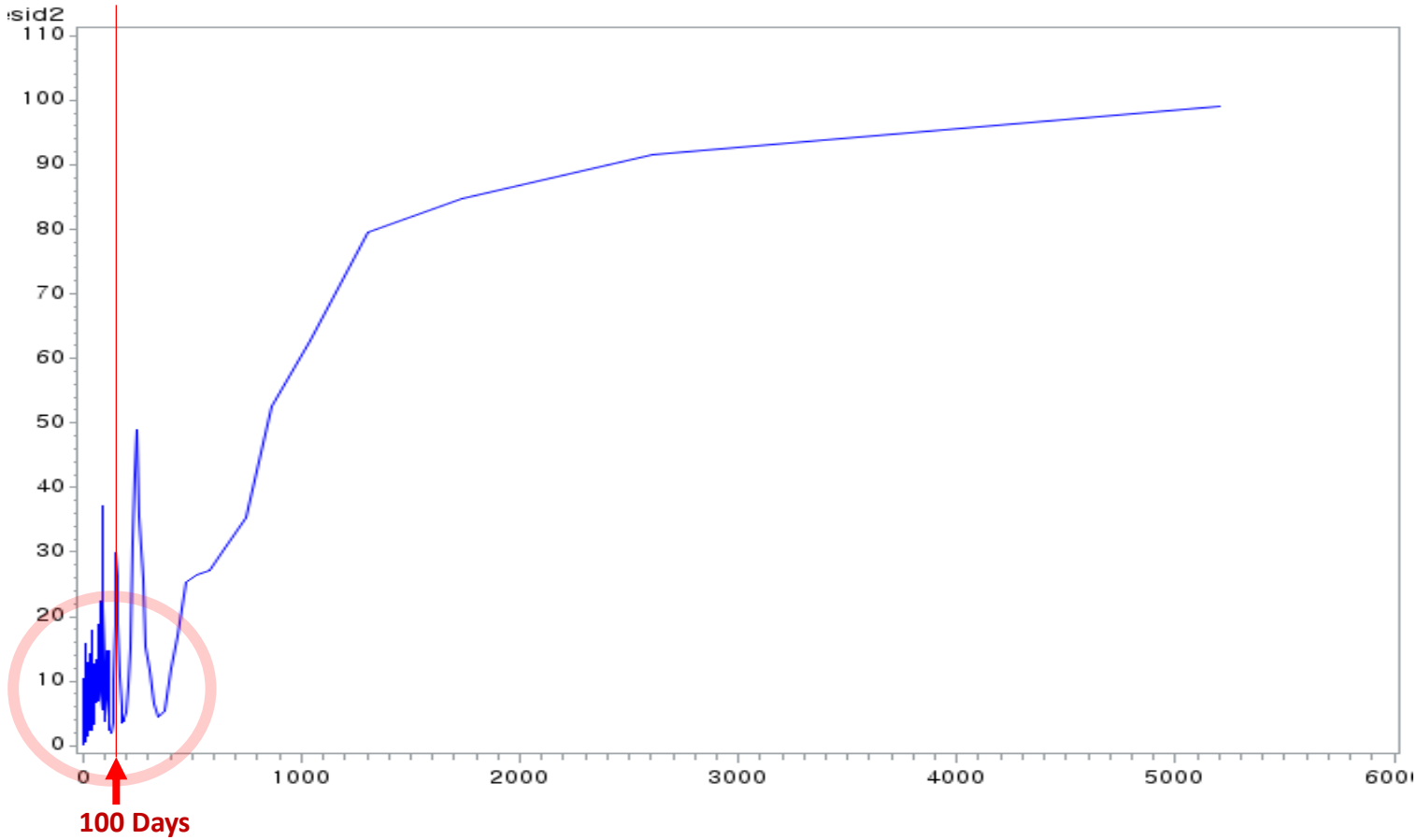
  - Applying low pass filter to eliminate frequency higher than **0.55 cycle/yr (period lower than 667 days)** and we are going to add this model to inter-annual trend

➢ Then why 50 days and 667 days?





We filtered that signal so eliminate period lower than 50 days.
But it might seem that it still has noise (in the red circle) at the high frequency



This is a result of some simulation and it might seem that there are noise in the period lower than 100 days.
And there are basic assumption that it takes 2 or 3 months (60~90 days) to mixing the background atmospheric concentration in the northern hemisphere

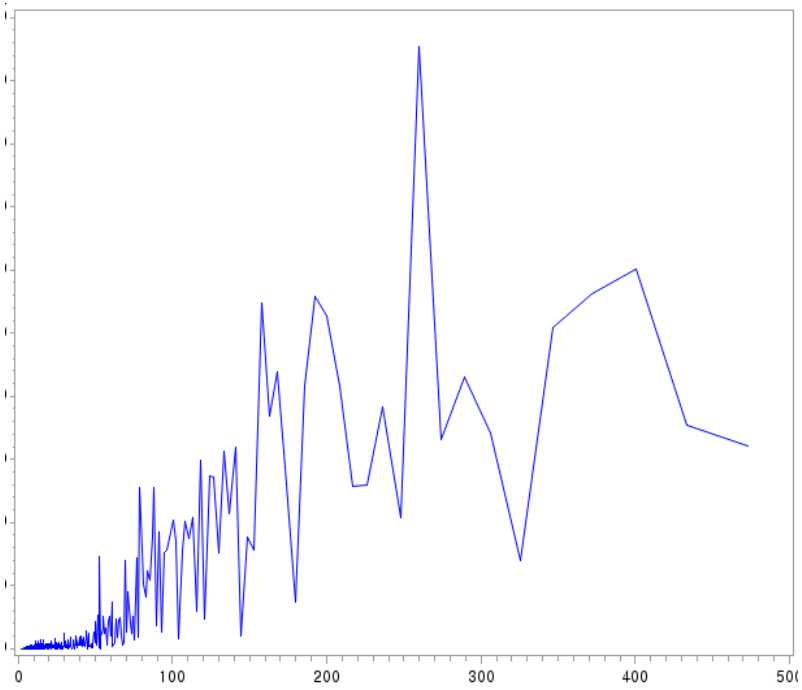
- There are no clear evidence to select cut-off of 667 days for long-term trend

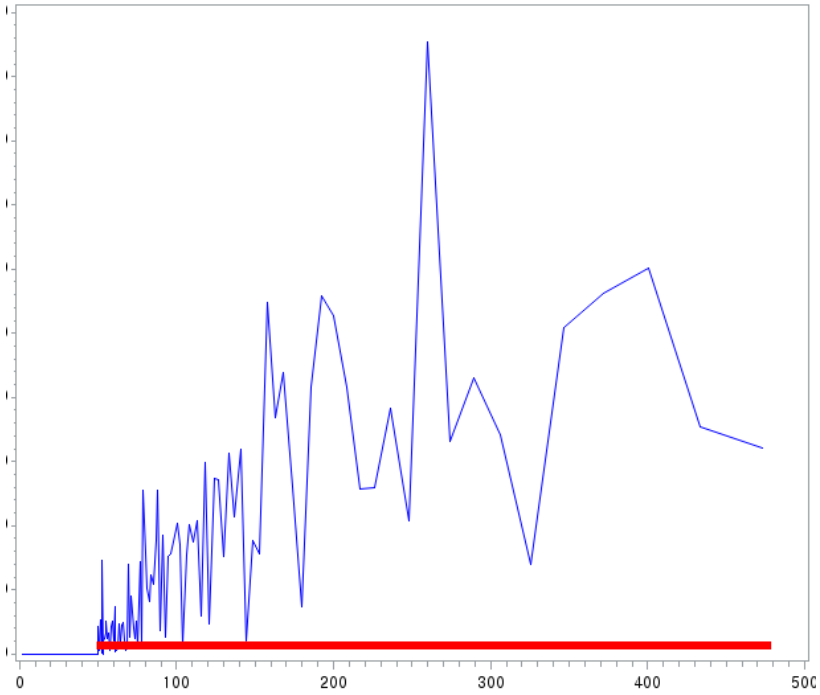
- Sufficient study about filtering threshold values are required to get more accurate result of frequency analysis

III. Appendix

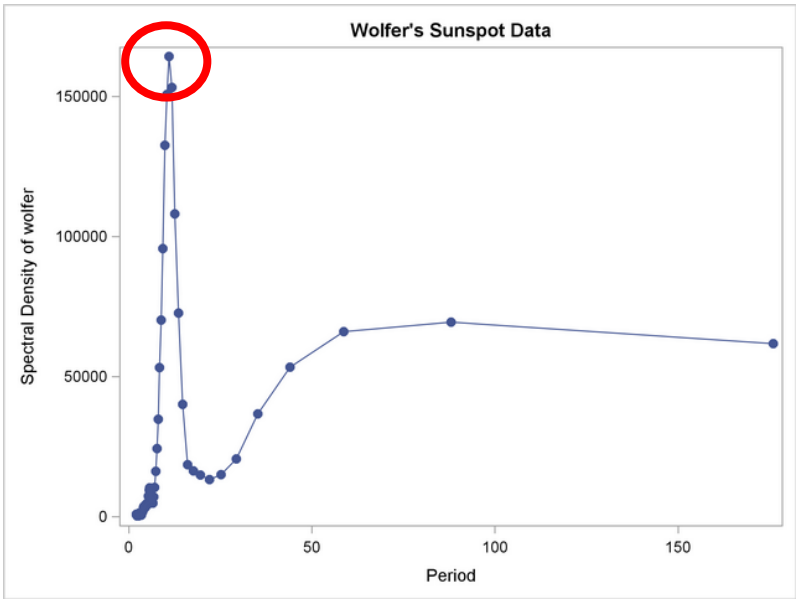
1. Threshold of Preprocessing
2. Interpolation Method for Missing Value
3. Threshold of Low Pass Filtering
4. Using Specific Period instead of LPF

- Researcher decides which period to use for spectral analysis
- We select every filtered frequency and use that for residual model by inverse transformation in this study
- **Instead, we can consider to select some powerful frequency because they can explain almost of the pattern of residuals**





- Now, we select all of these frequency to model



- But sometimes, 2 or 3 frequency can decide whole pattern of the data because there are powerful

➔ In this case there are no difference between the two methods (Using all frequency or Using 2 or 3 frequency)

- Select some frequency that can explain most of the pattern
- Try significance test about that frequency through Fisher's g-test
- If selected frequency has significance, then try the test for next frequency
- Iterate this process to select all of significance frequency

Using this method, we can expect to explain of specific frequency that we found.

That is, if the frequency is similar to the frequency of tide in the same area, then we estimate that this concentration should be affected by tide of that area.
And it means that our CO2 concentration is not pure.

So, we think in depth research into this method is require to find adequate CO2 concentration model

➔ We can consider all these matters in the view of statisticians.
That is, there are much more issues that we may research.