A Functional Data Analysis Registration Model and Its Application to CO_2 Flux Data

Fang He[1], Duncan Murdoch[2], and Reg Kulperger[2]

[1,2]Department of Statistical and Actuarial Sciences, University of Western Ontario

September 15, 2016

Abstract. Carbon dioxide (CO_2) flux is needed for assessing and monitoring the annual carbon cycle. Only a small proportion of the land is currently covered by proper equipment to directly collect CO_2 flux data. With the help of the Moderate Resolution Imaging Spectroradiometer (MODIS) which is carried by NASA satellites, corresponding data, such as normalized difference vegetation index (NDVI), is freely available from NASA. Our goal is building a model using MODIS data to predict the CO_2 flux data at any location. The CO_2 flux data has an obvious annual cycle with the phase changing from year to year. How to build a model to estimate the annual effect and seasonal dynamics is a challenging task. We use two models to analyze the data. The first model analyzes the data as a time series indexed by day, with additional covariates, such as species and NDVI. We use a generalized additive model (GAM). One defect of this model is its inability to capture the seasonal dynamics. In the second model we treat each year as an independent multivariate observation. We use functional data analysis (FDA) to smooth the CO_2 flux data for each year separately. Then we use the FDA registration to standardize the seasonal dynamics. On the registered time scale, we build a model of the registered curves using a multivariate normal distribution. We use Dirichlet regression to model the seasonal dynamics and map the registered curves back to the natural time scale. These two steps allow us to simulate CO_2 flux on the natural time scale. We use 4 locations to conduct the study. The simulation results demonstrate the potential of the second model based on the FDA registration method.

Contents

1 Introduction

Climate change has a big impact on plants and animal life. One thing that ecologists are interested in is phenology. i.e., the analysis of cyclic and seasonal natural phenomena, especially in relation to climate and plant and animal life. There are many ways to conduct the analysis. Current challenges include source data selection, gap filling techniques, and lack of comparison of current phenology study methods. There are usually two types of sources used in the phenology studies. One source is remote sensing data, the other is ground based data. Compared to the global wide remote sensing data, ground based data in a phenology study only covers a small portion of the land surface. Since it is costly to get ground based data at a new location, we are interested in building a model, and using remote sensing data to predict the ground based data.

How much carbon dioxide moves through a unit area per unit time is called carbon dioxide (CO_2) flux. CO_2 flux above-canopy/below-canopy measures the CO_2 consumption and production by plants. CO_2 flux is an important component of the total atmospheric carbon balance, and it is important for the study of global climate change.

Flux towers provide ground-based direct observations of carbon exchange. The tower is installed at a fixed location on the Earth. Each tower works locally. Currently there are over 650 tower sites in the world working together, and forming a network of regional networks. Based on where the CO_2 fluxes are measured, there are soil-surface CO_2 flux, below-canopy CO_2 flux, above-canopy CO_2 flux, etc. Figure 1 shows flux tower examples.

2 Introduction

Climate change has a big impact on plants and animal life. One thing that ecologists are interested in is phenology. i.e., the analysis of cyclic and seasonal natural phenomena, especially in relation to climate and plant and animal life. There are many ways to conduct the analysis. Current challenges include source data selection, gap filling techniques, and lack of comparison of current phenology study methods. There are usually two types of sources used in the phenology studies. One source is remote sensing data, the other is ground based data. Compared to the global wide remote sensing data, ground based data in a phenology study only covers a small portion of the land surface. Since it is costly to get ground based data at a new location, we are interested in building a model, and using remote sensing data to predict the ground based data.

How much carbon dioxide moves through a unit area per unit time is called carbon dioxide (CO_2) flux. CO_2 flux above-canopy/below-canopy measures the CO_2 consumption and production by plants. CO_2 flux is an important component of the total atmospheric carbon balance, and it is important for the study of global climate change.

Flux towers provide ground-based direct observations of carbon exchange. The tower is installed at a fixed location on the Earth. Each tower works locally. Currently there are over 650 tower sites in the world working together, and forming a network of regional networks. Based on where the CO_2 fluxes are measured, there are soil-surface CO_2 flux, below-canopy CO_2 flux, above-canopy CO_2 flux, etc. Figure 1 shows flux tower examples.

The distribution of the flux tower networks (Figure 2) is not uniform. A large proportion of the land is not covered by the flux tower network. Between two adjacent flux towers, the flux is unknown.

MODIS is the abbreviation for moderate-reslution imaging spectroradiometer. The following information of MODIS is summarized from NASA website `http://modis.gsfc.nasa.gov/about/design.php`. MODIS is an instrument currently on board two satellites, Terra (December 1999) and Aqua (May 2002). It has 36 spectral bands, the wavelength is from 0.4μm to 14.4μm. The resolutions are 1 km or finer (2 bands at 250 m, 5 bands at 500 m, 29 bands at 1 km). The altitude of MODIS is 705 km. It takes every 1 to 2 days for MODIS to scan the whole world. Figure 3 shows how MODIS works.

Four locations have been chosen in this study. Each location collected above-canopy CO_2 flux data from ground stations and corresponding normalized difference vegetation index (NDVI) from MODIS. Figure 4 shows negative dependency between the two measurements for each location.

NASA releases the global data collected by MODIS, such as NDVI, leaf area index, temperature and emissivity, and thermal anomalies and fire, free of charge for the public. Statistical models can be built

(a) Flux towers at MSU's Kellogg Biological Station. Photo credit: Bill Krasean.

(b) Flux tower that can measure below/above-canopy CO_2 flux.

Figure 1: Flux tower examples.

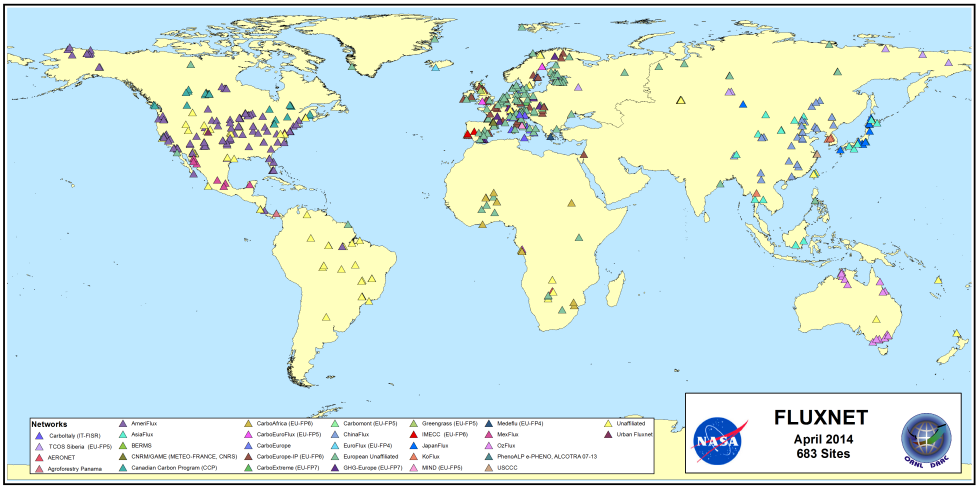
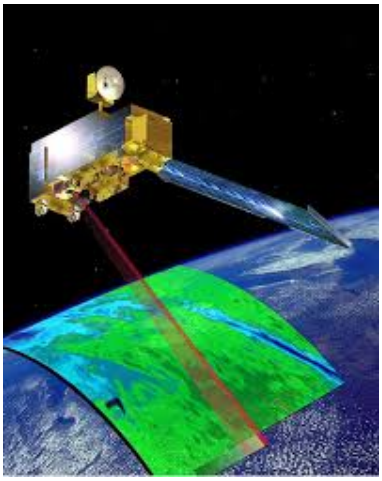


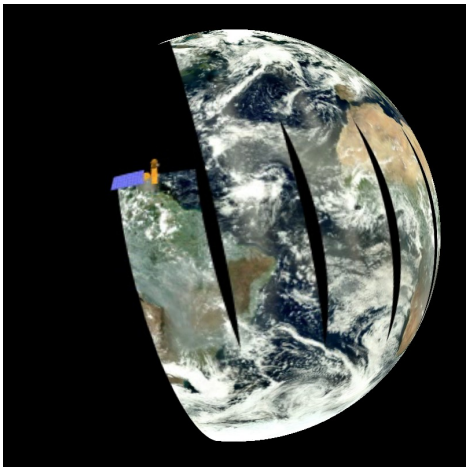
Figure 2: Distribution of flux tower networks. Photo credit: FLUXNET

combining both nearby CO_2 flux data and corresponding satellite data to predict the CO_2 flux data at a new location. With the help of proper statistical models, we can estimate the CO_2 flux at a wider area.

The following three chapters give the scientific background: a brief introduction of MODIS and carbon dioxide flux appears in Section 3, followed by relevant research that has been done by ecologists and statisticians (Section 4), and the display of the data we used in this proposal, including how flux data locations were selected and preprocessed (Section 5). We used two approaches to analyze our data. We describe our efforts to apply a class of additive models in Section 6.1, followed by sections on using functional data analysis. Section 6.2 shows how we apply spline smoothing. Functional registration is described in Section 6.3. Section 6.4 and Section 6.6 present how we apply multivariate normal and Dirichlet regression to further simulate CO_2 flux data at a specific location. To simulate the CO_2 flux data at a wider range, one commonly used spatial model is Kriging. A brief introduction of Kriging presents in Section 6.7. Section 7 gives a short conclusion, and summarizes the model we proposed. We present the future work in Section 8.



(a) Photo credit: NASA



(b) Photo credit: NASA

Figure 3: How MODIS works.

3 Background

In the Introduction we mentioned two data types, CO_2 flux and MODIS data. In this section we will discuss them in more detail. Section 3.1 will give a brief description of what NASA's Earth Observing System (EOS) is and the importance of having EOS. The relationship of EOS and Terra and Aqua will be shown. Section 3.2 will present basic information about different types of CO_2 flux together with eddy covariance which is currently the standard method used by ecologists to measure fluxes of trace gases between ecosystems and the atmosphere.

3.1 MODIS Data

According to NASA's EOS project science office `http://eospso.nasa.gov/` "EOS is a coordinated series of polar-orbiting and low inclination satellites for long-term global observations of the land surface, biosphere, solid Earth, atmosphere, and oceans". Figure 5 shows Terra and Aqua, two missions of EOS.

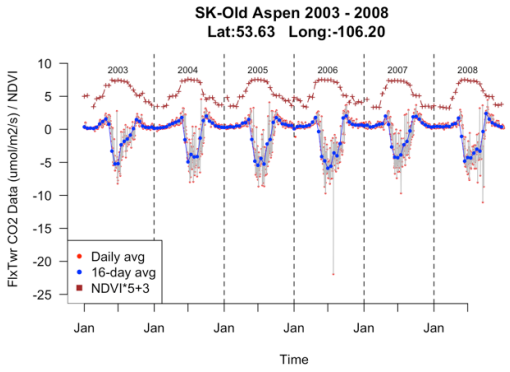
Table 1 is summarized from `http://science.nasa.gov/missions/terra/` and `http://atrain.nasa.gov/publications/Aqua.pdf`.

Table 1: Some details of Terra and Aqua satellites

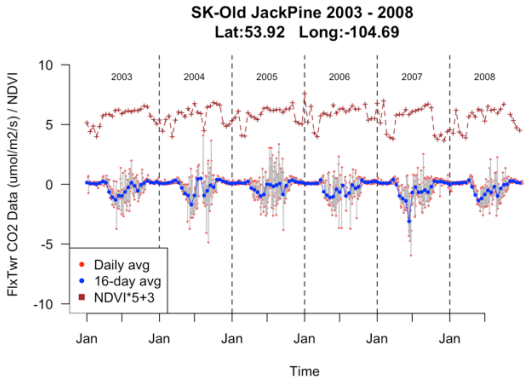
	Primary Mission	**Lauched Day**	**Passes equator**	
			Direction	**Time**
Terra	atmosphere, land, and oceans	Dec 18, 1999	North to South	Morning
Aqua	water, radiative energy flux, aerosols, etc.	May 4, 2002	South to North	Afternoon

Both satellites are in sun-synchronous orbits. The orbital plane rotates approximately one degree each day eastwards to keep pace with the Earth's movement around the sun. In this way, the satellite passes over any given point of the planet's surface at the same local solar time. The data can be transmitted directly from the spacecraft to ground stations equipped with an average 3m or larger X-band receiving system and appropriate hardware and software, though the raw data received from MODIS can not be used directly. It needs decoding, interpretation, scaling, and positioning. NASA processes the raw MODIS data in different forms and stages known as products. Table 2 shows definitions of MODIS products from level 0 to level 4.

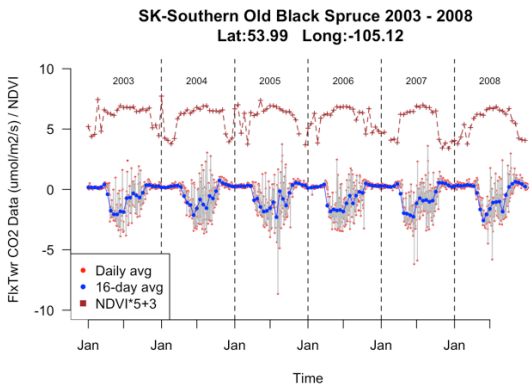
MODIS data used in this thesis was collected from the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) MODIS land product subsets.



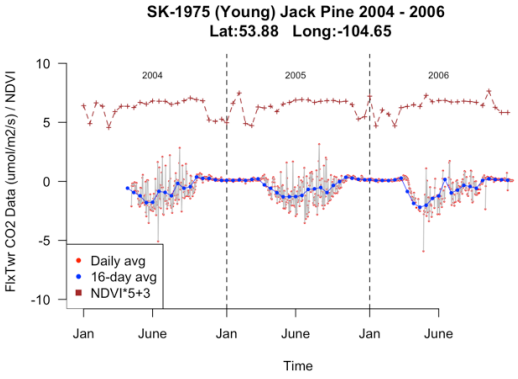
(a) Negative correlation between flux tower and MODIS data at SK-Old Aspen.



(b) Negative correlation between flux tower and MODIS data at SK-Old Jack Pine.



(c) Negative correlation between flux tower and MODIS data at SK-Southern Old Black Spruce.



(d) Negative correlation between flux tower and MODIS data at SK-Young Jack Pine.

Figure 4: Negative dependency between two measurements.

Table 2: Definitions of MODIS data processing levels.

Product Levels	Product Definitions
Level 0	Unprocessed instrument data.
Level 1A	Unprocessed instrument data alongside ancillary information.
Level 1B	Data processed to sensor units, e.g. brightness temperatures.
Level 2	Derived geophysical variables, e.g. sea ice concentration.
Level 3	Variables that are mapped on a grid, e.g. data using EASE-Grid.
Level 4	Modeled output or variables derived from multiple measurements.

3.2 Carbon Dioxide Flux Data

Flux measurements are widely used to estimate the exchange of heat, water, and carbon dioxide, as well as methane and other trace gases. CO_2 flux is dependent on the amount of CO_2 crossing an area, the size of the area being crossed, and the time it takes to cross this area.

Due to the complicated ecosystems of forests, different functional properties can be characterized by distinctive layers (**?**). Based on the CO_2 flux measuring height, there are 3 different types CO_2 flux data. They are Soil CO_2 flux, CO_2 flux below-canopy, and CO_2 flux above-canopy.

Soil CO_2 flux measures a physical process driven primarily by the CO_2 concentration diffusion gradient between the upper soil layers and the atmosphere near the soil surface. Figure 7 shows a device measuring soil CO_2 flux. Figure 1b is an example of how the below/above-canopy CO_2 flux gets measured.

The eddy covariance method is one of the most direct and defensible ways to measure such fluxes. The

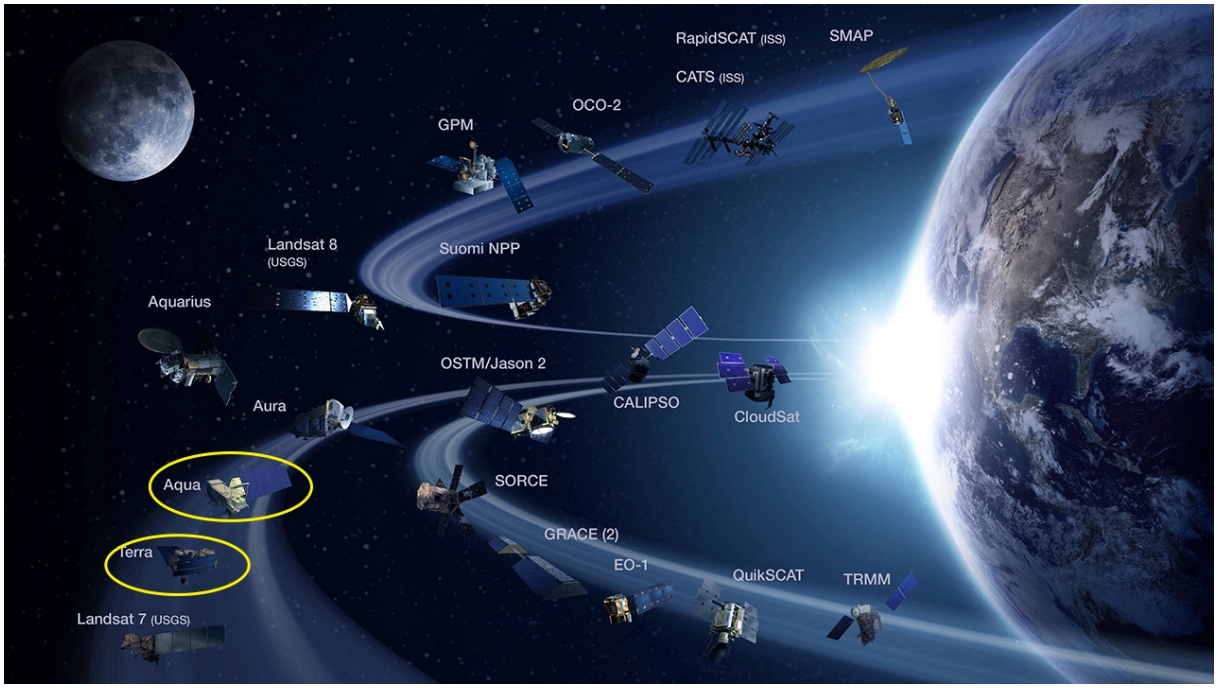


Figure 5: NASA Earth science division operating missions.

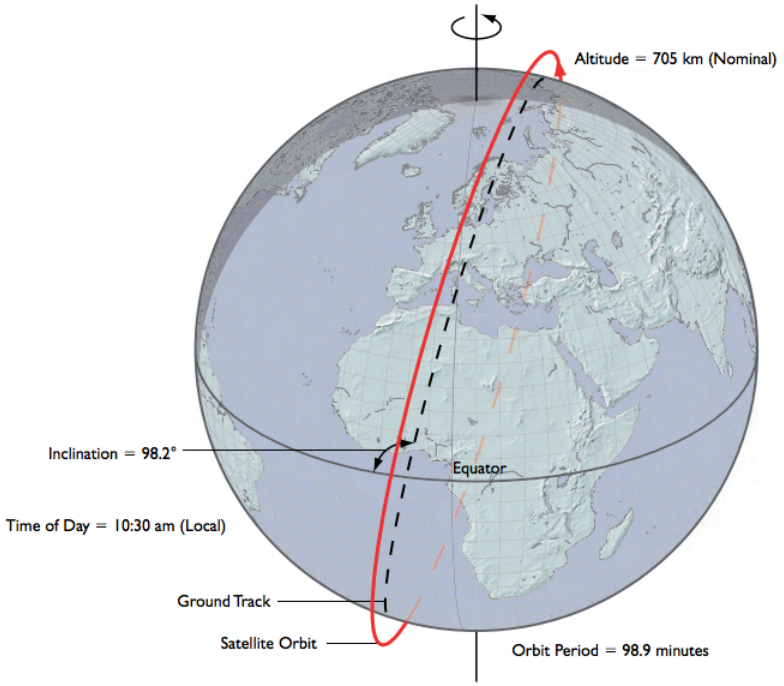


Figure 6: 3D model of Terra Satellite orbit.(**?**)

"covariance" in eddy covariance method is different from the definition of covariance in statistics. For more information, please see Appendix B. Uniform terminology and single methodology are still being developed for the eddy covariance method. Much of the effort is being done by network (e.g., FluxNet, ICOS, NEON, ect.) to unify various approaches. FLUXNET, a "network of regional networks", coordinates regional and global analysis of observations from micro meteorological tower sites. One of the conventional ways to

Figure 7: 6400-09 Soil CO_2 flux chamber.

calculate vertical fluxes is

$$F = \overline{\rho_a \omega s}, \tag{1}$$

where ω denotes the vertical wind speed, s represents the mixing ratio, the ρ_a means air density, and the bar over the product of the three parameters means average.

Flux tower data used in this thesis came from the Canadian Carbon Program (CCP). Figure 8 shows 8 years of CO_2 flux above-canopy data at Quebec City of harvested black spruce/jack pine. No obvious annual pattern showed up. The positive records of CO_2 flux occur during periods when the plants release more CO_2 than the amount they absorb, and records are negative when more CO_2 is absorbed by photosynthesis than is released by respiration.

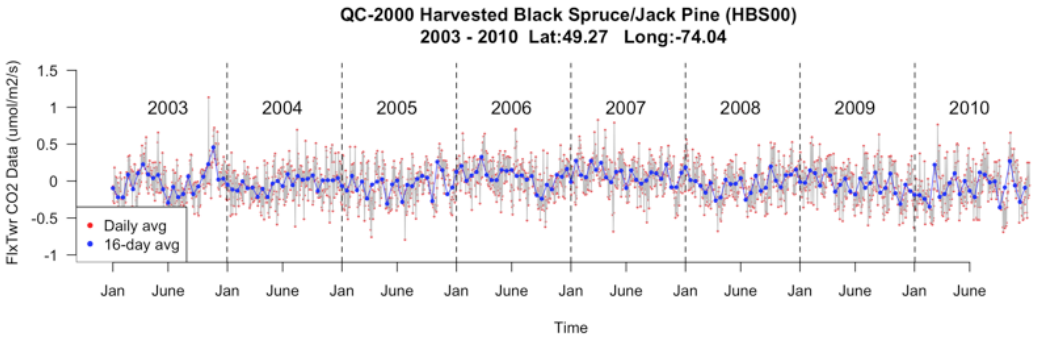


Figure 8: Quebec City harvested black spruce/jack pine CO_2 flux data from 2003 to 2010.

4 Literature Review

There are 44 MODIS products (MOD01 – MOD44) with different processing levels that can be used to study global climate change. The flux tower networks also collect ground-based data over the years. Many studies have been done that use either the MODIS data or the flux tower data, or use MODIS and flux tower data together.

4.1 Ecology

In ecology, many studies covered the following 3 topics.

1. Comparing the biomass calculated by MODIS and flux data separately. Biomass, such as regional evaporation and gross primary production (GPP) which is the amount of chemical energy produced in a given length of time are used in the studies. **?** and **?** described studies about calculating GPP from MODIS and flux data, and revealing the relationship of the estimations by linear regression. Similar analysis has been applied to evaporation (**?**).

2. Since the data collection of MODIS and flux tower are highly dependent on weather conditions, together with unexpected mechanical defects, gap filling techniques were developed. **?** showed comparisons of gap filling techniques for carbon flux data.

3. Using both MODIS data and flux data to study the phenology. **?** discussed using 3 different sigmoid functions to estimate phenology dates, and comparing the data driven outcomes with visual assessment. The challenges as summarized in **?** in the current study of phenology include lacking standard protocols of choosing proper biological events, lacking comparisons of current methods of transition dates estimation (e.g. the start of spring, the end of fall).

The studies using MODIS and flux data cover a variety of aspects. Besides the above three topics, machine learning techniques are used in the prediction of GPP (**?**) and evapotranspiration (**?**). A continental scale model was built by support vector machine (SVM).

4.2 Statistics

In statistics, functional data analysis is one of the appropriate tools to analyze the underlying patterns of our data. One assumption of FDA is that the underlying process generating the data is smooth. **?** mentioned "FDA was designed to take advantage of replications". "The primary advantage of FDA is that it allows the researcher to ask questions about when in a time series diffferences may exist between two or more sets of observations." (**?**)

A functional datum is not a single observation but a set of measurements over a continuum, usually time. A typical analysis of functional data begins with using smoothing techniques to represent each observation as a functional object. The original data are then set aside, and the smooth estimated curves are used for further analysis.

Application of FDA has appeared in a large number of publications across multiple fields, such as medicine, chemistry, managerial science, and ecology. Medical research including shape analysis and gene information extraction. **??** published their analysis of applying FDA in bone shape and hippocampal shape analysis by using registration tools and functional principle component analysis (fPCA). Besides shape analysis, **?** and **?** applied FDA in gene and genetic association studies. **?** mentioned functional data analysis in brain imaging studies. **?** introduced the application of FDA in chemical data. **?** and **?** gave application of FDA in business. **?** and **?** described the application of FDA in ecology.

?, **?**, and **?** not only used FDA in the analysis, but compared the results with traditional multivariate methods as well.

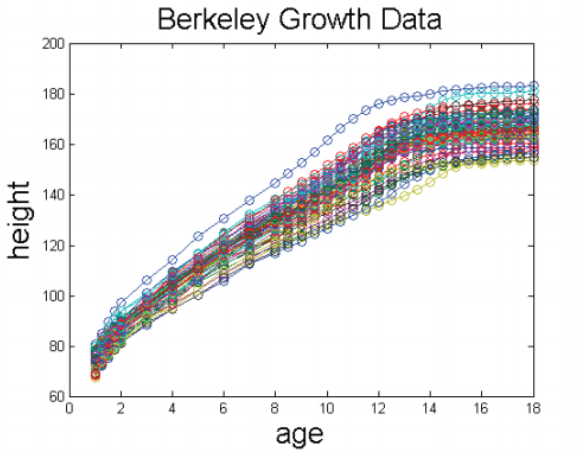


Figure 9: Functional data smoothing example. (**?**)

Figure 9 is an example of functional data smoothing. Heights of 20 girls were taken from ages 0 through 18 with unevenly spaced time points. The circles in Figure 9 represent the original data points. The smooth lines are the fitted curves for each girl. **?** also mentioned the first steps in a functional data analysis are smoothing and interpolation. After functional data smoothing, we can estimate the height of a given girl at

any age between 0 and 18. One of the advantages of functional data smoothing is that we can get continuous derivatives. The second derivatives of Figure 9 are shown in the left panel of Figure 10.

The individual differences of height acceleration curves makes the overall pattern not so obvious. It makes more sense that we compare "the pubertal growth spurts of two children at their respective ages of peak velocity rather than at any fixed age" (**?**). The feature alignment technique is called functional data registration. Figure 10 compares the original data with the data after registration.

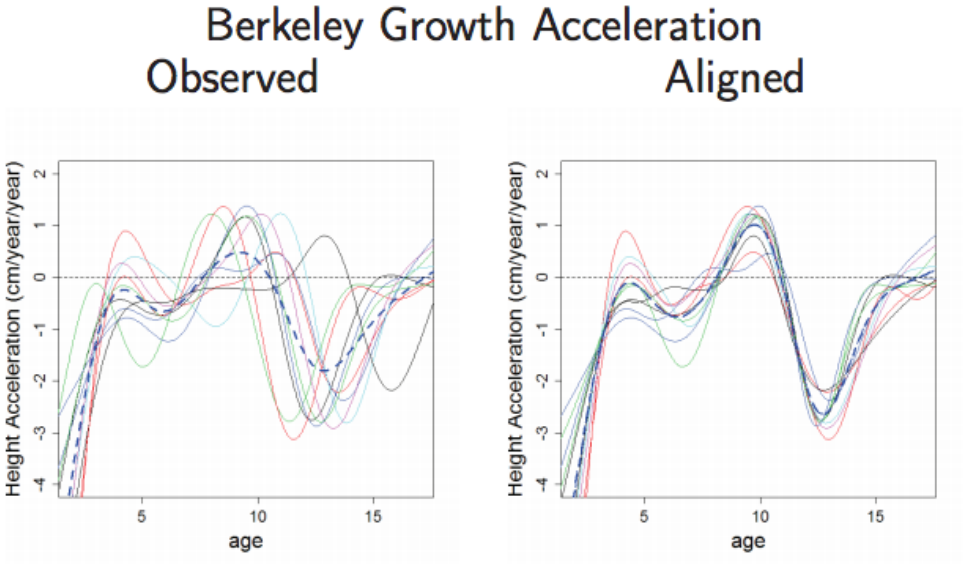


Figure 10: Functional data registration example. (**?**)

One way to conduct registration is landmark registration. A landmark or a feature of a curve is some characteristic that one can associate with a specific argument value. Landmarks are typically maxima, minima, or zero crossings of curves. They may be identified with zeros at the level of some derivatives.

The landmarks chosen in our study are bounded and ordered between 0 and 365, they are all positive and dependent with each other. The Dirichlet regression (**?**) turned out to be an appropriate tool to model their behaviour. Dirichlet regression always applies to compositional data (positive proportions sum up to one). Compositional data are used in chemistry and related fields. **?** applied Dirichlet regression to psychiatric data. **?** showed examples of how to use R to conduct Dirichlet regression. **?** introduced the application of Dirichlet regression in developmental rate isomorphy in ectotherms.

5 Data

The NDVI data we downloaded from ORNL DAAC MODIS land product subsets is level 3, 16-day average data. The quality of MODIS data is highly sensitive to the visibility of the day. The 16-day average data does not exactly take the average over 16 days of data, but may choose the most desirable record among the 16 day collections to represent it. The frequency of CO_2 flux above-canopy collected from CCP is every 30 minutes. For two different types of data, having the same frequency will be easier for later analysis. We took the daily average of CO_2 flux above-canopy data first. Then we took the 16 day average of the daily average of CO_2 flux above-canopy data. Since 365 is not a multiple of 16, the first 22 16-day average CO_2 flux above-canopy data were each calculated by 16 daily average CO_2 flux above-canopy data, while the last 16-day average data was calculated by only 13 daily average data. Figures 11 to 14 show the plots of daily average and 16-day average of CO_2 flux above-canopy data together for 4 different locations.

5.1 Location Selection

Flux towers give researchers the ability to watch as carbon dioxide concentrations vary between the earth and atmosphere, signalling the increase or decrease of the gas. There are 32 sites in the CCP. We wanted

our CO_2 flux above-canopy data to meet the following requirements.

1. Each site has at least 3 years high quality CO_2 flux above-canopy data showing an annual pattern.
2. Sites are geometrically close to each other.
3. Each site has only one species, but we want multiple species over the selected sites.

We checked the data at 32 CCP sites one by one.

Table 3: 18 sites that have CO_2 flux above-canopy data.

Site Name	Data Collection Timeframe
BC-Campbell River 1949 Douglas-fir	1997-2015
BC-Campbell River 1988 Douglas-fir	2001-2015
BC-Campbell River 2000 Douglas-fir	2000-2015
NB-Charlie Lake 1 1975 Balsam fir	2003-2005
ON-Borden Mixedwood	1995-2015
ON-Groundhog River Mixedwood	2003-2015
ON-Turkey Point 1939 White Pine	2003-2015
ON-Turkey Point 1974 White Pine	2003-2015
ON-Turkey Point 1989 White Pine	2003-2015
ON-Turkey Point Decidous	2012-2014
QC-2000 Harvested Black Spruce/Jack Pine (HBS00)	2001-2010
QC-Eastern Old Black Spruce (EOBS)	2003-2010
SK-1975 Young Jack Pine	2003-2015
SK-2002 Jack Pine	2003-2015
SK-1994 Jack Pine	2001-2015
SK-Old Aspen	1996-2015
SK-Old Jack Pine	1994-2015
SK-Southern Old Black Spruce	1994-2015

Table 3 shows that 18 out of 32 sites recorded CO_2 flux above-canopy data. The time frame listed in Table 3 is the data collecting time range of each site listed on CCP website. The actual time range of collecting CO_2 flux above-canopy data sometimes will be much shorter, and varies a lot from site to site.

Data collected at BC-Campbell River area has only one species, Douglas fir. NB-Carlie Lake area did not provide enough data. ON-Turkey Point White Pine has no exact canopy height information listed on line. Mixed wood has more than one species, we will consider this more complicated data later.

When all of these rules were applied, we ended up with 4 sites of data as shown in Table 4.

Table 4: CO_2 flux above-canopy data information.

Longitude	Latitude	Site Name	Time Range	Data Type
-106.1978	53.6289	SK Old Aspen	2003 - 2008	CO_2 Flux Above Canopy 39m
-104.6920	53.9163	SK Old Jack Pine	2003 - 2008	CO_2 Flux Above Canopy 28m
-105.1178	53.9872	Sk Southern Old Black Spruce	2003 - 2008	CO_2 Flux Above Canopy 25m
-104.6453	53.8758	Sk 1975 (Young) Jack Pine	2004 - 2006	CO_2 Flux Above Canopy 16m

5.2 Data View

The 16-day average data of SK Old Aspen (Figure 11) shows slightly increasing trend from January to April, slightly decreasing trend from October to December,and a sharp decreasing trend followed by a sharp increasing trend from April to October each year. There is one influential data point of the daily average in 2006.

The 16-day average data of SK Old Jack Pine (Figure 12) has narrower range of fluctuation, compared to SK Old Aspen. The underlying annual trend is not as obvious as SK Old Aspen. There is an interesting W shaped pattern in 2004.

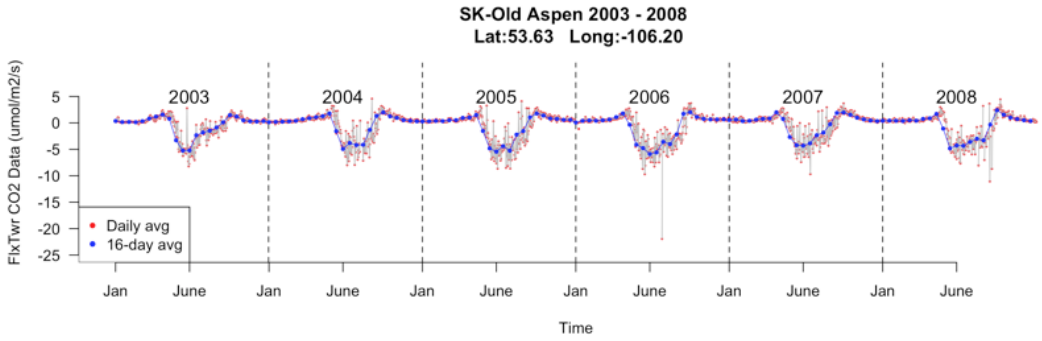


Figure 11: Saskatchewan old aspen CO_2 flux data from 2003 to 2008.

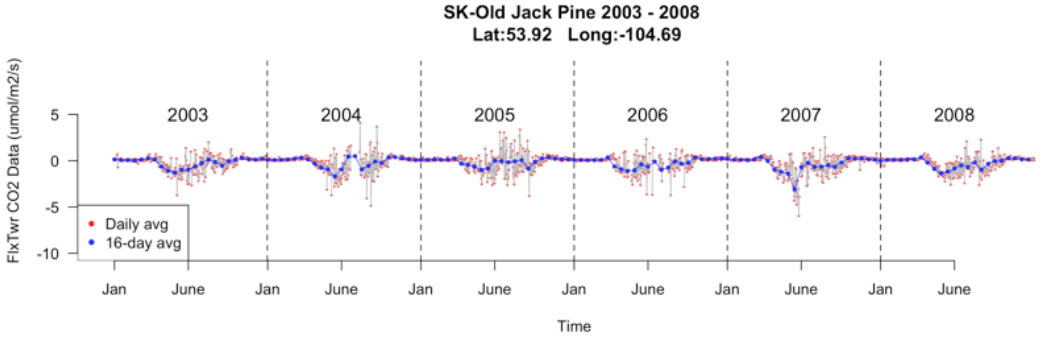


Figure 12: Saskatchewan old jack pine CO_2 flux data from 2003 to 2008.

The strength of the underlying trend of the 16-day average data of SK Southern Old Black Spruce (Figure 13) is between the SK Old Jackpine and SK Old Aspen.

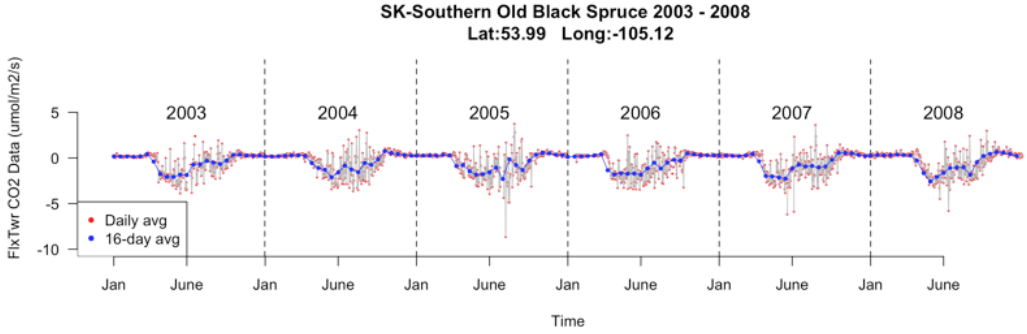


Figure 13: Saskatchewan southern old black spruce CO_2 flux data from 2003 to 2008.

SK 1975 (Young) Jack Pine site only has 3 years of CO_2 flux data (Figure 14).

The corresponding NDVI data for SK Old Aspen shows in Figure 15. The NDVI data for SK Old Jack Pine (Figure 16) has a lot more variation than other sites. The NDVI data for SK Southern Old Black Spruce (Figure 17) has stable performance in summer time. The NDVI of SK (Young) Jack Pine (Figure 18) has similar behaviour as SK Southern Old Black Spruce.

6 Methods

The class of additive models is a flexible and powerful tool. Section 6.1 describes how we applied a univariate additive model to 4 locations' CO_2 flux. Besides treating the CO_2 flux data as a single time

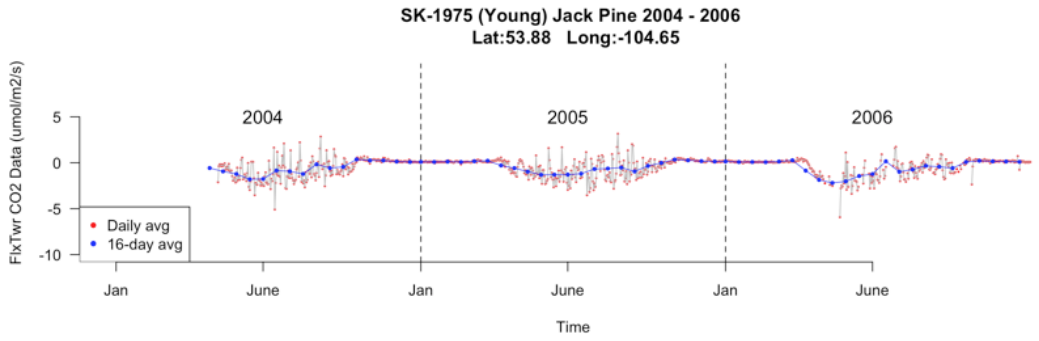


Figure 14: Saskatchewan young jack pine CO_2 flux data from 2004 to 2006.

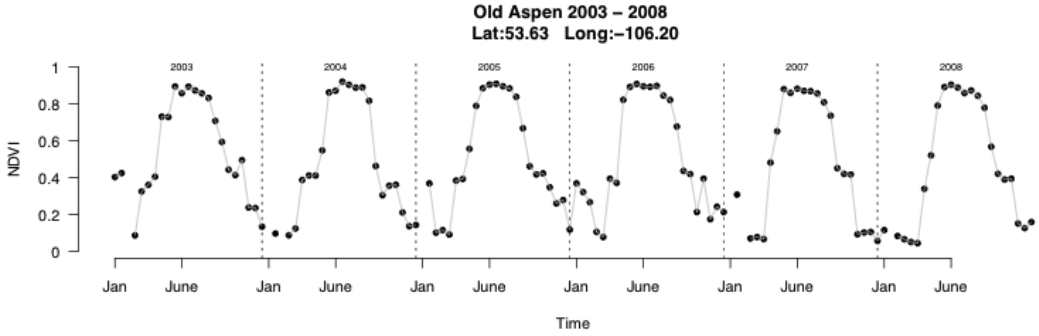


Figure 15: Saskatchewan old aspen NDVI data from 2003 to 2008.

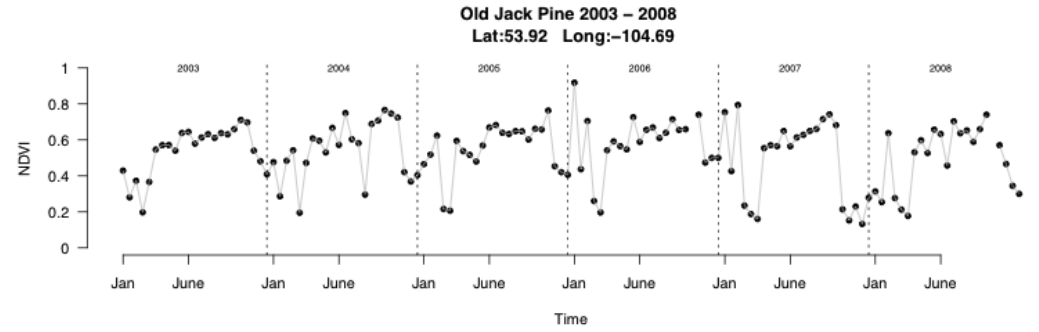


Figure 16: Saskatchewan old jack pine NDVI data from 2003 to 2008.

series, we can treat each year as independent multivariate observation. Since the shape of CO_2 flux data is similar at each location over the year, we use functional smoothing (Section 6.2) to capture the underlying pattern. Except for similarity in shape, the seasonal effect is quite different year to year. Thus we make use of the idea of registration (Section 6.3). As mentioned in Section 4, registration can help to reduce the individual difference, and better reveal the common features. With registration, simulations can be done with less individual variability. One important element for registration is the warping function (Section 6.3.1). The warping function is a bridge transformation between nature/real time scale and registered time scale. We used a multivariate normal (Section 6.4) to simulate new curves in the registered time scale. The next step is to transform the simulated curves back to real time scale by the inverse warping function for further study. To simulate random warping functions, the Dirichlet regression (Section 6.6) is applied, and simulated values are drawn from the fitted model.

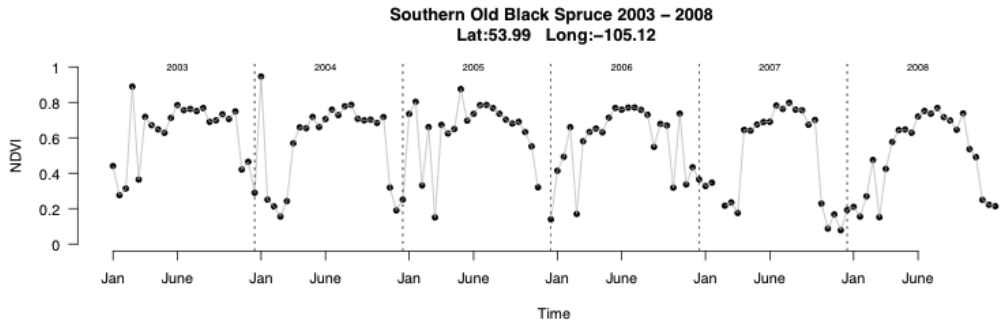


Figure 17: Saskatchewan southern old black spruce NDVI data from 2003 to 2008.

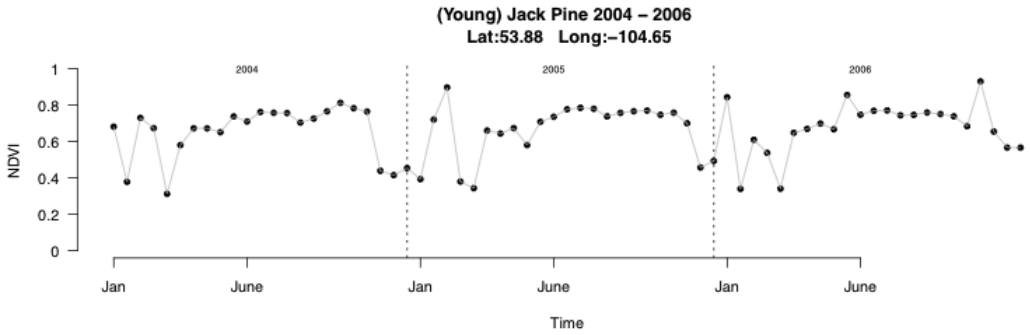


Figure 18: Saskatchewan young jack pine NDVI data from 2004 to 2006.

6.1 Additive Model

A generalized additive model (Hastie and Tibshirani, 1986, 1990) is a generalized linear model with a linear predictor involving a sum of smooth functions of covariates. It is known as GAM.

To fit a univariate additive model, flux tower data at each of the 4 locations was used as the dependent variable, while time (solar days), species of plants, and NDVI data are the independent variables used in the model. Each of the locations has been modelled in the following way.

$$\text{Flux} \sim f_1(\text{Time}) + \text{Species} + f_2(\text{Time,Species}) + \text{NDVI}, \tag{2}$$

where functions f_k $(k = 1, 2)$ are smooth functions of covariates; Time is every 16 day staring at the first day of each year; Species has 4 levels, old aspen, old jack pine, old black spruce, and young jack pine; (Time,Species) is the interaction between Time and Species; NDVI is a numeric vector. f_1(Time) is the average of CO_2 flux by the change of Time, and f_2(Time,Species) represents the departures for specific species. Throughout this section, we use f as a generic function, which will become clear in the context.

One downside of this approach is that the meaning of estimated coefficients is not straightforward. The residuals of the fitted model are shown in Figure 19. To better view the residuals, we plot the residuals at each location in Figure 20 with the corresponding QQ plot. Figure 20 shows that the residuals are not follow normal distribution, and summer has more variation than winter. Assuming the data over years has the same cycle is not appropriate. The phase of the underlying pattern is different. Other model fitting tools should be applied. We used spline smoothing (Section 6.2) and functional registration (Section 6.3) to solve this problem.

6.2 Spline Smoothing

A k^{th} order spline is a piecewise polynomial function of degree k, that is continuous and has continuous derivatives of orders $1, \ldots, k-1$, at its knot points.

Formally, a function $\phi : \mathbb{R} \to \mathbb{R}$ is a k^{th} order spline with knot points at $t_1 < \cdots < t_m$, if

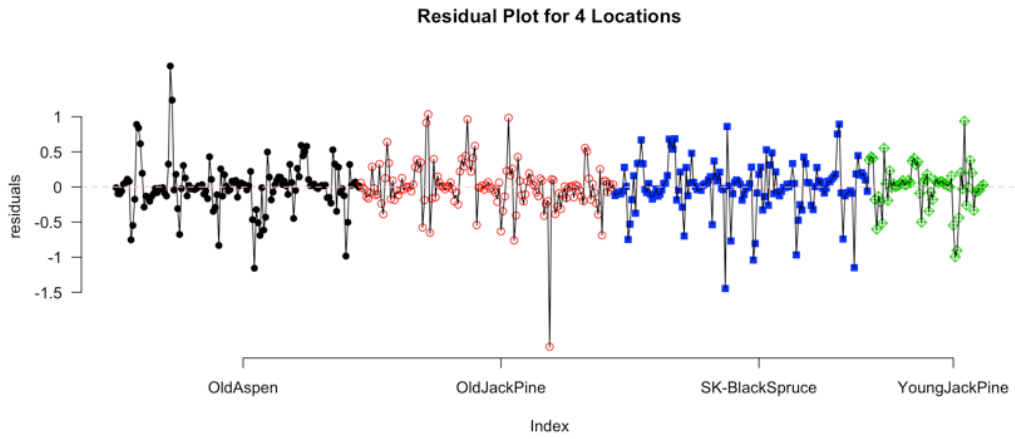


Figure 19: Residuals of additive model.

1. ϕ is a polynomial of degree k on each of the intervals $(-\infty, t_1], [t_1, t_2], \ldots, [t_m, \infty)$,
2. $\phi^{(j)}$, the j^{th} derivative of ϕ, is continuous at $t_1, \ldots, t_m$, for each $j = 0, 1, \ldots, k-1$.

The most common case considered is $k = 3$, i.e. cubic splines which have continuous first and second derivatives.

Our data have 23 data points (t_i, y_i), $i = 1, \ldots, 23$ each year. The regression model $f(t) = E(Y|T = t)$ can be estimated by fitting a k^{th} order spline with knots at some pre-specified locations $t_1, \ldots, t_m$. The means considering functions of the form $f(t) = \sum_{j=1}^{m+k+1} \beta_j \phi_j(t)$, where $\beta_1, \ldots, \beta_{m+k+1}$ are coefficients and $\phi_1, \ldots, \phi_{m+k+1}$ are the basis functions for k^{th} order splines over the knots $t_1, \ldots, t_m$. The coefficients $\beta_1, \ldots, \beta_{m+k+1}$ can be estimated by least squares, i.e. minimizing

$$\sum_{i=1}^{n} (y_i - f(t_i))^2, \tag{3}$$

and the fitted line is

$$\widehat{f}(t_i) = \sum_{j=1}^{m+k+1} \widehat{\beta}_j \phi_j(t_i). \tag{4}$$

The matrix form of spline regression is defined below:

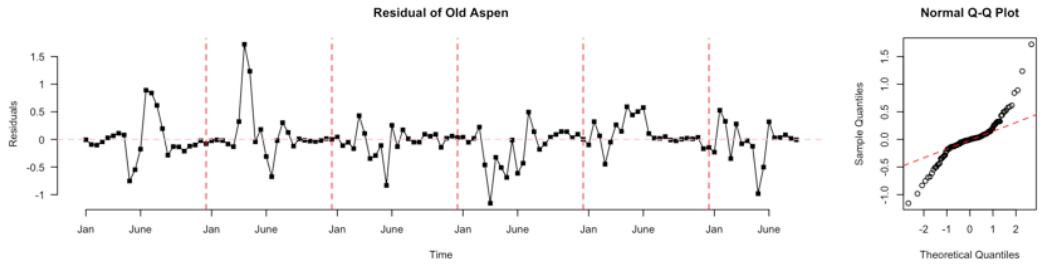
$$\underset{\sim}{y} = \Phi \underset{\sim}{\beta} + \epsilon, \tag{5}$$

where $\underset{\sim}{y} = (y_1, y_2, \ldots, y_n)^T$, and $\underset{\sim}{\beta} = (\beta_1, \beta_2, \ldots, \beta_{m+k+1})^T$, and

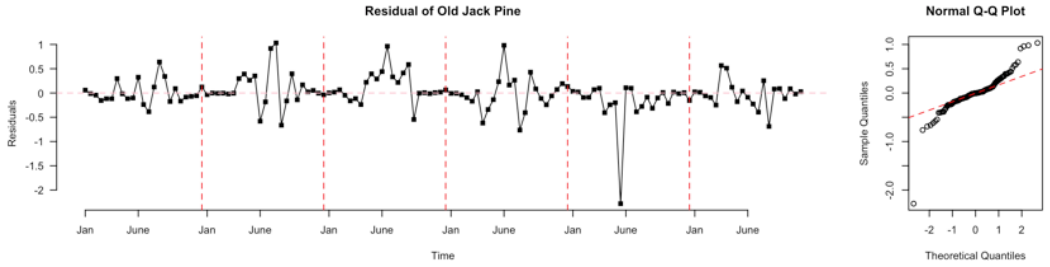
$$\Phi = \begin{pmatrix} \phi_1(t_1) & \phi_2(t_1) & \cdots & \phi_{m+k+1}(t_1) \\ \phi_1(t_2) & \phi_2(t_2) & \cdots & \phi_{m+k+1}(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(t_n) & \phi_2(t_n) & \cdots & \phi_{m+k+1}(t_n) \end{pmatrix}. \tag{6}$$

The estimated value for $\widehat{\underset{\sim}{\beta}}$ is $(\Phi^T \Phi)^{-1} \Phi^T \underset{\sim}{y}$.

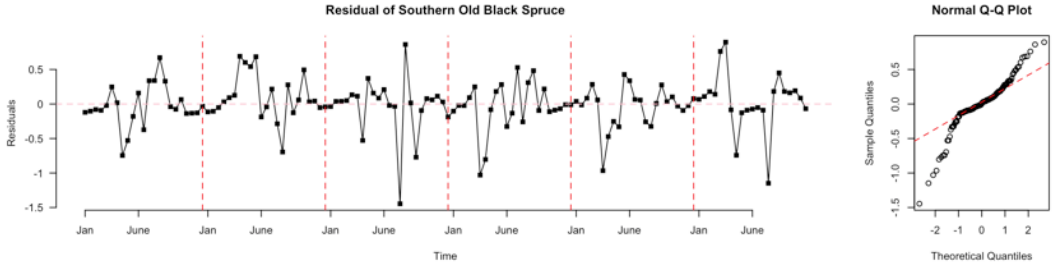
We fit the data year by year for each location. The knot points for old aspen data show in Table 1, 120, 130, 140, 170, 200, 220, 250, 260, 270, and 365 every year. The spline smoothing of old aspen data is shown in Figure 21.



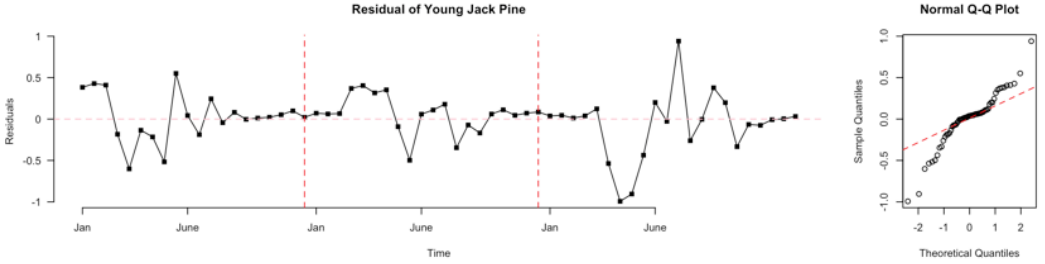
(a) Residuals of SK-Old Aspen location.



(b) Residuals of SK-Old Jack location.



(c) Residuals of SK-Southern Old Black Spruce location.



(d) Residuals of SK-Young Jack Pine location.

Figure 20: Residuals of each location of GAM and corresponding QQ plots.

Table 5: Knots used in old aspen data.

Day of year	1	120	130	140	170	200	220	250	260	270	365
Date	Jan 1	Apr 30	May 10	May 20	Jun 19	July 19	Aug 8	Sep 7	Sep 17	Sep 27	Dec 31

6.3 Registration

The additive model (Equation 2) assumes the underlying cycle is the same for every year, which may be inappropriate here. In the flux tower data of old aspen (Figure 21), the two peaks at around May and

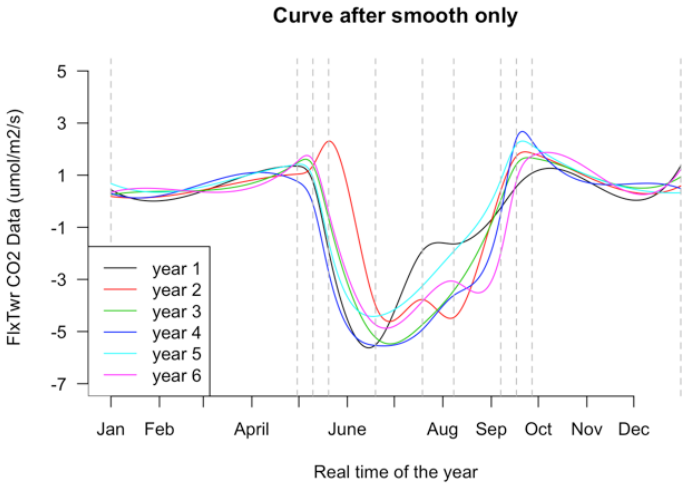


Figure 21: Spline smoothing of 6-year old aspen data. (The vertical dash lines indicate the location of knot points.)

October each year as well as the low point between the two peaks are not lined up. We can align curves by fixing the location of a feature, such as the summer minimum or winter maximum. A warping function is the connection of the original curve and the registered curve.

6.3.1 Warping Function

We have N years of data. For each year, we pick L number of landmarks. As mentioned in Section 4, landmarks $t_{p,v},\ p = 1, \ldots, N, v = 1, \ldots, L$ are typically maxima, minima, or zero crossings of curves. For 6-years old aspen data, we pick 3 time points, $t_{p,1}, t_{p,2}, t_{p,3}$, two at the peaks (around May and October), and one at the low point (near June) every year.

We want to construct a time transformation $h(t_{\text{regi}})$. For curve at year p time t_{regi} the transformation is $h_p(t_{\text{regi}})$ where $1 \leq t_{\text{regi}} \leq 365$ such that the registered curves f_p^* satisfy the following equations,

$$f_p^*(t_{\text{regi}}) = f_p(t_{\text{real}}), \quad \text{where } t_{\text{real}} = h_p(t_{\text{regi}}). \tag{7}$$

The time warping function has the properties: (i) $h_p(1) = 1$, (ii) $h_p(365) = 365$, (iii) between the adjacent landmarks, h_p is linear, (iv) we also require that h_p is strictly monotonic: $a < b$ implies that $h_p(a) < h_p(b)$.

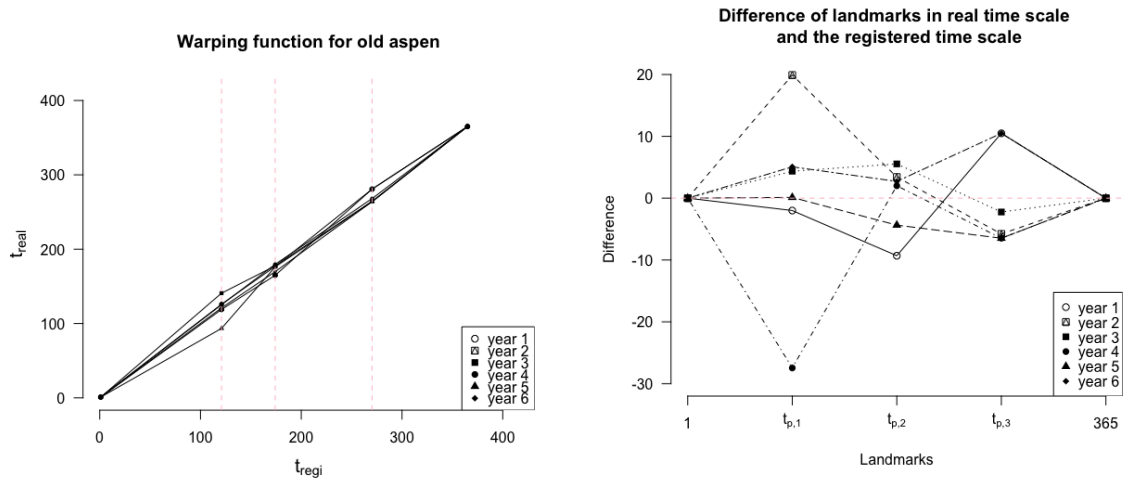
Figure 22a shows the warping function for the old aspen data. The x axis represents time on registered time scale, and the corresponding y axis is the time on real time scale. Over years of data, the first landmarks $t_{p,1},\ p = 1, \ldots, 6$ on real time scale has more variation than the other two. The 3 vertical dashed lines in Figure 22a show the landmarks, $t_v^*, v = 1, \ldots, 3$ on registered time scale. The value of $t_v^*, v = 1, \ldots, L$ is arbitrary, we calculate them by

$$t_v^* = \frac{1}{N} \sum_{p=1}^{N} t_{p,v}, \quad v = 1, \ldots, L. \tag{8}$$

Through the warping function, we know for a fixed year when time on registered time scale is given what's the corresponding time on real time scale should be. To better view the behaviour of landmarks, we plot the difference of landmarks in real time scale and the registered time scale in Figure 22b. For year p landmark v, we calculate the difference by

$$\text{Difference} = t_{p,i} - t_i^*. \tag{9}$$

By Equation 7, we get $f_p^*(t_{\text{regi}})$ every 16 days starting from the first day of each year. We apply spline smoothing (Section 6.2) to $f_p^*(t_{\text{regi}})$, where $t_{\text{regi}} = 1, 17, 33, \ldots, 353$ with the same basis when we fit the raw CO_2 flux data. The coefficients on the registered time scale are $\underset{\sim}{\widehat{\beta}}^* = (\widehat{\beta}_1^*, \widehat{\beta}_2^*, \ldots, \widehat{\beta}_{m+k+1}^*)^T$.



(a) SK-Old Aspen's warping function.

(b) Landmarks of each year minus the mean landmarks.

Figure 22: Old aspen's warping function and the difference of landmarks in real time scale and the registered time scale.

Curves in Figure 21 after registration are shown in Figure 23. After registration, the peaks and valleys are lined up. Compared to Figure 21, there is less variation among lines in Figure 23 at any given time of the year.

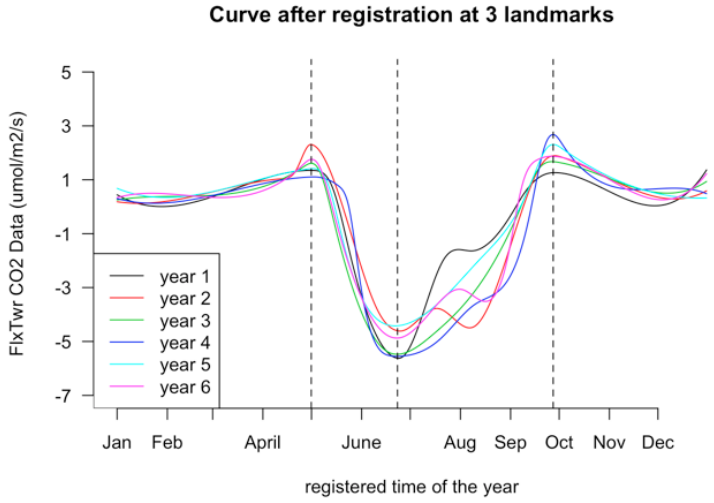


Figure 23: Registered curves of 6-year old aspen data.

In order to simulate random CO_2 flux processes, there are two steps: (i) random registered curve $f^*_{\texttt{sim}}(t_{\text{regi}})$, (ii) inverse warping function. In Section 6.4, we discuss how to generate random curves $f^*_{\texttt{sim}}(t_{\text{regi}})$. In Section 6.6, we present how to produce inverse warping function.

6.4 A Model of the Registered Curves: Multivariate Normal

We take the annual data, filtered to produce $\widehat{\beta}^*$ (Section 6.3.1). We now are looking for a distribution to model $\widehat{\beta}^*$. Simulated random year CO_2 flux data on the registered time scale can be expressed by

$$f^*_{\texttt{sim}}(t_{\text{regi}}) = \sum \beta^*_{\texttt{sim}} \phi(t_{\text{regi}}). \tag{10}$$

We can model $\beta^*_{\texttt{sim}}$ based on $\widehat{\beta}^*$ using a multivariate normal distribution. Then through Equation 10 we can get the simulated curves $f^*_{\texttt{sim}}$ on registered time scale. To describe a single curve we need $m+k+1$ coefficients, a multivariate normal distribution will be used to simulate the $m+k+1$ coefficients.

6.4.1 Non-degenerate Case

When the covariance matrix of $\widehat{\beta}^*$ is positive definite, the distribution of multivariate normal (MVN) is given below:

$$f_X(x_1,\ldots,x_q) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right), \tag{11}$$

where X is a real q-dimensional column vector, μ is a 1 by q vector which is the mean of each column vector of X, $|\Sigma|$ is a q by q covariance matrix of X, and $|\Sigma|$ is the determinant of Σ.

6.4.2 Degenerate Case

When the covariance matrix is not full rank, $m+k+1$ by $m+k+1$ symmetric matrix Σ of rank $d < m+k+1$. Σ has d positive eigenvalues ($\lambda_1 > \lambda_2 > \cdots > \lambda_d > \lambda_{d+1} = 0, \ldots, \lambda_{m+k+1} = 0$), and the Σ^{-1} is undefined. By eigen decomposition, we have

$$\Sigma = EDE', \tag{12}$$

where E is $m+k+1$ by d matrix, and has columns which are the eigenvectors of the positive eigenvalues of the d by d diagonal matrix $D = diag(\lambda_1, \lambda_2, \ldots, \lambda_d)$. The generalized inverse of Σ is defined as

$$\Sigma^- = ED^{-1}E'. \tag{13}$$

The p.d.f. of the singular MVN is defined over a lower dimensional subspace

$$f_X(x_1,\ldots,x_q) = ((2\pi)^d|D|)^{-\frac{1}{2}} \exp\left\{-\frac{(x-\mu)'\Sigma^-(x-\mu)}{2}\right\}. \tag{14}$$

6.5 One Simulation Method

When covariance matrix $|\Sigma| = 0$, the inverse of Σ is not defined. To simulate MVN under this condition, we need Cholesky decomposition and linear transformation. By Cholesky decomposition, there exists a lower triangular $m+k+1$ by $m+k+1$ matrix L, such that

$$\Sigma = LL'. \tag{15}$$

Each column of X is i.i.d. univariate standard normal. The number of columns of X depends on the number of simulations we need. The simulated value Y is

$$Y = \mu + LX, \tag{16}$$

where $Var(YY') = Var(LXX'L') = Var(LL')$, since XX' is identity matrix.

We apply R function mvrnorm(n=1,mu,Sigma) to conduct the simulation. We use the sample mean and the sample covariance to estimate μ and Σ. We simulate 10000 CO_2 flux data on the registered time scale. Figure 24 shows 5 simulations and 95% point-wise prediction intervals.

6.6 A Model for the Inverse Warping Function: Dirichlet Regression

We generate $\beta^*_{\texttt{sim}}$ by MVN, which can be used to further produce $f^*_{\texttt{sim}}(t)$, the random year CO_2 flux data on the registered time scale. Then we need to map the the simulated curve from the registered time scale back to the real time scale. Since we need to build a connection of registered and real time scale, we take advantage of the idea of warping function (Section 6.3.1). Compared to the previous registration which gives the landmarks on real time scale then we need to define the landmarks on registered time scale, now we know the landmarks on registered time scale, it is important to set up the landmarks t_v^{sim}, $i = 1, \ldots, L$ on real time scale for each simulated curve.

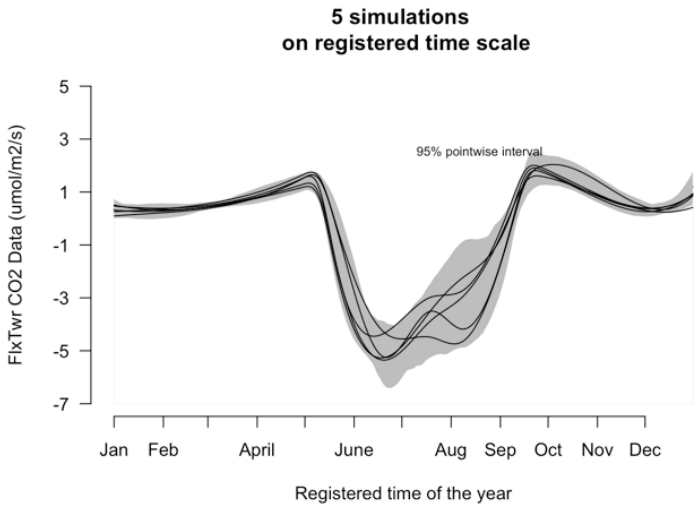
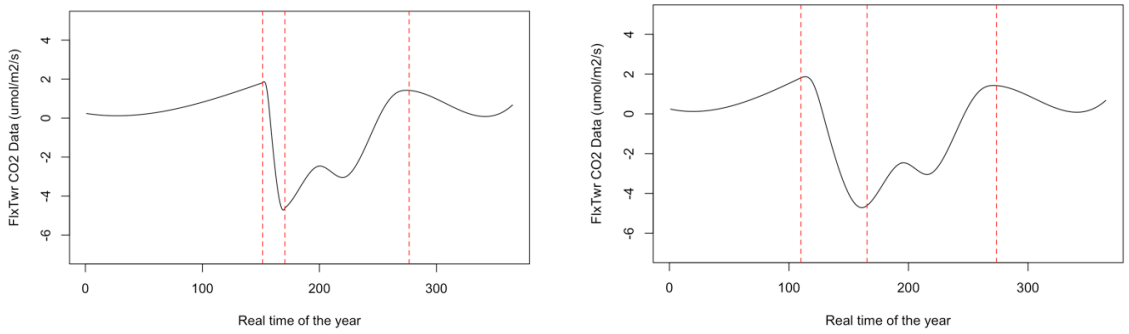


Figure 24: 5 simulated curves of old aspen data on the registered time scale.

We use two different methods to simulate the landmarks on real time scale. The first model we continue using MVN. It has the following problems,

1. We can't guarantee $t_1^{sim} < \cdots < t_L^{sim}$.
2. Sometimes t_v^{sim} is very close to t_{v+1}^{sim}.
3. We can't guarantee $1 \leq t_v^{sim} \leq 365$.

After 10000 simulation, over 95% of the times that all the landmarks are within [0,365]. We could discard the out of range landmarks in practice. As for the order problem, we can sort the simulated landmarks. But for the distance between landmarks, there is no easy way to solve this problem. Figure 25a is an example of two landmarks generated by MVN that look too "close" to each other. Model other than MVN should be considered at this point. Later in the section, we will introduce Dirichlet distribution and Dirichlet regression to overcome the problem.



(a) Landmarks simulated by multivariate normal.

(b) Landmarks simulated by Dirichlet regression.

Figure 25: Landmarks generated by two methods.

L landmarks of year p data divide $[0, 365]$ to $L+1$ intervals, $[0, t_{p,1}]$, $[t_{p,1}, t_{p,2}]$, $\ldots, [t_{p,L-1}, t_{p,L}]$. The length of each interval are $l_1 = t_{p,1}$, $l_2 = t_{p,2} - t_{p,1}$, $\ldots$, $l_{L+1} = 365 - t_{p,L}$, and satisfy following property:

1. $l_1, \ldots, l_{L+1} > 0$
2. $\sum_{i=1}^{L+1} l_i = 365$

Thought it is not easy to directly simulate the landmarks at the same time, there is a distribution to describe the behaviour of the intervals. It is Dirichlet distribution.

6.6.1 Dirichlet Distribution

The Dirichlet distribution of order $k \geq 2$ with parameters $\alpha_1, \ldots, \alpha_k > 0$ has a probability density function with respect to Lebesgue measure on the Euclidean space R^{k-1} given by

$$f(x_1, \ldots, x_k; \alpha_1, \ldots, \alpha_k) = \frac{1}{B(\underset{\sim}{\alpha})} \prod_{i=1}^{k} x_i^{\alpha_i - 1}, \tag{17}$$

on the open $(k-1)$ dimensional simplex defined by $x_1, \ldots, x_{k-1} > 0$, $x_1 + \cdots + x_{k-1} < 1$, $x_k = 1 - x_1 - \cdots - x_{k-1}$, and zero elsewhere.

The normalizing constant is the multivariate Beta function, which can be expressed in terms of the gamma function:

$$B(\underset{\sim}{\alpha}) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)}, \quad \underset{\sim}{\alpha} = (\alpha_1, \ldots, \alpha_k). \tag{18}$$

Compared to the properties that our $l_1, \ldots, l_{L+1}$ have, if we rescale the sum of $l_1, \ldots, l_{L+1}$ to 1, then the rescaled intervals exactly follow the Dirichlet distribution.

6.6.2 Dirichlet Regression

We assume that the i^{th} year length of adjacent landmarks $y_{i.}$ where $i = 1, \ldots, n$ follows a Dirichlet distribution with parameter $\alpha(x_{i.})$, where $\alpha(x_{i.}) = (\alpha_1(x_{i.}), \ldots, \alpha_k(x_{i.}))$, and each $\alpha_l(x_{i.})$ where $l = 1, \ldots, k$ is a linear combination of $x_{i.}$:

$$\alpha_l(x_{i.}) = x_{i,1}\psi_{1,l} + x_{i,2}\psi_{2,l} + \cdots + x_{i,p}\psi_{p,l} = x_{i.}\psi_{.l}. \tag{19}$$

The parameters to be estimated are $\psi = (\psi_{c,l}, c = 1, \ldots, p, l = 1, \ldots, k)$, subject to the constraint $\alpha(x_{i.}) > 0$.

Let $x_{i.}$ be the given covariates for year i where $i = 1, \ldots, n$, and consider the response variable $y_i = (y_{i,1}, \ldots, y_{i,k})$ to be a positive vector with a conditional Dirichlet distribution, $y_i | x_{i.} \sim \mathcal{D}(\alpha_1(x.), \ldots, \alpha_k(x.))$. Assuming $y_{1.}, \ldots, y_{n.}$ are i.i.d. Given ψ, the likelihood function is

$$l(\psi) = \prod_{i=1}^{n} \left[\Gamma(A_i(x_{i.})) \prod_{l=1}^{k} \frac{y_{il}^{\alpha_l(x_{i.})-1}}{\Gamma(\alpha_l(x_{i.}))} \right], \tag{20}$$

where $A_i(x_{i.}) = \sum_{l=1}^{k} \alpha_l(x_{i.})$.

The gradients of log-likelihood is

$$\frac{\partial \log l(\psi)}{\partial \psi_{c,l}} = \sum_{i=1}^{n} \left\{ [\Psi(A_i(x_{i.})) - \Psi(\alpha_l(x_{i.})) + \log y_{i,l}] x_{i,c} \right\}, \tag{21}$$

where $\Psi(u) = \frac{\partial log \Gamma(u)}{\partial u}$.

Numerical methods are required to compute the maximum likelihood estimate (MLE). **?** discussed, and gave numerical solution of Dirichlet regression.

Starting values and regularization policies must be carefully chosen for the optimization algorithm to converge. **?** proposed a method for choosing starting values.

With the simulated curves f^*_{sim} by MVN on registered time scale and the simulated landmarks by Dirichlet regression on real time scale, then we have

$$f^{sim}(t_{\text{real}}) = f^*_{sim}(t_{\text{regi}}), \text{ where } t_{\text{real}} = h^{-1}(t_{\text{regi}}). \tag{22}$$

We have t_i^* on registered time scale, and t_i^{sim} simulated by Dirichlet Regression to construct the inverse warping function. Compared to equation 7, model in equation 22, transforming simulated curves from registered time scale to real time scale, uses the inverse warping function. Figure 26 shows 10 simulated curves of old aspen data on real time scale.

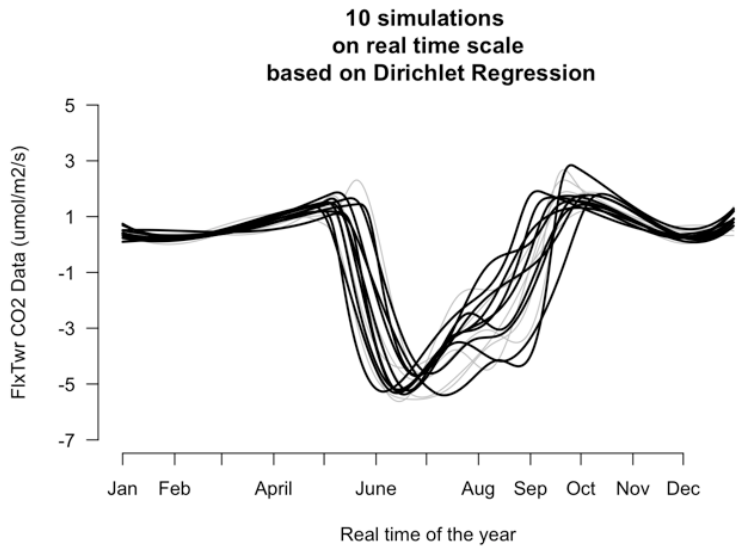


Figure 26: 10 simulated curves of old aspen data on real time scale.

6.7 Kriging

We show how to generate random curves for one selected location in previous sections. Currently we collect data from 4 different locations. As more and more data from different locations are going to be collected, we are interested in building a model to connect all the locations together. Kriging uses an interpolation method to connect different locations. In this section we show a brief introduction of Kriging.

A process of interest is observed at M locations $S_1, \ldots, S_M$. These n observations are denoted by $Z(S_1), \ldots, Z(S_M)$. If the process $Z(S)$ satisfies the following two equation,

$$E(Z(S)) = \mu \tag{23}$$

$$Var(Z(S+h) - Z(S)) = 2\gamma(h), \ \ \forall h \tag{24}$$

we say process $Z(S)$ is intrinsic stationary (IS). The difference between IS and second order stationary is presented in Appendix E.

We start by assuming the underlying mean function has the simplest structure which is constant. More complicated case will be studied later. Under our mean constant assumption, the process can be expressed as

$$Z(S) = \mu + \delta(S), \tag{25}$$

where μ is the constant, and $\delta(S)$ is intrinsically stationary.

Let S_0 denote an unsampled location. Our goal is to find a good estimation of $Z(S_0)$, $\widehat{Z}(S_0)$, based on $Z(S_1), \ldots, Z(S_M)$ (equation 26).

$$\widehat{Z}(S_0) = E(Z(S_0)|Z(S_1), \ldots, Z(S_n)) \tag{26}$$

The "good" means $\widehat{Z}(S_0)$ minimizes $MSE(\widehat{Z}(S_0))$.

$$MSE(\widehat{Z}(S_0)) = E((\widehat{Z}(S_0) - Z(S_0))^2) \tag{27}$$

Since the conditional expectation in equation 26 is not easy to calculate, one modest way to construct $\widehat{Z}(S_0)$ is the linear function of the observed value (equation 28).

$$\widehat{Z}(S_0) = \sum_{i=1}^{M} \lambda_i Z(S_i), \text{ where } \sum \lambda_i = 1 \tag{28}$$

Then problem is transferred to be

$$\text{minimize } E\left\{\left[\sum \lambda_i Z(S_i) - Z(S_0)\right]\right\}^2, \text{ subject to} \sum \lambda_i = 1. \tag{29}$$

After calculation and simplification (details shown in Appendix D), we get

$$\left[\sum \lambda_i Z(S_i) - Z(S_0)\right]^2 = \sum \lambda_i (Z(S_i) - Z(S_0))^2 - \frac{1}{2}\sum\sum \lambda_i \lambda_j (Z(S_i) - Z(S_0))^2. \tag{30}$$

The variogram function $2\gamma(h)$ is defined as

$$2\gamma(h) = E((Z(S+h) - Z(S))^2), \tag{31}$$

where $\gamma(h)$ is called the semi-variogram.

We take expectation on both side of equation 30, and combine the variogram function (equation 31), equation 27 can be expressed as

$$MSE(\widehat{Z}(S_0)) = E((Z(S_0) - \widehat{Z}(S_0))^2) = 2\sum \lambda_i \gamma(S_0 - S_i) - \sum\sum \lambda_i \lambda_j \gamma(S_i - S_j) \tag{32}$$

We use Lagrange multiplier to minimize $MSE(\widehat{Z}(S_0))$ (equation 32).

$$L(\lambda_1, \ldots, \lambda_n) = 2\sum \lambda_i \gamma(S_0 - S_i) - \sum\sum \lambda_i \lambda_j \gamma(S_i - S_j) - \theta\left(\sum \lambda_i - 1\right) \tag{33}$$

Differentiating L (equation 33) with respect to λ_i for $i = 1, \ldots, M$, and θ, we get

$$\frac{\partial L}{\partial \lambda_i} = 2\gamma(S_0 - S_i) - \sum \lambda_j \gamma(S_i - S_j) - \theta, \tag{34}$$

and

$$\frac{\partial L}{\partial \theta} = -\sum \lambda_i + 1. \tag{35}$$

We set all $M+1$ derivatives equal to 0. There are $M+1$ equations and $M+1$ unknown parameters ($\lambda_i, i = 1, \ldots, M$ and m).

$$\sum \lambda_j \gamma(S_i - S_j) + \frac{\theta}{2} = \gamma(S_0 - S_i), \quad i = 1, \ldots, M \tag{36}$$

and

$$\sum \lambda_i = 1. \tag{37}$$

If we define $\gamma_{ij} = \gamma(S_i - S_j)$, the matrix representation of the linear system shows below

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1M} & 1 \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2M} & 1 \\ \vdots & & & & \vdots \\ \gamma_{M1} & \gamma_{M2} & \cdots & \gamma_{MM} & 1 \\ 1 & 1 & \ldots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_M \\ \frac{\theta}{2} \end{bmatrix} = \begin{bmatrix} \gamma_{01} \\ \gamma_{02} \\ \vdots \\ \gamma_{0M} \\ 1 \end{bmatrix}. \tag{38}$$

We can get $\widehat{\lambda}_i,\ i = 1, \ldots, n$, by solving equation 38.

7 Conclusion

GAM is one way to model the CO_2 flux tower data, the detailed modelling process shows in Appendix C. GAM has the following problems.

- The residuals neither have homogeneous variance nor are normally distributed.
- The model interpretation is not straightforward.
- The behaviour of seasonal dynamics, such as the start of spring and the end of fall, remains unclear.

The steps of the other method to model the CO_2 flux tower data is shown in Table 6. The step 1 has already been used in multiple fields of researches. For those researches, after registration, the next step usually is using fPCA or other dimension reduction tools to find the main source of variation. In our case, to understand how the annual CO_2 flux data behaves on real time scale is the most important thing, because on registered time scale there isn't much information of phenology, such as when spring starts and when fall ends. We propose Step 2 and Step 3 to achieve our goal. We name Step 3 the FDA registration model. To capture the behaviour of a whole curve is not an easy task. With FDA registration model other steps, for a specific location, we can picture the behaviour of CO_2 flux data.

Table 6: Modelling steps

Steps		**Warping**	**Landmarks**
1	$f_p^*(t_{\text{regi}}) = f_p(t_{\text{real}})$	$t_{\text{real}} = h_p(t_{\text{regi}})$	$t_i^* = \frac{1}{N}\sum_{p=1}^{N} t_{p,i}$
2	$f_p^*(t_{\text{regi}}) \xrightarrow{MVN} f_{sim}^*(t_{\text{regi}})$		
3	$f^{sim}(t_{\text{real}}) = f_{sim}^*(t_{\text{regi}})$	$t_{\text{real}} = h^{-1}(t_{\text{regi}})$	$t_i^{sim} \sim$ Dirichlet Regression

8 Future Work

The study performed in this proposal provides basis for future research in several aspects. These aspects include:

1. Since the coefficients estimated by OLS follow MVN distribution, we suspect that $\widehat{\beta}^*$ follow asymptotic MVN distribution. In Section 6.4, we used MVN without further illustration to generate coefficients β_{sim}^* from $\widehat{\beta}^*$ under the registered time scale. We are going to detect the distribution of $\widehat{\beta}^*$, and gather more statistical support to make our steps more convincing.

2. In Section 6.6, we applied FDA registration model (equation 22) to the simulated curves under registered time scale. We plan to study the distribution of our predictions, and make inference including giving specific expression of the overall prediction band. Besides, we will apply goodness of fit test to check the compatibility of the sample with our proposed probability distribution function.

3. Spatio-temporal aspect should also be studied. With respect to CO_2 flux data, we will consider embedding MVN and FDA registration with additional information, such as latitude, longitude, and species. The GLM structure incorporates well with all the information. Another spatio-temporal model commonly used is Kriging.

4. We will also analyze the structure of NDVI data, and link it to CO_2 flux data in order to predict CO_2 flux data at location without towers.

5. We will apply the whole analysis to a wider range of CO_2 flux data. AmeriFlux networks offers a great planform. We are going to collect CO_2 flux data from the sites in Figure 27. More information of those sites can be found at the official website - AmeriFlux Site List.

Figure 27: All Ameriflux sites.

Appendix A Selected Notations

Table 7: Selected notations.

Symbols	Meaning
F	vertical fluxes
ρ_a	air density
ω	vertical wind speed
s	mixing ratio
N	number of years of data
n	number of observations each year
m	number of knot points
t_i	interior knot points $i = 1, 2, \ldots, m$
ϕ	spline with degree k
$\phi^{(j)}$	the j^{th} derivative of ϕ
L	length of landmarks
β $(\widehat{\beta})$	(fitted) coefficient of raw data on real time scale
β^* $(\widehat{\beta}^*)$	(fitted) coefficient of registered smooth data on registered time scale
β^*_{sim} $(\widehat{\beta}^*_{sim})$	(fitted) coefficient of simulated registered data on registered time scale
$f_p(t_{\text{real}})$	smooth curve of year p raw data
$f^*_p(t_{\text{regi}})$	registered smooth curve of year p
$f^*_{sim}(t_{\text{regi}})$	simulated registered curve
$f^{sim}(t_{\text{real}})$	simulated curve on real time scale
$t_{p,v}$	the vth landmark of year p smoothed curve
t^*_v	the vth landmark of registered curve
t^{sim}_v	the vth simulated landmark on real time scale
M	number of locations
S_i	location i, $i = 1, \ldots, M$
$Z(S)$	process at location S which satisfies intrinsic stationary

Appendix B Eddy covariance method to calculate vertical flux.

In turbulent flow, $\bar{\rho}_a$ is the mean air density, $\bar{\rho}'_a$ is the deviation of ρ_a from the mean air density $\bar{\rho}_a$, $\bar{\omega}$ represents the mean wind speed, ω' is the deviation in vertical wind speed ω from its mean value $\bar{\omega}$, and $\bar{s}$ is the mean value of the gas concentration, s' is the deviation of gas concentration from its mean $\bar{s}$.

Vertical flux can be presented as:

$$F = \overline{\rho_a \omega s} \tag{39}$$

By Reynolds decomposition that $\rho_a = \bar{\rho_a} + \rho'_a$, where $\rho'_a = \rho_a - \bar{\rho_a}$, $\omega = \bar{\omega} + \omega'$, where $\omega' = \omega - \bar{\omega}$, and $s = \bar{s} + s'$, where $s' = s - \bar{s}$.

$$F = \overline{(\bar{\rho}_a + \rho'_a)(\bar{\omega} + \omega')(\bar{s} + s')} \tag{40}$$

$$F = \overline{(\bar{\rho}_a\bar{\omega}\bar{s} + \bar{\rho}_a\bar{\omega}s' + \bar{\rho}_a\omega'\bar{s} + \bar{\rho}_a\omega's' + \rho'_a\bar{\omega}\bar{s} + \rho'_a\bar{\omega}s' + \rho'_a\omega'\bar{s} + \rho'_a\omega's')} \tag{41}$$

Since the averaged deviation is zero, i.e. $\bar{\rho}'_a = \bar{\omega}' = \bar{s}' = 0$ the above equation is simplified as:

$$F = \overline{(\bar{\rho}_a\bar{\omega}\bar{s} + \bar{\rho}_a\omega's' + \rho'_a\bar{\omega}s' + \rho'_a\omega'\bar{s} + \rho'_a\omega's')} \tag{42}$$

Then important assumption is make for conventional eddy covariance, which is density fluctuations are assumed negligible, i.e. $\rho'_a = 0$:

$$F = \overline{(\bar{\rho}_a\bar{\omega}\bar{s} + \bar{\rho}_a\omega's')} \tag{43}$$

Then another important assumption is made - mean vertical flow is assumed negligible for horizontal homogeneous terrain, i.e.$\bar{\omega} = 0$:

$$F = \overline{(\bar{\rho}_a\omega's')} = \bar{\rho}_a\overline{\omega's'} \tag{44}$$

In statistical language:

3 random variables ρ_a, ω and s, where $\rho_a = E(\rho_a) + \epsilon_{\rho_a}$, $\omega = E(\omega) + \epsilon_\omega$, and $s = E(s) + \epsilon_s$.

Assumptions are $\epsilon_{\rho_a} = E(\omega) = E(\epsilon_\omega) = E(\epsilon_s) = 0$.

$$F = E(\rho_a\omega s) \tag{45}$$

$$F = E[(E(\rho_a) + \epsilon_{\rho_a})(E(\omega) + \epsilon_\omega)(E(s) + \epsilon_s)] \tag{46}$$

By our assumptions, the equation can be simplified as follows:

$$F = E[E(\rho_a)\epsilon_\omega(E(s) + \epsilon_s)] \tag{47}$$

$$F = E[E(\rho_a)\epsilon_\omega E(s)] + E(E(\rho_a)\epsilon_\omega\epsilon_s) \tag{48}$$

$$F = E(\rho_a)E(\epsilon_\omega)E(s) + E(E(\rho_a)\epsilon_\omega\epsilon_s) \tag{49}$$

$$F = E(E(\rho_a)\epsilon_\omega\epsilon_s) \tag{50}$$

Discussion

The "covariance" in eddy covariance method is different from the covariance in statistics which is shown below:

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] \tag{51}$$

Appendix C R code of GAM model (equation 2)

```
library(mgcv)

Res.Var <- function(flux.data,NDVI.data,n.year){
NDVI <- NDVI.data
Indi0 <- c(rep("W",5),rep("S",15),rep("W",3))
Indi <- rep(Indi0, n.year)
Time <- rep(seq(1,366,by=16),n.year) ## here the time represent day
mod <- gam(flux.data ~ s(Time, bs="cc", k=12) + NDVI + Indi)
var(residuals(mod))
}

OldAspen_var <- Res.Var(OldAspen_16day,OldAspen_NDVI,6)
OldJackPine_var <- Res.Var(OldJackPine_16day,OldJackPine_NDVI,6)
Southern_var <- Res.Var(Southern_16day,Southern_NDVI,6)
YoungJackPine_var <- Res.Var(YoungJackPine_16day,YoungJackPine_NDVI,3)

FlxTwr <- c(OldAspen_16day,OldJackPine_16day,Southern_16day,YoungJackPine_16day)
ndvi <- c(OldAspen_NDVI,OldJackPine_NDVI,Southern_NDVI,YoungJackPine_NDVI)
time <- rep(seq(1,365,by=16),21)
Indi0 <- c(rep("W",5),rep("S",15),rep("W",3)); indi <- rep(Indi0,21)
spe <-c(rep("OldAspen",length(OldAspen_NDVI)),rep("OldJackPine",length(OldJackPine_NDVI)),
rep("BlackSpr",length(Southern_NDVI)),rep("YoungJackPine",length(YoungJackPine_NDVI)))
Weight <- c(rep(1/OldAspen_var,138),rep(1/OldJackPine_var,138),rep(1/Southern_var,138),
rep(1/YoungJackPine_var,69))
w <- Weight/mean(Weight)
D <- data.frame(Flx=FlxTwr,Time=time,NDVI=ndvi,Spe=spe,W=w,Indi=indi)
rownames(D) <- paste0("row",1:nrow(D))

Mix <- gam(Flx ~ s(Time,bs="cc",k=23)+s(Time,bs="cc",k=23,by=Spe)+ Spe + NDVI,
weights = W,data = D)

summary(Mix)
```

Appendix D The detail of equation 30

Appendix E Difference between second order stationary and intrinsic stationary

- Second order stationary (SOS)
  1. $$E(Z(S)) = \mu \tag{52}$$
  2. $$COV(Z(S), Z(S+h)) = COV(Z(0), Z(h)) = C(h), \ \ \forall h \tag{53}$$
- Intrinsic stationary (IS)
  1. $$E(Z(S)) = \mu \tag{54}$$
  2. $$Var(Z(S+h) - Z(S)) = 2\gamma(h), \ \ \forall h \tag{55}$$

Under IS assumption:

$$\begin{aligned} 2\gamma(h) &= Var(Z(S+h) - Z(S)) && (56)\\ &= E\{[(Z(S+h) - Z(S)) - E(Z(S+h) - Z(S))]^2\} && (57)\\ &= E\{[Z(S+h) - Z(S)]^2\}, \ \ \forall h && (58) \end{aligned}$$

If SOS holds,

$$\begin{aligned} 2[C(0) - C(h)] &= 2[Cov(Z(S), Z(S)) - Cov(Z(S+h), Z(S))] && (59)\\ &= 2[Var(Z(S)) - Cov(Z(S+h), Z(S))] && (60)\\ &= 2Var(Z(S)) - 2Cov(Z(S+h), Z(S)) && (61)\\ &= Var(Z(S+h)) + Var(Z(S)) - 2Cov(Z(S+h), Z(S)) && (62)\\ &= Var(Z(S+h) - Z(S)) && (63)\\ &= 2\gamma(h), \ \ \forall h, && (64) \end{aligned}$$

which satisfies the IS assumptions too.

Let's consider AR(1) process $Z_i = \phi Z_{i-1} + \epsilon_i$ with $\phi = 1$. That is

$$Z_i = Z_{i-1} + \epsilon_i. \tag{65}$$

By iteration,

$$Z_{i+j} = Z_i + \sum_{k=1}^{j} \epsilon_{i+k}. \tag{66}$$

So the covariance of Z_{i+j} and Z_i is

$$Cov(Z_{i+j}, Z_i) = Var(Z_i), \tag{67}$$

which is a function of i. This doesn't satisfy SOS. However,

$$\begin{aligned} Var(Z_{i+j} - Z_i) &= Var(Z_i + \sum_{k=1}^{j} \epsilon_{i+k} - Z_i) && (68)\\ &= Var(\sum_{k=1}^{j} \epsilon_{i+k}) && (69)\\ &= jVar(\epsilon_i), && (70) \end{aligned}$$

which IS holds.