

関数データに基づく統計的モデリング

松井 秀俊^{1,2}

(受付 2018 年 7 月 11 日；改訂 12 月 20 日；採択 12 月 21 日)

要 旨

農業をはじめとする多くの分野では、1つの個体が時間や位置の変化に伴い繰り返して計測されたデータが多く用いられる。本論文では、このような形式で与えられたデータを分析するための方法の1つである、関数データ解析に基づくさまざまな分析手法を紹介する。関数データ解析では、経時測定データを関数化処理し、得られた関数データ集合を対象として分析を行う。特にここでは、関数データ解析の枠組みで回帰分析、時系列解析、空間データ解析を行うための方法を紹介する。また、各手法に対して適用例を示し、どのようなデータに対して、どのような結果が得られるかについて説明する。

キーワード：回帰分析、関数データ解析、空間データ解析、経時測定データ、時系列解析。

1. はじめに

データ科学に基づく意思決定や価値創造が、さまざまな分野で行われるようになってきた。農業分野もその例に漏れず、ICT化による栽培データ管理によって、作物の増産に繋げる動きが強まっている。近年の計測・測定機器の発展に伴い、大量のデータが比較的安価で容易に取得できるようになったことも、この流れを後押ししている。測定されるデータの形式は多岐にわたるが、その中でも多いのが、1つの個体が時間や位置などの変化に応じて繰り返して測定される形式のデータである。例として、日ごとの気温や日照時間、深さに対する土壌成分、波長ごとのスペクトル強度など、さまざまなものが挙げられる。本論文ではこのような形式のデータを経時測定データとよぶ。

農業においては、作物の成長具合は、生育期間中の気温や日照時間といった環境要因が大きく関わっているとされ、これらの関係性を統計モデルによって明らかにすることができれば、収量増産のための栽培管理などに役立つと考えられる。環境要因に関するデータとしては、例えば気象庁のウェブサイト (<http://www.jma.go.jp/jma/index.html>) からダウンロードできる。これらのデータの多くは、各観測地点において経時測定データとして与えられている。このような形式のデータを直接多変量データとして扱い、回帰分析などの統計手法を適用することがまず考えられるが、次のような問題点が生じる。1点目に、観測データは一般的にはノイズが混入されており、経時測定データのばらつきが大きい場合は本質構造を捉えにくくなる。2点目として、観測時点数の増加はそのままデータの次元の増加につながるため、推定量が不安定になる可能性がある。観測時点数が多いデータほどより多くの情報をモデルに取り込むことが

¹ 滋賀大学 データサイエンス学部：〒522-8522 滋賀県彦根市馬場1丁目1-1

² 科学技術振興機構さきがけ：〒332-0012 埼玉県川口市本町4-1-8

できるが、その分解が困難になってしまう恐れがある。また、計測機器の故障などにより、本来計測されるはずの時点で計測が行われない(欠測する)状況が考えられる。あるいは、データの取得環境によっては、個体ごとに観測時点数が不均一であったり、観測時点が不揃いである状況もあり得る。このような場合、欠損値補間(高井 他, 2016)を適用する方法も考えられるが、特に観測時点が不揃いである場合は、一般的には古典的な多変量解析手法を直接適用することは困難となる。これが 3 点目の問題である。

このような形式のデータに対して、個々のデータを関数化し、得られた関数をデータとして扱う方法が Ramsay and Silverman (2005)によって確立され、関数データ解析(Functional Data Analysis; FDA)とよばれている。関数データ解析では、経時測定データを本論文の 2 節で述べる方法などを用いて関数化処理することで、観測ノイズを除去し、さらに、観測時点数の増大により引き起こされるデータの高次元化を抑えることができる。また、データを関数化することにより、個体ごとに観測時点や観測時点数が異なっていたとしても、連続時間上でデータを表現できるため、容易に分析が行えるようになる。これにより、上記の 3 点の問題点が解決される。関数データ解析について詳しく書かれた書籍としては、Ramsay and Silverman (2005)の他に、Ferraty and Vieu (2006), Horváth and Kokoszka (2012)で全般的な方法についての紹介がある。Hsing and Eubank (2015)は、関数データ解析の理論を詳しく述べている。Shi and Choi (2011)は、ベイズアプローチの観点から関数データ解析におけるモデル推定法を紹介している。Ramsay et al. (2009), Kokoszka and Reimherr (2017)では、関数データ解析に基づく手法と、それらを実行するための R などのプログラムがまとめられている。Ramsay and Silverman (2002)は関数データ解析の応用事例を幅広く紹介している。関数データ解析に関するレビュー論文も数多く報告されており、例えば Cuevas (2014), Morris (2015), Shang (2014), Jacques and Preda (2014), Wang et al. (2015), Reiss et al. (2017)が挙げられる。Ullah and Finch (2013)では、さまざまな分野への関数データ解析の応用研究に関する文献を紹介しているが、医学への応用が多く、農学への応用は僅かであることが分かる。さらに、関数データ解析を行うための R パッケージも数多く実装されており、その一覧が Scheipl (2018)にまとめられている。

関数データ解析では、スカラーデータに対する統計的分析手法を関数データの枠組みへ拡張したものが数多く提案されている。本論文では、経時測定データを関数データとして扱い、分析するための方法をいくつか紹介する。紹介する手法は農業データ分析に特化したものではなく、汎用的に用いることができるものであるが、農業データへの応用が有効であると考えられるものをまとめた。まず、説明変数と目的変数のいずれか一方、あるいは両方が関数データとして与えられたとき、両者の関係を表す関数回帰モデルについて説明する。次に、関数データに基づく手法として提案された、各個体が関数データとして与えられた時系列データの予測に用いられる関数時系列解析として、Hyndman and Ullah (2007)の方法を紹介する。さらに、複数地点で観測された関数データから、未観測地点の曲線を予測する関数空間データ解析として、Giraldo et al. (2011)の方法についても紹介する。いずれの手法に対しても、適用例を通して、どのようなデータに対して、どのような分析を行うことができるかについて説明する。

本論文の構成は次の通りである。2 節では、関数データに基づく分析手法について紹介する前に、取得された経時測定データを関数データ化するための方法について紹介する。続く 3 節では、関数データに基づく回帰モデルである関数回帰モデルについて、モデルおよび推定方法の概略を紹介する。4 節では、関数データの枠組みでの時系列解析の方法について紹介する。そして 5 節では、空間データ解析に基づき未観測地点の関数データを予測するための方法について解説する。3 節から 5 節については、各手法の適用例についても紹介する。なお、4 節と 5 節の適用例については、Kokoszka and Reimherr (2017)に掲載されている R のプログラムを参

考にした．最後に，6 節でまとめと今後の展望について述べる．

2. データの関数化

本節では，1 年間の日別平均気温や平均湿度といった経時測定データを，関数データとして表すための方法について述べる．後の節で紹介する，関数回帰分析をはじめとした関数データ解析の多くの方法に対しては，まず第 1 段階で経時測定データの関数化を行い，続く第 2 段階で関数データ集合に対して分析を行うという 2 段階法が用いられている．

いま，第 i 番目の個体が，第 α 時点 $t_{i\alpha}$ で観測値 $x_{i\alpha}$ を得たとする．ここで，個体番号は $i = 1, \dots, n$ ，時点番号は $\alpha = 1, \dots, n_i$ の値を取るものとする．時点 $t_{i\alpha}$ やその数 n_i は，個体ごとに異なっているもよい．また，ここでは分かりやすさのために t を時間とみなして説明しているが， t は位置や深さ，波長など，本来は連続的に変化するようなものであればよい．経時測定データの例として，図 1 左は，世界のいくつかの都市で観測された月別平均気温のデータを図示したものである．このようなデータを関数化するために，各個体に対して平滑化を適用する．すなわち，第 i 個体のデータ $x_{i\alpha}$ ($\alpha = 1, \dots, n_i$) は，次式のように関数 $x_i(t)$ が時点 $t_{i\alpha}$ において誤差 $e_{i\alpha}$ が加わって得られたものと仮定し， $x_i(t)$ を推定する方法である．

$$x_{i\alpha} = x_i(t_{i\alpha}) + e_{i\alpha}.$$

関数 $x_i(t)$ を推定する方法の 1 つとして， $x_i(t)$ は基底関数展開，すなわち，基底関数とよばれる既知の関数系 $\phi_1(t), \dots, \phi_M(t)$ の線形結合によって表されると仮定する．

$$(2.1) \quad x_i(t) = \sum_{m=1}^M w_{im} \phi_m(t) = \mathbf{w}_i^T \boldsymbol{\phi}(t).$$

ただし $\mathbf{w}_i = (w_{i1}, \dots, w_{iM})^T$ は係数パラメータで，これを推定することで関数データ $x_i(t)$ が得られる．また， $\boldsymbol{\phi}(t) = (\phi_1(t), \dots, \phi_M(t))^T$ とする．基底関数の種類としては，さまざまなものが提案されている．汎用的に用いられているものとしては， B -スプライン (井元・小西, 1999; de Boor, 2001) や動径基底関数 (Bishop, 1995; 安道 他, 2001) などが挙げられる．また，1 年間の日別平均気温のデータのように，周期性をもつデータに対してはフーリエ級数が，心電図のようにスパイク状の変動に重要な意味をもつようなデータにはウェーブレット (Donoho and Johnstone, 1994) が有効である．基底関数展開以外の方法としては，局所線形回帰などのノンパラメトリック回帰 (Härdle, 1990) に基づく平滑化によって関数データを得る方法も考えられている (Yao et al., 2003)．平滑化の詳細に関しては Green and Silverman (1994)，小西 (2010) などを，関数データ解析の枠組みでの関数化については Ramsay and Silverman (2005) の 4 章および 5 章，荒木・小西 (2004)，Araki et al. (2009b)，Fujii and Konishi (2006) などを参照されたい．図 1 左における各個体のデータに対して，基底関数展開に基づく平滑化を行うことで，図 1 右のような関数データ (曲線) が得られる．

関数データを表現するもう 1 つの方法として，関数主成分分析 (Besse and Ramsay, 1986; 茅野 他, 2006) を用いるものがある．これは，確率的に変動するランダム関数 $X(t)$ が，適当な仮定の下で次のように表されることを利用する．

$$(2.2) \quad X(t) = \mu(t) + \sum_{l=1}^{\infty} \xi_l v_l(t).$$

ただし， $\mu(t)$ は平均関数とする．ランダム関数の厳密な定義や仮定の詳細については，Ferraty and Vieu (2006)，Horváth and Kokoszka (2012) などを参照されたい．ここで， ξ_l ($l = 1, 2, \dots$)

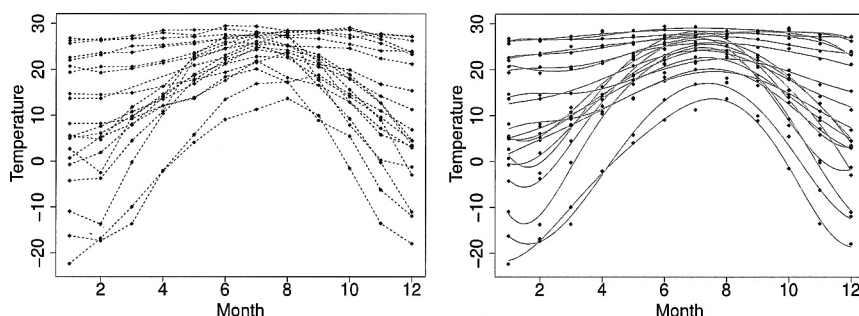


図 1. 左：世界の各都市の月別平均気温データ。破線でつながった点は同じ都市のデータを表す。右：左のデータに対して基底関数展開に基づく平滑化を適用し得られた関数データ。

は平均 0, 分散 λ_l ($\sum_{l=1}^{\infty} \lambda_l < \infty$, $\lambda_1 > \lambda_2 > \dots$) をもつ確率変数, $v_l(t)$ ($l = 1, 2, \dots$) は正規直交関数とする。この表現は Karhunen-Loève 展開としても知られている。実際には, 上述した基底関数展開やノンパラメトリック回帰に基づく平滑化によって得られた関数データ $x_i(t)$ を, 次のように $v_l(t)$ の有限個の線形和で表現する。

$$(2.3) \quad x_i(t) = \bar{x}(t) + \sum_{l=1}^L \xi_{il} v_l(t).$$

ここで $\bar{x}(t) = \frac{1}{n} \sum_{i=1}^n x_i(t)$ とし, λ_l (ξ_{il} の分散) と $v_l(t)$ はそれぞれ標本共分散関数 $c(s, t) = \frac{1}{n-1} \sum_{i=1}^n (x_i(s) - \bar{x}(s))(x_i(t) - \bar{x}(t))$ の第 l 固有値, 固有関数として求められる。また, ξ_{il} は $x_i(t) - \bar{x}(t)$ と $v_l(t)$ の内積を計算することで得られる。(2.1) 式で用いた基底関数 $\phi_m(t)$ は汎用的に用いられる形状の関数であったのに対して, 固有関数 $v_l(t)$ は直交性の条件の下で, 与えられたデータを $l = 1, 2, \dots$ の順に最もよく表現する関数を与える。このことから, 一般的に固有関数の個数 L は (2.1) 式の基底関数の個数 M よりも少なくなり, 時点数の多い経時測定データ分析における次元縮小には特に有効である。固有関数の個数 L の決定に関しては, 累積寄与率や, 交差検証法, AIC などを用いた方法が考えられている (Yao et al., 2005a; Crainiceanu et al., 2009)。なお, 関数主成分分析は, 一度関数化したものを主成分によって表現するため, 関数主成分による関数データの近似精度は, 関数化の精度に依存する。

データによっては, 各個体が非常にまばらな時点でのみ観測されるような場合もある。例えば図 2 左は, 図 1 左のデータの一部だけを人工的に取り出したもので, 各個体が限られた時点でのみ観測されていることがわかる。このような形式のデータはスパース経時測定データとよばれている。スパース経時測定データに対して, 前述のように個体ごとに独立して平滑化を行うと, 観測時点数が少ないために適切な曲線推定ができない可能性が高い。そこで, 個体それぞれではなく, 全個体の観測情報を使ってデータを関数化する方法が考えられている。1 つは, データ集合 $\{x_{i\alpha}; i = 1, \dots, n, \alpha = 1, \dots, n_i\}$ を局所線形回帰などを用いて平滑化する方法である (Yao et al., 2005a, 2005b)。また, 基底関数展開に基づく非線形混合効果モデルを適用することで, 全個体の観測情報を使って個体 1 つ 1 つを関数化することもできる (James et al., 2000; Rice and Wu, 2001; 松井 他, 2016)。非線形混合効果モデルを適用することで, 全ての個体の情報を共有し, 図 2 左のような, 1 つの個体がまばらな時点で観測されたデータであっても, 図 2 右のように, データが観測されている時間の範囲で各個体の曲線を当てはめることができる。Zhang and Wang (2016) は, 時点数がスパースな場合とそうでない場合の経時測定

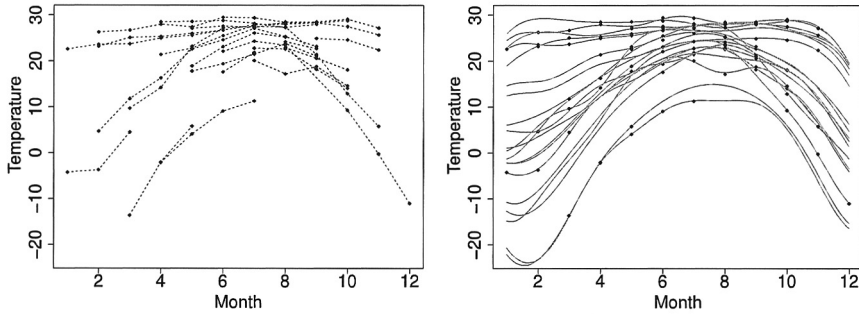


図 2. 左：図 1 左のデータの一部のみを人工的に抜き取ったもの。右：左のデータに対して非線形混合効果モデルを適用し得られた関数データ。

データについて、局所線形回帰により関数化した際の、推定値の理論的な性質を統一的に示している。

3. 関数回帰モデル

説明変数，あるいは説明変数と目的変数の両方が関数データとして与えられたとき，これらの関係をモデル化する関数回帰モデルの概要を説明する．例えば，作物の生育期間中の気温や日照時間の推移を説明変数，作物の生育状況のデータを目的変数とし，これらの関係をモデル化することを考える．年に 1 度だけ収穫される作物の収量を考えたい場合は，目的変数は一般の回帰モデルと同様にスカラーとして与えられるだろう．一方で，トマトなどのように特定の期間で毎日のように収穫される作物に対しては，目的変数も経時測定データとなるため，関数データとして扱うこともできる．本節では，目的変数がスカラーおよび関数データで与えられた場合のモデルと，その推定方法についてそれぞれ解説する．

3.1 スカラー目的変数に対するモデル

いま，目的変数と説明変数に対して， n 組の観測値 $\{(y_i, x_i(t)); t \in \mathcal{T} \in \mathbb{R}, i = 1, \dots, n\}$ を得たとする．ここで， y_i はスカラーとして得られた目的変数， $x_i(t)$ は関数として表された説明変数とする．また， $\mathcal{T}$ は関数 $x_i(t)$ の定義域とする．説明変数に関するデータについては，実際には離散時点で観測されたものであるため，2 節で述べた平滑化手法などを用いて関数化されたものとする．このとき， y_i と $x_i(t)$ との関係を表す関数線形回帰モデルは，次で与えられる（荒木・小西, 2004; Ramsay and Silverman, 2005; Araki et al., 2009b）．

$$(3.1) \quad y_i = \beta_0 + \int_{\mathcal{T}} x_i(t) \beta_1(t) dt + \varepsilon_i.$$

ただし β_0 は切片， $\beta_1(t)$ は回帰係数関数， ε_i は平均 0，分散 σ^2 を持つ観測誤差とする．

関数線形回帰モデル (3.1) の推定問題は，未知パラメータである β_0 ， $\beta_1(t)$ ， σ^2 を推定することに対応する．しかし， $\beta_1(t)$ は無限次元のパラメータとなるため，これを直接推定することは困難である．これらのパラメータを推定するためのアプローチの 1 つとして，基底関数展開を用いる方法がある．まず，説明変数 $x_i(t)$ が，(2.1) のように基底関数展開によって表されるものとする．ここで， w_i は関数化の段階で推定されたものであり，ここでは既知とする．基底関数の代わりに固有関数を用いてもよい．そして，係数関数 $\beta_1(t)$ も同様に，次のように基底関

数展開によって表されるとする.

$$(3.2) \quad \beta_1(t) = \sum_{m=1}^M b_{1m} \phi_m(t) = \mathbf{b}_1^T \boldsymbol{\phi}(t).$$

ここで, $\mathbf{b}_1 = (b_{11}, \dots, b_{1M})^T$ は未知パラメータとする. また, (2.1) 式の基底関数と (3.2) 式の基底関数について, ここでは共通のものを用いているが, その個数や種類は異なっているがよい. 基底関数展開 (2.1), (3.2) の仮定により, 関数線形回帰モデル (3.1) は次のように変形される.

$$\begin{aligned} y_i &= \beta_0 + \int_{\mathcal{T}} \mathbf{w}_i^T \boldsymbol{\phi}(t) \boldsymbol{\phi}^T(t) \mathbf{b}_1 dt + \varepsilon_i \\ &= \beta_0 + \sum_{m=1}^M \mathbf{w}_i^T \Phi \mathbf{b}_1 + \varepsilon_i \\ (3.3) \quad &= \mathbf{z}_i^T \mathbf{b} + \varepsilon_i. \end{aligned}$$

ここで, $\Phi = \int_{\mathcal{T}} \boldsymbol{\phi}(t) \boldsymbol{\phi}^T(t) dt$ は基底関数ベクトル $\boldsymbol{\phi}(t)$ の各成分の積の積分を要素にもつ $M \times M$ 行列, $\mathbf{z}_i = (1, \mathbf{w}_i^T \Phi)^T$ はここでは既知のベクトル, そして $\mathbf{b} = (\beta_0, \mathbf{b}_1^T)^T$ は未知パラメータベクトルとする. 特に, $\boldsymbol{\phi}(t)$ がフーリエ級数や固有関数などのように正規直交基底の場合, Φ は単位行列となる. (3.3) 式の形から, 最小 2 乗法や正則化法といった, 古典的な線形回帰モデルに対する推定法と同様の方法を用いて, パラメータを推定できる. 例えば, (3.3) 式のモデルのパラメータ β を正則化最小 2 乗法により推定する場合は, 次の正則化誤差 2 乗和を最小にする $\mathbf{b}$ を求める.

$$\sum_{i=1}^n (y_i - \mathbf{z}_i^T \mathbf{b})^2 + n\lambda \mathbf{b}^T \Omega \mathbf{b}.$$

ただし, $\lambda > 0$ は正則化の度合いを調整する正則化パラメータで, Ω は非負値定符号行列とする. 正則化誤差 2 乗和の最小化により, 次の推定量が得られる.

$$\hat{\mathbf{b}} = (Z^T Z + n\lambda \Omega)^{-1} Z^T \mathbf{y}.$$

ただし, $Z = (\mathbf{z}_1, \dots, \mathbf{z}_n)^T$, $\mathbf{y} = (y_1, \dots, y_n)^T$ とする. パラメータの推定値 $\hat{\mathbf{b}} = (\hat{\beta}_0, \hat{\mathbf{b}}_1^T)^T$ を用いて, 新たな観測 $x_0(t)$ が得られたときの目的変数の値を予測したり, 係数関数の推定値 $\hat{\beta}_1(t) = \hat{\mathbf{b}}_1^T \boldsymbol{\phi}(t)$ から, 説明変数の任意の時点 t における寄与を定量化できる. Ogden et al. (2002) は, 水田の俯瞰撮影により得られた画像のデータを関数データとして表現し, イネの倒伏度合いを表す指標を予測するために関数回帰モデルを適用した.

関数線形回帰モデルは, 説明変数が複数の場合へも容易に拡張できる. それだけでなく, 説明変数として関数データだけではなく, スカラーデータを説明変数に加えることもできる. すなわち, p_1 変数のスカラー説明変数 x_{ij} ($j = 1, \dots, p_1$) と, $p - p_1$ 変数の関数説明変数 $x_{ij}(t)$ ($j = p_1 + 1, \dots, p$) が与えられたとき, 関数線形回帰モデルは

$$(3.4) \quad y_i = \beta_0 + \sum_{j=1}^{p_1} x_{ij} \beta_j + \sum_{j=p_1+1}^p \int_{\mathcal{T}} x_{ij}(t) \beta_j(t) dt + \varepsilon_i$$

のように与えられる. なお, 関数データ $x_{ij}(t)$ の係数がスカラー β_j として与えられる, すなわち, 関数説明変数 $x_{ij}(t)$ の目的変数 y_i への寄与が t によらない場合, 対応する項は $\int x_{ij}(t) dt \cdot \beta_j$ となり, $x_{ij}(t)$ の $\mathcal{T}$ 上の積分値 $\int x_{ij}(t) dt$ を説明変数としたものとして (3.4) 式右辺第 2 項に含

めることができる．(3.4)式に対して，説明変数が1つのみのモデルに対するものと同様の仮定を置くことで，このモデルも(3.3)式と同様のベクトル表記を行うことができ，やはり一般の線形回帰モデルと同様に推定できる．

複数の説明変数の中から，目的変数に実際に寄与している変数の組み合わせを選びたい場合は，変数選択問題を考える必要がある．例えば，気温，日射量，大気中のCO₂濃度といった複数の環境要因の経時変化のうち，どの組み合わせが作物の収量に寄与しているかを調べるという問題に対応する．変数選択の方法としては，古典的なステップワイズ法などを関数線形回帰モデルに適用することができる．一方で，スパース推定 (Hastie et al., 2015; 川野 他, 2018) を利用することで，ある j について，説明変数 $x_j(t)$ に対する係数関数を $\hat{\beta}_j(t) \equiv 0$ と推定することができる．このとき， $x_j(t)$ は目的変数に寄与していないものとみなすことができ，変数選択が行われたことになる．関数回帰モデルに対するスパース推定については，Matsui and Konishi (2011)，Gertheiss et al. (2013)，Lian (2013)，Zhao et al. (2015)などを参照されたい．また，James et al. (2009)，Lee and Park (2012)は，スパース推定の考えを利用して，係数関数 $\beta(t)$ の定義域 $\mathcal{T}$ の全区間ではなく一部区間のみを0と推定することで，目的変数に寄与する説明変数の区間の選択を行っている．

線形モデルの枠組みを拡張した関数回帰モデルの1つとして，次の関数2次回帰モデルが提案されている (Yao and Müller, 2010; Fuchs et al., 2015; Usset et al., 2016)．

$$(3.5) \quad y_i = \beta_0 + \sum_{j=1}^p \int_{\mathcal{T}} x_{ij}(t) \beta_j(t) dt + \sum_{j,k} \iint_{\mathcal{T} \times \mathcal{T}} x_{ij}(s) x_{ik}(t) \beta_{jk}(s, t) ds dt + \varepsilon_i.$$

ここで， $\beta_{jk}(s, t)$ は2変数の係数関数である．すなわち，2次回帰モデルは，説明変数 $x_{ij}(t)$ の2次の情報だけでなく，単一および2つの説明変数の，異なる2時点 s, t 間の交互作用まで考慮に入れたモデルとなっている．Wei et al. (2014)は，スカラーデータとして与えられた遺伝型と，経時測定データとして与えられた環境情報の相互作用(遺伝子-環境相互作用)を関数回帰モデルの枠組みで表現している．

本項で述べたモデルはさらに，目的変数がスカラーであることから，一般化線形モデル (McCullagh and Nelder, 1989) の枠組みへも容易に拡張できる (James, 2002)．例えば，気温や日照時間を説明変数としたとき，目的変数として年間の総収量といった連続量ではなく，生育に成功したか否か，あるいは病害の有無といった2値のデータの判別を行いたい場合や，生育に成功した株の割合といった比率のデータの推定および予測を行いたい場合は，関数ロジスティック回帰モデルを用いることが考えられる (Araki et al., 2009a; Matsui et al., 2011; Kayano et al., 2016)．あるいは，病害が発生した回数などのカウントデータを目的変数として扱いたい場合は，関数ポアソン回帰を用いることもできる．関数回帰モデルはこの他にも，適応モデル (James and Silverman, 2005) やニューラルネットワークモデル (Rossi et al., 2005)，加法モデル (Müller and Yao, 2008; Müller et al., 2013; Zhu et al., 2014)，一般加法モデル (McClean et al., 2014)，ノンパラメトリックモデル (Ferraty et al., 2007; Rachdi and Vieu, 2007) といった非線形モデルなどへ拡張されている．さらに，Reiss and Ogden (2007, 2010)は主成分回帰の枠組みでの関数回帰モデルの推定法を提案している．このように，目的変数がスカラーの場合は，従来の回帰モデルと同様の拡張が行われている．

本項では，関数化を行った後に関数回帰モデルを推定するという，2段階法に基づく方法を紹介した．一方で，Crainiceanu et al. (2009)は，関数回帰モデルの推定において，2段階法ではなく，1段階で同時にモデルを推定する方法を提案している．そして，真の係数関数の値が大きいときや，説明変数に対応する経時測定データの誤差分散が大きいとき，1段階法とは異なり2段階法では係数関数の推定値にバイアスが生じることを示している．しかし Crainiceanu

et al. (2009)は同時に、2段階法はロバストな推定結果を与える上、1段階法と比べて計算コストが少なく、推定結果に大きな差がないことから、必ずしも2段階法を否定しておらず、両手法を適用し結果を比較するのがよいと述べている。

3.2 関数目的変数に対するモデル

3.1項では、目的変数 y_i は古典的な回帰モデルと同様にスカラーを想定した。一方で、トマトなどのように年間の一定期間を通して繰り返して収穫される作物に対して、気象情報と収量との関係を回帰モデルで表現する場合は、目的変数も説明変数と同様に関数データとして扱う方法が考えられる。本項では、説明変数および目的変数が共に関数データで与えられた場合のモデルについて紹介する。

目的変数と説明変数に関する n 組の観測値 $\{(y_i(t), x_i(s)); t \in \mathcal{T} \subset \mathbb{R}, s \in \mathcal{S} \subset \mathbb{R}, i = 1, \dots, n\}$ が得られたとする。なお、 $x_i(s)$, $y_i(t)$ はいずれも中心化されたもの、すなわち、それぞれの標本平均 $\frac{1}{n} \sum_{i=1}^n x_i(s)$, $\frac{1}{n} \sum_{i=1}^n y_i(t)$ を引いたものを扱うこととする。このとき、関数線形回帰モデルは次で与えられる (Ramsay and Dalzell, 1991; 下川 他, 2000; Matsui et al., 2009)。

$$(3.6) \quad y_i(t) = \int_{\mathcal{S}} x_i(s) \beta_1(s, t) ds + \varepsilon_i(t).$$

ここで、 $\beta_1(s, t)$ は説明変数 $x_i(s)$ の時点 s による、目的変数 $y_i(t)$ の時点 t への寄与を表す2変数の係数関数、 $\varepsilon_i(t)$ は平均0の誤差関数とする。

係数関数 $\beta_1(s, t)$ を推定する方法の1つとして、積分誤差2乗和 $\sum_{i=1}^n \int_{\mathcal{T}} \varepsilon_i^2(t) dt$ を最小にするものがある。推定量を求めるにあたり、次の仮定をおく。まず、 $x_i(s)$ および $y_i(t)$ が、次のように基底関数 $\phi(s) = (\phi_1(s), \dots, \phi_{M_x}(s))^T$, $\psi(t) = (\psi_1(t), \dots, \psi_{M_y}(t))^T$ に基づく線形結合によって表されたとする。

$$x_i(s) = \sum_{m=1}^{M_x} w_{im} \phi_m(s) = \mathbf{w}_i^T \boldsymbol{\phi}(s), \quad y_i(t) = \sum_{m=1}^{M_y} v_{im} \psi_m(t) = \mathbf{v}_i^T \boldsymbol{\psi}(t).$$

ここで、 $\mathbf{w}_i = (w_{i1}, \dots, w_{iM_x})^T$, $\mathbf{v}_i = (v_{i1}, \dots, v_{iM_y})^T$ であり、これらは2節で述べた方法であらかじめ関数化の段階で推定されているものとする。さらに、 $\beta_1(s, t)$ は2種類の基底関数 $\phi(s)$, $\psi(t)$ を用いて次のように表されたと仮定する。

$$\beta_1(s, t) = \sum_{m=1}^{M_x} \sum_{m'=1}^{M_y} \phi_m(s) b_{mm'} \psi_{m'}(t) = \boldsymbol{\phi}^T(s) B \boldsymbol{\psi}(t).$$

ここで、 $B = (b_{mm'})_{mm'}$ はパラメータからなる行列とする。これらの仮定より、(3.6)式は次のように表される。

$$\begin{aligned} \mathbf{v}_i^T \boldsymbol{\psi}(t) &= \int_{\mathcal{S}} \mathbf{w}_i^T \boldsymbol{\phi}(s) \boldsymbol{\phi}^T(s) B \boldsymbol{\psi}(t) ds + \varepsilon_i(t) \\ &= \mathbf{w}_i^T \Phi B \boldsymbol{\psi}(t) + \varepsilon_i(t). \end{aligned}$$

これを用いて、積分誤差2乗和は次で与えられる。

$$\begin{aligned} \sum_{i=1}^n \int_{\mathcal{T}} \varepsilon_i^2(t) dt &= \sum_{i=1}^n \int_{\mathcal{T}} (\mathbf{v}_i^T - \mathbf{w}_i^T \Phi B) \boldsymbol{\psi}(t) \boldsymbol{\psi}^T(t) (\mathbf{v}_i^T - \mathbf{w}_i^T \Phi B)^T dt \\ &= \text{tr}\{(V - ZB) \Psi (V - ZB)^T\}. \end{aligned}$$

ここで $W = (w_1, \dots, w_n)^T$, $V = (v_1, \dots, v_n)^T$, $\Psi = \int_T \psi(t)\psi^T(t)dt$, $Z = W\Psi$ とする. これを B について最小化することで, 推定量

$$\text{vec}(\hat{B}) = (\Psi \otimes Z^T Z)^{-1} \text{vec}(Z^T V \Psi)$$

を得る. ただし, vec は行列の列ベクトルを全て 1 列に並べる作用素で, $\otimes$ はクロネッカー積を表す. モデル (3.6) を推定するための他の方法として, 誤差関数 $\varepsilon_i(t)$ にガウス過程を仮定し, バイズアプローチに基づきパラメータを推定する方法 (Shi and Choi, 2011) も考えられる.

ここで, 説明変数および目的変数が時間の関数データである場合に (3.6) 式のモデルを適用する場合, 注意が必要である. (3.6) 式では, $s > t$ を満たす範囲の係数関数 $\beta_1(s, t)$ は, 時点 s の説明変数が, それよりも過去の時点 t の目的変数への寄与を表していることになり, 矛盾した関係性を表す可能性がある (ただし例外として, データに周期性を仮定している場合は, 例えば「12 月の気温が翌年 1 月の収量に影響を及ぼす」というように, 次の周期の時点への寄与を表していると解釈できる). この問題点を解消するために, Malfait and Ramsay (2003) は, 時間の前後関係を考慮に入れた関数線形回帰モデルを構築し, その推定方法を提案した. この他にも, 同様の目的に対する研究が Harezlak et al. (2007), Şentürk and Müller (2010), Kim et al. (2011) によって行われている.

説明変数と目的変数間の関係性を調べるにあたって, 説明変数と同じ時点での目的変数への影響にのみ興味がある場合, すなわち $s = t$ での関係性のみを考える場合は, 回帰モデルは次で与えられる.

$$(3.7) \quad y_i(t) = \beta_0(t) + x_i(t)\beta_1(t) + \varepsilon_i(t).$$

ここで, $\beta_0(t)$ は説明変数に関する定数項に対応するベースライン関数, $\beta_1(t)$ は $x_i(t)$ の係数関数とする. このモデルは関数同時モデルとよばれているほか, Hastie and Tibshirani (1993) によって提案された変化係数モデルの特別な場合とみなすこともできる (Hoover et al., 1998). 関数同時モデルの推定方法については, Fan and Zhang (2008), Manrique et al. (2018) などを参照されたい.

3.3 適用例

3.1 節で紹介した関数線形回帰モデルを用いて, 作物のデータを分析した例を紹介する. ここでは, 兵庫県神戸市にある農場で 2012 年から 2014 年の間の 2 期で計測されたトマトの単位区画あたりの収量を目的変数, 農場で計測された環境情報の経時測定データを関数データの説明変数として扱い, これらの関係をモデル化する.

収量のデータは 10 月から翌 7 月まで日ごとに測定されているが, ここでは各週で総和をとったものを 1 個体, つまりスカラーデータとして扱った. 収量が計測された期間は 1 期目, 2 期目ともに 39 週分であった. つまり, 分析に用いる週の数(サンプルサイズ)は 78 である. トマトの収量については, その実が開花してから結実するまでのおよそ 60 日間の環境要因が関係していると考えられている. そこで, 週次の単位区画あたり収量 [kg] を目的変数, その週から遡った 60 日間の気温 [°C] および日射量 [W/m²] のデータを説明変数とした. 説明変数に対応するデータは毎分測定されているが, ここでは 1 日のデータの平均値をとり, さらに関数データ化したものを扱った. 図 3 は, 各週(個体)から遡った 60 日間の気温および日射量を, 関数データとして表現したものである. また, 図 4 に, 2 期分の単位区画あたりの週次収量を示す. 収量は春から初夏にかけて大きく伸びる時期があるなど, 年間の時期にも依存している. そこで, 上記の変数に加えて, 季節成分も説明変数に加えた. 具体的には, 第 i 週の初日の年間での通算日数 $d_i \in \{1, \dots, 365\}$ を, B-スプライン基底関数展開によって変換した

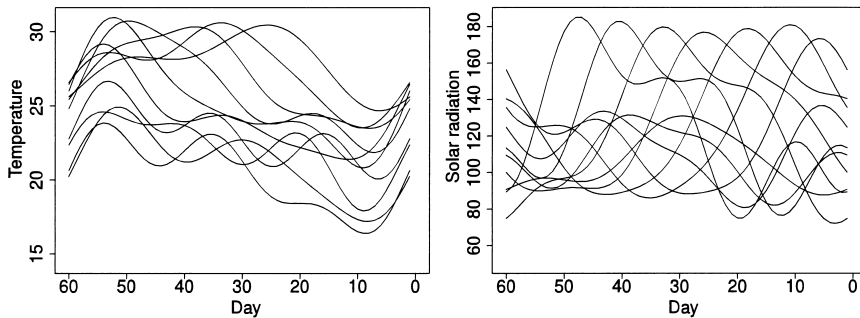


図 3. 週次収量を計測した週から遡った 60 日間の環境情報のデータ. 左: 気温, 右: 日射量を関数データ化したもの (一部を抜粋). 各グラフの左側ほど過去の日付を表す.

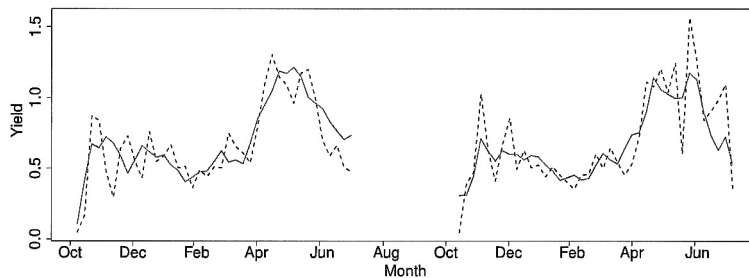


図 4. 2012 年 10 月から 2014 年 7 月の間に計測されたトマトの週次収量 (破線) および, 関数回帰モデルにより得られた収量の予測値 (実線).

$\psi(d_i) = (\psi_1(d_i), \dots, \psi_{M'}(d_i))^T$ を説明変数としてモデルに加える. 基底関数の個数 M' は 30 と固定した. 以上をまとめて, 収量を推定するためのモデルを次で与える.

$$y_i = \sum_{j=1}^2 \int_{\mathcal{T}} x_{ij}(t) \beta_j(t) dt + \psi(d_i)^T \gamma + \varepsilon_i.$$

ここで, γ は季節成分 $\psi(d_i)$ の係数, $x_{i1}(t)$, $x_{i2}(t)$ はそれぞれ気温, 日射量の関数データ, $\beta_1(t)$, $\beta_2(t)$ は対応する係数関数である. 2 種類の説明変数については, フーリエ級数による基底関数展開によって表現し, 基底関数の個数は共に 7 個とした. また, 関数回帰モデルを正則化最小 2 乗法で推定し, 推定に伴う正則化パラメータの値を, 交差検証法を用いて選択した.

図 4 に, 関数回帰モデルによる収量 y_i の予測値を示している. 収穫期間を通じて大まかな変動は捉えられていることが分かるが, 各期の収穫初めの時期や, 春から夏にかけての収量増産期の局所的な変動までは捉えられていない. 図 5 左は, 季節成分 $\psi(d)^T \gamma$ の推定値である. 上述の通り, 春から初夏にかけて収量が伸びている様子が捉えられていることがわかる. また, 図 5 中央と右に, それぞれ係数関数 $\beta_1(t)$, $\beta_2(t)$ の推定曲線を示す. 例えば, 気温については収穫のおよそ 40 日前に比較的大きな重みがかかっていることから, この期間の気温が高いほど収量が増加すると解釈することができる. しかし, 2 つの係数関数の信頼帯がほぼすべての区間で 0 を含んでいることから, 今回の方法では, 収量・環境要因間の関係性を適切に捉えているとは言い難い. 果実の生育には環境情報だけでなく様々な要因が複雑に絡み合っていると

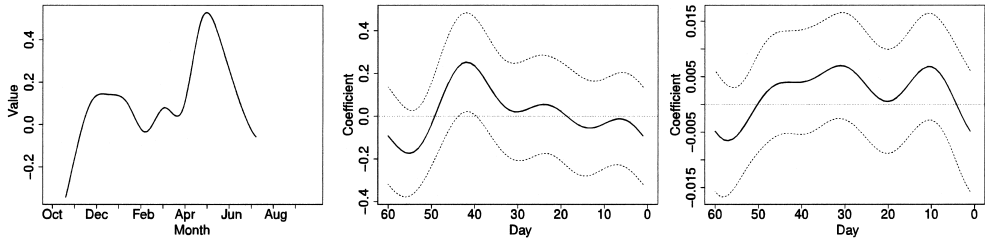


図 5. 左：季節成分 $\psi(t)^T\gamma$ の推定値。横軸は 10 月 1 日から翌 9 月 30 日までの期間に対応している。中央，右：それぞれ気温と日射量の係数関数。左側ほど過去の日付を表し，破線は 95% 信頼帯を表す。

考えられるため，それらのデータを含めた分析を行う必要がある．さらに，今回の分析では説明変数の時間区間を収穫週から遡った 60 日間と固定したが，この区間も不変の事実ではなく，実際には前後している可能性がある．そのため，例えば James et al. (2009) の方法に応用して，何日前までの環境要因が収量に影響を及ぼしているかを推定するといった方法も考えられるが，この点は今後の課題とする．

4. 関数時系列解析

時系列解析は，データが経時的に観測されたとき，その傾向を捉えたり，将来の時点の値を予測することが目的である（たとえば 北川, 2005）．主な時系列解析のための方法は，一連の時系列データに基づいて次の 1 点，または複数の点を予測するものであった．これに対して，時系列解析を関数データの枠組みへ発展させ，将来の点ではなく関数(曲線)を予測するための研究も行われている (Bosq, 2000; Horváth and Kokoszka, 2012; Hörmann and Kokoszka, 2012; 北川 他, 2016)．このような解析は関数時系列解析とよばれている．

関数時系列解析を，気温のデータを例に挙げて説明する．いま，ある地点の 20 年分の月別平均気温のデータが観測されたとする．1 年間のデータを 1 個体の関数データとして捉えると，20 個体の関数データが与えられていることになる．このデータ集合を用いて，次の年の (1 時点ではなく) 年間の気温曲線を予測することが，関数時系列解析の目的となる．

ここでは，Hyndman and Ullah (2007) の方法に基づく関数時系列の予測方法について紹介する．この方法は，アメリカにおける各年齢層の死亡率の毎年の推移のデータから，将来の死亡率を予測することを目的に提案されたものだが，一般的な関数時系列データに対しても適用できる．

4.1 固有関数展開による予測

いま，第 i 期の関数データ $x_i(t)$ からなる関数時系列データ集合 $\{x_i(t); i = 1, \dots, n\}$ が与えられたとする．このデータを用いて， h 期先のデータ $x_{n+h}(t)$ を予測することを考える． $x_i(t)$ は，各 i に対して (2.2) 式のように，関数主成分分析に基づいて次のように表されると仮定する．

$$x_i(t) = \hat{\mu}(t) + \sum_{l=1}^L \hat{\xi}_{il} \hat{v}_l(t).$$

ここで，平均関数の推定値 $\hat{\mu}(t)$ としては，例えば標本平均 $\frac{1}{n} \sum_{i=1}^n x_i(t)$ などが考えられる．一方で，Hyndman and Ullah (2007) は外れ値にロバストな関数を得るために， $x_1(t), \dots, x_n(t)$ の

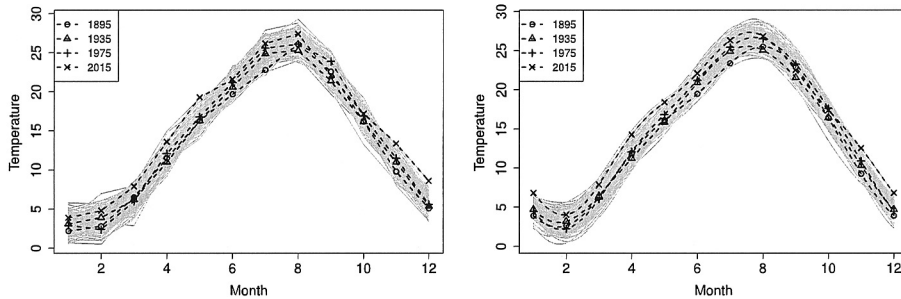


図 6. 左：1894 年から 2017 年までの、彦根市の月別平均気温データ。右：観測データを関数データ化したもの。

中央値曲線を推定値として用いている．続いて，各 l について， $\hat{\xi}_{1l}, \dots, \hat{\xi}_{nl}$ にスカラーデータに対する時系列解析を適用し，第 n 期に対する h 期先の係数 $\xi_{n+h,l}$ の予測値 $\hat{\xi}_{n+h|n,l}$ を得る．ここで，固有関数 $\hat{v}_1(t), \dots, \hat{v}_L(t)$ は正規直交であることから， $\hat{\xi}_{i1}, \dots, \hat{\xi}_{iL}$ は無相関となる．したがって， $l = 1, \dots, L$ それぞれに対して，単変量の時系列解析手法を適用する．こうして得られた予測値 $\hat{\xi}_{n+h|n,1}, \dots, \hat{\xi}_{n+h|n,L}$ を用いて， h 期先の関数データ $x_{n+h}(t)$ の予測値 $\hat{x}_{n+h|n}(t)$ を次で与える．

$$(4.1) \quad \hat{x}_{n+h|n}(t) = \hat{\mu}(t) + \sum_{l=1}^L \hat{\xi}_{n+h|n,l} \hat{v}_l(t).$$

この方法により，関数時系列データの傾向を，固有関数とその係数を使って表現できる．

4.2 適用例

滋賀県彦根市にある地上気象観測装置で観測された，1894 年から 2017 年までの月別平均気温のデータ（気象庁ウェブサイトより取得）に対して，関数時系列解析を適用する．図 6 左は，1894 年から 2017 年までの年間月別平均気温を図示したものである．このデータを用いて，2018 年の年間月別平均気温曲線の予測を試みる．

まず，R パッケージ `fda` を用いて，月別平均気温のデータを関数データ化した．ここでは各個体の曲線が周期性を持つことから，基底関数としてフーリエ級数を用いた．基底関数の個数は 7 個とした．図 6 右は，各年の気温のデータを関数化したものである．続いて，得られた関数データに対して関数主成分分析を適用し，平均関数 $\hat{\mu}(t)$ および固有関数 $\hat{v}_l(t)$ ，その係数 $\hat{\xi}_{il}$ を導出した．第 1 主成分から第 4 主成分の寄与率はそれぞれ 46.8%，16.7%，10.2%，9.1% で，第 4 主成分までで累積寄与率が 80% を超えたため，ここでは第 4 主成分までを採用した．図 7 は，関数主成分分析により得られた平均関数の推定値および，第 4 主成分までの固有関数と各年の係数を並べたものである．固有関数から，第 1 主成分は年間を通じた気温の高さを表しており，係数の推移を見ると年の経過に伴い緩やかに増加していることが分かる．また，第 2 主成分は冬が寒く（暖かく），夏が暑い（涼しい）年の特徴を表していると考えられる．

次に，図 7 右下に示す第 1～第 4 主成分の係数をそれぞれ時系列データとみなして，時系列解析を適用した．ここでは R の `forecast` パッケージを利用して自己回帰和分移動平均 (ARIMA) モデルを適用し，2018 年以降に対応する点の予測を行った．その結果を図 7 右下に併せて示している．この結果から，第 1 主成分である「年間を通じた気温の高さ」の係数の予測値としては若干の上昇傾向が見られ，第 2～第 4 主成分の係数の予測値も僅かな変化はあるが，いずれ

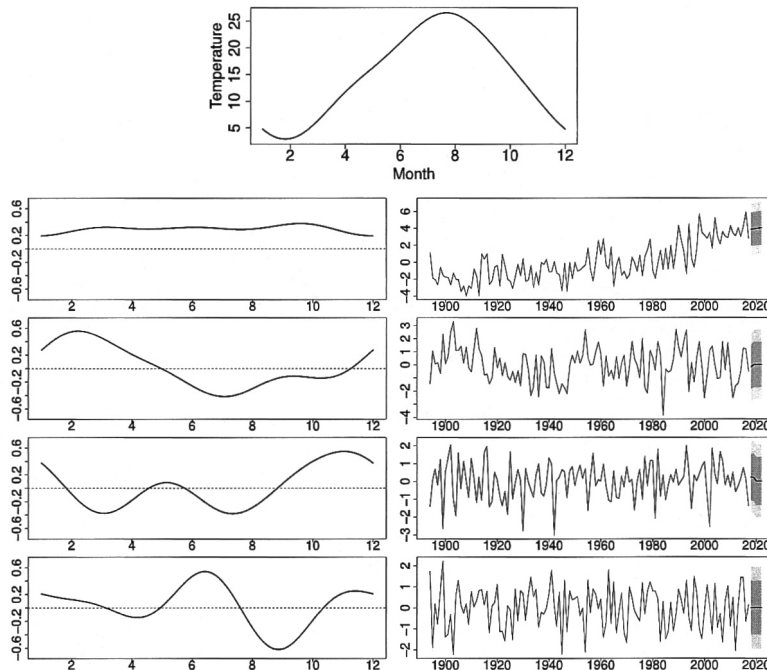


図 7. 気温の関数データ集合に対して関数主成分分析と関数時系列解析を適用し得られた結果. 上: 年間の平均関数. 左下: 上から第 1~第 4 主成分の固有関数. 横軸は時間(月)を表す. 右下: 各主成分に対応する各年の関数主成分. 横軸は個体(年)を表す. 右側にある太線および帯(濃色・淡色)は, それぞれ時系列解析によって得られた将来の係数の予測値および 80%, 95% 信頼帯を表す.

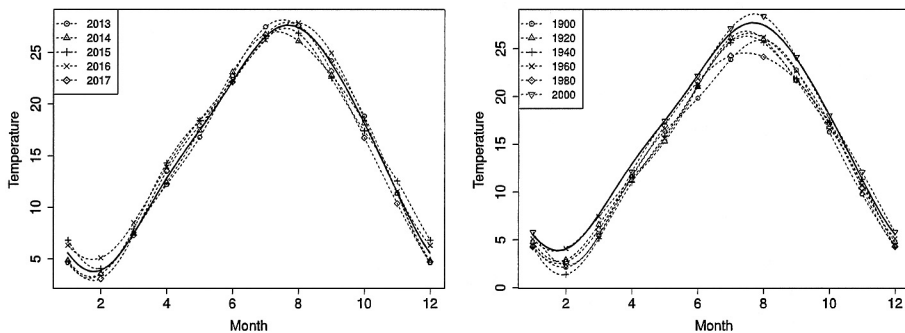


図 8. 左: 2013 年から 2017 年までの月別平均気温曲線(破線)と, 2018 年の予測気温曲線(実線). 右: 1900 年から 2000 年までの 20 年ごとの月別平均気温曲線(破線)と, 2020 年の予測気温曲線(実線).

も信頼帯を考慮に入れると有意な変化であるとは言い難い. 図 8 は, 過去の年の月別平均気温と, 予測した将来の年の月別平均気温を図示したものである. この結果から, 将来の予測値は直近の気温と比べても大きな変化は無いように見える.

4.3 その他の関数時系列解析手法

4.1 節で述べた方法以外にも、関数データの枠組みで自己回帰モデルを構築し、パラメータを推定することで関数時系列データを予測する方法などが考えられている。いま、関数時系列データ $x_i(t) (i \in \mathbb{Z})$ が与えられたとき、1 次の関数自己回帰モデル FAR(1) は、次で与えられる (Bosq, 2000)。

$$x_i(t) = \Phi(x_{i-1}(t)) + \varepsilon_i(t).$$

ここで、 Φ は関数の L^2 空間上への作用素、 $\varepsilon_i(t)$ は平均 0 の誤差関数とする。 Φ としては、実際は積分作用素 $\Phi(x)(t) = \int \varphi(t, s)x(s)ds$ が多く用いられる。すなわち、FAR(1) モデルは次で表される。

$$x_i(t) = \int \varphi(t, s)x_{i-1}(s)ds + \varepsilon_i(t).$$

この他に、関数データに対する ARCH モデルや ARMA モデルなども提案されている (Hörmann et al., 2013; van Delft and Eichler, 2018)。FAR モデルの一致性や収束率といった理論的な性質は、Horváth and Kokoszka (2012) にまとめられている。しかし、FAR モデルについては既存のパッケージでは実装が容易ではないこと (FAR(1) を推定するための R プログラムは Kokoszka and Reimherr (2017) に掲載されている) や、FAR(1) を一般化した p 次の関数自己回帰モデル FAR(p) に関する研究はあまり報告されていないという指摘がある (Aue et al., 2015)。

その一方で、4.1 節で述べた Hyndman and Ullah (2007) の方法は、FAR モデルに比べて定常性や特定の自己相関構造の仮定を必要とせずさまざまなデータに応用できる上、実装も非常に簡単である。しかし、理論的な最適性などの保証はされていない。Aue et al. (2015) は、Hyndman and Ullah (2007) の方法のように係数に対する単変量時系列解析を適用する方法では、ラグを考慮した主成分間の相関を考慮できず情報の損失が起こると指摘している。その上で、係数に対して単変量ではなく多変量時系列解析を適用する方法を提案し、予測量の一致性などの理論的性質を示している。

5. 空間関数データ解析

複数の地点で得られた観測値について、計測された位置情報を保持した形式のデータは空間データとよばれる。空間データ解析は、その位置情報を用いて、距離に応じたデータの関係性の分析や、未観測地点における予測を行うことなどが目的である。空間データ解析に関連する文献としては、例えば Gelfand et al. (2010)、間瀬 (2010)、堤・瀬谷 (2012)、瀬谷・堤 (2014) が挙げられる。また、時間と空間の変化に応じて計測された時空間データの解析法については、矢島・平野 (2012) による詳細な解説がある。

ここでは、複数の地点でそれぞれ経時測定データが観測された状況を考える。例えば、全国のさまざまな都市における年間の月別平均気温の推移や、農場内におけるさまざまな区域で生育されている同一品種の作物の草丈といった生育過程のデータなどが考えられる。このような、各地点で計測された経時測定データを関数データとして扱ったものは空間関数データとよばれ、空間関数データを対象とした分析は空間関数データ解析とよばれている。空間関数データを用いて、データが計測されていない地点の関数データの予測を行う問題などがこれに該当する。空間関数データ解析に関しては、Delicado et al. (2010) によって最近の研究がまとめられているほか、Menafoglio et al. (2013)、Aguilera-Morillo et al. (2017) などによる解析方法の研究がある。応用例として、Giraldo et al. (2011) は土壌の硬度を表す貫入抵抗性を、複数地点それぞれでさまざまな深度で計測したデータに対して適用し、任意の地点での、深度に対する

貫入抵抗性を表す曲線の予測を行っている．また，Hörmann and Kokoszka (2013)は空間関数データの推定量に対する漸近的な性質について述べている．

本節では，時点 t および d 次元空間上の地点 s で観測された空間ランダム関数 $\{X(s; t); s \in \mathcal{D} \subset \mathbb{R}^d, t \in \mathcal{T} \subset \mathbb{R}\}$ を考え， $\mathcal{D}$ 上の複数の地点それぞれで関数データが与えられたとき，新たな地点における関数データを予測する方法について紹介する．なお，ここでは，空間ランダム関数に 2 次定常性および等方性を仮定する．すなわち，任意のベクトル $h \in \mathcal{D}$ と， s に依存しない関数 $\mu(t), \sigma^2(t), C(h; t, u)$ を用いて，

$$E(X(s; t)) = \mu(t), \quad V(X(s; t)) = \sigma^2(t),$$

$$\text{Cov}(X(s; t), X(s + h; u)) = C(h; t, u)$$

と表すことができるとする．つまり，データの平均，分散関数は地点 s に依存せず，共分散関数は 2 地点間の距離 $r = \|h\|$ のみに依存するものとする．2 次定常性の仮定により， $\gamma(r; t) = \frac{1}{2}E[x(s + h; t) - x(s; t)]^2 = \frac{1}{2}V[x(s + h; t) - x(s; t)]$ も s に依存しない関数となる．これは時点 t における(セミ)バリオグラムとよばれている．空間関数データ解析についても，関数回帰モデルや関数時系列解析と同様に，スカラーデータに対する空間データ解析の枠組みへ帰着させることができる．

5.1 関数クリギング

いま， n 個の観測地点 s_i ($i = 1, \dots, n$) において関数データ $x(s_i; t)$ を得たとする．このとき，新たな地点 s_0 における関数データの予測値 $\hat{x}(s_0; t)$ は，次のように $x(s_i; t)$ ($i = 1, \dots, n$) の線形結合

$$(5.1) \quad \hat{x}(s_0; t) = \sum_{i=1}^n \lambda_i x(s_i; t)$$

によって表されるとし，未知パラメータ λ_i ($i = 1, \dots, n$) を推定する問題を考える (Goulard and Voltz, 1993; Giraldo et al., 2011)．ここで， $x_0(s; t)$ の 2 次定常性を保証するために， $\sum_{i=1}^n \lambda_i = 1$ という条件が必要となる．この条件を踏まえて，パラメータ λ_i ($i = 1, \dots, n$) は，ラグランジュの未定乗数法を用いて，次の積分予測 2 乗誤差の制約付き最小化問題によって推定される．

$$(5.2) \quad \min_{\lambda_1, \dots, \lambda_n, \mu} \int_{\mathcal{T}} E[x(s_0; t) - \hat{x}(s_0; t)]^2 dt + \mu \left(\sum_{i=1}^n \lambda_i - 1 \right).$$

ただし μ はラグランジュの未定乗数とする．2 次定常性および(5.1)式の仮定を利用することで，パラメータ $\lambda_1, \dots, \lambda_n, \mu$ の推定量は，次の方程式の解として与えられる．

$$\begin{pmatrix} \gamma(\|s_1 - s_1\|) & \cdots & \gamma(\|s_1 - s_n\|) & -1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(\|s_n - s_1\|) & \cdots & \gamma(\|s_n - s_n\|) & -1 \\ -1 & \cdots & -1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma(\|s_0 - s_1\|) \\ \vdots \\ \gamma(\|s_0 - s_n\|) \\ -1 \end{pmatrix}.$$

ここで， $\gamma(r) = \int_{\mathcal{T}} \gamma(r; t) dt$ はトレースバリオグラムとよばれる．また，積分予測 2 乗誤差は，次のように表すことができる．

$$\sigma_{s_0}^2 = \sum_{i=1}^n \lambda_i \gamma(\|s_i - s_0\|) + \mu.$$

パラメータの推定量や積分予測 2 乗誤差の最小値の導出については、スカラーデータに対するクリギングとはほぼ同様の流れで行うことができる。

実際には、トレースバリオグラム $\gamma(r)$ は未知であるため、これを推定する必要がある。トレースバリオグラムは、観測された 2 地点間の距離が r となるようなデータの分散を求めることで得られるが、距離がちょうど r となるようなデータは存在しないか、あったとしてもごく僅かである。そこで、2 点間の距離が r に近い点の集合 $N_r = \{(s_i, s_j) : \|s_i - s_j\| \approx r\}$ を考え、次のようにトレースバリオグラムを推定する。

$$(5.3) \quad \hat{\gamma}(r) = \frac{1}{2|N_r|} \sum_{i,j \in N_r} \int_{\mathcal{T}} [x(s_i; t) - x(s_j; t)]^2 dt.$$

ただし、 $|N_r|$ は集合 N_r の元の数とする。また、(5.3) 式の積分については、関数データ $x(s_i; t)$ が (2.1) 式と同様に基底関数展開 $x(s_i; t) = \mathbf{w}_i^T \phi(t)$ によって表されると仮定することで、次のように表せ、計算できるようになる。

$$\int_{\mathcal{T}} [x(s_i; t) - x(s_j; t)]^2 dt = (\mathbf{w}_i - \mathbf{w}_j)^T \Phi (\mathbf{w}_i - \mathbf{w}_j).$$

2 節で述べた関数化の結果は、関数クリギングの予測に影響を与える（このことは、3 節の回帰分析や 4 節の時系列解析でも同様である）。より具体的には、トレースバリオグラムの推定値 (5.3) は、基底関数展開に基づく関数化における平滑化パラメータや基底関数の個数といった調整パラメータに依存する。Giraldo et al. (2011) は、関数クリギングの予測精度の観点から、関数データ $x(s_i; t)$ を得る際の調整パラメータを選択するための基準として、次の関数交差検証法を提案した。

$$(5.4) \quad FCV = \sum_{i=1}^n \sum_{\alpha=1}^{n_i} \{x(s_i; t_{i\alpha}) - \hat{x}(s_i; t_{i\alpha})^{(-i)}\}^2.$$

ここで $\hat{x}(s_i; t_{i\alpha})^{(-i)}$ は、 n 地点のデータのうち第 i 地点を抜き取った $n-1$ 地点のデータを用いて、第 i 地点のデータを (5.1) 式で予測したものである。(5.4) 式を最小にするような、関数化における調整パラメータの値を、パラメータの最適値とみなして選択する。これにより、関数クリギングの予測の観点から、経時測定データの関数化が行われることになる。

5.2 適用例

地上気象観測装置が設置されている日本国内の観測所のうち、各都道府県で 1 か所の 2017 年の月別平均気温のデータ（気象庁ウェブサイトより取得）を用いる。ただし、沖縄県のデータに関しては、他都道府県に比べて大きく距離が離れているためここでは除外し、46 地点のデータを扱った。観測地点は、図 9 左の地図上に点で示した場所にある。各地点における月別平均気温のデータは、図 9 右で与えられる。ここでは、東京以外の 45 地点のデータに対して関数クリギングを適用することで、東京の年間気温曲線を予測することを試みる。

データの関数化については、R パッケージ `fda` を使い、基底関数としてフーリエ級数を用いた。また、関数クリギングについては R パッケージ `geofd` を適用した。関数データ集合に対して、トレースバリオグラム (5.3) の推定値を求めた結果が図 10 左である。図 10 左中の黒点は、集合 N_r の中心が等間隔に 13 個となるように選択されたものである。この黒点に対して、指数型を仮定してトレースバリオグラムの当てはめを行った。日本の都道府県は位置が細長く分布しているため、距離が大きく離れた 2 地点のデータが少ない。このこともあり、2 点間の距離が離れたデータについては当てはまりがよくない。推定されたトレースバリオグラム $\hat{\gamma}(r)$ を

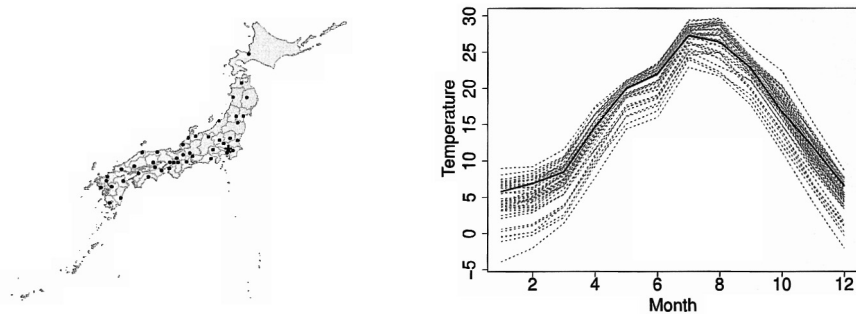


図 9. (左) データが観測された地上気象観測装置の位置。予測地点である東京は「+」で示している。(右) 各地点の月別平均気温のデータ。実線は東京のデータ。

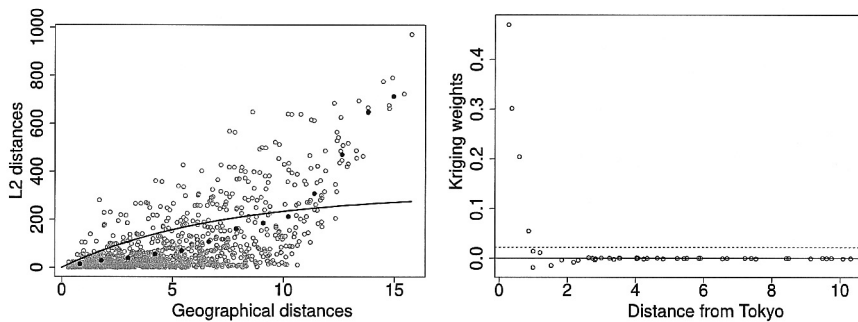


図 10. 関数クリギングの推定。左：各 2 地点間の距離に対するデータの分散(灰丸)および近傍でまとめられた点(黒丸)と、当てはめられたトレースバリオグラム(曲線)。右：東京からの距離(横軸)に対する、重みパラメータ λ_i の推定値(縦軸)。破線は全地点に対する重み λ_i を全て等しくした $(1/45)$ としたものの。

用いて得られた、(5.1)式の重みパラメータ λ_i の推定値を図 10 右に示す。この図より、東京から近い地点ではより強い重みが与えられている一方で、そこからやや離れた地点の重みはほぼ 0 であり、それらの情報はほぼ加味されていないことが分かる。(5.1)式に基づいて東京の気温 $x(s_0, t)$ を予測した結果を、図 11 に示す。破線で示した、45 地点の気温の平均関数を用いた予測に比べて、よい予測ができていることが分かる。なお、関数クリギングを適用するにあたり、(5.4)式の関数交差検証法を用いて基底関数の個数を選択した結果、3 個が選択された。

6. まとめと今後の展望

本論文では、農業データの分析に有用であると考えられる関数データ解析について、さまざまなデータ形式や分析目的に応じた手法を紹介した。ここでは特に、関数データに基づく回帰分析および時系列解析、空間データ解析について説明した。加えて、それぞれの手法に対して実データを含む適用例を通して、各手法がどのようなデータに対して、どのように分析できるのかについて議論した。

3 節では、気温といった環境情報を時間の関数データとみなし、これらを説明変数として関数回帰モデルを適用することで、トマトの収量の予測を試みた。しかし、作物の成長に関わる

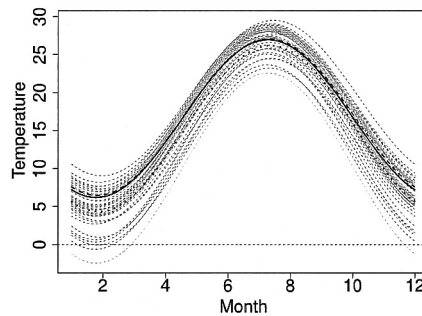


図 11. 関数クリギングによる東京の気温の予測結果. 灰破線は 45 地点の実測値, 黒破線は東京の気温の実測値を関数化したもの, 黒実線は関数クリギングによる東京の予測気温曲線, 黒点線は 45 地点の平均関数を表している.

環境要因は、気温や日射量以外のものも含めて複雑に絡み合っていると考えられるため、作物成長のメカニズムに応じたより適切なモデルを構築する必要があると考えられる。今回用いたモデルに関しても、収量に関連すると考えられる環境情報やその期間を主観で与えたが、これらについてもデータから推定、選択するための方法を適用することが望ましい。4 節では、関数時系列解析について説明し、日本の気温の年間推移の予測に用いた。今回は予測結果だけを示したが、より厳密には予測精度などの検証を行う必要がある。5 節では、日本全国の気温のデータに対して関数クリギングを適用した。しかし、日本の各都道府県の空間データに対して、関数クリギングの適用において用いた 2 次定常性の仮定は強いと考えられる。この仮定を緩めた方法の検討が必要である。

そのほかの課題として、2 節で述べた関数データ化の方法によるその後の関数データ解析への影響や、関数データ解析を見越した関数化の方法について言及された研究はあまり報告されていない。3.1 節の最後でも述べたように、関数回帰モデルに対して、2 段階推定では推定量にバイアスが生じてしまうが、その一方で 1 段階法では計算コストが問題になる。関数クリギングに対する関数交差検証法 (5.4) のように、関数データ化およびその後の分析までを包括したモデル評価を行うことでも、関数データ解析に適した関数データを得られると考えられるが、関数回帰モデルなどに対してはやはり計算コストが問題になることが予想される。また、関数主成分表現 (2.3) における、固有関数の個数の選択についても検討の余地があると考えられる。Yao et al. (2005a) や Crainiceanu et al. (2009) の方法では、固有関数の数を交差検証法や AIC、累積寄与率を使って決定しており、これによって良い推定結果が得られているようであるが、固有関数の個数による、関数回帰モデルなどへの推定精度の影響の理論的評価についても、今後調査すべき課題であると思われる。

本論文で述べた方法以外にも、関数データ解析についてはさまざまなものが提案されている。例えば、関数データに対する判別分析 (Hastie et al., 1995; James and Hastie, 2001)、クラスター分析 (James and Sugar, 2003; Misumi et al., 2019)、統計的仮説検定 (Fan and Lin, 1998; Hall and Van Keilegom, 2007) などの研究がある。さらに、関数データに対する微分を対象とした分析 (Ramsay, 1996) や、微分方程式の解の導出への応用 (Ramsay et al., 2007) など、関数データならではの研究もある。1 節でも述べたように、関数データ解析の農業分野への応用はまだ少ない。今後、農業分野の様々な経時測定データに対する関数データ解析の適用が期待される。

謝 辞

本研究を遂行するにあたり、データを提供していただいた東馬場農園(兵庫県神戸市)に感謝いたします。また、本論文の執筆にあたり、文献調査等に協力いただいた横浜市立大学の三角俊裕氏に感謝いたします。本稿の改訂にあたり、大変貴重なご指摘ならびに適切なご意見をいただいた査読者の方々へ御礼申し上げます。本研究は、若手研究(B)(課題番号:16K16020)および戦略的創造研究推進事業さきがけ(課題番号:JPMJPR16O6)の支援によって行われたものです。

参 考 文 献

- Aguilera-Morillo, M. C., Durbán, M. and Aguilera, A. M. (2017). Prediction of functional data with spatial dependence: A penalized approach, *Stochastic Environmental Research and Risk Assessment*, **31**, 7-22.
- 安道知寛, 井元清哉, 小西貞則 (2001). 動径基底関数ネットワークに基づく非線形回帰モデルとその推定, *応用統計学*, **30**, 19-35.
- 荒木由布子, 小西貞則 (2004). 動径基底関数展開に基づく関数回帰モデリング, *応用統計学*, **33**, 243-256.
- Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2009a). Functional logistic discrimination via regularized basis expansions, *Communications in Statistics. Theory and Methods*, **38**, 2944-2957.
- Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2009b). Functional regression modeling via regularized Gaussian basis expansions, *Annals of the Institute of Statistical Mathematics*, **61**, 811-833.
- Aue, A., Norinho, D. D. and Hörmann, S. (2015). On the prediction of stationary functional time series, *Journal of the American Statistical Association*, **110**, 378-392.
- Besse, P. and Ramsay, J. (1986). Principal components analysis of sampled functions, *Psychometrika*, **51**, 285-311.
- Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*, Oxford University Press, New York.
- Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*, Springer, New York.
- Crainiceanu, C. M., Staicu, A. M. and Di, C. Z. (2009). Generalized multilevel functional regression, *Journal of the American Statistical Association*, **104**, 1550-1561.
- Cuevas, A. (2014). A partial overview of the theory of statistics with functional data, *Journal of Statistical Planning and Inference*, **147**, 1-23.
- de Boor, C. (2001). *A Practical Guide to Splines*, Springer, New York.
- Delicado, P., Giraldo, R., Comas, C. and Mateu, J. (2010). Statistics for spatial functional data: Some recent contributions, *Environmetrics*, **21**, 224-239.
- Donoho, D. and Johnstone, J. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, **81**, 425-455.
- Fan, J. and Lin, S.-K. (1998). Test of significance when data are curves, *Journal of the American Statistical Association*, **93**, 1007-1021.
- Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models, *Statistics and Its Interface*, **1**, 179-195.
- Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*, Springer, New York.
- Ferraty, F., Mas, A. and Vieu, P. (2007). Nonparametric regression on functional data: Inference and practical aspects, *Australian & New Zealand Journal of Statistics*, **49**, 267-286.
- Fuchs, K., Scheipl, F. and Greven, S. (2015). Penalized scalar-on-functions regression with interaction term, *Computational Statistics & Data Analysis*, **81**, 38-51.
- Fujii, T. and Konishi, S. (2006). Nonlinear regression modeling via regularized wavelets and smoothing parameter selection, *Journal of Multivariate Analysis*, **97**, 2023-2033.

- Gelfand, A. E., Diggle, P., Guttorp, P. and Fuentes, M. (2010). *Handbook of Spatial Statistics*, CRC Press, Boca Raton, Florida.
- Gertheiss, J., Maity, A. and Staicu, A.-M. (2013). Variable selection in generalized functional linear models, *Stat*, **2**, 86-101.
- Giraldo, R., Delicado, P. and Mateu, J. (2011). Ordinary kriging for function-valued spatial data, *Environmental and Ecological Statistics*, **18**, 411-426.
- Goulard, M. and Voltz, M. (1993). Geostatistical interpolation of curves: A case study in soil science, *Geostatistics Tróia '92*, 805-816, Springer, Dordrecht.
- Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall/CRC, London.
- Hall, P. and Van Keilegom, I. (2007). Two-sample tests in functional data analysis starting from discrete data, *Statistica Sinica*, **17**, 1511-1531.
- Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.
- Harezlak, J., Coull, B., Laird, N., Magari, S. and Christiani, D. (2007). Penalized solutions to functional regression problems, *Computational Statistics & Data Analysis*, **51**, 4911-4925.
- Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models, *Journal of the Royal Statistical Society Series B*, **55**, 757-796.
- Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis, *The Annals of Statistics*, **23**, 73-102.
- Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalization*, Chapman & Hall/CRC, Boca Raton, Florida.
- Hoover, D., Rice, J., Wu, C. and Yang, L. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data, *Biometrika*, **85**, 809-822.
- Hörmann, S. and Kokoszka, P. (2012). Functional time series, *Handbook of Statistics*, Vol. 30, 157-186, Elsevier, Amsterdam.
- Hörmann, S. and Kokoszka, P. (2013). Consistency of the mean and the principal components of spatially distributed functional data, *Bernoulli*, **19**, 1535-1558.
- Hörmann, S., Horváth, L. and Reeder, R. (2013). A functional version of the ARCH model, *Econometric Theory*, **29**, 267-288.
- Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*, Springer, New York.
- Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, Wiley, Chichester.
- Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach, *Computational Statistics & Data Analysis*, **51**, 4942-4956.
- 井元清哉, 小西貞則 (1999). B-スプラインによる非線形回帰モデルと情報量規準, *統計数理*, **47**, 359-373.
- Jacques, J. and Preda, C. (2014). Functional data clustering: A survey, *Advances in Data Analysis and Classification*, **8**, 231-255.
- James, G. (2002). Generalized linear models with functional predictors, *Journal of the Royal Statistical Society Series B*, **64**, 411-432.
- James, G. and Hastie, T. (2001). Functional linear discriminant analysis for irregularly sampled curves, *Journal of the Royal Statistical Society Series B*, **63**, 533-550.
- James, G. and Silverman, B. (2005). Functional adaptive model estimation, *Journal of the American Statistical Association*, **100**, 565-576.
- James, G. and Sugar, C. (2003). Clustering for sparsely sampled functional data, *Journal of the American Statistical Association*, **98**, 397-408.
- James, G., Hastie, T. and Sugar, C. (2000). Principal component models for sparse functional data, *Biometrika*, **87**, 587-602.
- James, G., Wang, J. and Zhu, J. (2009). Functional linear regression that's interpretable, *The Annals*

- of *Statistics*, **37**, 2083-2108.
- 川野秀一, 松井秀俊, 廣瀬慧 (2018). 『スパース推定法による統計モデリング』, 共立出版, 東京.
- 茅野光範, 小西貞則, 平川英樹, 久原哲 (2006). 正則化基底展開法に基づく関数主成分分析とその応用, 応用統計学, **35**, 1-17.
- Kayano, M., Matsui, H., Yamaguchi, R., Imoto, S. and Miyano, S. (2016). Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to timedependent biomarker detection, *Biostatistics*, **17**, 235-248.
- Kim, K., Sentürk, D. and Li, R. (2011). Recent history functional linear models for sparse longitudinal data, *Journal of Statistical Planning and Inference*, **141**, 1554-1566.
- 北川源四郎 (2005). 『時系列解析入門』, 岩波書店, 東京.
- 北川源四郎, 田中勝人, 川崎能典 (2016). 『時系列分析ハンドブック』, 朝倉書店, 東京.
- Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*, CRC Press, Boca Raton, Florida.
- 小西貞則 (2010). 『多変量解析入門』, 岩波書店, 東京.
- Lee, E. R. and Park, B. U. (2012). Sparse estimation in functional linear regression, *Journal of Multivariate Analysis*, **105**, 1-17.
- Lian, H. (2013). Shrinkage estimation and selection for multiple functional regression, *Statistica Sinica*, **23**, 51-74.
- Malfait, N. and Ramsay, J. (2003). The historical functional linear model, *The Canadian Journal of Statistics*, **31**, 115-128.
- Manrique, T., Crambes, C. and Hilgert, N. (2018). Ridge regression for the functional concurrent model, *Electronic Journal of Statistics*, **12**, 985-1018.
- 間瀬茂 (2010). 『地球統計学とクリギング法・R と geoR によるデータ解析』, オーム社, 東京.
- Matsui, H. and Konishi, S. (2011). Variable selection for functional regression models via the L1 regularization, *Computational Statistics & Data Analysis*, **55**, 3304-3310.
- Matsui, H., Kawano, S. and Konishi, S. (2009). Regularized functional regression modeling for functional response and predictors, *Journal of Math-for-Industry*, **1**, 17-25.
- Matsui, H., Araki, T. and Konishi, S. (2011). Multiclass functional discriminant analysis and its application to gesture recognition, *Journal of Classification*, **28**, 227-243.
- 松井秀俊, 三角俊裕, 横溝孝明, 小西貞則 (2016). 非線形混合効果モデルに基づく関数データクラスタリング, 応用統計学, **45**, 25-45.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Model*, Chapman & Hall/CRC, London.
- Mclean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F. and Ruppert, D. R. (2014). Functional generalized additive models, *Journal of Computational and Graphical Statistics*, **23**, 249-269.
- Menafoglio, A., Secchi, P. and Dalla Rosa, M. (2013). A universal kriging predictor for spatially dependent functional data of a Hilbert space, *Electronic Journal of Statistics*, **7**, 2209-2240.
- Misumi, T., Matsui, H. and Konishi, S. (2019). Multivariate functional clustering and its application to typhoon data, *Behaviormetrika* (to appear).
- Morris, J. S. (2015). Functional regression, *Annual Review of Statistics and Its Application*, **2**, 321-359.
- Müller, H. and Yao, F. (2008). Functional additive models, *Journal of the American Statistical Association*, **103**, 1534-1544.
- Müller, H., Wu, Y. and Yao, F. (2013). Continuously additive models for nonlinear functional regression, *Biometrika*, **100**, 607-622.
- Ogden, R. T., Miller, C. E., Takezawa, K. and Ninomiya, S. (2002). Functional regression in crop lodging assessment with digital images, *Journal of Agricultural, Biological, and Environmental Statistics*, **7**, 389-402.
- Rachdi, M. and Vieu, P. (2007). Nonparametric regression for functional data: Automatic smoothing parameter selection, *Journal of Statistical Planning and Inference*, **137**, 2784-2801.
- Ramsay, J. (1996). Principal Differential Analysis, *Journal of the Royal Statistical Society Series B*,

- 58, 495-508.
- Ramsay, J. and Dalzell, C. (1991). Some tools for functional data analysis, *Journal of the Royal Statistical Society Series B*, **53**, 539-572.
- Ramsay, J. and Silverman, B. (2002). *Applied Functional Data Analysis: Methods and Case Studies*, Springer, New York.
- Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*, 2nd ed., Springer, New York.
- Ramsay, J. O., Hooker, G., Campbell, D. and Cao, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach, *Journal of the Royal Statistical Society Series B*, **69**, 741-796.
- Ramsay, J. O., Hooker, G. and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*, Springer, New York.
- Reiss, P. T. and Ogden, R. T. (2007). Functional principal component regression and functional partial least squares, *Journal of the American Statistical Association*, **102**, 984-996.
- Reiss, P. T. and Ogden, R. T. (2010). Functional generalized linear models with images as predictors, *Biometrics*, **66**, 61-69.
- Reiss, P. T., Goldsmith, J., Shang, H. L. and Ogden, R. T. (2017). Methods for scalar-on-function regression, *International Statistical Review*, **85**, 228-249.
- Rice, J. and Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves, *Biometrics*, **57**, 253-259.
- Rossi, F., Delannay, N., Conan-Guez, B. and Verleysen, M. (2005). Representation of functional data in neural networks, *Neurocomputing*, **64**, 183-210.
- Scheipl, F. (2018). *CRAN Task View: Functional Data Analysis*, <https://cran.r-project.org/web/views/FunctionalData.html>.
- Şentürk, D. and Müller, H. (2010). Functional varying coefficient models for longitudinal data, *Journal of the American Statistical Association*, **105**, 1256-1264.
- 瀬谷創, 堤盛人 (2014). 『空間統計学：自然科学から人文・社会科学まで』, 朝倉書店, 東京.
- Shang, H. L. (2014). A survey of functional principal component analysis, *Advances in Statistical Analysis*, **98**, 121-142.
- Shi, J. Q. and Choi, T. (2011). *Gaussian Process Regression Analysis for Functional Data*, CRC Press, Boca Raton, Florida.
- 下川真由子, 水田正弘, 佐藤義治 (2000). 関数データ解析における関数回帰分析の拡張, *応用統計学*, **29**, 27-39.
- 高井啓二, 星野崇宏, 野間久史 (2016). 『欠測データの統計科学』, 岩波書店, 東京.
- 堤盛人, 瀬谷創 (2012). 応用空間統計学の二つの潮流：空間統計学と空間計量経済学, *統計数理*, **60**, 3-25.
- Ullah, S. and Finch, C. F. (2013). Applications of functional data analysis: A systematic review, *BMC Medical Research Methodology*, **13**:43.
- Usset, J., Staicu, A. M. and Maity, A. (2016). Interaction models for functional regression, *Computational Statistics and Data Analysis*, **94**, 317-330.
- van Delft, A. and Eichler, M. (2018). Locally stationary functional time series, *Electronic Journal of Statistics*, **12**, 107-170.
- Wang, J.-L., Chiou, J.-M. and Mueller, H.-G. (2015). Review of functional data analysis, *arXiv preprint arXiv:1507.05135*.
- Wei, P., Tang, H. and Li, D. (2014). Functional logistic regression approach to detecting gene by longitudinal environmental exposure interaction in a case-control study, *Genetic Epidemiology*, **38**, 638-651.
- Yao, F. and Müller, H.-G. (2010). Functional quadratic regression, *Biometrika*, **97**, 49-64.
- Yao, F., Müller, H.-G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A. and Vogel, J. S. (2003). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate, *Biometrics*, **59**, 676-685.

- Yao, F., Müller, H. and Wang, J. (2005a). Functional data analysis for sparse longitudinal data, *Journal of the American Statistical Association*, **100**, 577-590.
- Yao, F., Müller, H. and Wang, J. (2005b). Functional linear regression analysis for longitudinal data, *The Annals of Statistics*, **33**, 2873-2903.
- 矢島美寛, 平野敏弘 (2012). 時空間大規模データに対する統計的解析法, *統計数理*, **60**, 57-71.
- Zhang, X. and Wang, J. L. (2016). From sparse to dense functional data and beyond, *The Annals of Statistics*, **44**, 2281-2321.
- Zhao, Y., Chen, H. and Ogden, R. T. (2015). Wavelet-based weighted lasso and screening approaches in functional linear regression, *Journal of Computational and Graphical Statistics*, **24**, 655-675.
- Zhu, H., Yao, F. and Zhang, H. (2014). Structured functional additive regression in reproducing kernel Hilbert spaces, *Journal of the Royal Statistical Society Series B*, **76**, 581-603.

Statistical Modeling via Functional Data Analysis

Hidetoshi Matsui^{1,2}

¹Faculty of Data Science, Shiga University

²Japan Science and Technology Agency, PRESTO

The development of technologies for measurement devices enables us to obtain high-throughput data. In agriculture in particular, more and more repeated measured data with respect to time, position, depth, etc. for individual have been obtained. However, sometimes it is difficult to analyze such data for the following three reasons. First, they usually contain observational errors, making it difficult to reveal unknown structures. Second, repeated measurements lead to the high-dimensional data, to which it is hard to apply traditional statistical analyses. Third, the observed numbers and time points may differ for individual. This leads to difficulties in applying the traditional multivariate analysis. Functional data analysis is one of the most useful tools for resolving the aforementioned problems. In this approach, longitudinal data are expressed as a smooth function for each individual, and then information is drawn from the collection of functions. In this paper, we introduce several techniques for functional data analysis that would be useful for analyzing agricultural data. In particular, we describe regression analysis, time series analysis, and spatial data analysis for functional data. For each method, we provide an empirical example in which the technique is used to analyze real data.