

Mathematica を用いての NDVI に対する 非線形カーブフィッティング

日本大学・生物資源科学部 五十嵐 正夫 (Masao IGARASHI)*
日本大学大学院・生物資源科学研究科 塩谷 悠 (Haruka SHIOTANI)*
東京情報大学・総合情報学部 布広 永示 (Eiji NUNOHIRO)**
東京情報大学・総合情報学部 朴 鍾杰 (Jong Geol Park)**
College of Bioresource Sciences, Nihon University*
Tokyo University of Information Sciences**

1 はじめに

農学・生命系の学部では、非線形なデータに対してそれに良く適合する曲線が必要とされる場合が多い。例えば植物や動物の生長や温度データなどは、多くの例において非線形な曲線となる。この場合、次の3つの問題点が生じる。

- (1) 農学・生命系の学生は、曲線を求めるための反復法やそのプログラムを書くことに慣れていないために、データの解析に時間がかかる場合がある。
- (2) 反復法を用いる場合、反復停止則の設定方法に慣れていないために反復が停止せず、必要とする非線形カーブが得られない場合がある。
- (3) データと得られた曲線の適合度合いの統一的な判定がないために、それに対して、的確な判断が下せない場合がある。

そこで本稿では、Mathematica(Ver.7)のFindFitの関数を用いて与えられた非線形データに対するモデル曲線、そのモデルの係数決定、及び得られた曲線の精度推定について考察してみる。例題としてはNDVI、地表の温度、海水の表面温度を取り上げる¹。

2 準備

ここでは次の4つの事柄についてあらかじめ説明しておく。

- (1) Mathematica Ver.7 の FindFit 関数
- (2) 重相関係数 R^2 とここでの相対誤差 CR
- (3) ここで利用するデータの種類

¹この研究は平成20年度の「私立大学戦略的研究基盤形成支援事業、東京情報大学、研究代表、新沼勝利」の支援を受けた。

(4) データ解析の目的

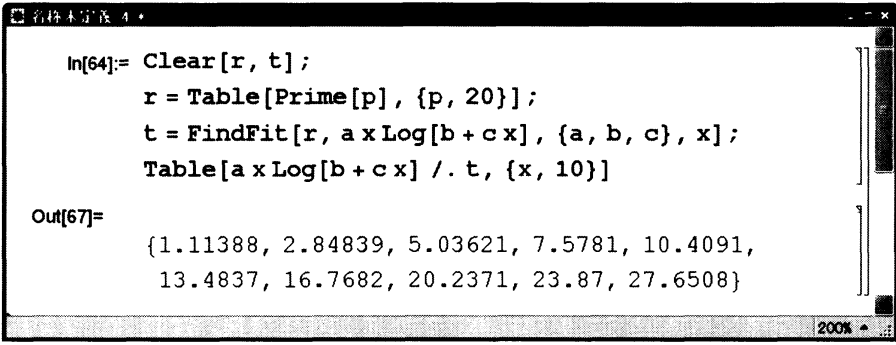
FindFit の一つの利用例を図1に示す。リストのrとtを省略するにはそれぞれを%に置き換えればよい。

2.1 FindFit

まずデータr(素数)を用意して、それに適合すると推定されるモデル関数、ここではxを変数としa,b,cを定数とする次の関数を考えている。

$$ax \log_e(b + cx) \quad (1)$$

モデル関数の係数が決定した後、FindFit関数を用いてのデータの推定値が図1のTableのところでも出力されている。



```

In[64]:= Clear[r, t];
r = Table[Prime[p], {p, 20}];
t = FindFit[r, a x Log[b + c x], {a, b, c}, x];
Table[a x Log[b + c x] /. t, {x, 10}]

Out[67]:=
{1.11388, 2.84839, 5.03621, 7.5781, 10.4091,
13.4837, 16.7682, 20.2371, 23.87, 27.6508}
  
```

図1: FindFit の利用法

2.2 相対誤差とCR

データが与えられて、それに適合する回帰直線や重回帰直線に対する適合度合いの尺度としては様々な指標がある。なかでも(重)相関係数 R^2 は良く用いられる。しかしながら非線形な曲線に対して、その指標は適用できないとされている。ここでは(3)式のような数値解析における相対誤差と類似した指標を導入する。

まず、真値を y 、その近似値を \hat{y} とする。相対誤差REは

$$RE = \left| \frac{y - \hat{y}}{y} \right| \quad \text{または} \quad \left| \frac{y - \hat{y}}{\hat{y}} \right| \quad (2)$$

で定義される[3]。 y と \hat{y} は近接した値であり、例えば $RE=0.001$ となれば $-\log_{10}(0.001) = 3$ であるから、近似値 \hat{y} と真値 y は10進数でおおよそ3桁一致することになる。そこで計算途中で分母がゼロとなったときの危険を避けるために、分母を $\max(|y|, |\hat{y}|)$ で置き換えることにする。

m 個のデータを $\{y_k\}, k = 1, 2, 3, \dots, m$ として, それに対応する推定値を $\{\hat{y}_k\}, k = 1, 2, 3, \dots, m$ とする. それらの値に対して CR を次のように定義する. この CR はデータと推定値の相対誤差の平均値と見ることができる.

$$CR = \frac{\sum_{k=1}^m \frac{|y_k - \hat{y}_k|}{\max(|y_k|, |\hat{y}_k|)}}{m} \quad (3)$$

2.3 データ

エルニーニョ現象, ラニーニャ現象, あるいは森林の面積やバイオマス評価を目的として, 衛星から様々なデータが送られて来ている. その中には正規化植生指数 (NDVI), 海面温度 (SST), 陸地温度 (LST) などがある. ここではそれらの3種類のデータを取り扱うことにする. データの観測地点を次の図2に示す. NDVIの観測点は1点, 丸印のつ

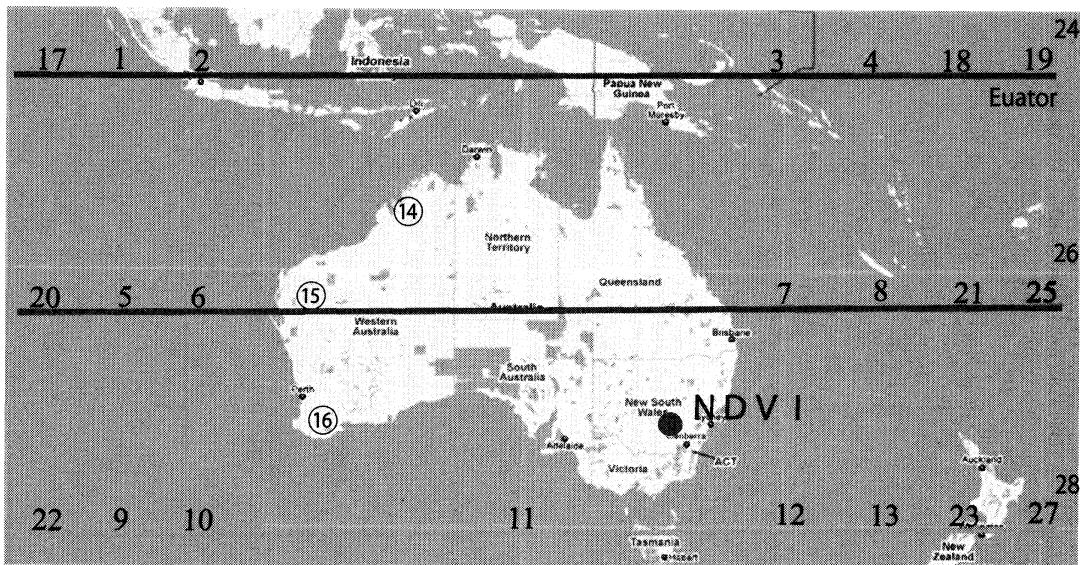


図2: Observation points of NDVI

いた⑭, ⑮, ⑯は陸地温度の観測点, それ以外は海面温度の観測地点である. ここで用いているデータは1984年から1999年までのデータである. データは8bits情報(0から255)として衛星から受け取っている [1].

3 NDVIモデルと結果

このデータの季節変化と作物の作付け, 収穫時期を考慮して, 試行錯誤的に次のモデルを利用することにした [2].

$$f(x) = a_1 + a_2x + a_3x^2 + (a_4 + a_5x) \sin\left(\frac{\pi}{n}x + a_6\right), \quad 0 \leq x \leq 52 \quad (4)$$

変数 x はデータを週単位で 1 年間追跡するのでその変域は $0 \leq x \leq 52$ とした。 a_1, a_2, \dots, a_6 は FindFit² で求める未知数、 n は 10 から 30 までパラメータとして変化させ、前に定義した CR が最小となる n を採用した。

3.1 Mathematica のプログラム

Mathematica は Ver.7 を使い、プログラムは図 3 のように FindFit を用いた。この例では 1999 年の NDVI を用いている。

```

In[1]:= Clear[q, g, sgn, s, cr] (*99*)
q = NDVI[99] = {196, 183, 176, 175, 174, 174, 174, 173, 173, 173, 172,
  171, 184, 196, 196, 196, 197, 197, 198, 198, 197, 196, 195, 194,
  193, 196, 200, 203, 206, 207, 208, 210, 211, 212, 211, 210, 210,
  209, 209, 209, 206, 203, 199, 194, 187, 179, 178, 178, 177, 176, 175, 175};
q = Log[q];
t = Length[q];
f1 = ListPlot[q, PlotRange -> {5, 6}];
Do[{Clear[g],
  g[x_] = (a1 + a2 x) Sin[ $\frac{\pi x}{\pi} + a3$ ] + a4 x2 + a5 x + a6 /.
  FindFit[q, (a1 + a2 x) Sin[ $\frac{\pi x}{\pi} + a3$ ] + a4 x2 + a5 x + a6,
  {a1, a2, a3, a4, a5, a6}, {x}],
  p = Table[g[x], {x, 1, t}],
  s = Abs[p - q] / Max[Abs[p], Abs[q]],
  sgn = Sum[s[[i]], {i, 1, t}] / t,
  Print["n=", pi, " REM=", sgn], cr[pi] = sgn},
{pi, 10, 30}];
r0 = Table[cr[i], {i, 10, 30}];
r1 = Table[k, {k, 10, 30}];
r10 = Transpose[{r1, r0}];
ListPlot[r10, PlotRange -> {0.0, 0.006}]
(*ListPlot[Table[cr[i], {i, 10, 30}], PlotRange -> {0.001, 0.01}]*)
f2 = Plot[g[x], {x, 1, t}];
Show[f1, f2]

```

図 3: サンプルプログラム

結果として x 軸が n 、 y 軸が (3) 式の CR となっている図 4 が出力される。この図 4 より $n = 13$ の時に CR が最小となることが読みとれる。この CR は 99 年の NDVI の結果なので R_{99} と書くことにする。

3.2 R^2 との比較

非線形曲線近似に対してのデータの当てはまり具合の指標は少ないようである [?]。便宜的に線型近似に対して用いられる次の R_j^2 (j 年の重相関係数) が代用されているよう

²最小 2 乗法と Newton 法で構成されていると思われる

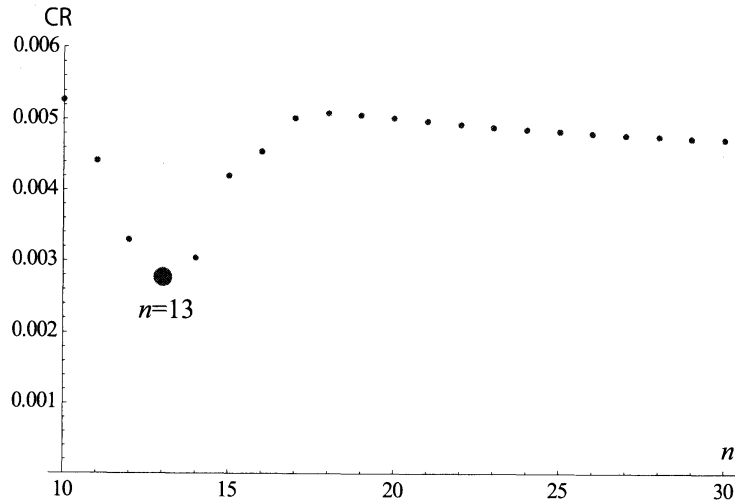


図 4: n と CR 値の図

である. ここで $\{y_{i,j}\}$ は i 週, j 年のデータ, \bar{y}_j は j 年の平均値, $\hat{y}_{i,j}$ は推定値である.

$$R_j^2 = \frac{\sum_{i=1}^{52} (\log_e \hat{y}_{i,j} - \log_e \bar{y}_j)^2}{\sum_{i=1}^{52} (\log_e y_{i,j} - \log_e \bar{y}_j)^2} = \frac{j \text{ 年の残差の平方和}}{j \text{ 年の平方総和}} \quad (5)$$

そこで, $n=10$ から 30 までのうち CR が最小となる n と R^2 が最大となる n の比較を表にすると表 1 となる. 便宜的に用いられている R^2 の評価と CR の評価が良く一致して

表 1: CR の最小値を与える n と R^2 の最大値を与える n

年	'82	'83	'84	'85	'86	'87	'88	'89	'90
最小 CR を与える n	11	12	11	11	11	11	10	12	12
最大 R^2 を与える n	11	11	11	11	11	11	10	12	12

年	'91	'92	'93	'94	'95	'96	'97	'98	'99
最小 CR を与える n	18	11	16	×	16	16	21	12	12
最大 R^2 を与える n	17	11	16	×	15	16	21	12	13

注) 94 年はデータの取れなかった年である.

いることがわかる.

4 CR の海面と陸地の温度への適用

観測点 20 における 99 年の海面の温度に関して, CR と R^2 の週変化を図 5 に示す. 海面の温度に関する CR の最小値を与える n と R^2 の最大値を与える n が一致している

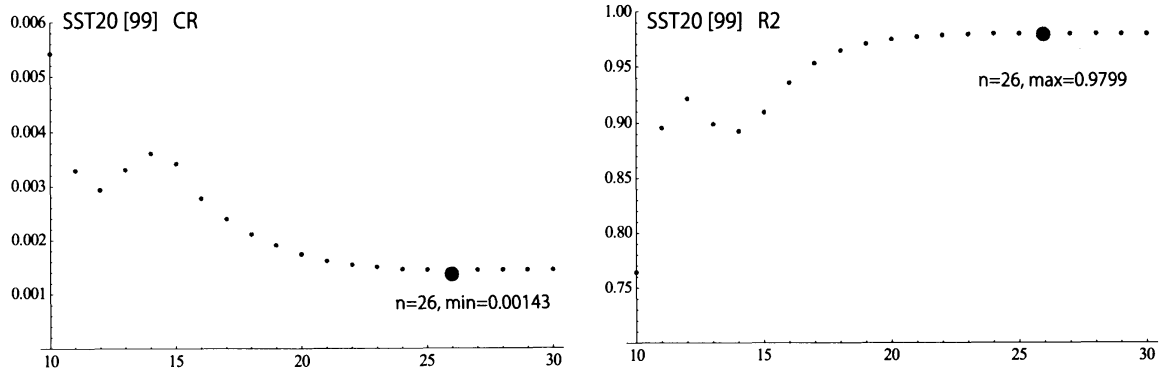


図 5: SST20 における n と CR, n と R^2 の変化と最良 n

ことがわかる. 次に観測点 15 における 99 年の地表温度に関して CR と R^2 の週変化を図 6 に示す. この場合でも CR の最小値を与える n と R^2 の最大値を与える n は大きく異なっているが, 図からもわかるように R^2 の値は $n=18$ で 0.981, それ以降 $n=30$ まで数値 0.98 の部分の変化はない. 従って $n=18$ でも $n=30$ でも R^2 の値は同程度と見ることができる.

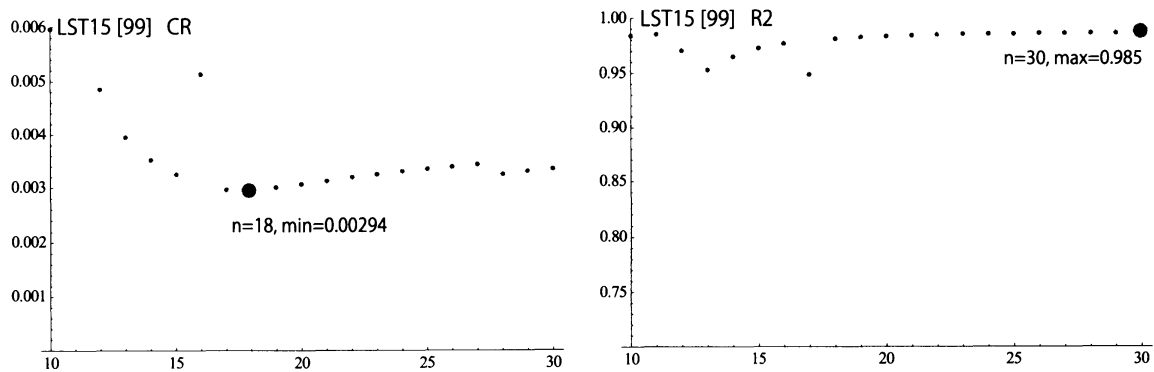


図 6: LST15 における n と CR, n と R^2 の変化と最良 n

5 結論

ここでは直接非線形連立方程式を解くことなく, Mathematica の関数 FindFit を用いて非線形モデルに対する係数を求めた. 提案したモデルは陸地や海面の温度データに対しても適用したが, 特定の n を除いて, 係数が求まらないと言った事例は生じなかった. この意味で FindFit は安定していると推定され, 数値計算や数式処理に不慣れが学生にとって満足度の高い関数を思われる.

また時系列データに対する非線形モデルの適合度合いに関しては, 従来 (6) 式の RMS (最小 2 乗) 評価が用いられていたようである [4]. ここでは (3) 式の相対誤差平均和を

考え、データの大小によらない評価指標の導入を試み、その数値的実験を Mathematica を用いて行った。RMS が絶対評価であるのに対して CR は相対評価になっている。このことは、例えば A データに対する非線形曲線 f_A の CR_A と、B データに対する非線形曲線 f_B の CR_B を比較することにより、その適合度合いの比較ができることを意味する。

$$\text{RMS} = \sqrt{\frac{1}{m} \sum_{k=1}^m (y_k - \hat{y}_k)^2} \quad (6)$$

更に、表 1 に示されているように、従来用いられている指標 R^2 における最適 n と CR の最適 n がほぼ一致することより、CR 指標は本実験データにおいては満足できるものであった。

ここでは 6 変数のモデルを設定したが、より変数の少ないモデルを考える必要がある。係数の数に配慮して、データと推定値の「適合度合い」を考える必要があると考えている。

参考文献

- [1] Kenneth J. Mackin, *et al*, Land Surface Cover Classification by Soft Computing Methods using MODIS Satellite Data, INFORMATION, 13,3(B), pp.1013-1018(2010)
- [2] Masao IGARASHI, *et al*, A Fitness Criterion on Finding Nonlinear Curve for the NDVI Data, INFORMATION, 13,3(A)601-606(2010)
- [3] S.D. コンテ, C. ドボアー著, 吉澤正訳, 電子計算機による数値解析と算法入門, ブレイン図書出版, 1980.
- [4] 松葉育雄著, 非線形時系列解析, 朝倉書店, 2000.