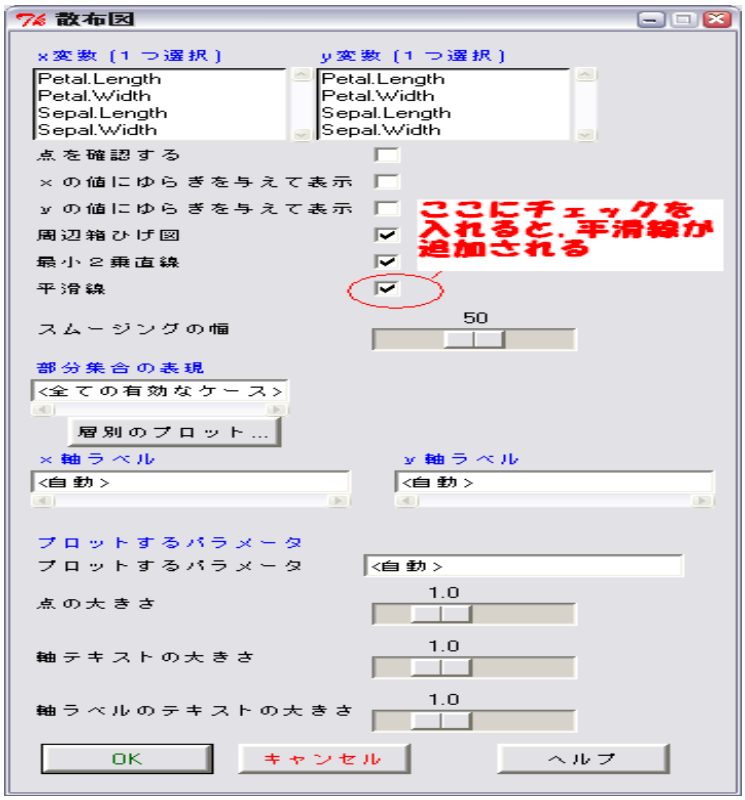
「平滑化」についての資料　　08/05/07

一、注目した理由：

「R Commander ハンドブック」の第 3 章を勉強するときに，散布図を描かせる部分で，初めて見た単語「平滑線」ということに疑問を持った。

二、参考資料：

1.「R Commander ハンドブック」（40 ページ）に書かれている内容：



スムージング【smooth：平滑化する】の幅：平滑線のスムージングの幅を指定する。

2.「R と S-PLUS による多変量解析」（22 ページ中段）に書かれている内容：

単純な線形回帰モデルを当てはめた結果と同時に，**局所重みつき回帰モデル（lowest）**を当てはめた結果を加えるとより効果的である。（中略）当てはめに採用するモデルでは：

$$y_i = g(x_i) + \epsilon_i$$

を仮定している。ここで**g**は「平滑化」関数であり，　ϵ_i　は平均がゼロで分散は定数のランダム変数である。

3.「R help」

3.1 scatterplot 関数（散布図を描かせるための関数）についての「R help」

scatterplot(car) R Documentation

Scatterplots with Boxplots

Description

Makes fancy scatterplots, with boxplots in the margins, a lowess smooth, and a regression line; `sp` is an abbreviation for `scatterplot`.

余白に箱ひげ図，lowess による局所回帰平滑化や回帰直線などで，凝った scatterplot ができる。「sp」は scatterplot の省略形である。

Usage

```
scatterplot(x, ...)

## S3 method for class 'formula':
scatterplot(formula, data, xlab, ylab, legend.title, subset, labels=FALSE, ...)


デフォルト:
scatterplot(x, y, smooth=TRUE, span=0.5, ...)

sp(...)
```

Arguments

`formula`	``model" formula, of the form `y ~ x` or (to plot by groups) `y ~ x | z`, where `z` evaluates to a factor or other variable dividing the data into groups.
`data`	data frame within which to evaluate the formula.
`x`	vector of horizontal coordinates.
`y`	vector of verical coordinates.
`smooth`	**if `TRUE` a lowess nonparametric regression line is drawn on the plot.** **もし `TRUE` であれば，プロットに平滑線が描かれる**
`span`	**span for the lowess smooth.** **平滑化のための平滑化パラメータ** δ 値が大きい程より滑らかになる **!?**
`reg.line`	function to draw a regression line on the plot or `FALSE` not to plot a regression line.
`boxplots`	if `"x"` a boxplot for `x` is drawn above the plot; if `"y"` a boxplot for `y` is drawn to the right of the plot; if `"xy"` both boxplots are drawn.
`lwd`	width of plotted lines.
`groups`	a factor or other variable dividing the data into groups; groups are plotted with different colors and plotting characters.
`legend.title`	title for legend box; defaults to the name of the groups variable.

3.2 lowess 関数についての「R help」

3.3 loess 関数についての「R help」

４．RjpWiki より：lowess() による平滑化

5.『日本統計学会誌』（第 34 巻シリーズ J 第 2 号 pp.187-207）論文「局所回帰による時系列の分解から明らかになった野鳥羽数の環境要因変化との関連性」-島津　秀康・柴田　里程

『日本統計学会誌』（第34巻シリーズJ第2号pp.187-207）

局所回帰による時系列の分解から明らかになった野鳥羽数の環境要因変化との関連性†

島津　秀康 *　　柴田　里程 **

Analysis of Bird Count Series by Local Regression to Explore Environmental Changes

Hideyasu SHIMADZU * and Ritei SHIBATA **

4　モデル

時系列データに対する有効なアプローチの１つとしてオリジナル時系列の分解が挙げられる．得られたいくつかの成分を実際の現象と照らし合わせることで，現象の背後にある構造を明らかにするのが目的である．事実，経済時系列の解析によく用いられる季節調整法は，あらかじめ季節効果，月効果など，解釈のしやすい特定のサイクルを仮定してオリジナル時系列をそれらの成分へ分解していく手法である．仮定したサイクルがデータに適当であればよいが不適当な場合，当然のことながらモデルの当てはめは不適切な結果を導くことになる．今回の場合は現象の構造が明確でないうえに解析の目的が羽数変動に注目し，その原因を探ることにあることから，あらかじめ特定のサイクルを仮定しないノンパラメトリックな平滑化を複数回適用し，オリジナル時系列をいくつかの成分に分解する方法を採用することにした．

ノンパラメトリックな平滑化は一般に

- シュプライン（区分多項式）平滑化
- 局所回帰平滑化

に大別できる．シュプライン平滑化は，あらかじめ定めた節点（ふしてん）で区分された区間それぞれに異なる多項式を当てはめ，それらの多項式を繋ぎ合わせて平滑曲線を求める．従って，定めた節点の意味，区分ごとに異なる多項式の意味を解釈する

必要がある．しかし今回の場合，節点をどのように決めるのが適当かすら不明であるので，シュプライン平滑化による時系列の分解は困難である．一方，局所回帰平滑化は，目的とする平滑曲線が局所的に多項式で近似できると想定した平滑化を行う．一般に平滑化の手法としてよく知られている核型平滑化は，局所回帰平滑化で 0 次多項式を想定することに相当するが，4.1 節で示すように端点での扱いは必ずしも同等ではない．注目する局所におけるデータの振る舞いによっては高次の多項式を想定した方が自然であることも多いので，端点での扱いの統一性も考慮し，ここでは核型平滑化ではなく局所回帰平滑化を用いることにした．具体的には，局所回帰平滑化のアルゴリズム`lowess`(Cleveland 1979) の発展版である`loess`(Cleveland and Devlin 1988) によって，与えられた羽数系列を分解する．局所回帰平滑化を用いたモデル構築に関してはチェンバース・ヘイスティ（1994），Fan and Gijbels (1996) を参照されたい．

本研究同様に野鳥羽数時系列データに対して局所回帰を用いた解析を行っている James *et al.* (1996) は，あらかじめトレンドの関数を定めることなくアプローチできる手法として局所回帰を挙げ，1966 年から 1992 年までの 26 年間分 BBS データにもとづき，北アメリカ中央部から東部にかけて観察されたアメリカムシクイの仲間（American Warblers）26 種の観測羽数に手法を適用している．その結果，抽出したトレンドの増減が地理的な高度の差異に強く依存している様子を示し，その原因として大気汚染による食資源劣化の可能性を指摘している．しかしその議論はモデルを通して羽数増減の具体的な要因にまで言及しているわけではない．これに対して本研究の目的は，あくまでも局所回帰を用いた平滑化よる野鳥観測羽数時系列の分解から羽数変化の背後にある構造と環境要因のと関係をモデルによって表現することにある．

4.1　局所回帰平滑化

一般に局所回帰は所与のデータ $(t_j, y_j), j = 1, \ldots, n$ に対してその背後に滑らかな関数 $f(t)$ の存在を仮定し，この $f(t)$ を p 次多項式によって局所的に近似する．具体的には，以下のように重み関数 $w(\cdot)$ によって定められる近傍に重みをつけた最小二乗法によって求める．つまり，

$$\sum_{j=1}^{n} w\left(\frac{|t_j - t|}{d_\delta(t)}\right)\{y_j - f_t(t_j)\}^2 \underset{f_t}{\rightarrow} \min \tag{1}$$

となる f_t を求める．ただし，$d_\delta(t) = \max_{j;t_j \in U_\delta(t)} |t_j - t|$ であり，$U_\delta(t)$ は t の $[n\delta]$ 最近隣近傍，$[n\delta]$ は $n\delta$ を越えない最大整数である．δ は平滑化パラメータ（`span`）と呼ばれ，最近隣近傍内のデータ数が全データ数 n に対して占める割合を定めている．さらに，このとき f_t としては

$$f_t(s) = \sum_{k=0}^{p} \beta_k(t)(s - t)^k$$

のような p 次局所多項式を想定する．今回，用いたデータ解析ソフトウェア S-PLUS の関数`loess` では，w として以下のような 3 乗 3 次の重み関数

$$w(x) = \begin{cases} (1-x^3)^3, & 0 \le x < 1 \\ 0, & \text{その他} \end{cases}$$

を採用しており，関数`loess` の引数`degree` によって多項式の次数 p を指定できる．結果としての平滑曲線は各近傍ごとに当てはめられた多項式の定数項部分の係数 $\hat{\beta}_0(t)$ によって与えられる．具体的には

- $p = 0$

$$\hat{\beta}_0(t) = \sum_j \frac{1}{s_0(t)} w\left(\frac{|t_j - t|}{d_\delta(t)}\right) y_j$$

- $p = 1$

$$\hat{\beta}_0(t) = \sum_j \frac{s_2(t) - (t_j - t) s_1(t)}{s_0(t) s_2(t) - s_1(t)^2} w\left(\frac{|t_j - t|}{d_\delta(t)}\right) y_j$$

- $p = 2$

$$\begin{aligned} \hat{\beta}_0(t) &= \frac{1}{c(t)} \sum_j \Big[\left\{ s_2(t) s_4(t) - s_3(t)^2 \right\} - (t_j - t) \left\{ s_1(t) s_4(t) - s_2(t) s_3(t) \right\} \\ &\quad + (t_j - t)^2 \left\{ s_1(t) s_3(t) - s_2(t)^2 \right\} \Big] w\left(\frac{|t_j - t|}{d_\delta(t)}\right) y_j \end{aligned}$$

となる．ただし，$s_r(t) = \sum_j (t_j - t)^r w\left(|t_j - t| / d_\delta(t)\right)$，$c(t) = 2 s_1(t) s_2(t) s_3(t) + s_0(t) \left\{ s_2(t) s_4(t) - s_3(t) \right\} - s_1(t)^2 s_4(t) - s_2(t)^3$ である．

このように，局所回帰平滑化によって得られる平滑曲線も核型平滑化によって得られる平滑曲線と同様，$y_1, \ldots, y_n$ の局所重みつき平均の形をしている点では類似している．しかし，先にも触れたように，核型平滑化は端点付近で有効データ数が次第に減少するために特異な挙動を示すことがあるのに対して，局所回帰平滑化では常に近傍内に $[n\delta]$ 個のデータが含まれるように平滑化が行われるため，端点での特異な挙動は起きにくい．なお，等間隔観測ならば $t = t_j$ での $p = 1$ としたときの平滑値は端点が近傍に含まれない限り，重み関数の対称性から $p = 0$ のときの平滑値と一致する．

いずれにしろ，局所回帰平滑化のよさは最近隣近傍を採用していることによる端点での挙動の自然さと，同一の重み関数と平滑化パラメータであっても，p を変えることにより局所的な挙動を統一的に制御できるところにある．

4.2 局所回帰による時系列データの分解

ここでは，オリジナル時系列に局所回帰を一度施し，さらにその残差に対して平滑化パラメータ δ（`span`）を変え，再度，局所回帰を適用することで，2 本の平滑曲線（ト

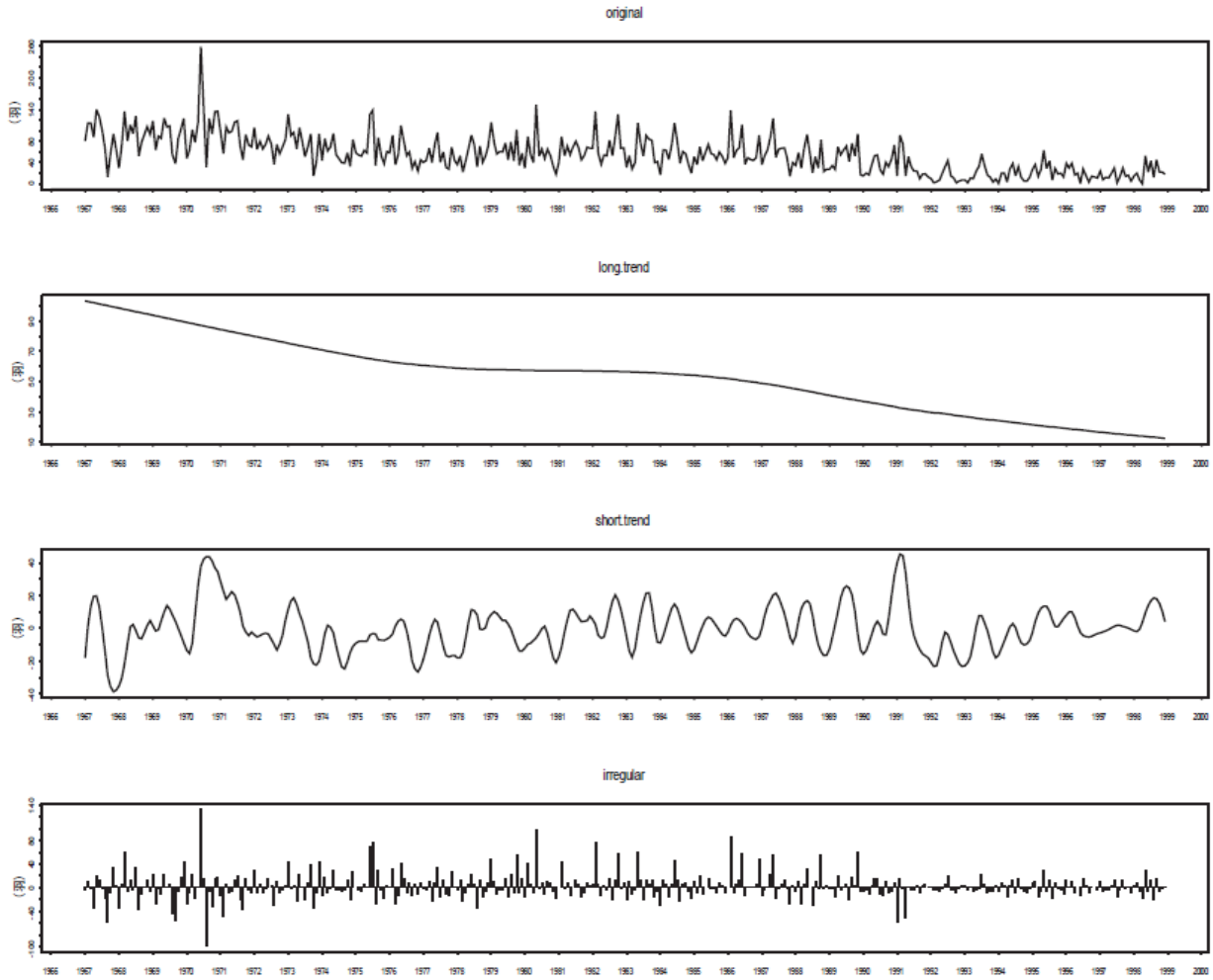


図 3: 観測羽数分解の例（スズメの場合）

レンド）と 1 本の残差系列（イレギュラー系列）を抽出する 2 段階平滑化により時系列を分解した．分解のアルゴリズムについては付録 B を参照されたい．このように時系列データに対して局所回帰による平滑化を複数回施す手法は Shibata and Miura(1997) によって時系列データの新しい分解手法として提案され，日本の日次金利時系列データの解析に導入された．現在ではさまざまな金融時系列にも適用されている．

図 3 には S-PLUS で観測羽数時系列データの分解を行った結果をスズメを例として示した．x 軸には 1967 年 1 月から 1998 年 12 月までの暦日が示されているが，計算上は通算月 $t = 1, \ldots, 384$ として考える．最上段がスズメの観測羽数時系列，2 段目が緩やかな変化を見せる長期トレンド，3 段目が短期トレンド，4 段目が残りのイレギュラー系列である．一般的には，種 i のオリジナル時系列が以下のように 3 本の時系列の和に分解される．

$$Z_i(t) = L_i(t) + S_i(t) + I_i(t), \ t = 1, \ldots, 384.$$

この分解は，時系列にトレンド，季節サイクルといった既存のモデルを仮定せず，局所回帰による平滑化を用いて，長期トレンド $L_i(t)$，短期トレンド $S_i(t)$ の抽出を目的

としている．ここで $L_i(t)$ は数十年の長期的スケールで緩やかに変化するトレンドである．また，$S_i(t)$ は1年レベルの比較的短期で変化するトレンドではあるが，あらかじめサイクルの仮定された季節トレンドではない．あくまでも平滑化によりデータから自然に抽出されたトレンドである．これが特定のサイクルを仮定した時系列の分解とは大きく異なる点である．$L_i(t), S_i(t)$ がモデルの非確率的な部分であり，$I_i(t)$ が確率的に振舞うランダムな部分に相当する．

今回の分解ではオリジナル時系列を順次，長期トレンド $L_i(t)$，短期トレンド $S_i(t)$，イレギュラー系列 $I_i(t)$ と分解していくため，得られるトレンドは分解の手順に依存する．しかし，短期的なトレンドを抽出した上で長期的なトレンドを抽出するという逆の手順は不自然であり，ここでは考えない．もちろんこの分解系列の数は今回のように3本に限る必要はなく，必要に応じて適切に分解すればよい．このとき，各トレンドがそれぞれ意味のあるものとして解釈可能で，かつ，残りのイレギュラー系列が特徴のないランダムな時系列になっていることが重要となる．

局所回帰による平滑化はS-PLUSでは関数`loess`を用いれば容易に行えるが，その際に平滑化パラメータ δ を引数`span`で，多項式の次数 p を引数`degree`で指定する必要がある．平滑化パラメータおよび多項式の次数の選択については様々な議論がなされているが（Fan and Gijbels 1996），解釈可能なトレンドを得るためには，データから探索的に設定する必要がある．長期，短期トレンドの抽出に用いた平滑化パラメータおよび多項式の次数については，各トレンドの節で述べることにする．