

Rにおける時系列解析

2002MM064 村田 友恵

指導教員 田中 豊

1 はじめに

現在、工学・科学・経済学などに関する統計的データ解析の分野において、過去のデータ解析や将来予測を行う方法として、時系列解析は重要になっている。時系列解析は、時間によって変動するデータを分析する手法一般を意味するものである。この研究の目的は、さまざまな時系列解析の方法論を理解して統計解析ソフト R を使って、大気中の二酸化炭素濃度データの時系列解析を行い、今後 10 年間の予測を行うことである。

2 使用したデータについて

この研究では、統計解析ソフト R に組み込まれているハワイ島マウナロアで観測された 1959 年 1 月から 1997 年 12 月の大気中の二酸化炭素濃度 (ppm) の月別データと、参考文献 [1] より入手した 1998 年 1 月から 2003 年 12 月の大気中の二酸化炭素濃度 (ppm) の月別データを加えた合計 540 観測値を使用した。

3 時系列データの形式および時系列解析の方法

3.1 時系列データの形式

時系列データとは時間の流れとともに順番に観測されたデータのことであり、大別して統計的な性質が時点に依存せず一定であるものを定常時系列と呼び、時間の経過とともに変化するものを非定常時系列と呼ぶ。

3.2 標本自己相関関数と標本偏自己相関関数

ラグ (時間差) h の標本自己相関関数は h だけ離れた点の値の間の相関係数を示し、標本偏自己相関関数はラグが h より小さい標本自己相関関数で説明される部分を取り除いたラグ h の相関係数である。

3.3 自己回帰移動平均 (ARMA) モデルと自己回帰和分移動平均 (ARIMA) モデル

p 次の自己回帰過程において Z_t が q 次の移動平均過程であるとき、 X_t は

$$X_t = \sum_{i=1}^p \phi_i X_{t-i} + Z_t + \sum_{i=1}^q \theta_i Z_{t-i}, \{Z_t\} \sim WN(0, \sigma^2)$$

を満足する。これらを次数 (p, q) の自己回帰移動平均過程あるいは ARMA 過程という。

ARIMA (p, d, q) 過程は、 d 階の差分をとることで $\nabla_d X$ が ARMA (p, q) となる確率過程を示す。

3.4 季節 ARIMA (SARIMA) モデル

B をラグ演算子とすると、 d と D が非負の整数であるとき、 $\{X_t\}$ が周期 s の季節 ARIMA $(p, d, q) \times (P, D, Q)_s$ 過程であるとは、その差分系列 $Y_t = (1-B)^d(1-B^s)^D X_t$

が次のような性質を満たす ARMA 過程であることである。

$$\phi^*(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t, \{Z_t\} \sim WN(0, \sigma^2)$$

(ただし、 $\phi(z) \neq 0$ と $\theta(z) \neq 0$ が $|z| \leq 1$ で成り立つ場合である。実際には D は 2 以上となることはまれであり、 P と Q は 3 未満が普通である。)

3.5 モデルの診断と予測

一般にデータへのモデルの適合度は、観測値とあてはめたモデルから得られる予測値を比較することにより判定する。モデルの適合には、赤池の情報量基準 (AIC) が最小の AIC 値をもつモデルを見つけることや、残差の自己相関に関する Ljung-Box による統計量を用いて、与えられた時系列データの残差の自己相関が有意ではないことを判定する。また、累積ピリオドグラムを用いて、モデルの推定した残差について系列相関の有無を検定する。データに対して適切なモデルを見つけたら、そのモデルを用いて将来予測を行う。

4 大気中の二酸化炭素濃度データの解析

4.1 データのプロットと標本自己相関関数

図 1 は大気中の二酸化炭素濃度データのプロットで、図 2 はその標本自己相関関数である。標本自己相関関数は、正の値でゆっくりと減少するため滑らかなトレンド成分と季節成分が存在すると考えられる。

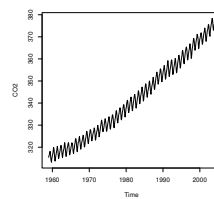


図 1 X_t のプロット

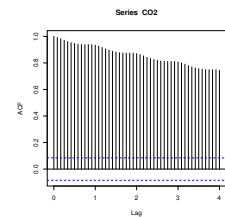


図 2 X_t の標本自己相関関数

4.2 データ変換と標本自己相関関数

もとのデータは、時間が経過するにつれて上昇していることがわかり、このデータを差分をとる方法で変換する。季節成分を取り除くためにラグ 12 差分をとり、さらに線形的なトレンド成分を取り除くためにラグ 1 差分を取った。得られたデータのプロットが図 3 になる。差分の取り方の順序をかえても最終的に同様な結果が得られた。データ変換した値の標本自己相関関数は図 4 で、ラグが大きくなるにつれて値は減少し、12 期のラグにおける値が過剰な負の相関を示した。

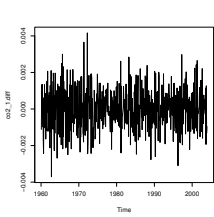


図3 $\nabla\nabla_{12}X_t$ のプロット

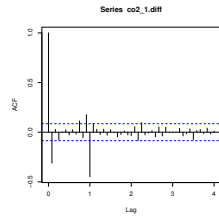


図4 標本自己相関関数

4.3 季節 ARIMA(SARIMA) モデルのあてはめ

大気中の二酸化炭素濃度データの系列 $\{X_t\}$ のラグ 12 とラグ 1 の差分をとった系列は、 $Y_t = (1 - B^{12})(1 - B)X_t$ である。 $\hat{\rho}(12) = -0.451, \hat{\rho}(24) = -0.023, \hat{\rho}(36) = -0.018$ から年次モデル (すなわち $P=0$ かつ $Q=1$) として一次の移動平均を考える。さらに $\hat{\rho}(1), \hat{\rho}(2), \dots, \hat{\rho}(11)$ の結果より $\rho(1)$ が唯一の非ゼロの短期の相関係数となるので、月次モデル (すなわち $p=0$ かつ $q=1$) としてやはり一次の移動平均が選ばれる。したがって、差分系列 $\{Y_t\}$ の標本平均 (0.002) を考慮して、 $\{Y_t\}$ のモデルとして

$$Y_t = 0.002 + (1 + \theta_1 B)(1 + \Theta_1 B^{12})Z_t$$

が得られた。得られたモデルは

$$\hat{\theta}_1 = -0.3520, \hat{\Theta}_1 = -0.8535$$

および

$$\hat{\sigma}^2 = 0.0825$$

であり、AIC 値は 202.6 となる。対応する $\{X_t\}$ のあてはめたモデルは SARIMA(0, 1, 1) \times (0, 1, 1)₁₂ 過程 (MA(1) モデル)

$$\nabla\nabla_{12}X_t = 0.002 + (1 - 0.3520B)(1 - 0.8535B^{12})Z_t$$

である。ここで $\{Z_t\} \sim WN(0, 0.0825)$ である。

また、MA(13) モデルを系列 $\{Y_t\}$ にあてはめ、その係数の推定値の標準誤差を検討した。その結果、

$$\begin{aligned} \nabla\nabla_{12}X_t = & 0.002 + Z_t - 0.3296Z_{t-1} \\ & + 0.0356Z_{t-10} + 0.0569Z_{t-11} - 0.9441Z_{t-12} \\ & + 0.3405Z_{t-13}, \{Z_t\} \sim WN(0, 0.07684) \end{aligned}$$

となる。その AIC 値は 202.39 で MA(1) モデルの 202.6 と非常に近い値になる。モデルは両方ともあてはまりが良いと言える。

4.4 予測

もとのデータから 1959 年 1 月から 1997 年 12 月のデータを抜き出して、1998 年 1 月から 2003 年 12 月のデータを予測し、観測値と比較した。図 5 は MA(1) モデルであてはめを行い、1998 年 1 月から 2003 年 12 月のデータの予測を行い、もとの 1998 年 1 月から 2003 年 12 月の

データと比較したものである。同様に MA(13) モデルのあてはめを行い、予測値と観測値の比較を行ったプロットが図 6 である。二つのモデルとも、値は周期性を示し上昇した。あてはめた二つのモデルの平均二乗誤差が 1.297 と 1.510 で、MA(1) モデルでの値が小さかったため、MA(1) モデルの方が良いと言える。

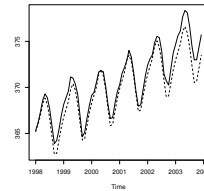


図 5 MA(1) モデルの予測値 (点線) と観測値 (実線)

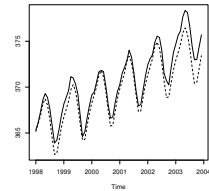


図 6 MA(13) モデルの予測値 (点線) と観測値 (実線)

4.5 予測

もとのデータを用いて、10 年後である 2004 年 1 月から 2013 年 12 月のデータを予測した。図 7 は、MA(1) モデルを用いて 10 年間の予測を行った予測値のプロットである。同様に MA(13) モデルを用いて 10 年間の予測を行った予測値のプロットが図 8 である。二つのモデルとも、値は周期性を示しながら上昇した。

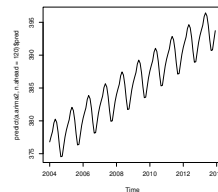


図 7 MA(1) モデルの予測値

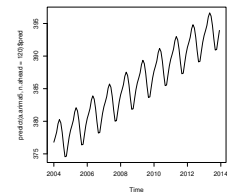


図 8 MA(13) モデルの予測値

5 おわりに

今回の研究で、一から時系列解析の方法論と統計解析ソフト R を使って時系列解析を行う方法を理解するのにとても苦労した。この研究を進めるにつれて、時系列解析の理論について一つ一つ理解することの難しさや理解したときの喜びを得ることができた。また、時系列解析について学ぶところが多いが、実際に大気中の二酸化炭素濃度データを使って、データの解析とあてはまりの良いモデルを見つけ、原系列の 10 年後の将来予測を行うことができたのでよかった。

参考文献

- [1] maunaloa.co2 : ftp://cdiac.esd.ornl.gov/pub/.
- [2] P.J. ブロックウェル, R.A. デービス著 : 入門時系列解析と予測, CAP 出版, (2000).