時系列・時空間データからの情報抽出

川崎　能典　　モデリング研究系　准教授

2010年7月9日　統計数理研究所　オープンハウス

はじめに: 統計的モデリングによる時系列・時空間データからの情報抽出の研究事例として，ここでは琵琶湖水質データの長期トレンド解析を紹介する．本研究は，河合研一（別府大学），大久保卓也（滋賀県琵琶湖環境科学研究センター），金藤浩司（統計数理研究所・データ科学研究系）との共同研究である．

1. 問題の背景とデータ取得の状況

日本最大の淡水湖である琵琶湖は「故障水出保全対策特別措置法」が指定する湖沼のひとつであり，5年ごとの湖沼水質保全計画作成の基礎データとして，滋賀県と国土交通省が琵琶湖の水質を定期的にモニタリングしている．

実際に利用可能なデータ(測定項目)は多様であるが，透明度，水温，pH等の物理量に加え，クロロフィル，フェオフィチン，全窒素，全リン，塩素などの化学物質が，1979年以降49の固定観測地点(図3内に赤で打点した地点)において，月1回測定されている．

2. データ解析の目的

クロロフィルa濃度の月次データに基づき，過去30年の傾向変化から湖沼水質改善の証拠を探る．クロロフィルに着目する理由として，(1) クロロフィル濃度は植物プランクトンの指標，値が高いほど水質が悪いこと，(2) 赤潮，アオコが発生すると高くなること，(3) 近年は下水道の整備を中心に排水対策が進んでおり，その効果を確認したいこと，が挙げられる．

最大の問題点は，季節変動が大きく時系列を図示しただけでは傾向変化が明らかでないこと，欠測がおびただしいことである．この両方に対処できるのが，平滑化事前分布を利用した時系列モデルである．

3. 平滑化事前分布モデル：データの前処理

地点ごとの時系列データy_tが，トレンド成分μ_t，季節変動成分s_t，不規則変動成分ϵ_tによって$y_t = \mu_t + s_t + \epsilon_t$ と表されると仮定する．

これは，ϵ_tを誤差項，μ_tとs_tを未知係数とする線形回帰モデルと見なせるが，μ_tとs_tは固定係数ではなく時変であるから，観測値の個数Tに対して未知量はϵ_tの分散も含めて$2T+1$となり，このままでは最小二乗問題は解けない．

そこで，トレンドと季節成分に時間変化の滑らかさを仮定し，次の制約付き最小二乗問題

$$\sum_{t=1}^{T}(y_t - \mu_t - s_t)^2 + \lambda_1 \sum_{t=3}^{T}(\Delta^2 \mu_t)^2 + \lambda_2 \sum_{t=13}^{T}(\Delta_{12} s_t)^2 + \lambda_3 \sum_{t=12}^{T}(\sum_{j=0}^{11} s_{t-j})^2$$

解としてμ_tおよびs_tを得る．平滑化の度合いλ_iは最尤法で決められる．

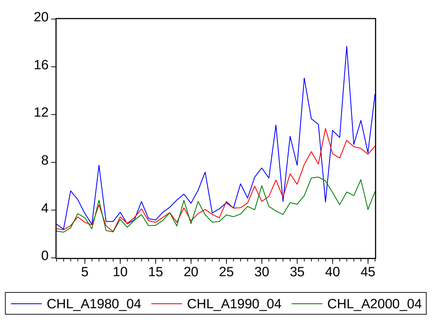

図1: トレンド推定値に基づくクロロフィルa濃度の経年変化

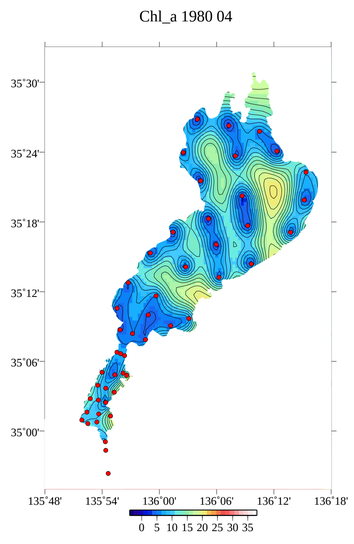

図2(a): クロロフィルa濃度の等高線プロット(1980年4月)

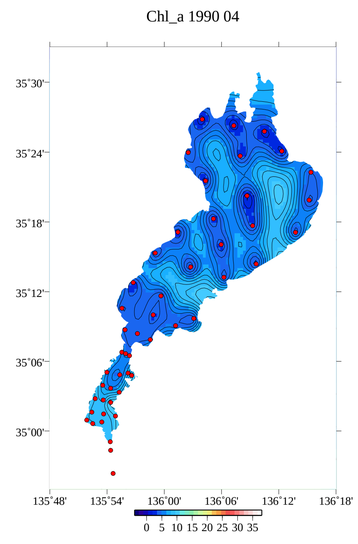

図2(b): クロロフィルa濃度の等高線プロット(1900年4月)

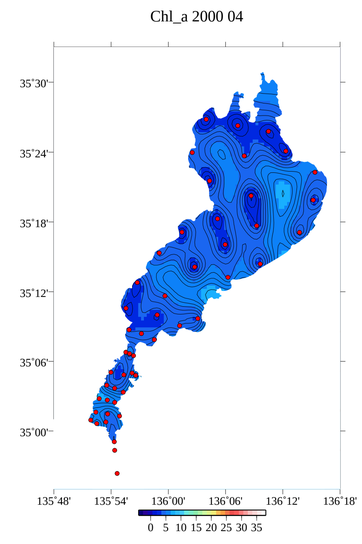

図2(c): クロロフィルa濃度の等高線プロット(2000年4月)

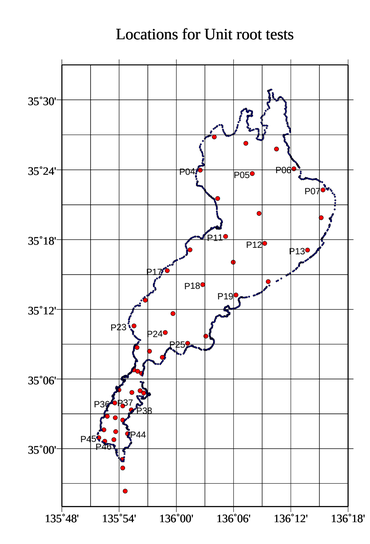

図3: 観測地点と選定地点

4. 前処理結果の視覚的チェック

図1は，横軸を便宜上観測ステーションの番号順に取り，各地点でトレンド成分として推定された値のプロットを，1980年4月(青)，1990年4月(赤)，2000年4月(緑)と，10年おきに観察したもの．(時系列プロットではない．) 横軸を左から右に辿るにつれて，観測地点は北から南に移動する．曲線が全体として下方にシフトしていることから，近年はクロロフィルaの濃度が減少していることがわかる．

各観測ステーションごとに得られたトレンド値を元に空間平滑化を行い，クロロフィル濃度の等高線図を各時点ごとに作成したのが，図2(a)–(c)である．青みが強くなるほど，水質が改善されていることを示しており，この20年の間に徐々に水質の改善が進んでいることが視覚的に捉えられている．

5. 単位根検定によるトレンドの有無の検証

分析をさらに一歩進め，単位根検定と呼ばれる検定手法を用いてトレンドの有無をややフォーマルに検証する．対象地点は，北湖13地点，南湖6地点に間引いた．(図3参照．) まず，

$$y_t = \alpha + \delta t + \rho y_{t-1} + \epsilon_t$$

の推計に基づくAugmented Dickey-Fuller(ADF)検定を実施する．帰無仮説は非定常だが，$\rho = 1$(確率的トレンド)が棄却されてもδ有意なら決定論的トレンドありと結論される．

結果の補強・確認のため，帰無仮説に定常を取るKPSS検定(Kwiatkovski et al. 1992) も行う．すなわち以下の推計式

$$y_t = \alpha + \delta t + \epsilon_t + u_t,\ u_t = u_{t-1} + \xi_t$$

に基づき，$\sigma_\xi = 0$を検定する．(ここでϵ_tは任意の定常過程である．) 確率的トレンド(u_t)が存在しないならば，それを生成するイノベーション過程が退化している$(\sigma_\xi = 0)$，というのがこの推計式の基本的アイディアである．

6. 単位根検定に基づく結論

確率的，決定論的，いずれに意味にせよ殆どの地点でトレンドの存在が確認される．詳細は紙幅の関係で省略，論文を参照されたい．(Kawasaki, et al. 2009) ただし，ADFで帰無仮説が受容(KPSSの場合なら棄却)され，δの係数も有意でない(しかも正)場合は，方向(下降トレンドかどうか)が見えにくいので視覚的チェックが必要である．

季節調整を行って，トレンド系列に対して検定を行った場合には，概して決定論的トレンドの存在が指示されており，確率的トレンドは重要ではない．(平滑化が効いていて，変動性は少ない．) ADFとKPSSとで結論が矛盾するのは，トレンド系列の場合はP44だけだが，季調済系列ではP07, P18, P19, P23, P38, P44の6地点．

季調済系列をもちいると，ノイズ項も足すので，外れ値等の影響を受けてしまう．結果的に，単位根が好まれてしまう傾向にある．外的要因に関してはintervention termで対応するか，トレンドのイノベーションを非ガウス化する方法も考えられる．外れ値や構造変化の要因となりうる既知の事象はある．(例：1994年の大渇水．)

これまでの琵琶湖の水質保全対策は，富栄養化防止に重点が置かれていたが，競合リスクの観点から悪化している指標もあると思われる．今後は，生態系保全，景観保全，水産資源管理，健康リスク・生態リスク管理などの視点を重視した施策が必要であり，そのための統計的モデリングが望まれる．

参考文献

Kawasaki, Y., Kawai, K., Okubo, T., Kanefuji, K. (2009): Long term trend analysis of water quality in Lake Biwa, Proceedings of 18th World IMACS Congress and MODSIM09 International Congress on Modeling and Simulation, 3172-3178.