# 位置情報を用いた変化係数にもとづく回帰分析手法の実装について

滋賀大学　佐藤健一

## 1．はじめに

GIS(Geographical Information System)の普及により，放射線の健康影響を評価するデータにも，地理的な位置情報が付与される機会が増えている．ここでは，簡単に線形重回帰モデル，いわゆる一般的な回帰モデルを利用した時空間データの統計解析手法を紹介する．例として，図 1 に示すカナダの 35 都市における 1 年間，365 日分の日平均気温データを扱う．このデータは統計ソフト R の fda ライブラリに CanadianWeather として含まれている．ここでの関心は，カナダの他の都市，あるいは任意の地点の 365 日の気温の予測にある．


**図 1**．365 日の気温が観測されたカナダの 35 都市.

## 2．スプライン基底

カナダの最北に位置する Resolute の 365 日の気温データを図 2 に示す．気温は 1 月 1 日から少しずつ上昇し，およそ 200 日後，つまり 7 月下旬頃に最も高くなり，また減少に転じる．回帰分析では直線が仮定されることが多いが，このような非線形傾向には適さない．また，初等数学で学習する多項式を利用した回帰を考えることも

できるが，高次の多項式は数値計算が破綻しやすいという短所がある．そこで，線形の枠組みの中で非線形曲線を表すことができるスプライン基底の利用を考える．

スプライン基底は，平たく言えば折れ線回帰であり，節点κをもつスプライン基底は$(t-\kappa)_+ = t-\kappa \ \ (t-\kappa > 0),\ \ 0\ (otherwise)$，とかける．365 日の観測値に対して，時間に関する直線と，50 日から 300 日まで 50 日刻みで節点を配置すれば，非線形曲線を 8 個の基底からなる線形一次結合で表現できる．節点の尖りが気になる場合には，スプライン基底の 2 乗を考えればよい．

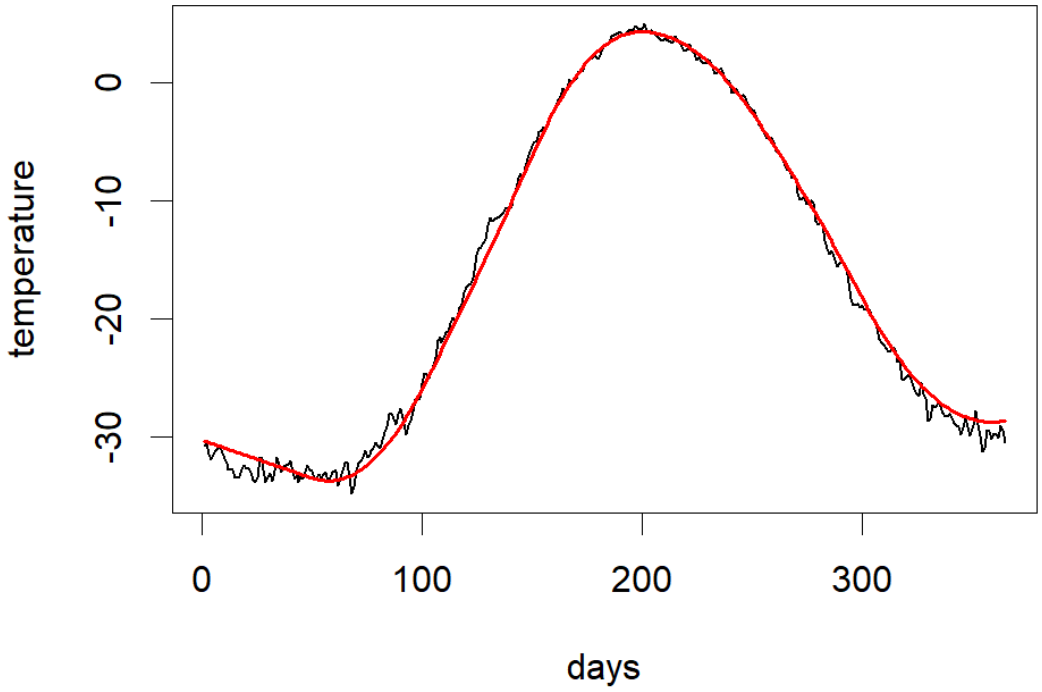


**図 2.** 最北に位置する都市 Resolute（西経-94.54，北緯 74.41）における 1 月 1 日から 12 月 31 日までの 365 日の日平均気温．黒い実線は観測値を，赤い実線は重回帰モデルによる予測値を示す．

## 3．変化係数

このようにして，35 都市それぞれの気温データに対して重回帰モデルを適用することは可能である．しかしながら，我々の目的は任意の地点における気温の予測であるため，位置情報を利用した重回帰モデルを構成したい．そこで，時間や位置によって回帰係数が変化することを許す変化係数の考え方を用いる．すなわち，時間に関して仮定した 8 個の回帰係数$\theta_j$，$\mathrm{j} = 1,\cdots,8$が位置$(u,v)$によって変化するとする．つまり，$\theta_j = \theta_j(u,v)$ となる．さらに，$u$と$v$についても線形性を仮定すれば，例えば，

$$\theta_j(u,v) = \beta_{j,0} + \beta_{j,1}u + \beta_{j,2}v + \beta_{j,3}uv$$

のように，$(u,v)$上で変化する曲面として表すことができる．今，非線形曲線の基底のうち，時間$t$に関する直線に着目すれば，

$$\theta_1(u,v) + \theta_2(u,v)t = \beta_{1,0} + \beta_{1,1}u + \beta_{1,2}v + \beta_{1,3}uv + (\beta_{2,0} + \beta_{2,1}u + \beta_{2,2}v + \beta_{2,3}uv)t$$

となる．右辺の括弧を展開すれば時間$t$と位置$(u,v)$の積，あるいは交互作用項で書けることが分かる(Tonda and Satoh, 2017; Satoh and Tonda, 2016, Satoh and Tonda, 2014)．

## 4．推定と変数選択

結果的に非常に沢山の説明変数と回帰係数を準備することになるが，R において交互作用項は次のように，用いるすべての説明変数の積として自動で生成されるため，容易に実装できる．実際，この場合は 8*3*3=72 個の

説明変数が使われる．

lm(y~(1+t+sp50+sp100+sp150+sp200+sp250+sp300)*(1+u+I(u^2))*(1+v+I(v^2)))

多くの説明変数を用いた場合には，現在の観測値に対する過剰適合が懸念される．これを防ぐために，回帰係数のパラメータ空間を縮小させるリッジ回帰，Lasso 回帰，あるいは Elastic Net 回帰などの罰則付パラメータ推定が使われる．ここでは，変量選択基準 AIC を使った変数減少法による変数選択によって説明変数そのものを，68 個まで減らした(Fujikoshi and Satoh, 1997, Satoh, Fujikoshi and Kobayashi, 1997; Satoh, 1997)．なお，68 個の説明変数も多く感じるが，この解析におけるデータ数は 365*35=12,775 と十分に多い．68 個のうち P 値が 0.1 以下の回帰係数は 64 個あった．また，観測値と予測値の相関係数の 2 乗となる重相関係数は 0.983，自由度調整済重相関係数も 0.983 となっており，当てはまりは良好であった．推定された適合値は，35 都市の 365 日のデータを滑らかに説明する時空間曲面を構成する．図 2 に Resolute での経時曲線を示す．観測値との当てはまりも良いことが見て分かる．

また，図 3 に 1 月 1 日と 7 月 20 日における気温の空間曲面を等高線と色で示す．なお，この推定曲面は 365 日の任意の日にちにおいて，地図上の任意の地点に対して得られるものであり，解析の目的である任意の地点における 365 日の気温の予測を可能にする．

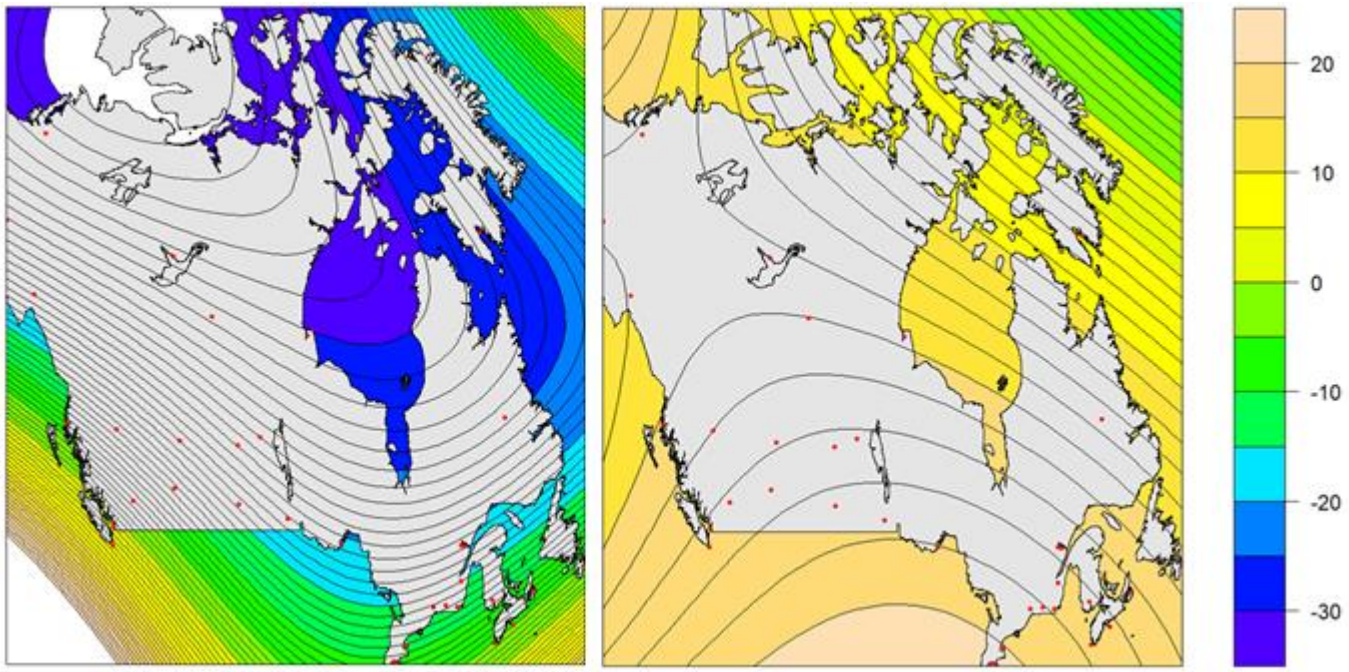


**図 3.** 重回帰モデルによって推定された（左）1 月 1 日，（右）7 月 20 日の気温の空間曲面.

## 5．おわりに

本手法以外でも，指定した地点における気温の予測は位置情報を用いた重み付平均などで実装できるが，線形回帰に比べると平面平滑化の作業はバンド幅の最適化などのなど計算機的な負荷も小さくない．また，このように位置情報を持つデータにおいては，近いほど観測値の相関係数が高くなることが知られており，これをバリアグラムとしてモデルに組み込むことがある．さらに，地点を固定した場合にも観測時点が近ければ，経時的にも相関係数が高くなることも知られており，これらも自己相関係数としてモデルに考慮されることがある．一方で，提案手法のように時間的および空間的に独立性を仮定した場合であっても，一般化推定方程式の理論によれば作業相関行列の与え方に依らず，回帰係数の推定量としては一致性を持つことが知られている．つまり，推定量が真の回帰係数に収束するスピードが遅いものの，収束すること自体が保証される．また，本稿では重回帰モデル

を用いて説明したが，変化係数を用いた空間的な回帰手法は適用範囲が広く，二値データを目的変数とするロジスティック回帰や，計数データに対するポアソン回帰、センサリングを含む生存時間データに対するコックス回帰などにも同様に利用できる(Tonda, Satoh and Kamo, 2015; Tonda, Satoh, et. al, 2012)．放射線のリスク評価においても，時間や空間の情報がある場合には，積極的に利用を試みたい．

**参考文献**


1. T. Tonda and K. Satoh: Estimating varying coefficients for a balanced growth curve model without specifying spatial-temporal baseline trend, Journal of The Japan Statistical Society, 47, 1-12, 2017.
2. K. Satoh and T. Tonda: Estimating regression coefficients for balanced growth curve model when time trend of baseline is not specified, American Journal of Mathematical and Management Sciences, 35(3), 183-193, 2016.
3. T. Tonda, K. Satoh and K. Kamo: Detecting a local cohort effect for cancer mortality data using a varying coefficient model, Journal of Epidemiology, 25 (10), 639-646, 2015.
4. K. Satoh and T. Tonda: Estimating semiparametric varying coefficients for geographical data in a mixed effects model, Journal of The Japan Statistical Society, 44(1), 25-41, 2014.
5. T. Tonda, K. Satoh, K. Otani, Y. Sato, H. Maruyama, H. Kawakami, S. Tashiro, M. Hoshi and M. Ohtaki: Investigation on circular asymmetry of geographical distribution in cancer mortality of Hiroshima atomic bomb survivors based on risk maps: analysis of spatial survival data, Radiation and Environmental Biophysics, 51(2), 133-141. 2012.
6. Y. Fujikoshi and K. Satoh: Modified AIC and Cp in Multivariate Linear Regression, Biometrika, 84, 707-716, 1997.
7. K. Satoh, Y. Fujikoshi and M. Kobayashi: Variable Selection for the Growth Curve Model, Journal of Multivariate Analysis, 60, 277-292, 1997.
8. K. Satoh: AIC-type Model Selection Criterion for Multivariate Linear Regression with a Future Experiment, Journal of Japan Statistical Society, 27, 135-140, 1997.