MA4　化学輸送モデルを用いた非線形 Land Use Regression モデルによる大気汚染の時空間変動推計

Estimating spatiotemporal variation of air pollution by statistical model combined with chemical transport model

指導教員　近藤明教授・共生環境評価領域
28H19015　衛藤信之介（Shinnosuke ETO）

Abstract: A land use regression (LUR) model with a linear regression approach is widely used to estimate air pollutant concentrations in epidemiological studies, but it has difficulty in reproducing short-term variability of the concentrations. This study proposed a new LUR model which is able to reproduce spatiotemporal variation with high accuracy. This model integrates two approaches: one is a nonlinear regression technique with a machine learning algorithm called random forests; the other is a chemical transport model called CMAQ for preparation of predictor variables that represent short-term variation. The new LUR model was applied to estimates of annual and daily NO_2 and $PM_{2.5}$ concentrations in the Kinki region in the year 2014. Although the nonlinear regression technique did not improve the accuracy in terms of spatial variations in the annual mean concentrations because of overfitting, the new model showed much higher accuracy than the existing model in terms of spatial and temporal variations in the daily concentrations. This result illustrates the advantage of the new approaches to represent the spatiotemporal variation of air pollutants.

Keywords: Air pollution, Random forests, Linear regression, LUR, CMAQ

１．はじめに

大気汚染物質への曝露評価を行う際には，その濃度変動を正確に推計することが不可欠となる．そこで大気汚染物質の濃度を土地利用割合，気象データ，道路延長，衛星観測データなどの周辺の環境情報から回帰する Land Use Regression (LUR) モデルが，その推計誤差の小ささから広く用いられている．しかし回帰に線形回帰を用いた場合，年平均値などの中長期間の平均濃度を正確に推計することが可能である一方，日平均値などの短期間の濃度変動を正確に再現することが困難とされている．そこで本研究では，回帰に機械学習アルゴリズムであるランダムフォレストを用いる非線形 LUR モデルを使用し，また濃度の短期変動が再現可能な化学輸送モデル CMAQ の出力値を新たな説明変数として組み合わせることで，大気汚染濃度の時空間変動を正確に再現できるモデルを構築することを目的とした．

２．方法

大気汚染の時空間変動の推計対象は，2014 年の近畿圏における年・日平均 NO_2 および $PM_{2.5}$ の濃度とした．推計モデルとして以下の四つの LUR モデルを構築し、精度を比較した．

1. LM (説明変数に CMAQ を加えていない, 既存の LUR モデルと同じ説明変数による線形回帰モデル)
2. LM+CMAQ (説明変数の一つに CMAQ の出力値を加えた線形回帰モデル)
3. RF (説明変数に CMAQ を加えていない，既存の LUR モデルと同じ説明変数による非線形回帰 (ランダムフォレスト) モデル)
4. RF+CMAQ (説明変数の一つに CMAQ の出力値を加えた非線形回帰 (ランダムフォレスト) モデル)

これらのモデルの精度検証として，データを 6 分割して 5 つのデータで学習し，残りの 1 つで検証を行う 6 分割交差検証を行った. この交差検証においては, ある地点の全期間のデータを抜く空間検証と，ある期間の全地点のデータを抜く時間検証を行い，R^2 および RMSE の観点からモデルの検証を行った．

３．結果と考察

表 1 に NO_2 および $PM_{2.5}$ の年・日平均値に対する各モデルの交差検証結果を示す．年平均値の推計において，両物質で LM+CMAQ が最も良い精度で推計した．LM，RF の両方でモデルに CMAQ を加えることで精度は向上したが，非線形モデルの導入効果は確認できなかった．これは濃度と変数の線形関係にランダムフォレストが非線形に適合してしまい，過学習を起こしたためだと推測される．日平均値の推計においては、RF+CMAQ が最も良い精度を示し，非線形モデルを用いたことと化学輸送モデルを用いたことの有効性が示された．LUR モデルに CMAQ を統合したことで，NO_2 についてはその日変動の要因である排出量の日変動や化学反応が CMAQ を通じて再現可能になり，$PM_{2.5}$ については日本における主要な汚染源である越境汚染の反映が可能になったため，精度は向上したものと考えられる．**図 1** に RF+CMAQ による日平均値の推計値と観測値の散布図を示す. **図 1** と, **表 1 から,** NO_2 に対する RF+CMAQ では時間・空間検証の精度の差はさほど確認できないが，$PM_{2.5}$ に対する RF+CMAQ では時間検証結果が空間検証結果と比較して劣った．これは，$PM_{2.5}$ は空間変動よりも時間変動の方が大きく，分割データによっては高濃度あるいは低濃度となる時間変動条件を学習するためのデータが不足していたことが考えられる．

表 1　各モデルによる NO_2 および $PM_{2.5}$ の年・日平均値の交差検証結果

		年平均		日平均			
				空間検証		時間検証	
		R^2	RMSE	R^2	RMSE	R^2	RMSE
NO_2	LM	0.64	2.54	0.44	5.45	0.45	5.39
	LM+CMAQ	0.76	2.00	0.63	4.44	0.63	4.44
	RF	0.69	2.26	0.65	4.23	0.59	4.71
	RF+CMAQ	0.74	2.06	0.73	3.77	0.73	3.81
$PM_{2.5}$	LM	0.46	1.55	0.27	7.71	0.21	7.59
	LM+CMAQ	0.46	1.32	0.71	4.89	0.65	5.01
	RF	0.31	1.89	0.19	8.72	0.19	8.70
	RF+CMAQ	0.42	1.59	0.86	3.55	0.64	5.00

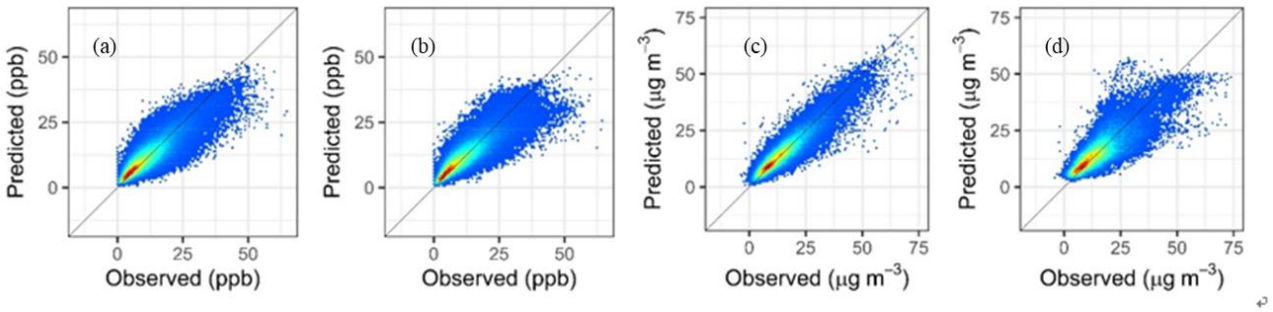

図 1　RF＋CMAQ による (a) NO_2 日平均値の空間検証，(b) NO_2 日平均値の時間検証，(c) $PM_{2.5}$ 日平均値の空間検証，(d) $PM_{2.5}$ 日平均値の時間検証結果の散布図

４．結論

本研究の結論を，以下にまとめる．

- 年平均値推計では LM+CMAQ が，日平均値推計では RF+CMAQ が濃度変動を正確に再現した．
- 既存の LUR モデルと比較して，非線形回帰モデルと化学輸送モデルを用いた LUR モデルは濃度の時空間分布を正確に再現できる．