３ 調 査 研 究

３・１ 報 文

１）ノンパラメトリックデータの回帰：一般化加法モデルによる大気環境データの非線形回帰モデル構築に関する検討

古澤 尚英

要 旨

ノンパラメトリックデータの回帰・推論・予測を行う場合、一般化加法モデルを用いると、線形回帰モデルに比べて予測精度が高く、モデルの説明性が高いと言われている。本報では、ノンパラメトリックデータである光化学オキシダント濃度について、一般化加法モデルで日射量、気温及び風速データを用いた回帰・推論・予測を行い、最適なモデルを得ることができた。また、一般化加法モデルは線形回帰モデルより高い精度で回帰・推論・予測を行うことができ、環境データを解析する際の重要な選択肢の一つとなることが示された。

キーワード：一般化加法モデル，光化学オキシダント，予測，推論，R 言語

は じ め に

大気汚染物質は，大気汚染防止法第 22 条に基づき都道府県知事が監視することとなっており [1]，大気汚染防止法第 22 条の規定に基づく事務の処理基準 [2] に基づき常時監視項目や観測局数等の体制整備に関することが定められている。 大気汚染物質のうち，自動測定機によりモニタリングされている窒素酸化物，硫黄酸化物，微小粒子状物質等の物質は，令和 4 年度の全国調査結果において環境基準をほぼ 100%達成していたが，一方で，光化学オキシダント (以下「Ox」という。) は環境基準達成率が 0.1%）であった。

このような状況から，Weather Research and Forecasting model (WRF) や (CMAQ) 等の大気環境シミュレーション等を用いて，大気汚染物質の空間濃度分布の予測，大気汚染物質の生成，飛来予測などの研究が進められている [4) -6)]。 また，大気汚染物質の空間濃度分布予測は，大気汚染常時監視測定局の配置検討や光化学オキシダント注意報等発令地域区分の検討等，行政的な活用がしやすい有用なデータであるが，一方で，大気環境シミュレーションには専用のソフトウェアの準備が必要であり，これらのソフトウェアの構築には多大なコストがかかるため地方自治体職員が単独で実施することは難しく，シミュレーションモデルを用いた検討は一部の研究者が先行して実施している状況である。

そこで筆者を含む検討メンバーは，国立環境研究所と地方環境研究所とのⅡ型実施共同研究 [7] の枠組みにおいて，常時監視測定局の観測体制見直し及び発令地域区分等の検討に使用する目的で，大気汚染物質の空間濃度分布を統計・機械学習的に求めるライブラリの開発を行った [8) , 9)]。 このライブラリに導入した空間濃度分布を計算する Regression Kriging 法 (以下「RK 法」という。) は，土地利用回帰モデル (Land Use Regression: LUR) と地球統計学手法である Ordinary Kriging (OK) を組み合わせた手法であり，大気汚染物質の空間濃度分布を精度良く予測することができると報告されている [10) , 11)]。 筆者らは既報 [11)] で微小粒子状物質を対象とした検討を行ったが，このライブラリでは予測できる大気汚染物質の対象を拡張して構築した。 また，LUR の計算方法に，線形回帰モデル以外にもランダムフォレスト，サポートベクトルマシン，一般化線形モデル，一般化加法モデル及びニューラルネットワークの全 6 モデルを利用できるようにした [8)]。 このライブラリを利用して Ox の高精度な空間濃度分布を計算する条件検討を進めているが，これら 6 モデルのうち一般化加法モデル (以下「GAM」という。) を実装するための日本語情報が少なく，また，大気環境データのような時間-空間情報を持つノンパラメトリックデータに適応させた事例は更に情報が少ない状況であった。 そのため，本報では Ox 空間濃度分布を予測する GAM の構築・精度検証の検討を行い，GAM の有用性について実証するとともに，今後の環境データ解析における基礎検討資料とすることを目的とする。

方　法

1. 線形回帰モデルについて

従来, 環境分野における事象の解析や予測には線形回帰モデルが広く利用されてきたため, GAM と比較するためにもまずは線形回帰モデルの構築と解釈の方法を解説する。

式 1 に線形回帰式を示す[12]。 p > 1 である場合は重回帰式とも呼ばれる。

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j \qquad (式\ 1)$$

式 1 において, $\hat{Y}$ は予測変数, $\hat{\beta}_0$ は切片, X_j は説明変数, $\hat{\beta}_j$ は各説明変数の重み (直線の傾き) を示している。この線形回帰モデルは, 説明変数の効果が重み $\hat{\beta}_j$ として表現され, 事象に対する説明性が高く解釈も容易であることから広く利用されている。 R による線形回帰モデルを構築する lm() 関数では, 直線以外にも対数, 累乗等の曲線を当てはめたモデルを構築することもできる。 一方で, 回帰直線・曲線に当てはまりが悪い事象は予測精度が著しく悪くなることが知られており, 正規分布に従わない, 大気汚染物質の観測データのようなノンパラメトリックデータはその傾向が大きい。 このことから, 近年はランダムフォレストやニューラルネットワーク等の機械学習モデルを用いた解析により, 推論・予測が行われている事例が増えている[13]。 高い精度が得られる機械学習モデルであるが, ランダムフォレスト, サポートベクトルマシン及びニューラルネットワーク等のモデルは, モデル内部はブラックボックスと呼ばれ, 説明変数がどのように作用して予測を行っているのかは説明が難しいとされている[14]。 そのため, 解釈などの説明性を求める場合は線形回帰, 精度の高い回帰や予測を行う合は機械学習を使う等, 目的に応じて使い分けられている。

2. 一般化加法モデルについて

一般化加法モデル (Generalized Additive Model: GAM) は, もともと一般化線形モデル (Generalized Liner Model: GLM) で非線形性を扱うことができるよう改良されたモデルであり, 説明性と予測精度の両方を備えたモデルと言われている[14]。 式 2 に GAM のモデル式を示す。

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon \qquad (式\ 2)$$

式 2 で, ϵ は平均 0 の誤差項を表す。 式 1 と異なるのは, 説明変数がそれぞれ非線形な平滑化関数 f_j に適用されていることである。 また, GAM では 1 つの関数を全データに対してそのまま適応させるのではなく, はじめにデータを複数の区間に分割し, それぞれの区間にパラメータを調整した平滑化関数を設定し, 各平滑化関数の合計値が回帰モデルとなる回帰スプラインが代表的な手法となっている[14]。 区間数や規定関数を適切に選択することで, どのようなデータに対しても回帰が可能と言われている。 GAM に関する詳細は参考文献[12),15)]を参照してほしい。

3. 利用するデータの選定及び計算条件

Ox 生成の起源として, 日射量, 気温, 風速及び VOC 等の前駆物質の存在が挙げられる[16),17)]。 九州では春先 4~5 月にかけてこれらの気象条件が揃うことが多く, 熊本県で過去に光化学スモッグ注意報等を発令したのもこの時期であった[18)]。 本報で構築する Ox の回帰モデル構築の説明変数には, 信頼性の高いデータが手に入りやすい日射量, 気温及び風速を用いることとした。 また, GAM では時空間属性を含めてモデル化することが可能であり[19)], 本報ではこれらの検討も行うことから, 緯度, 経度及び時

表 1 使用したデータの一覧

利用データ	目的変数・説明変数名	概要
常時監視測定局観測データ	Ox	光化学オキシダント 1 時間値濃度 [ppb]
地上付近気象庁気象データ (MSM-S)	temperature	気温 1 時間値 [℃]
	wind-speed	風速 1 時間値 [m/s]
地上付近気象庁気象データ (MSM-r1h)	solar-radiation	日射量 1 時間値 [W/m^2]
位置情報	lon	経度 [°]
	lat	緯度 [°]
時間	date	通算時間 [-]
	time	0 ~ 23 [hour]

間 (通算時間及び日内変動時間 0 ~ 23 時) も説明変数として準備した。 データの詳細は表 1 に示す。

学習及び予測を行う期間は，九州の広い範囲で Ox 濃度が高濃度となり, 熊本県でも注意報の発令が行われていた 2019 年 5 月 23 日 ~ 5 月 24 日を含む 10 日間とした。 また，予測地点には九州域を 5 × 5 [km] の格子領域に分割した地点を設定した。 表 2 に計算条件の一覧を示す。

4. 解析を実行するためのツールの選択

解析には R version 4.4.1 (2024-06-14) 及び mgcv パッケージ (ver. 1.9.1) を用いた。 R は回帰モデルの構築が容易であり，かつ，構築できるモデルの種類が多いことから，経済，医療，環境など様々な分野の解析に利用されている。 また， mgcv パッケージは R で GAM を実行するための標準的なパッケージであり, 利用実績も多いことから採択した。

以降では，GAM を用いて非線形モデルによる一般的なモデル (BASE) ，時間効果モデル (temporally vary coefficient model: TVC) ，空間効果モデル (spatially varying coefficient model: SVC) 及び時空間効果モデル (Spatially and Temporally Varying Coefficient models: STVC) による GAM を構築する。 なお，比較として線形回帰モデル (Linear Regression model: LM) による回帰モデルも構築する。

回帰モデルの構築と予測

1. 一般的な回帰モデル

時間や空間効果を検討する前に, 本章では一般的な回帰モデルの構築について解説する。

1.1 線形回帰モデルによる予測

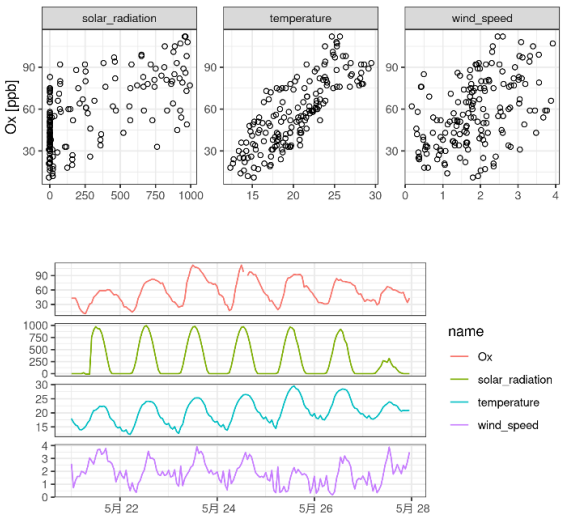

図 1 宇土運動公園における Ox 濃度と日射量，気温及び風速の関係。各変数の散布図及び時系列図。

R での回帰モデル構築について簡単に紹介する。 回帰モデル解説には, 分かりやすいように宇土運動公園 1 局のデータを使ってモデルを構築し，目的変数 (Ox) と説明変数 (日射量, 気温及び風速) の関係を確認する。 宇土運動公園局の Ox 濃度，日射量，気温及び風速については図 1 に示す。 図 1 によると，Ox 濃度と日射量及び気温は正の相関関係が見られている。 また，時系列データから日射量，気温及び Ox 濃度が日内変動で関連した挙動を示していることが示されている。

続いて，リスト 1 に線形回帰によるモデル化の結果 (m1) を示す。以下に, 出力結果について簡単に解説する。

表 2 計算条件

計算対象	概要	
学習データ	対象期間:	2019-05-21 0h ~ 2019-05-27 23h
	予測変数:	九州域の常時監視測定局 Ox 1 時間値データ (Ox)
	説明変数:	日射量 (solar-radiation)
		気温 (temperature)
		風速 (wind-speed)
予測データ	対象期間:	2019-05-23 6h ~ 2019-05-24 23h
	計算領域:	5 × 5 [km] (九州域)
	説明変数:	日射量 (solar-radiation)
		気温 (temperature)
		風速 (wind-speed)

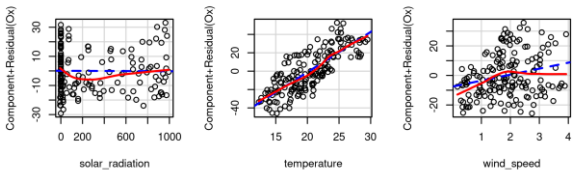

図 2 宇土運動公園における Ox 濃度の線形回帰分析結果
青破線:回帰直線 赤曲線:平滑化曲線

Call: には線形回帰モデルの構築方法が表示されている。formula が実際のモデル式を表し，予測変数 Ox を説明変数である日射量 (solar_radiation)，気温 (temperature) 及び風速 (wind_speed) で説明している。

Coefficients: には線形回帰式の切片 (Intercept) と各説明変数の重み (Estimate) が示されており，標準誤差 (Std.Error)，*t* 値 (t value) 及び *p* 値 (Pr(>|t|)) が更に続いている。また, 最後列には *p* 値の優位性 (Signif. codes) が記号 (*, .,) で示されている。 それ以降は決定係数 (*R^2* 値: Multiple R-squared) や *F* 検定の結果から求めた *p* 値 (p-value) が示されている。

リスト 1 から，線形回帰では R^2 値が 0.67 であり，予測に用いた説明変数のうち気温及び風速の影響が大きかった。なお，日射量の影響は小さかった。

図 2 に線形回帰で得られた Ox と各説明変数との関係を示す。 図の x 軸は説明変数の値，y 軸は各説明変数による Ox の予測値を示している。 また，青破線は線形回帰により得られた回帰直線を示したもので, 赤実線は平滑化曲線を示す。 図 2 から，説明変数のうち気温は回帰直線と平滑化曲線がほぼ一致しており, 当てはまりが良い結果であった。 一方で，日射量と風速はデータのばらつきが大きく, 直線性が見られず十分にモデル化されていなかった。

1.2 一般化加法モデルによる予測

次に, 同じデータを用いて GAM による回帰の方法を示す。 GAM を使った計算には， mgcv パッケージの gam()

リスト 2 宇土運動公園における Ox 濃度の一般化加法モデル結果

```
(m2 <- gam(formula = Ox ~ s(solar_radiation) + s(temperature) + s(wind_speed), data = na.omi
t(uto))) |> summary()

Family: gaussian
Link function: identity

Formula:
Ox ~ s(solar_radiation) + s(temperature) + s(wind_speed)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.1617     0.9612   59.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                    edf Ref.df      F  p-value
s(solar_radiation) 2.210  2.732  1.676 0.227818
s(temperature)     6.949  8.040 22.072  < 2e-16 ***
s(wind_speed)      3.645  4.528  4.979 0.000663 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.743   Deviance explained = 76.3%
GCV = 168.19  Scale est. = 154.28     n = 167
```

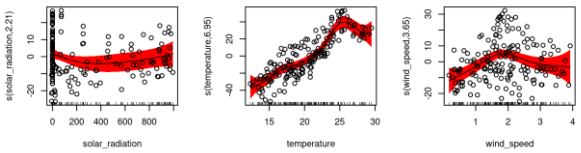

図 3 宇土運動公園における Ox 濃度の一般化加法モデル結果
黒破線:回帰曲線 赤色領域:95%信頼区間

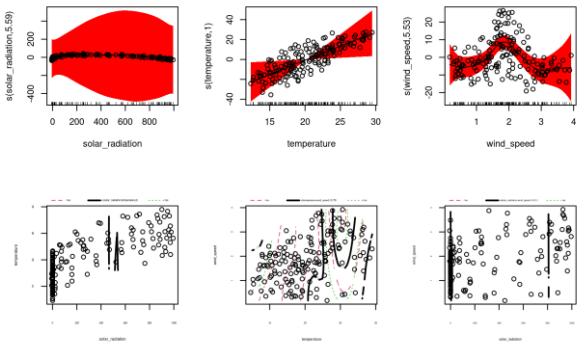

図 4 宇土運動公園における Ox 濃度の交互作用項を導入した GAM 予測結果
黒破線:回帰曲線 赤色領域:95%信頼区間

関数を用いる。 また, GAM ではスプライン関数を適応させるために s() 関数で説明変数を囲ってモデル式を構築する。 モデル化した GAM の結果 (m2) をリスト 2 に示す。 線形回帰と似た出力となっているが, 切片や説明変数の重みを示す coefficients: はパラメトリックとノンパラメトリック (Approximate significance of smooth terms:) に分けて表示されている。 また, 結果に逸脱度の説明性 (Deviance explained) [%] と一般化クロスバリデーション (Generalized Cross Validation: GCV) 誤差が表記されている点は, 線形回帰の結果の表示と異なっている。 ここで, 逸脱度の説明性 (Deviance explained) とは, 説明変数を用いずに計算した際の逸脱度 (Null Deviance) と説明変数を用いて計算した際の逸脱度 (Residual Deviance) との比をとっており, 値が大きいほどモデルの適合度は高いと言える (式 3)。

$$Deviance\ \ Explained = \left(1 - \frac{Null\ \ Deviance}{Residual\ \ Deviance}\right) \times 100 \quad (式\ 3)$$

m2 の結果では, R^2 値は 0.74 であり, 線形回帰よりも良い結果であった。 また, 逸脱度の説明性は 76.3 [%] であり, 総合的には m1 よりも m2 が良い結果となっていた。

次に, 説明変数の効果を図 3 に示す。 図の黒実線は回帰曲線, 赤色領域は 95%信頼区間を表している。 これによると, 気温は 25 [℃] , 風速は 2 [m/s] 付近で Ox 濃度が高い関係性が見られた。 説明変数の効果についても線形回帰の結果と異なっていた。

1.3 説明変数の選択と交互作用について

交互作用とは, 説明変数間で何らかの関係性が見られることを指し, この交互作用をモデルに組み込み事で予測精度が向上することがある。GAM では説明変数の交互作用を適応させた GA2M が考案されており, mgcv パッケージでは説明変数の交互作用を標準で扱うことができる [20]。本章では, GAM の交互作用の確認方法と説明変数の選択方法について確認する。

はじめに, 交互作用の確認のため説明変数間の相関係数を確認する (表 3) 。 変数間に相関が見られる場合, 交互作用がある可能性は高い。 表 3 によると, 日射量, 気温及び風速の間には一定以上の相関があり, 特に日射量と気温の間に強い相関が見られた。 このように, 説明変数間で交互作用が見られる可能性が高いことから, m2 モデルにすべての変数同士の交互作用を追加したモデル (m3) を定義した。 m3 の結果をリスト 3 及び図 4 に示す。 交互作用を定義していない m2 モデルと比べ, 交互作用を定義した m3 は R^2 値及び逸脱度の説明性が向上しており, m3 は m2 より明らかに良い結果となった。一方で, m3 の説明変数の中には m2 より p 値が低く, 精度が悪くなる変数も見られていた。 本検討では上記手順のように全ての説明変数間の組み合わせに相関が見られたため全ての組み合わせを適応させた計算を行ったが, 本来, 交互作用を検討する場合は, 計算負荷軽減のためにも, 事前に関連性の高い説明変数の組み合わせを検討し, その中から特に効果が期待される組み合わせを取捨選択して計算させるべきであり, このようにすることで回帰モデルの予測精度向上が期待できるようになると思われる。

一方で, 上記のように説明変数が複数存在する場合に説明変数間の関係を個別に確認することは, 時間に余裕がある場合を除き非効率的であり, 期待どおりの結果が得られない場合もある。 そ

表 3 説明変数間の相関係数

	solar_radiation	wind_speed	temperature
solar_radiation	1	0.4575	0.7593
wind_speed	0.4575	1	0.4405
temperature	0.7593	0.4405	1

リスト 3 交互作用を追加した GAM の構築

```
(m3 <- gam(formula = Ox ~ s(solar_radiation) + s(temperature) + s(wind_speed) +
             s(solar_radiation, temperature) + s(temperature, wind_speed) + s(solar_radiatio
n, wind_speed), data = na.omit(uto))) |> summary()

Family: gaussian
Link function: identity

Formula:
Ox ~ s(solar_radiation) + s(temperature) + s(wind_speed) + s(solar_radiation,
    temperature) + s(temperature, wind_speed) + s(solar_radiation,
    wind_speed)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.1617     0.7618   75.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                                   edf Ref.df     F p-value
s(solar_radiation)             5.585e+00  5.604 0.022 0.99994
s(temperature)                 1.000e+00  1.000 5.224 0.02401 *
s(wind_speed)                  5.526e+00  6.630 3.864 0.00121 **
s(solar_radiation,temperature) 1.071e-06 27.000 0.000 0.22728
s(temperature,wind_speed)      1.279e+01 27.000 2.800 < 2e-16 ***
s(solar_radiation,wind_speed)  1.951e+01 27.000 2.540 < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.839   Deviance explained = 88.2%
GCV = 133.12  Scale est. = 96.923     n = 167
```

のため, ステップワイズ法のように説明変数の取捨選択を自動化する手法も広く利用されており, 計算時間の短縮化も期待できる。 R の mgcv パッケージでは, GAM のステップワイズ法は定義されていないため, 本報では単純な GAM の前方ステップワイズ法を新たに定義し, 説明変数の自動化による取捨選択の手法を確認することとする。 なお, 説明変数の取捨選択の判断は *BIC* 基準[21]として実行した (リスト 4; 全出力は文末に記載)。

リスト 4 の結果を確認すると, 気温及び風速を説明変数としたモデルが *BIC* 基準で最も当てはまりの良い結果である結果となった。 一方で, R^2 値及び逸脱度の説明性を基準とすると, 気温, 風速及び日射量と風速の交互作用項を説明変数とすると, m3 よりもモデルの当てはまりが良くなっていた。 これは *BIC* 基準がパラメータの数 (説明変数の数) に対してペナルティを取る方法に起因していると思われ, ステップワイズ法を適用する場合は基準とする統計量の選択にも注意を払う必要がある。 なお, R の step() 関数にも使われている *AIC* 基準で試したところでは, 全ての説明変数及び交互作用を含めることが, 最も良い当てはまりとなっていた。

1.4 交差検証法による精度管理

続いて, モデルの精度検証に広く利用されている交差検証法 (クロスバリデーション) を用いた検証を行う。 交

リスト 4 前方ステップワイズ法による交互作用項を導入した GAM の構築（結果のみ表示）

```
Family: gaussian
Link function: identity

Formula:
Ox ~ s(temperature) + s(wind_speed)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.1617     0.9672    59.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                edf Ref.df      F  p-value
s(temperature) 7.468  8.429 38.812  < 2e-16 ***
s(wind_speed)  3.870  4.797  5.983 8.29e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =   0.74   Deviance explained = 75.8%
GCV = 168.67  Scale est. = 156.21    n = 167
```

差検証法では,予測に用いるデータを学習用と予測用に分け,学習用データのみを使って構築したモデルで予測用データを再現する作業を行う。 一般的な交差検証法では,データを 10 分割し, 9/10 のデータを用いて回帰モデルを構築し,残り 1/10 のデータを予測する。この計算を全ての組み合わせ (10 パターン) に対して行う (*10-fold-cross-validation*)。また,分割数を **k** で表して *k-fold-cross-validation* とも呼ばれている。

リスト 5 に新たに GAM 用の交差検証法を定義した。この関数を用いて,ここまでに構築したモデル m1 , m2 及び m3 の予測精度評価を行った。 交差検証法の評価には R^2 値及び *RMSE* 値を用いた。

図 5 に交差検証法による精度評価の結果を散布図で示す。 この図より, 3 モデルの中では m3 で R^2 値が最も大きく, *RMSE* 値が最も小さい値となり, 交互作用項を導入したモデルの予測精度が最も良い結果となった。 一方で, 線形回帰モデルの m1 は, 3 モデルの中で最も予測精度が悪い結果となった。 このように, 交差検証法を用いると予測精度を簡単に比較することができ,ここまでに構築した m1 ~ m3 では, GAM は線形回帰よりも予測精度が高いことが示された。

1.5 モデルの再構築

前節までの検討により,宇土運動公園 1 局の観測データを用いた回帰分析では, 線形回帰モデルよりも GAM の予測精度が高く,かつ,回帰の説明性が高いことが示された。次節からは,九州全域の観測データを用いて回帰モデルを構築・比較していくため,すでに構築したモデルの m1 及び m2 を, 九州全域のデータを使ってモデルを再構築した (m1 : LM, m2 : BASE, リスト 6)。 次節以降では新しい m1 (LM) 及び m2 (BASE) を用いて, 時間及び空間属性を導入した GAM との比較を行う。

なお, m3 で構築した交互作用項ありのモデルについては,次節からは扱うデータの次元が増えることで回帰モデルの構築の難易度が上がることが予想されるため,これ以上の検討は本報では扱わないこととする。 検討方法は他の研究論文等を参照してほしい [22)]。

2. 時間効果を考慮した GAM

前章では, 一般的な事象である, 無次元のデータを用いた GAM の構築について検討した。 本章からは, 対象期間の九州全域の常時監視測定局 Ox 濃度 1 時間値を用いて, 時間-空間変動を考慮した GAM の構築について検討

リスト 5 交差検証法の定義

```
cross_validation <- function(data = NULL,
                             formula = NULL,
                             nfold = 10,
                             rep = 10, ...) {
  stopifnot(nfold > 0)
  nfold <- floor(nfold)
  num <- seq(1, rep, by = 1)
  apply(do.call(cbind, lapply(num, function(n) {
    id <- sample(nrow(data))
    fold_list <- split(id, ceiling(seq_along(id) / floor(length(id) / nfold)))
    ret <- matrix(NA, nrow = length(id), ncol = 1)
    for (fn in fold_list) {
      fit_model <- mgcv::gam(formula = formula, data = data[-fn, ], ...)
      ret[fn] <- predict(fit_model, newdata = data[fn, ])
    }
    return(ret)
  })), 1, mean)
}
RMSE <- function(x = double(),
                 y = double()) {
  round(sqrt(mean((x - y) ^ 2, na.rm = TRUE)), digits = 2)
}
R2 <- function(obs = double(),
               predict = double()) {
  round(1 - mean((obs - predict) ^ 2, na.rm = TRUE) / var(obs, na.rm = TRUE), digits = 2)
}
```

リスト 6 m1 (LM) 及び m2 (BASE) モデルの再構築

```
m1 <- lm(formula = Ox ~ solar_radiation + temperature + wind_speed, data = Kyushu)
m2 <- gam(formula = Ox ~ s(solar_radiation) + s(temperature) + s(wind_speed), data = Kyushu)
```

リスト 7 時間効果を導入した GAM (m4) の構築

```
m4 <- gam(formula = Ox ~ s(time, bs = "gp") + s(date, bs = "gp"),
          data = Kyushu)
```

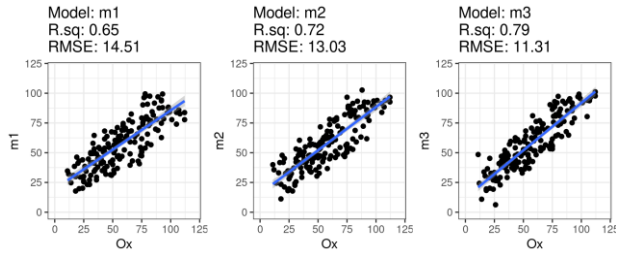

図 5 宇土運動公園のデータを用いた交差検証法による精度管理の結果
青線:回帰直線

する。 本節では時間効果を考慮した GAM (TVC) の検討を行う。

GAM の構築には, 引き続き mgcv パッケージの gam() 関数を用いる。 mgcv パッケージで時間変動を扱うためには説明変数に時間情報を追加する必要があり, 本報では通算時間 (date) による s(date, bs = "gp") , または, 日内変動時間 (0 ~ 23 時: time) による s(time, bs = "gp") をモデル式に追加する必要がある。 ここで bs = "gp" はスプライン関数にガウスモデルを適用させる方法であり,

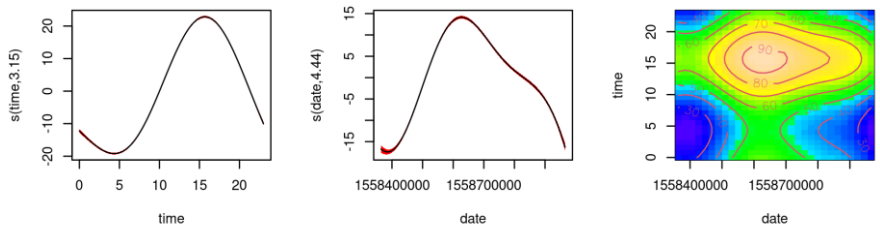

図 6 時間効果を用いた GAM (m4) の結果
黒破線:回帰曲線 赤色領域:95%信頼区間
コンター図:x 軸及び y 軸の説明変数の関係が見られる場合の Ox 濃度の平均濃度 [ppb]

リスト 8 時間効果及びその他説明変数を用いた GAM (m5) の構築

```
(m5 <- gam(formula = Ox ~ s(time, bs = "gp") + s(date, bs = "gp") + s(solar_radiation) + s(t
emperature) + s(wind_speed),
           data = Kyushu)) |> summary()

Family: gaussian
Link function: identity

Formula:
Ox ~ s(time, bs = "gp") + s(date, bs = "gp") + s(solar_radiation) +
    s(temperature) + s(wind_speed)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 57.56980    0.08307     693   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                    edf Ref.df       F p-value
s(time)           3.230  3.528  902.01  <2e-16 ***
s(date)           4.743  4.959 1939.84  <2e-16 ***
s(solar_radiation) 3.921  4.831   55.97  <2e-16 ***
s(temperature)    5.179  6.363  651.66  <2e-16 ***
s(wind_speed)     8.552  8.943   87.63  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Rank: 37/50
R-sq.(adj) =  0.708   Deviance explained = 70.8%
GCV =  161.4   Scale est. = 161.22    n = 23361
```

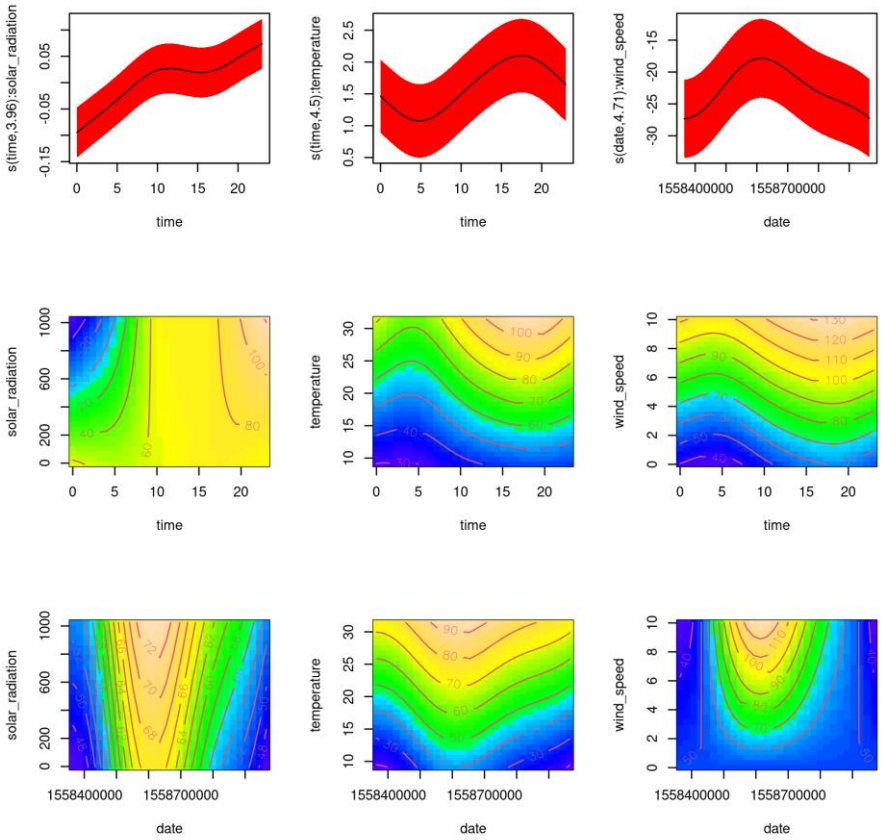

図 7 時間効果を説明変数に追加した GAM (m5) の結果
黒破線:回帰曲線 赤色領域:95%信頼区間
コンター図:x 軸及び y 軸の説明変数の関係が見られる場合の Ox 濃度の平均濃度 [ppb]

このオプションをつけることで時間及び空間効果が考慮されるようになる [19]。

はじめに，説明変数に時間効果のみで構築したモデル (リスト 7, 図 6: m4) について検討する。なお, 図 6 には，これまで同様 plot() 関数による説明変数 (単独または複数) 毎の Ox 濃度の予測値に加え， vis.gam() 関数による，2 つの説明変数間の関係をコンター図にして示している。

m4 の結果から，24 時間のうちで Ox 濃度が最も高くなる時間帯は 15 時付近，通算時間では 3 日 ~ 4 日目付近 (2019/5/23 ~ 5/24) であり, 図 1 に示した時系列図と一致する結果が得られた。

この m4 モデルの説明変数に日射量，気温及び風速を追加したモデルの結果をリスト 8 及び図 7 に示す (m5) 。m5 では，日射量が 600 [W/m^2] 付近で最低値となっていた。 気温は 20 [℃] 以上で Ox 濃度との関係が最大になっていた。 風速は 2 ~ 6 [m/s] 付近でプラスであり，それ以上風速が早くなっても影響度はほぼ変わらなかった。 時間情報と日射量, 気温及び風速の関係では, 日射量が多く，気温も高くなりやすい 15 時~20 時の間で Ox 濃度が高くなる傾向を示していた。 なお，風速は 15 時~20 時の間や 5/3 ~ 5/4 付近で，6 [m/s] の風があるときに Ox 濃度が最大値となっていた。

次に，日射量，気温及び風速を時間情報の因子とするモデルを構築する。 この場合，説明変数を s(date, bs = "gp", by = temperature) のように記述して GAM を構築する [23]。by = temperature 部分が気温のデータを時間情報の因子として取り込む設定を表している。 なお，説明変数の冒頭に 0 + を入れると，計算の過程で自動的に決定される切片係数を明示的に初期化し，各説明変数 (本報の場合，日射量，気温及び風速) の値を切片係数に割り当てるようになる [23]。

日射量及び気温を日内変動時間 (0 ~ 23 時) ，風速を通算時間の変動モデルとして構築した結果をリスト 9 及び図 8 に示す (m6) 。 この結果， 各説明変数の効果が m6 では m5 に比べて時間情報に適合しており，モデルが改善されていることが分かる。 また， m6 の逸脱度の説明性は 95.7 [%] で， m5 の 70.8 [%] に比べて大幅に向上していた。

リスト 9 時間効果を説明変数の因子として導入した GAM (m6) の構築

```
(m6 <- gam(formula = Ox ~ 0 +
             solar_radiation + s(time, bs = "gp", by = solar_radiation) +
             temperature + s(time, bs = "gp", by = temperature) +
             wind_speed + s(date, bs = "gp", by = wind_speed), data = Kyushu)) |> summary()

Family: gaussian
Link function: identity

Formula:
Ox ~ 0 + solar_radiation + s(time, bs = "gp", by = solar_radiation) +
    temperature + s(time, bs = "gp", by = temperature) + wind_speed +
    s(date, bs = "gp", by = wind_speed)

Parametric coefficients:
               Estimate Std. Error t value Pr(>|t|)
solar_radiation -0.01999    0.02369  -0.844  0.39868
temperature      0.83077    0.28834   2.881  0.00397 **
wind_speed      25.72510    3.09354   8.316  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                          edf Ref.df    F p-value
s(time):solar_radiation 3.964  4.595  108  <2e-16 ***
s(time):temperature     4.496  5.081  539  <2e-16 ***
s(date):wind_speed      4.709  5.220 1755  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Rank: 26/39
R-sq.(adj) =  0.696   Deviance explained = 95.7%
GCV = 167.99  Scale est. = 167.88    n = 23361
```

以上の結果から，時間変動を考慮したモデルでは m6 が最も良い結果となったため，以降では m6 を TVC の代表として比較を行う。

補足だが，日内変動時間のように循環する数値情報を扱う場合，GAM のモデル式では bs = "cc" とすることで，スプラインが循環する情報を扱うことができるようになる[20]。 このように設定するほうが良い結果となる場合もあるため，適用するスプラインや説明変数の関係を詳しく見ながらモデルを構築する必要がある。

3. 空間効果を考慮した GAM

本節では，空間効果を取り入れた GAM (SVC) の構築を行う。 位置情報の経度 (lon) と 緯度 (lat) を交互作用とし，スプラインはガウシアンモデルを設定して s(lon, lat, bs = "gp") のようにしてモデルを構築すると，空間効果を取り入れたモデルができる。

まず，リスト 10 及び図 9 に空間効果のみで構築した GAM の結果 (m7) を示し，Ox 濃度との関連度を確認する。 m7 の結果 (図 9) では，$X-Y$ 平面上に経度及び緯度の効果が等高線で示されている。 これによると，福岡県の北九州市周辺や大分県の一部で Ox 濃度が高くなる傾向にあり，長崎県や鹿児島県では Ox 濃度が低い傾向にあった。 なお，海上で Ox の生成にプラスの影響が見られている領域は，もともとデータが存在しない範囲であるた

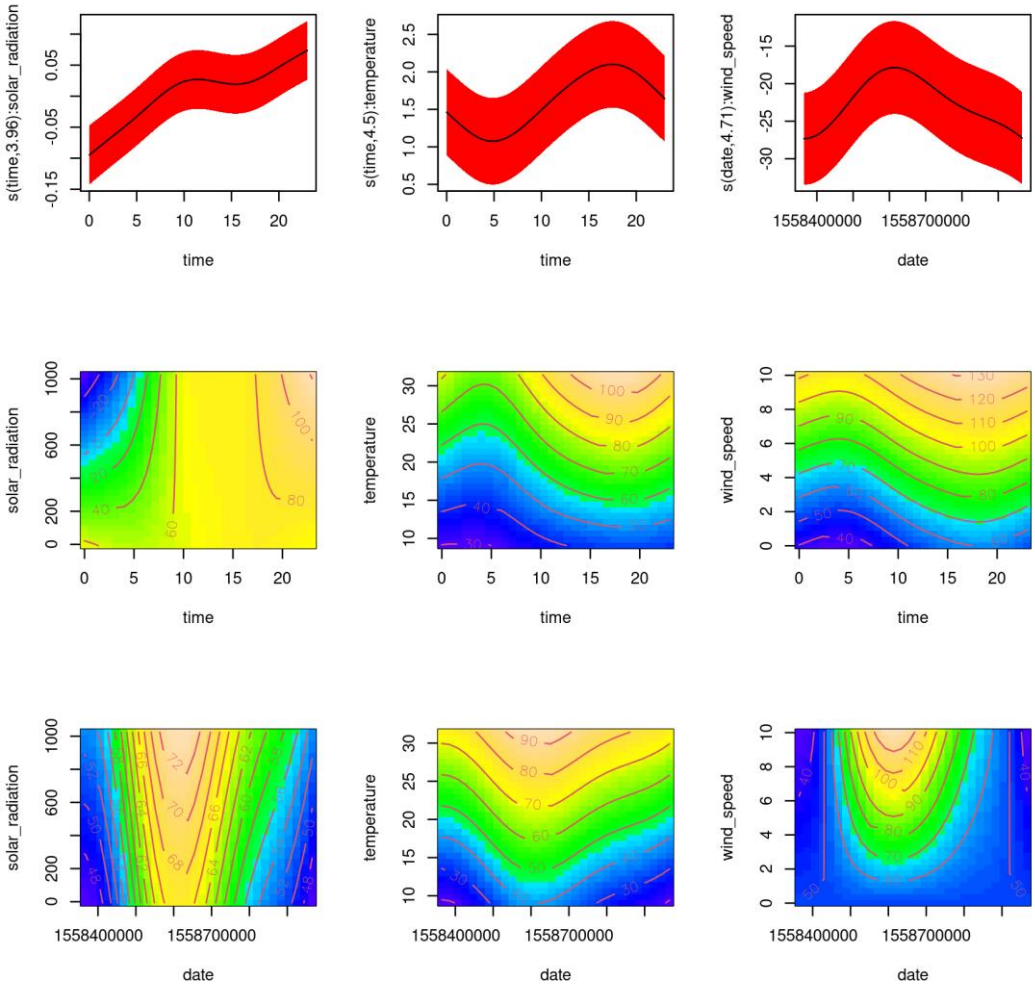

図 8 時間効果を説明変数の因子として導入した GAM (m6) の結果
黒破線:回帰曲線 赤色領域:95%信頼区間
コンター図:x 軸及び y 軸の説明変数の関係が見られる場合の Ox 濃度の平均濃度 [ppb]

リスト 10 空間効果を導入した GAM (m7) の構築

```
m7 <- gam(formula = Ox ~ s(lon, lat, bs = "gp"), data = Kyushu)
```

めこのようになっているが, 本検討では海上の予測は行わないため検討は省略する。

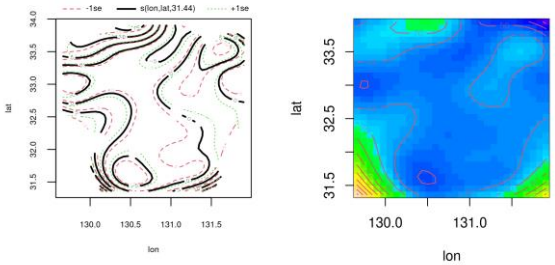

図 9 空間効果を導入した GAM (m7) の結果
黒破線:回帰曲線 赤色領域:95%信頼区間
コンター図:x 軸及び y 軸の説明変数の関係が見られる場合の Ox 濃度の平均濃度 [ppb]

リスト 11 空間効果を説明変数に追加した GAM (m8) の構築

```
(m8 <- gam(formula = Ox ~ s(lon, lat, bs = "gp") + s(solar_radiation) + s(temperature) + s(w
ind_speed), data = Kyushu)) |> summary()

Family: gaussian
Link function: identity

Formula:
Ox ~ s(lon, lat, bs = "gp") + s(solar_radiation) + s(temperature) +
    s(wind_speed)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.5698     0.1017   565.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                     edf Ref.df       F p-value
s(lon,lat)         31.702 31.965   60.43  <2e-16 ***
s(solar_radiation)  8.875  8.995   36.08  <2e-16 ***
s(temperature)      8.833  8.992 1359.28  <2e-16 ***
s(wind_speed)       5.626  6.823   82.29  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.562   Deviance explained = 56.3%
GCV = 242.39  Scale est. = 241.81     n = 23361
```

続いて， m7 に日射量, 気温及び風速を説明変数として追加した結果をリスト 11 及び図 10 に示す (m8) 。m8 では，経度 131 度~131.5 度付近及び緯度 33.5 度以上での領域で，日射の強さに関係なく Ox 濃度が高くなる傾向にあった。 気温及び風速効果は，位置情報に関係なく，気温が高く風速が遅いほど Ox 濃度が高くなる傾向にあった。また， m8 の逸脱度の説明性は 56.3 [%] であり， m6 と比べて極端に低い値となっていた。

最後に，日射量, 気温及び風速を空間情報の因子に設定して構築したモデルの結果をリスト 12 及び図 11 に示す (m9) 。 m9 結果では， m8 の結果に比べて日射量及び風速の効果がより現実的なものとなっており， m8 で見られた極端な地域性は見られなかった。 なお， m9 の逸脱度の説明性は 93.1 [%] となっており， m7 や m8 と比べて良い結果であった。

以上の結果から，空間変動を考慮したモデルでは m9 が最も良い結果となり，以降では m9 を SVC の代表として比較を行う。

4. 時空間効果を考慮した GAM

最後に，本節では時空間効果を考慮したモデル (STVC) を構築する。 時空間効果によるモデル化では，位置情報 lon 及び lat に加え，時間情報 date または time を交互作用とし， s(lon, lat, date, bs = "gp") のようにモデルを構築する [19) 23)]。

まずは時空間情報のみで構築したモデルを構築し (m10) ,その結果をプロットする (リスト 13 及び図 12) 。なお，これまで各説明変数による Ox 濃度の予測値をプロットしていたが，時空間効果の情報量が多くなり，結果の表示が見にくくなってしまうため，本節では vis.gam() 関数だけで結果を表示する。

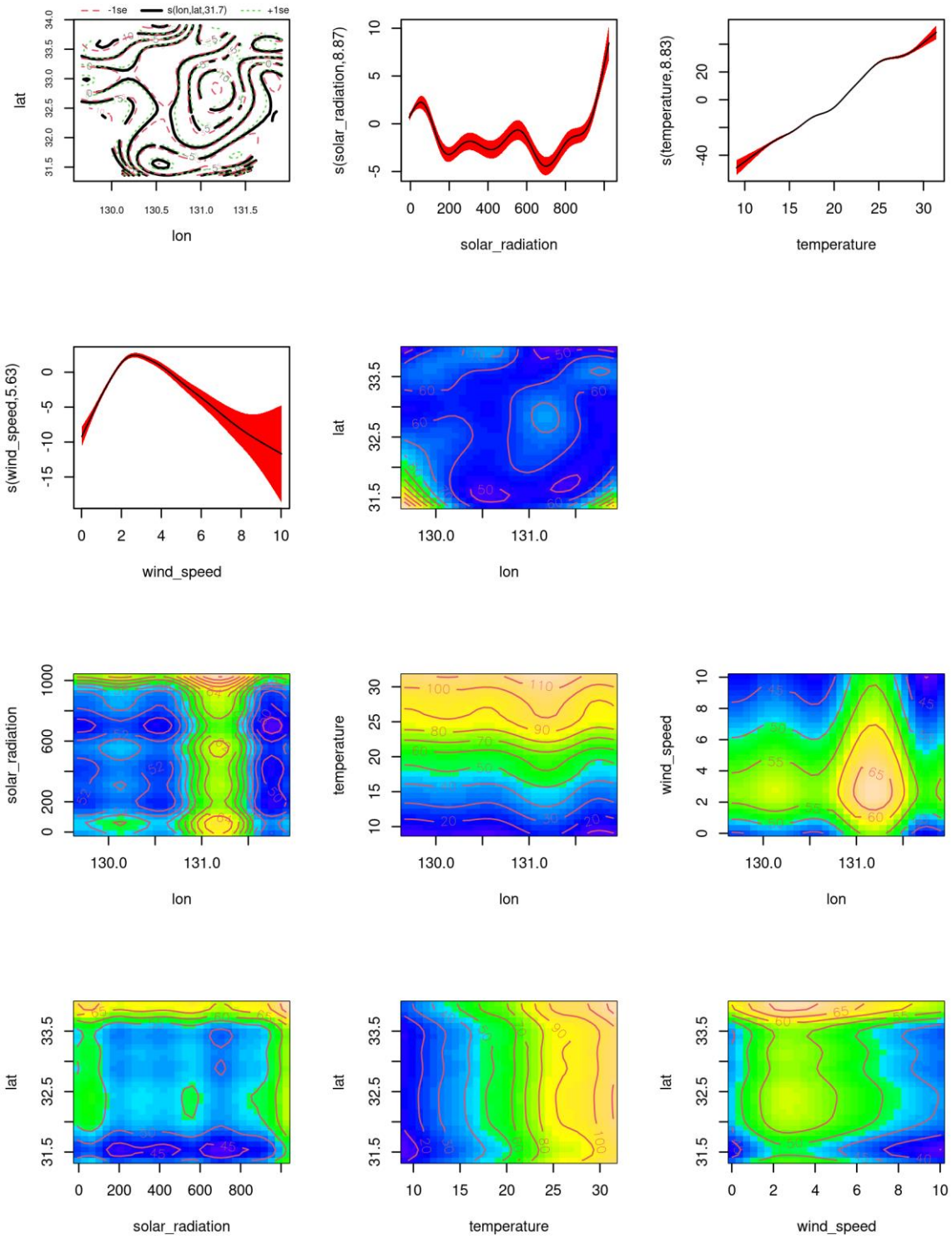

図 10 空間効果を説明変数に追加した GAM (m8) の結果

黒破線:回帰曲線 赤色領域:95%信頼区間

コンター図:x 軸及び y 軸の説明変数の関係が見られる場合の Ox 濃度の平均濃度 [ppb]

リスト 12 空間効果を説明変数の因子として導入した GAM (m9) の構築

```
(m9 <- gam(formula = Ox ~ 0 +
             solar_radiation + s(lon, lat, bs = "gp", by = solar_radiation) +
             temperature + s(lon, lat, bs = "gp", by = temperature) +
             wind_speed + s(lon, lat, bs = "gp", by = wind_speed), data = Kyushu)) |> summar
y()

Family: gaussian
Link function: identity

Formula:
Ox ~ 0 + solar_radiation + s(lon, lat, bs = "gp", by = solar_radiation) +
    temperature + s(lon, lat, bs = "gp", by = temperature) +
    wind_speed + s(lon, lat, bs = "gp", by = wind_speed)

Parametric coefficients:
                Estimate Std. Error t value Pr(>|t|)
solar_radiation   0.04438    0.09031   0.491    0.623
temperature      16.20156   10.78568   1.502    0.133
wind_speed      -70.56022   74.90184  -0.942    0.346

Approximate significance of smooth terms:
                             edf Ref.df      F p-value
s(lon,lat):solar_radiation 26.60  29.02  8.614  <2e-16 ***
s(lon,lat):temperature     30.75  31.44 20.808  <2e-16 ***
s(lon,lat):wind_speed      30.42  31.14 16.827  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Rank: 99/102
R-sq.(adj) =  0.514   Deviance explained = 93.1%
GCV = 270.07  Scale est. = 269.04    n = 23361
```

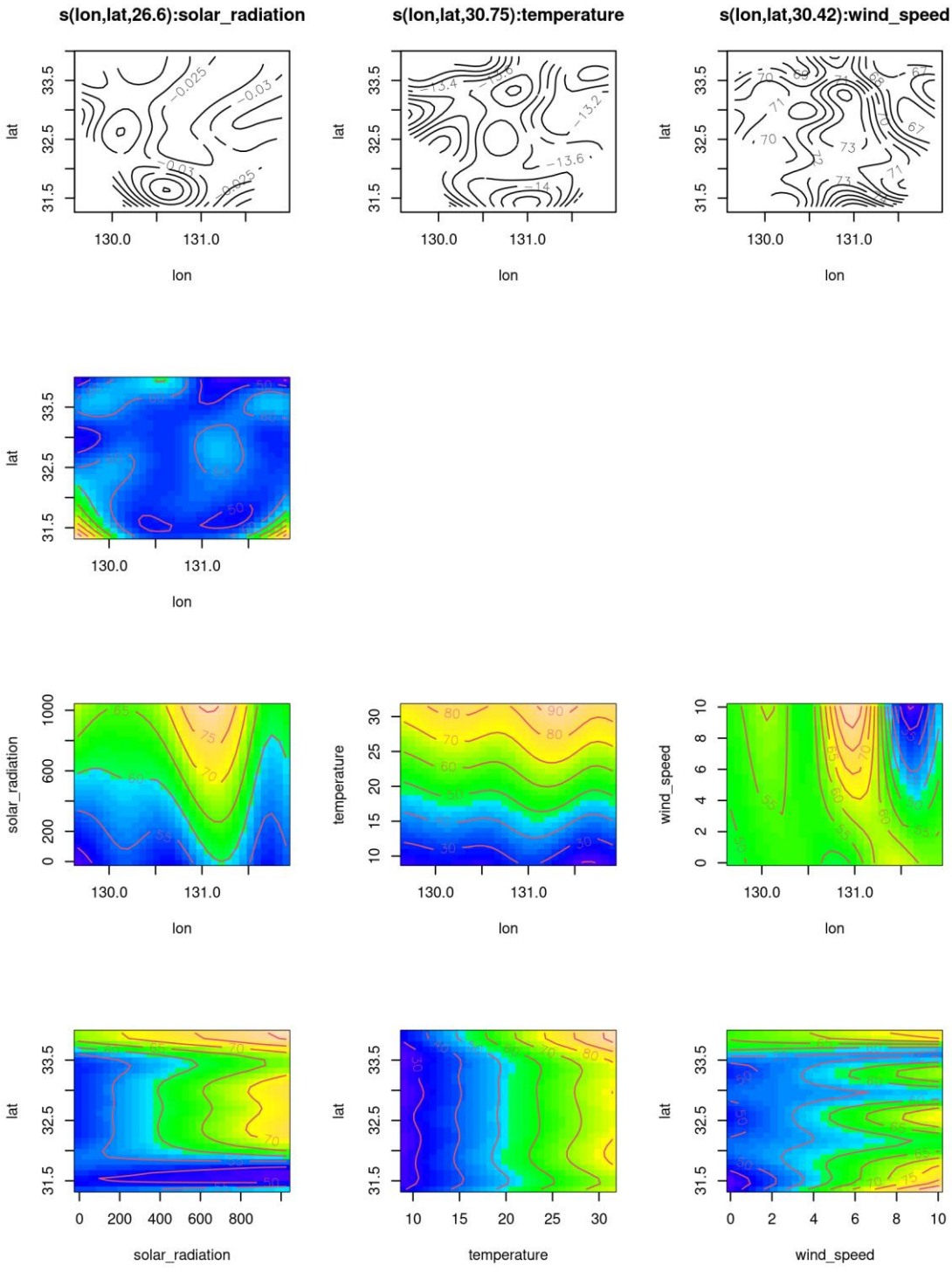

図 11 空間効果を説明変数の因子として導入した GAM (m9) の結果

黒実線:回帰曲線

コンター図:x 軸及び y 軸の説明変数の関係が見られる場合の Ox 濃度の平均濃度 [ppb]

リスト 13 時空間効果を導入した GAM (m10) の構築

```
m10 <- gam(formula = Ox ~ s(lon, lat, time, bs = "gp") + s(lon, lat, date, bs = "gp"), data = Kyushu)
```

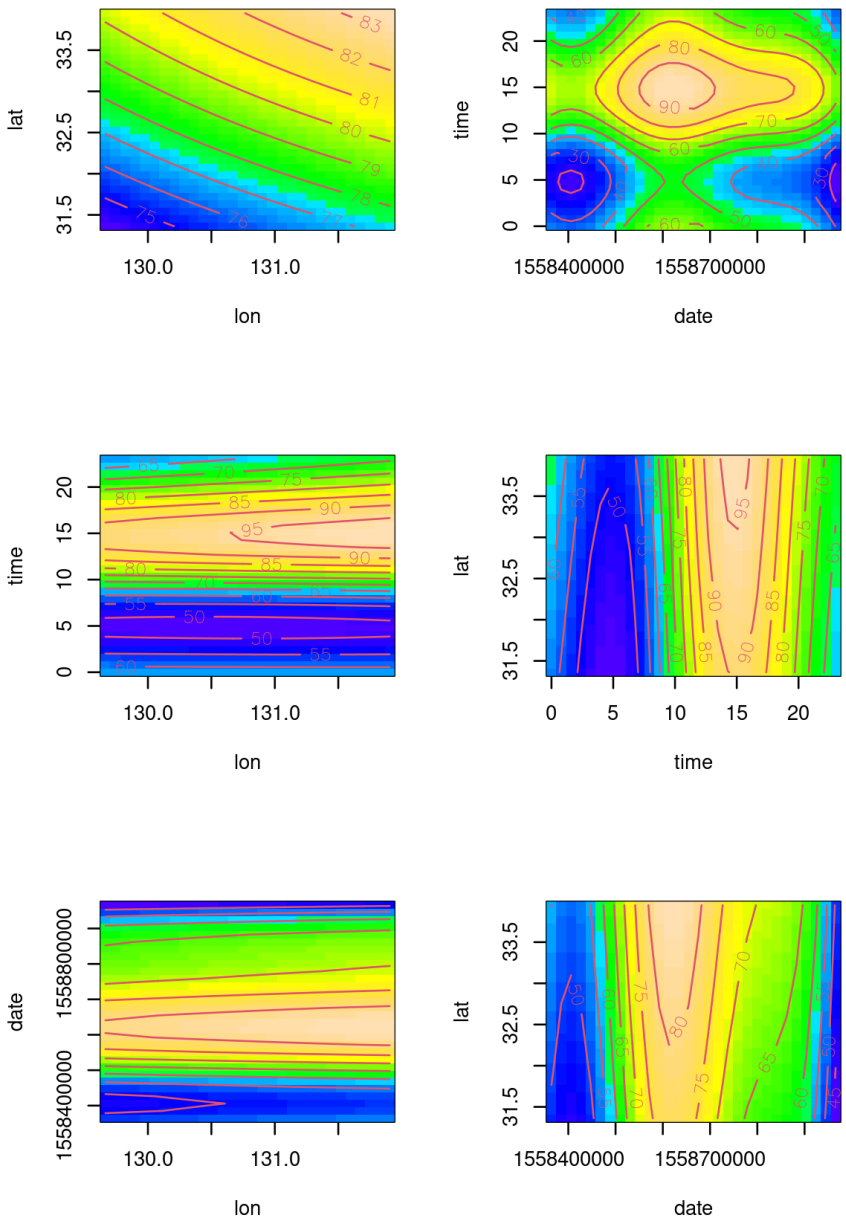

図 12 時空間効果を導入した GAM (m10) の結果

黒破線:回帰曲線 赤色領域:95%信頼区間

コンター図:x 軸及び y 軸の説明変数の関係が見られる場合の Ox 濃度の平均濃度 [ppb]

m10 の結果では，時間及び空間変動効果は m6 (TVC) 及び m9 (SVC) と同様の結果となっていたが，空間変動において m8 及び m9 で見られた地域差がほぼなくなり，北東から南西にかけて徐々に Ox 濃度が低くなる傾向になっていた。

次に m11 に日射量，気温及び風速を説明変数として追加したモデルの結果をリスト 14 に示す。 m11 では，各説明変数の p 値が 0.01 以下で当てはまりは良かったが，モデル結果のプロット (vis.gam()) がエラーで出力されなかった。そのため， m11 のモデルの評価は省略することとする。

最後に，日射量，気温及び風速を時空間効果の因子に設定して構築したモデルの結果をリスト 15 及び図 13 に示す (m12) 。 なお， m12 ではプロットの数が多いため，時間属性を上 2 段，空間属性を下 2 段に分けて表示している。m12 の時間変動は m6 とほぼ同じ結果となっていた。また，空間変動は m9 とは異なり，空間依存性をほとんど示さなっていた。 日射量のみ空間的な変動が確認されたが，空間効果は小さかった。

なお， m12 モデルの R^2 値は 0.72 ，逸脱度の説明性は 96.0 [%] であり，これまでに構築したモデルの中でも良い結果となった。

このように，GAM では空間効果より時間効果の影響度が強く現れる結果となった。

リスト 14 時空間効果を説明変数に追加した GAM (m11) の構築

```
(m11 <- gam(formula = Ox ~ s(lon, lat, date, bs = "gp") + s(lon, lat, time, bs = "gp") + s(s
olar_radiation) + s(temperature) + s(wind_speed),
            data = Kyushu)) |> summary()

Family: gaussian
Link function: identity

Formula:
Ox ~ s(lon, lat, date, bs = "gp") + s(lon, lat, time, bs = "gp") +
    s(solar_radiation) + s(temperature) + s(wind_speed)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 57.56980    0.08764   656.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                     edf Ref.df       F p-value
s(lon,lat,date)    6.963  6.999 1534.95  <2e-16 ***
s(lon,lat,time)    2.000  2.000   64.51  <2e-16 ***
s(solar_radiation) 8.897  8.997  132.95  <2e-16 ***
s(temperature)     6.849  7.980 1610.95  <2e-16 ***
s(wind_speed)      7.486  8.441   68.40  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Rank: 35/131
R-sq.(adj) =  0.675   Deviance explained = 67.6%
GCV = 179.69  Scale est. = 179.45    n = 23361
```

リスト 15 時空間効果を説明変数の因子として導入した GAM (m12) の構築

```
(m12 <- gam(formula = Ox ~ 0 +
              solar_radiation + s(lon, lat, time, bs = "gp", by = solar_radiation) +
              temperature + s(lon, lat, time, bs = "gp", by = temperature) +
              wind_speed + s(lon, lat, date, bs = "gp", by = wind_speed), data = Kyushu)) |>
summary()

Family: gaussian
Link function: identity

Formula:
Ox ~ 0 + solar_radiation + s(lon, lat, time, bs = "gp", by = solar_radiation) +
    temperature + s(lon, lat, time, bs = "gp", by = temperature) +
    wind_speed + s(lon, lat, date, bs = "gp", by = wind_speed)

Parametric coefficients:
                Estimate Std. Error t value Pr(>|t|)
solar_radiation   0.8676     1.4072   0.617   0.5375
temperature       9.8369     4.1690   2.360   0.0183 *
wind_speed        7.7110     0.7122  10.827   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                                   edf Ref.df       F p-value
s(lon,lat,time):solar_radiation 31.042 32.086   25.11  <2e-16 ***
s(lon,lat,time):temperature     13.848 14.336  180.45  <2e-16 ***
s(lon,lat,date):wind_speed       7.917  7.924 1307.07  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Rank: 58/315
R-sq.(adj) =   0.72   Deviance explained =   96%
GCV = 154.97  Scale est. = 154.61    n = 23361
```

以上の結果から，時空間変動を考慮したモデルでは m12 が最も良い結果となり，以降では m12 を TSVC の代表として比較を行う。

5. 最適な回帰モデル式の選択

これまで構築した m1 (LM)， m2 (BASE)， m6 (TVC)， m9 (SVC) 及び m12 (STVC) について比較検証を行う。モデルの比較には， R^2 値， *RMSE* 値， *AIC* ， *BIC* 及び逸脱度の説明性 (Dev.exp) を用いる。 また，予測精度の比較には， 10-fold-cross-validation で求めた R^2 値及び *RMSE* 値を用いる。

各モデルの統計量を表 4 に示す。 この結果によると， R^2 値， *RMSE* ， *AIC* 及び *BIC* は， m6 (TVC) 及び m12 (SVC) が他のモデルに比べて良い結果となっていた。 また，逸脱度の説明性については， m2 (BASE) が最も低く，それ以外の m6 (TVC)， m9 (SVC) 及び m12 (STVC) で高い値を示していた。

続いて，交差検証法による予測精度検証の結果を図 14

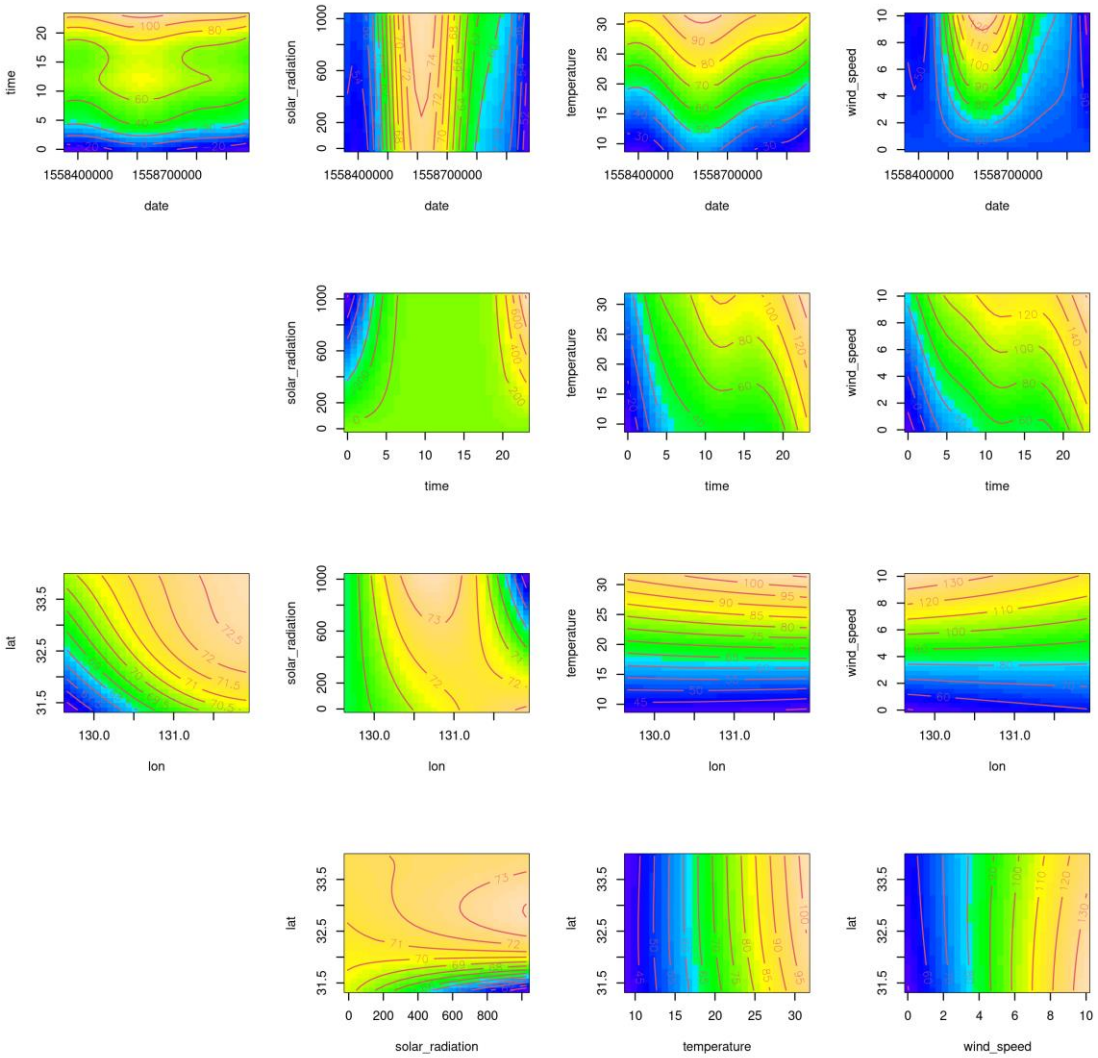

図 13 時空間効果を説明変数の因子として導入した GAM (m12) の結果
黒破線:回帰曲線 赤色領域:95%信頼区間
コンター図:x 軸及び y 軸の説明変数の関係が見られる場合の Ox 濃度の平均濃度 [ppb]

表 4 GAM の予測精度の比較 R2 値 (R.sq) , AIC, BIC 及び逸脱度の説明性 (Dev.exp)

	R.sq	AIC	BIC	Dev.exp
m1 (LM)	0.5016	197540.5	197580.8	−
m2 (BASE)	0.5265	196364.1	196565.3	0.5269
m6 (TVC)	0.6961	185996.5	186122.3	0.9566
m9 (SVC)	0.5136	197087.9	197815.3	0.9307
m12 (STVC)	0.7201	184112.0	184554.0	0.9601

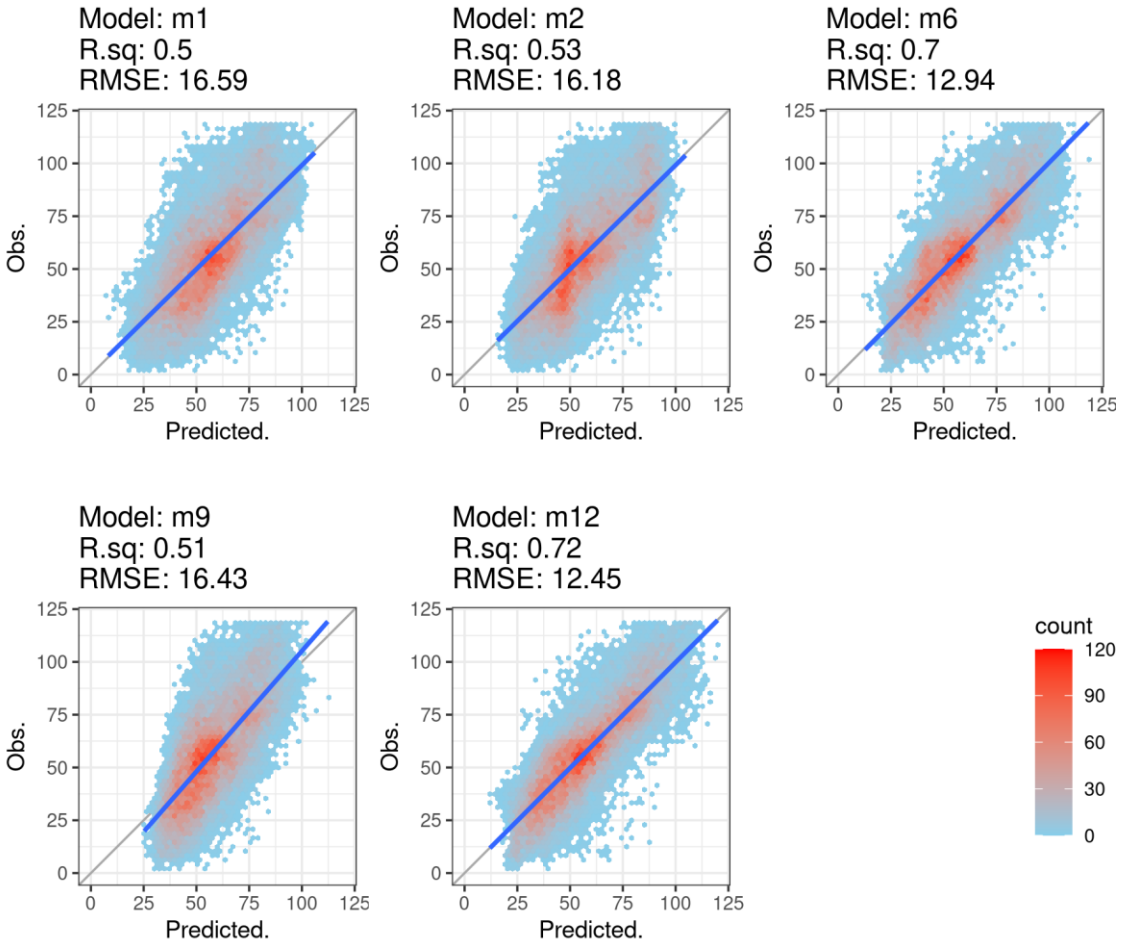

図 14 時空間効果を導入した GAM の結果比較
黒線:観測値 (Obs.) 及び予測値 (Predicted.) の同値線 青実線:回帰直線

に示す。 これらの散布図では，y 軸に観測値，x 軸に交差検証法の結果を示しており，データの重なり部分の色を変えて表示できるよう密度分布で示している。

図 14 によると， m6 (TVC) 及び m12 (STVC) で精度が高い結果となり，それ以外の m1 (LM)， m2 (BASE) 及び m9 (SVC) とは R^2 値及び *RMSE* 値に大きな差が見られた。 m6 (TVC) 及び m12 (STVC) では，データの密度が大きい領域が回帰直線付近に集まっており，かつ，線形に伸びていることから，これらのモデルの予測精度が高いことが示された。 なお， m6 (TVC) と m12 (STVC) との予測精度の差は小さく，空間変動の効果は時間変動の効果に比べて小さいことが示された。 しかしながら，時間効果だけでは説明ができなかった不確実性を空間効果が補って GAM が構築されていると思われるため，時間効果単独よりも時空間効果を取り込んだモデルを構築することでモデルの説明性が高くなっていると思われる。

これらの結果から，時空間効果を組み込んだ m12 (STVC) の精度及び回帰の説明性が，これまで構築したモデルの中で最も良い結果であることが示された。

6. 時空間データの予測

m12 (STVC) を使って空間濃度分布の予測を行う。 表 2 に示す条件で予測地点のデータを作成し， R の predict.gam() 関数で空間濃度分布の予測計算を行った (図 15) 。 結果は 2019-05-23 6h～2019-05-24 23h までの時間帯について示している。 なお，図 15 では m12 (STVC) による Ox 空間濃度分布予測値を地図上にコンター図として表示し，更に，常時監視測定局の位置及び観測値をコンター図に重ねて表した。

図 15 から， m12 (STVC) の予測値は観測値を概ね再現できていることが示された。 特に，Ox 濃度の予測値が低濃度になりやすい夜間～早朝の時間帯の予測精度につい

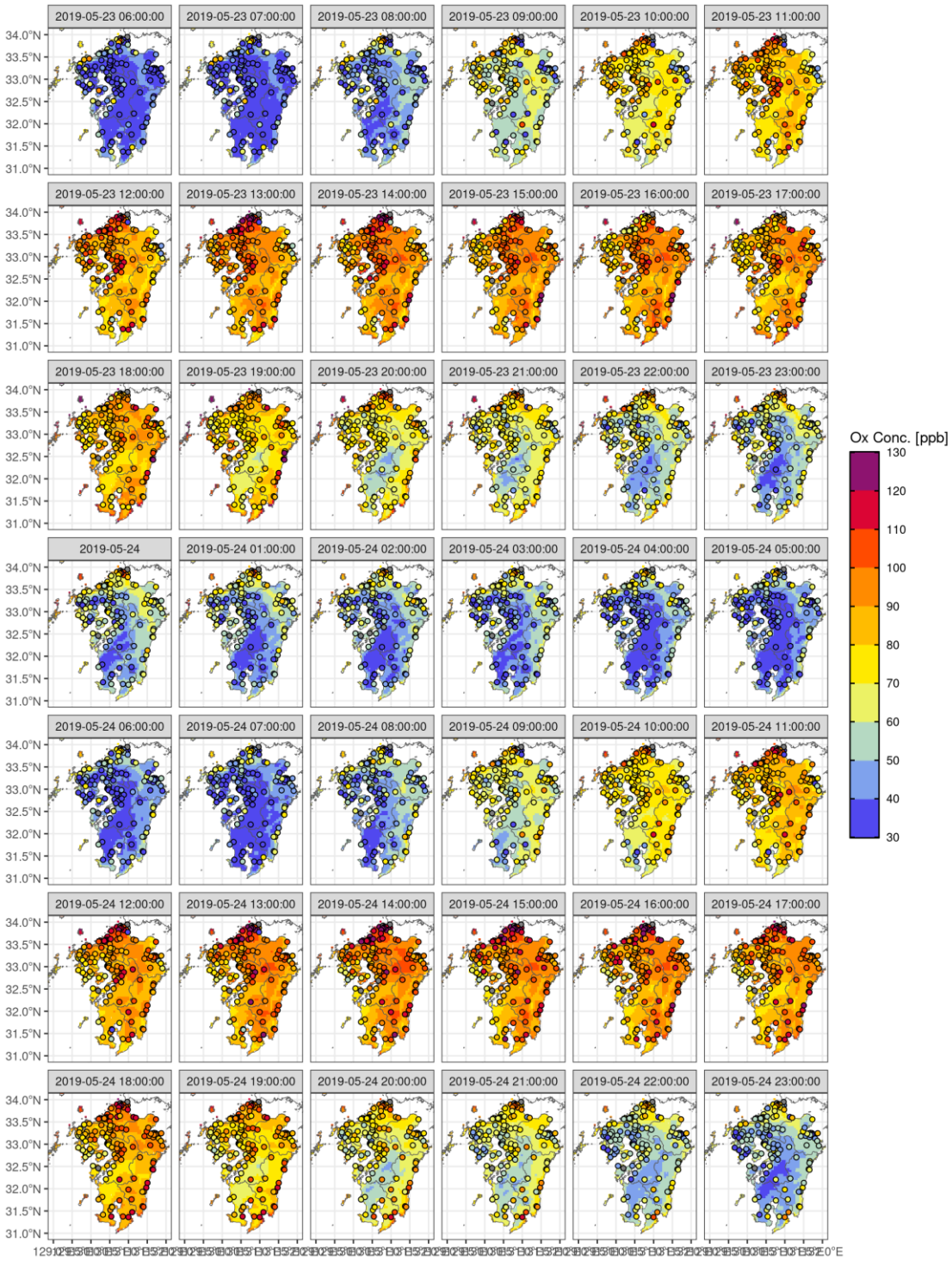

図 15 m12 (STVC) モデルを用いた空間濃度分布の予測結果
ポイントデータ:常時監視測定局の位置及び観測値
コンター図:m12 モデルを用いた Ox の空間濃度分布

ては高い予測精度となっていた。 一方で，日中に Ox 観測値が 120 [ppb] を超過するような極端に高い濃度を観測していた時間，2019/5/23 6h 付近で日射量が少ない時間帯に既に Ox 濃度が上昇している時間，2019/5/24 0h 付近の

ように日中に上昇した Ox 濃度が夕方～夜間でも低下しなかった時間帯について， m12 (STVC) では十分な予測ができなかった。

予測が十分ではなかった理由として，以下の 2 つの要因が考えられる。 1 つ目の要因は， m12 のモデルに組み込んだ通算時間 (date) の変動が実際の Ox 濃度の変動を捉えきれていなかったことである。 m6 ~ m9 において，date の効果は 3~4 日目の濃度が最も高くなるように，なだらかな曲線としてモデル化がされていた。 観測値であれば日内変動が見られるように，Ox 濃度の増減が表されているはずであるが，GAM を構築する際の設定でこのような結果となった。 機械学習では，予測を行う場合に学習データからの過学習が度々問題になっていることを考えると，この程度の予測誤差は起こりうることと思われる。一方で，日射量や気温の変動は毎日同じでは無いため，時間効果を time として構築した GAM では，時間変動の情報だけでは極端な観測値の予測は難しいと思われる。

2 つ目の要因は，説明変数とし使用した日射量，気温及び風速では説明できない事象が実際に起きていたことである。 Ox の生成に大きく影響している日射量と気温は，日の出から日の入りの時間帯に大きく依存しており，日の入り後の時間は日射の影響がないことから，Ox の生成量が 0 になる。 更に，Ox の主成分である O_3 は，NO による分解で減少することが知られている[24]。 そのため，Ox の国内における生成量は日射量に大きく影響を受けていると言えるが，国外で生成された Ox，又は Ox の前駆物質が国内に流入する現象である越境移流が起きていた場合，日射量，気温，風速及びこれらの相互関係からだけでは説明することができない濃度の変動を示すこととなり，通常であれば Ox 濃度が低下している時間帯に Ox 濃度が上昇する可能性がある。 本報の計算対象期間では，夜間に観測値で Ox 濃度が高い時間及び場所が確認でき，国外からの越境移流があった可能性が高い。 これらの事象の詳細な解析は本報の検討内容から逸脱するため追加検討は行わないこととするが， 越境移流を説明することができる説明変数には，化学輸送モデルによるシミュレーションの結果等，明らかに国外からの影響と識別することができるデータに限られている[6]。 他にも，回帰の結果に地球統計学手法である Kriging を組み合わせて，予測誤差を補間する手法の検討が行われており[11]，この手法を取り入れて予測精度を高めることも改善策の一つと言える。

ま と め

GAM により構築される予測精度の高いモデルは，その結果から多くの情報を得ることができ，説明変数及び予測変数間の関係性を詳しく確認することができた。 構築したモデルの中では，時空間効果を導入したモデルで最も高い予測精度及び説明性を示していたが，説明性が高い反面，m12 (STVC) による空間濃度分布の予測結果のように，空間的な変動は観測値と異なる部分もあった。 そのため，計算結果をそのまま受け入れるのではなく，GAM の結果を確認し，必要に応じてモデルを改善していく必要があることが示された。加えて，本報では時空間効果を定義する際に，説明変数を因子として指定する方法が最も精度が高い結果となっていたが， m5 ， m8 及び m11 のように " 説明変数 + 時空間効果の交互作用" とする方法が良い場合もある。 予測変数の種類でもモデルの構築方法は異なってくるため，最低限，本報で解説した方法を試して欲しい。

文 献（MS ゴシック 9pt）

1) e-Gov ポータル． https://elaws.e-gov.go.jp/ (2024 年 7 月閲覧)．
2) 環境省 ： 大気汚染防止法第２２条の規定に基づく大気の汚染の状況の常時監視に関する事務の処理基準 ． https://www.env.go.jp/air/osen/law22_kijun.html (2024 年 7 月閲覧)．
3) 環境省 ： 令和４年度 大気汚染状況について． https://www.env.go.jp/press/press_03287.html (2024 年 7 月閲覧)．
4) 菅田誠治， 大原利眞， 黒川純一， 早崎将光 ： 大気汚染予測システム (VENUS) の構築と検証， *J. Jpn. Soc. Atmos. Environ.*， **46(1)**， 49-59 (2011)．
5) 山村由貴， 新谷俊二， 力寿雄， 中川修平， 王哲， 鵜野伊津志 ： 夏季の太平洋高気圧条件下における高濃度 PM2.5 に対する火山の寄与解析， *J. Jpn. Soc. Atmos. Environ.*， **55(4)**， 169-180 (2020)．
6) 古澤尚英， 板橋秀一， 豊永悟史， 村岡俊彦 ： 2014 年冬季の熊本県中心部における微小粒子状物質の発生源感度解析， *J. Jpn. Soc. Atmos. Environ.*， **53(5)**， 194-205 (2018)．
7) 国立環境研究所： 地方環境研究所等との共同研究． https://www.nies.go.jp/kenkyu/chikanken/ (2024 年 7 月閲覧)．
8) 古澤尚英， 曽我稔， 豊永悟史， 荒木真， 菅田誠司 ： 大気環境学会九州支部第 24 回研究発表会講演要旨集， p14， (2024)．
9) 古澤尚英， 豊永悟史， 荒木真， 曽我稔， 菅田誠司 ： 第 65 回大気環境学会年会講演要旨集， p306， (2024)．
10) 辻本昌礼， 山本浩平， 亀田貴之 ： 気象モデル推定値を取り入れた Land Use Regression モデルによ

る国内大気汚染物質濃度分布推定, *J. Jpn. Soc. Atmos. Environ.*, **57(1)**, 1-14 (2022).

11) 小原大翼, 豊永悟史, 古澤尚英, 荒木真, 山本裕典, 矢野弘道, 山崎文雅 : 地方自治体における PM2.5 常時監視ネットワークの効率化の検討（I） -Regression Kriging 法による空間濃度分布予測-, *J. Jpn. Soc. Atmos. Environ.*, **57(2)**, 53-65 (2022).

12) 杉山将, 井手剛, 神嶌敏弘, 栗田多喜夫, 前田 英作 監訳 : 統計的学習の基礎 データマイニング・推論・予測, p14, p337-348 (2014), (共立出版) ; {Trevor Hastire, Robert Tibshirani, Jerome Friedman: ″The Elements of Statistical Learning -Data Mining, Inference, and Prediction Second Edition-″, (2009), (Springer New York) }.

13) Araki, S., Shima, M., Yamamoto, K : Spatiotemporal land use random forest model for estimating metropolitan NO2 exposure in Japan, *Science of The Total Environment*, **634**, 1269-1277 (2018).

14) 松田晃一 訳 : 解釈可能な AI -機械学習モデルの解釈手法を実践的に理解する-, p24-61 (2023), (株式会社マイナビ出版) ; {Ajay Thampi: Interpretable AI -Building Explainable Machine Learning Systems-, (2022), (Manning Publications) }.

15) 辻谷将明, 外山信夫 : R による GAM 入門, *行動計量学*, **34(1)**, 111-131 (2007).

16) 大原利眞, 若松伸司, 鵜野伊津志, 安藤保, 泉川碩雄 : 関東・関西地域における光化学オキシダントの経年動向に関する解析, *J. Jpn. Soc. Atmos. Environ.*, **30(2)**, 137-148 (1995).

17) 星純也, 石井康一郎 : 関東地域における揮発性有機化合物 (VOC) 排出量の変化と光化学オキシダント生成の関係について, *J. Jpn. Soc. Atmos. Environ.*, **48(5)**, 215-222 (2013).

18) 熊本県 : 熊本県の大気汚染の状況. https://kumamoto-taiki.jp/index.html (2024 年 7 月閲覧).

19) 村上大補 : 実践 Data Science R ではじめる地理空間データの統計解析入門, (2022), (講談社).

20) Simon Wood. : Mixed GAM Computation Vehicle with Automatic Smoothness Estimation (CRAN `mgcv` package manual.). https://cran.r-project.org/web/packages/mgcv/index.html (2024 年 7 月閲覧).

21) Schwarz Gideon. : Estimating the dimension of a model, *The annals of statistics*, 461-464 (1978).

22) Simon N. Wood : Generalized additive models: an introduction with R, p40 (2017), (chapman and hall/CRC).

23) Lex Comber. : Modelling Space and Time with GAMS: spatially and temporally varying coefficient models. https://bookdown.org/lexcomber/AGILE2024/ (2024 年 8 月閲覧).

24) 秋元肇 : 大気反応化学, p167-168 (2014), (朝倉書店).

リスト 4 前方ステップワイズ法による交互作用項を導入した GAM の構築

```
stepwise_gam <- function(target = character(), ex = character(), int_lev = NULL, FUN = BIC, da
ta) {
  FUN <- substitute(FUN)
  message_gam <- function(model) {
    message("==================")
    message("formula: ", deparse(formula(model)))
    message(deparse(FUN), ": ", eval(bquote(round(.(FUN)(model), digits = 2))))
    message("R.sq: ", round(summary(model)$r.sq, digits = 4))
    message("Deviance explained: ", round(summary(model)$dev.expl * 100, digits = 2), "%")
    message("==================¥n")
  }
  create_formula <- function(target, ex) {
    parse(text = paste(target, "~", paste(paste0("s(", ex, ")"), collapse = " + ")))[[1]] |> a
s.formula()
  }

  ## calc normal gam
  ms <- lapply(ex, function(x) {
    mgcv::gam(create_formula(target, x), data = data)
  })
  num <- eval(bquote(which.min(Map(.(FUN), ms))))
  m <- ms[[num]]
  message_gam(m)
  for (i in setdiff(seq(1, length(ex), by = 1), num)) {
    m2 <- mgcv::gam(create_formula(target, ex[c(num, i)]),
                    data = data)
    message_gam(m2)
    if (eval(bquote(.(FUN)(m2))) <= eval(bquote(.(FUN)(m)))) {
      num <- c(num, i)
      m <- m2
    }
    rm(m2)
  }

  ## calc interactions
  if (!is.null(int_lev)) {
    pat <- combn(ex, int_lev)
    for (i in seq(1, ncol(pat), by = 1)) {
      formula <- parse(text = paste(paste(deparse(formula(m)), collapse = ""), paste0(" + s(",
 paste(pat[, i], collapse = ", "), ")")))[[1]] |> as.formula()
      m2 <- mgcv::gam(formula,
                      data = data)
      message_gam(m2)
      if (eval(bquote(.(FUN)(m2))) <= eval(bquote(.(FUN)(m)))) {
```

```
        num <- c(num, i)
        m <- m2
      }
      rm(m2)
      rm(formula)
    }
  }

  message("Completed.")
  message("==================¥n")
  return(m)
}

stepwise_gam(target = "Ox", ex = c("solar_radiation", "temperature", "wind_speed"), int_lev =
2, data = uto, FUN = BIC) |> summary()
==================
formula: Ox ~ s(temperature)
BIC: 1380.16
R.sq: 0.6928
Deviance explained: 70.44%
==================

==================
formula: Ox ~ s(temperature) + s(solar_radiation)
BIC: 1380.45
R.sq: 0.7103
Deviance explained: 72.56%
==================

==================
formula: Ox ~ s(temperature) + s(wind_speed)
BIC: 1372.93
R.sq: 0.7397
Deviance explained: 75.75%
==================

==================
formula: Ox ~ s(temperature) + s(wind_speed) + s(solar_radiation, temperature)
BIC: 1413.78
R.sq: 0.79
Deviance explained: 82.89%
==================

==================
formula: Ox ~ s(temperature) + s(wind_speed) + s(solar_radiation, wind_speed)
```

```
BIC: 1420.05
R.sq: 0.8161
Deviance explained: 85.82%
==================

==================
formula: Ox ~ s(temperature) + s(wind_speed) + s(temperature, wind_speed)
BIC: 1381.82
R.sq: 0.7576
Deviance explained: 78.17%
==================

Completed.
==================



Family: gaussian
Link function: identity

Formula:
Ox ~ s(temperature) + s(wind_speed)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.1617     0.9672    59.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                 edf Ref.df      F  p-value
s(temperature) 7.468  8.429 38.812  < 2e-16 ***
s(wind_speed)  3.870  4.797  5.983 8.29e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =   0.74   Deviance explained = 75.8%
GCV = 168.67  Scale est. = 156.21    n = 167
```