卒業論文概要

LightGBMによる気象衛星データを用いた可降水量推定に関する研究

1240102 寺井ひびき

高知工科大学 システム工学群 建築・都市デザイン専攻

Email: 240102i@ugs.kochi-tech.ac.jp

本研究では，気候変動の監視において大気中の水蒸気量を把握することが重要であることから，機械学習を用いた可降水量推定手法の検討を行った．より精度良く可降水量を推定するために，モデル構築に使用する機械学習手法や説明変数の検討，学習曲線を用いた過学習の検証などを行った．この結果，ランダムフォレストで可降水量推定をした時よりも，LightGBMで重要度の低い説明変数は使用せずにモデルを構築した時の方が，推定精度が向上した．また，月毎に可降水量推定モデルを構築する場合，月毎に説明変数を選定しモデルを構築するのではなく，1年間の訓練データから構築した可降水量推定モデルで重要度が高いとされた上位20個の説明変数を用いても問題ないことが示唆された．さらに，精度比較の結果，ある年の各月データから構築した月毎のモデルを他の年のデータに適用して可降水量を高精度で推定することは困難であることが分かった．

Key Words ： *温暖化，機械学習，過学習，気象衛星ひまわり*

1. はじめに

(1) 背景

水蒸気は赤外放射をよく吸収し，最大の温室効果を持つ気体であるとされている．人間活動が人為的に水蒸気量を変化させることはないため，直接的な地球温暖化の原因ではない．しかし，他の温室効果ガスによって引き起こされた温暖化を増幅させる作用があるため，水蒸気は温暖化に対して大きな影響を及ぼす因子であると考えられている[1)]．このことから，大気中に含まれる水蒸気量を把握することは，気候変動を監視するうえで重要である．可降水量とは，地上の単位面積に立てた鉛直気柱内に含まれる全水蒸気量のことであり，大気中の水蒸気量を示す指標である．先行研究では，様々な手法を用いた可降水量の推定が行われてきた．赤塚らはGMS-5用に開発された赤外スプリットウィンドウチャンネルを用いた可降水量推定アルコリズムをMTSATデータに適用し，MTSATデータ用の可降水量推定手法の開発を行った[1)]．その後，赤塚は気象衛星ひまわり8号の観測データを用い，ランダムフォレスト(これ以降，RFとする)の回帰分析を用いた推定手法の検討を行った[2)]．また，谷村は赤塚同様，RFの回帰分析による検討を行い，可降水量の推定モデルは各年，各月それぞれで構築される必要があるとしている[3)]．

(2) 目的

本研究は，気象衛星ひまわりの観測データ[4)]を用い，機械学習の回帰分析を用いて高精度に可降水量を推定する手法を検討することを目的とする．本研究の独自性として，使用する機械学習手法の変更と，モデルを構築する際に使用する説明変数の検討，構築した可降水量推定モデルの過学習に関する検証を行う．また，先行研究同様に，他の年への適用可能性の検討や可降水量推定における重要変数の検討も併せて行う．

2. 手法

(1) 使用データ

本研究では，目的変数として各ラジオゾンデ観測点で観測された2016年の可降水量，説明変数として各ラジオゾンデ観測点の緯度，標高，2016年のひまわり赤外バンドの輝度温度等(**表-1**)を用いた．

表-1 説明変数

① 緯度	Lat
② 標高	Ele
③ 衛星天頂角	VZA
④ 観測バンドの輝度温度	(例) B08, B09
⑤ 輝度温度差	(例) B09-B08, B10-B08
⑥ 輝度温度差の2乗	(例) (B09-B08)^2
⑦ 輝度温度の鉛直成分	(例) B08cosVZA
⑧ 輝度温度差の鉛直成分	(例) (B09-B08)cosVZA
⑨ ⑥の鉛直成分	(例) ((B09-B08)^2)cosVZA
⑩ 輝度温度の自然対数	(例) ln(B08)
⑪ 輝度温度差の自然対数	(例) ln(B09-B08)
⑫ ⑥の自然対数	(例) ln((B09-B08)^2)
⑬ ⑩の鉛直成分	(例) (ln(B08))cosVZA
⑭ ⑪の鉛直成分	(例) (ln(B09-B08))cosVZA
⑮ ⑫の鉛直成分	(例) (ln((B09-B08)^2))cosVZA

使用する説明変数は，GMS-5やMTSATによる赤外スプリットウィンドウチャンネルを用いた可降水量推定アルゴリズム[1),5)]を参考にした．また，モデルの構築において，2016年の1年間のデータを訓練データ8割，テストデータ2割に分割して用いた．さらに検証データとして，2017年の可降水量データを用いた．

(2) 可降水量推定手法

可降水量の推定は，LightGBM(Light Gradient Boosting Machine)と呼ばれる機械学習手法の回帰分析によって行った．LightGBMとは，ツリーベースの学習アルゴリズムを使用する勾配ブースティングフレームワークである[6)]．この機械学習手法を選定した理由は，**2章3節a項**，**3章1節**にて後述する．

(3) 研究の流れ

本研究は以下に示す**図-1**の流れで行った．

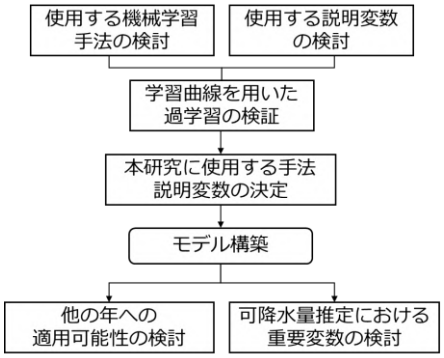

図-1 研究の流れ

a) 使用する機械学習手法の検討

本研究で使用する機械学習手法の検討は，決定係数とRMSE値を参考にして行った．RF，CatBoost，LightGBM，XGBoost，GBDTの5種類の機械学習手法を用いて，全観測点における2016年の1年間分の全データを結合したデータから可降水量推定モデル(これ以降，年間モデルと呼ぶ)を構築し，精度比較を行った．そして，精度が最も高い機械学習手法を本研究で用いることとした．このとき，より精度の良いモデルを構築するため，説明変数の数に関する検討も併せて行った．

b) 使用する説明変数の検討

先行研究[3)]では，全75個の説明変数を使用して可降水量推定モデルの構築が行われた．本研究では，モデル構築に使用する説明変数の数を変化させることで，可降水量推定の精度に違いが生じるのか検証を行った．この時，モデルに与える影響度はpermutation importanceを参考にした．これは，モデルが各特徴量にどの程度依存しているかを示すことのできる測定方法[7)]である．全説明変数を用いて年間モデルを構築したとき，影響度が大きかった説明変数から順に，10,20,～,60個とモデル構築に使用する説明変数の数を変化させ，精度比較を行った．また，同様に説明変数の数を変化させ，全観測点における2016年の月毎のデータを結合したデータから月毎に可降水量推定モデル(これ以降，各月モデルと呼ぶ)を構築し，精度比較を行った．

年間・各月モデルにおいて，それぞれ説明変数をいくつ使用する場合のモデル精度が最も良くなるのか検討し，この結果に基づいて**2章3節d項**以降のモデル構築で使用する説明変数の数を決定した．

c) 学習曲線を用いた過学習の検証

2016年の訓練データを用いて構築したモデルが過学習の状態になっていないか確認するため，学習曲線を描いて検証を行った．学習曲線とは，訓練データのサンプル数と予測性能の関係を示したグラフ[7)]のことである．過学習に関する検証を行うことで，モデルが学習時のデータに過度に適合し，未知のデータに対する予測精度が低下してしまうことを防ぐことができる．

d) モデル構築

2章3節a項から**2章3節c項**までを踏まえ，2016年の訓練データを用いて可降水量推定モデルの構築を行った．その後，2016年のテストデータと2017年の検証データを使用して，年間モデル・各月モデルそれぞれを用いた場合のRMSE・MAPEを計算・比較した．MAPEとは，予測値と観測値の差の絶対値を観測値で割ったものの平均であり，RMSE同様，モデル精度の評価をする指標として用いられる．

e) 他の年への適用可能性の検討

各月モデルにおける2016年，2017年のRMSE・MAPEの値を比較することによって，構築した可降水量推定モデルが他の年にも適用できるか検討した．

f) 可降水量推定における重要変数の検討

先行研究より，月毎にそれぞれモデルを構築する必要があると指摘されている[3)]ことから，各月モデルの構築において，年間を通じて同じ説明変数を使用するよりも，月毎に説明変数を選定しモデルを構築する方が，可降水量の推定精度が良くなると考えた．そこで，全説明変数を用いて各月モデルを構築し，特徴量重要度に基づいて，月毎に説明変数の選定を行った．このようにして選定した説明変数から各月モデルを構築し，**2章3節d項**で構築した各月モデルと推定精度を比較した．これにより，可降水量推定における重要変数の検討を行った．このとき，説明変数の数は**2章3節b項**の結果をもとに決定した．

3. 結果及び考察

(1) 使用する手法・説明変数の決定

検証の結果，LightGBMにより30個の説明変数を用いて年間推定モデルを構築し可降水量を推定した時，最も良い精度で推定することができた．この結果から，本研究では機械学習手法としてLightGBMを用い

ることとした．次に学習曲線を描き，推定モデルが過学習の状態になっていないことを確認しながら，LightGBMのパラメータを調整し，年間モデル及び各月モデルを構築した．この時，使用する目的変数の数を変化させ，年間モデル及び各月モデルの決定係数とRMSEの値を計算した(**表-2**)．また，モデル構築時に必要なパラメータ調整に関しては，決定木の1本の木の分岐の数を12，木の数を150としてモデル構築を行った．

年間モデルにおいて最も良い精度が得られたのは，パラメータ調整前と同様，説明変数を30個使用してモデルを構築したときで，RMSEの値は6.50㎜となった．また，各月モデルでは20個の説明変数を用いたとき，RMSE値が3.68㎜から6.80㎜となり，30個の説明変数を使用する場合よりも高い精度が得られた．

表-2 年間モデル構築に使用する説明変数の数の違いによる決定係数・RMSE値の比較

説明変数(個)	決定係数	RMSE(mm)
10	0.88	6.73
20	0.88	6.61
30	0.88	6.50
40	0.87	6.93
50	0.88	6.90
60	0.87	6.92

この結果より，本研究では年間モデル構築に使用する説明変数の数を30個，各月モデル構築に使用する説明変数の数を20個とした．このとき使用した説明変数を以下に示す(**表-3**)．年間モデルでは**表-3**に示した全ての説明変数を用い，各月モデルでは**表-3**の左2列に示した説明変数を用いてモデル構築を行った．また，**図-2**，**図-3**のように，年間・各月モデルにおける学習曲線を描いた．

表-3 モデル構築に使用した説明変数
(上からモデルに与える影響が大きい順)

Lat	((B13-B15)^2)cosVZA	((B13-B14)^2)cosVZA
(B13-B15)cosVZA	B15	ln(B13)
Ele	ln(B15)	((B09-B08)^2)cosVZA
(B13-B14)cosVZA	B10	B13cosVZA
(B14-B15)cosVZA	((B14-B15)^2)cosVZA	(ln(B14-B15))*cosVZA
B13	B08	(ln(B15))*cosVZA
ln(B10)	B10-B09	(ln(B09-B08))*cosVZA
(ln(B13-B15))*cosVZA	VZA	(ln(B09))*cosVZA)
ln(B10-B09)	B13-B14	B14
B09-B08	ln(B08)	B13-B15

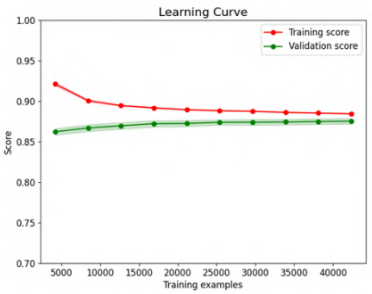

図-2 年間モデルの学習曲線

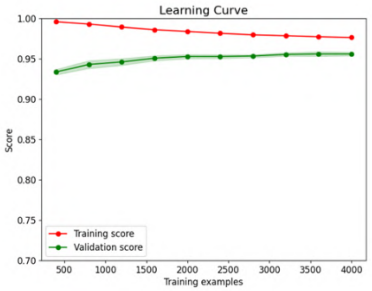

図-3 各月モデルの学習曲線(1月)

訓練データに対する予測精度と，テストデータに対する予測精度が両方高い精度で収束していることから，構築したモデルが過学習の状態ではないことを確認した．

(2) 可降水量推定結果

谷村の先行研究[3]において，2016年のRFによる各月モデルを用いたRMSEの値は4.02㎜から7.26㎜であったのに対し，本研究では3.68㎜から6.80㎜となり，全ての月において精度の向上が見られた(**表-4**)．また，2017年のデータと各月モデルを用いた検証でも，先行研究ではRMSEの値が5.26㎜から8.23㎜であったのに対し，本研究では4.71㎜から7.43㎜となり，精度の向上が見られた．また，年間モデルから得られた2016年のRMSEの値は6.50㎜となった．次に，年間・各月モデルにおけるMAPEを求めた(**表-5**)．

表-4 LightGBM 年間・各月モデルにおけるRMSE

	2016 RMSE(mm)	2017 RMSE(mm)	RMSEの差(mm)
1	3.68	4.71	1.03
2	4.60	5.02	0.42
3	5.13	4.72	-0.41
4	4.51	6.26	1.75
5	5.85	5.89	0.04
6	5.34	6.27	0.93
7	5.81	6.08	0.27
8	6.80	6.70	-0.10
9	5.57	6.44	0.87
10	5.36	7.43	2.07
11	4.64	5.55	0.91
12	4.26	4.96	0.70
年間	6.50	6.74	0.24

表-5 LightGBM 年間・各月モデルにおけるMAPE

	2016 MAPE(%)	2017 MAPE(%)	MAPEの差(%)
1	31.58	33.53	1.95
2	36.03	33.81	-2.22
3	39.59	32.04	-7.55
4	26.37	36.02	9.65
5	25.68	26.58	0.90
6	17.18	26.41	9.23
7	15.28	17.80	2.52
8	17.87	19.55	1.68
9	17.60	23.69	6.09
10	25.67	34.61	8.94
11	28.12	36.98	8.86
12	33.34	41.86	8.52
年間	32.60	34.98	2.38

2016年の各月のMAPEの値は15.28から39.59%となり，3月に最も大きい値となった．2017年の各月のMAPEの値は17.80から41.86%となり，12月に最も大きい値となった．また，年間モデルから得られた2016年のMAPEの値は32.60%となった．これらの結果より，各月モデルにおいて，RFで全説明変数を用いてモデルを構築する場合よりも，LightGBMで説明変数20個を用いて構築したモデルの方が，可降水量推定精度が向上することが分かった．

また，RMSEの値は冬季よりも夏季に大きくなる傾向にあるが，MAPEの値は夏季よりも冬季に大きくなる傾向にあることが分かった．これは，元々の可降水量の観測量が，冬季よりも夏季に多いためであると考えられる．RMSEの結果のみで考えると夏季の方が可降水量の推定精度が悪いといえるが，各月の可降水量の観測量を考慮すると冬季の方が推定モデルの精度が悪くなるといえる．

a） 他の年への適用可能性の検討

また，**表-4**，**表-5**より，2016年と2017年のRMSE・MAPEの差をそれぞれ求めた．RMSEの差の大きさは，0.04から2.07mmとなり，10月のモデルで最も差が大きくなった．MAPEの差の大きさは，0.90から9.65%となり，4月のモデルで最も差が大きくなった．また，RMSE・MAPEは月毎に異なる値を取ることから，年間モデルではなく，各月モデルを構築することが可降水量推定精度の向上において重要である．

5月の推定モデルでは2016年と2017年のRMSE・MAPEの差はほとんど見られなかったが，他の月では差が見られたため，ある年の各月データから構築した各月モデルを他の年のデータに適用して可降水量を高精度で推定することは困難であると考えられる．

b） 可降水量推定における重要変数の検討

表-6は，各月モデルにおいてモデルに与える影響度が大きいとされた上位5個の説明変数である．各月共通してモデルに及ぼす影響が大きいとされたのは，観測点の緯度(Lat)，標高(Ele)のデータであった．その他の重要度が高い説明変数については，月毎に異なる結果となった．

次に，年間モデルにおいて重要度が高いとされた上位20個の説明変数を用いて各月モデルを構築した場合と，月ごとの特徴量重要度に基づいて20個の説明変数を決定し各月モデルを構築した場合のRMSE値の比較を行った．その結果，ほとんどの月では年間モデルの特徴量重要度に基づいた20個の説明変数を用いて各月モデルを構築した場合の方が，推定精度が良くなることが分かった．このとき，2,6,8,11月は，月毎に説明変数を選定したときよりも推定精度が低下したが，RMSEの値で表すと0.05mmから0.56mm程度の低下であった．以上から，年間を通じて同じ説明変数を選定し，各月モデルを構築しても精度低下は小さいため，年間モデルにおいて重要度が高いとされた上位20個の説明変数を用いて各月モデルを構築しても問題ないことが示唆された．

表-6 各月モデルにおける重要変数

1月	2月	3月
Lat	Lat	Lat
Ele	Ele	Ele
(B13-B14)cosVZA	(B13-B14)cosVZA	(B13-B14)cosVZA
B15	B10	((B14-B15)^2)cosVZA
ln(B15)	B15	((B13-B14)^2)cosVZA
4月	**5月**	**6月**
Lat	Lat	Lat
Ele	Ele	Ele
(B14-B15)cosVZA	((B13-B15)^2)cosVZA	(B13-B15)cosVZA
(B13-B14)cosVZA	(B13-B15)cosVZA	(B13-B14)cosVZA
B15	B15	VZA
7月	**8月**	**9月**
Lat	Lat	Lat
Ele	(B13-B14)cosVZA	Ele
(B13-B14)cosVZA	Ele	(B13-B14)cosVZA
VZA	(B14-B15)cosVZA	VZA
(ln(B13-B15))*cosVZA	VZA	(ln(B13-B15))*cosVZA
10月	**11月**	**12月**
Lat	Lat	Lat
Ele	Ele	(ln(B13-B15))*cosVZA
(B13-B14)cosVZA	((B13-B14)^2)cosVZA	Ele
(B13-B15)cosVZA	B10	(B13-B14)cosVZA
ln(B10)	(B14-B15)cosVZA	B10

4. おわりに

本研究では，気象衛星ひまわりの観測データを用い，機械学習の回帰分析によってより精度良く可降水量を推定するための検討を行った．使用する機械学習手法と説明変数の数を検討し，モデルが過学習の状態でないことを確認した．その結果，以下のことが分かった．機械学習手法にはLightGBMを使用し，モデルの構築に用いる説明変数を20個に変更することで，可降水量の推定精度が向上した．このとき，全観測点の1年間の訓練データから構築した可降水量推定モデルにおいて重要度が高いとされた上位20個の説明変数を用いて，月毎のモデルを構築しても問題ないことが示唆された．また，ある年の各月データから構築した月毎のモデルを他の年のデータに適用して可降水量を高精度で推定することは困難であることが分かった．

参考文献

1) 赤塚慎，大吉慶，竹内渉：運輸多目的衛星 MTSAT 観測による可降水量推定手法の開発，日本リモートセンシング学会誌，Vol.31，No.5，pp. 481-489，2011
2) 赤塚慎：気象衛星ひまわりによる可降水量推定手法の検討，日本写真測量学会平成 31 年度秋季学術講演会発表 論文集，2019
3) 谷村朋哉：気象衛星データを用いた機械学習による可降水量推定手法の検討，2022年度高知工科大学システム工学群卒業研究概要書
4) ひまわり 8/9 号フルディスク(FD) gridded data 公開について：http://www.cr.chiba-u.jp/databases/GEO/H8_9/
5) 操野年之：可降水量，気象衛星センター技術報告特別号，pp.89・94，1996
6) LightGBM 公式ドキュメント：https://lightgbm.readthedocs.io/en/stable/
7) scikit-learn Machine Learning in Python：https://scikit-learn.org/stable/index.html