卒業論文概要

相対湿度分布の再現手法の検討

1240075　下村梨香子

高知工科大学　システム工学群　建築・都市デザイン専攻

E-Mail：240075c@ugs.kochi-tech.ac.jp

近年，熱中症の危険性がさらに高まっている．そこで環境省は熱中症の危険性を判断する数値として暑さ指数を公開している．相対湿度の観測が行われていない地点では相対湿度の予測値を用いて暑さ指数を計算している．本研究では機械学習手法を用いて面的に相対湿度を高精度に推定する手法を検討した．その結果, RandomForest は過学習の傾向があるため，本研究では CatBoost を用いることで精度が向上することが分かった．また土地利用形態の割合を説明変数として追加すると精度が向上した．しかし，気象観測データを入力データとしても，相対湿度推定モデル構築に利用した観測点以外の場所では精度よく推定することは困難であり，さらに, MSM データを用いて面的に相対湿度を高精度に推定することも困難であることが分かった．

Key word：機械学習、学習曲線、土地利用形態

1．序論

近年，地球温暖化に伴い気温が上昇しており，熱中症の危険性がさらに高まっている．その対策として環境省は熱中症の危険度を判断する数値として各気象観測点において暑さ指数を計算，公開している．暑さ指数とは，人間の熱バランスに影響の大きい気温，湿度，輻射熱の 3 つの指標を取り入れた温度の指標であり，影響の割合として湿度が高くなっている [1]．しかし相対湿度を観測している地点は限られている．よって面的に相対湿度を推定することで相対湿度が観測されている地点以外の場所でも暑さ指数の計算が可能となる．先行研究である川上の研究では，機械学習 RandomForest の回帰により推定モデルを構築し，数値予報データを用いて絶対湿度と相対湿度の分布を推定する手法を検討した [2]．結果として数値予報データを用いた場合，良い精度が得られないという結果となった．そこで本研究では相対湿度の推定精度向上のために機械学習手法の再検討，過学習の考慮，説明変数の再検討を行うことを目的とする．

2．手法

(1) 対象地域及び使用データ、期間

図-1 対象地点 [2]

本研究の対象地域は川上の研究の対象地域と同様に図-1 とした．

データは GPS 可降水量データ [3], 2021 年の気象官署の気象観測データ [4], 2021 年の 12 月から相対湿度の観測が行われている AMeDAS 観測点の気象観測データ [4]を用いる．また，数値予報データとして 2021 年の MSM データ [5]，さらに，国土数値情報の土地利用細分メッシュデータ [6]を用いた．

(2) 研究の手法

(a) 本研究の流れ

先行研究 [2]では四国の全気象官署における気象観測データ（気温，気圧，降水量，風速）と GPS 可降水量データを用いて機械学習の回帰により，相対湿度推定モデルの構築，精度検証を行った．次に推定モデルを用いて，各気象官署にて精度検証を行った，次に面的に推定を行うため，MSM データを入力データとして相対湿度を推定，マップ化及び各気象官署において RMSE 値を計算し精度検証を行った．次に 2021 年 12 月から相対湿度の観測が行われており，モデルの構築に用いていない AMeDAS 観測点を対象に気象観測データを用いて相対湿度を推定し，RMSE 値を計算し精度検証を行った．

本研究では機械学習手法の再検討，過学習を考慮し，土地利用形態割合を新たな説明変数として追加を行い，新しく相対湿度推定モデルの構築を行う．その後，先行研究と同様にして精度検証を行う．精度検証では基準値を推定精度の目標とする．基準値とは，気象観測データを真値としたときの，ＭＳＭデータの相対湿度のＲＭＳＥ値である(表-1)．

表-1 相対湿度推定モデルの評価に用いる RMSE 基準値

月	RMSE基準値(%)								
	高知	清水	宿毛	室戸岬	徳島	高松	多度津	松山	宇和島
1月	10.8	11.6	11.4	11.5	8.08	7.33	8.48	8.29	10.4
2月	10.8	11.4	12.4	9.59	9.04	10.6	9.16	9.33	10.6
3月	10.3	10.5	9.35	13.1	9.34	8.75	20.1	8.47	9.43
4月	12.3	11.4	10.7	11.8	12.2	16.6	19.5	11.6	12.3
5月	10.4	9.14	9.92	12.1	11.7	12.9	11.9	8.63	10.5
6月	8.92	8.92	8.37	13.00	8.92	12.0	15.6	8.28	8.92
7月	9.04	7.06	6.99	11.2	9.46	12.3	13.1	14.9	8.7
8月	8.73	8.22	7.57	11.1	8.73	11.3	16.6	9.20	7.83
9月	9.97	9.19	8.31	11.4	7.11	9.53	13.7	7.65	7.61
10月	10.8	8.80	9.92	9.64	8.88	12.1	16.8	7.67	9.39
11月	13.2	9.20	11.7	8.83	7.31	8.01	14.8	9.63	10.8
12月	11.5	9.39	9.96	9.57	7.38	7.42	7.73	7.86	9.44

(b) 機械学習の再検討

先行研究[2]では機械学習 RandomForest の回帰により推定モデルを構築した．本研究ではまず GBDT, RandomForest, CatBoost, XGBoost, LightGBM の 5 つの機械学習を用いて推定モデルを構築し，四国全体の月毎の相対湿度を推定，精度検証を実施する．

この時，説明変数に気温，気圧，降水量，風速（X, Y 方向），可降水量，目的変数に相対湿度を用いる．

(c) 過学習の考慮

先行研究 [2]では過学習の考慮がなされていなかった．本研究では過学習を考慮するために学習曲線を作成して過学習の発生状況を確認する．学習曲線とはモデルの過学習を判断することができるグラフである．過学習とは未知のデータに対してうまく汎化できない状態を指す [7]．2(2)(b)の結果から精度が良いと判断した機械学習の推定モデルの学習曲線を作成し，それぞれに過学習が発生していないかを確認していく．

(d) 土地利用形態の考慮

本研究では相対湿度推定モデル構築の際の説明変数として気象観測点周辺の土地利用の割合を追加する．まず，四国地方の土地利用細分メッシュデータから QGIS を用いて各観測地点から半径 5 ㎞の土地利用割合を集計する．集計された土地利用形態を植生，建築物，水域の 3 種類に統合し，それぞれ割合を計算する．それらの結果を説明変数として追加し，推定モデルを新たに構築する．同様に MSM データを用いた推定の際の入力データとして，1km メッシュ毎に植生，建築物，水域の割合を集計したデータをそれぞれ作成する．

３．結果及び考察

(1) 機械学習手法の再検討

四国地方全体で 5 つの機械学習を用いて相対湿度推定モデルを構築し，精度検証を行った．月毎に最も精度の高い機械学習を調べた結果，先行研究で用いられた RandomForest と新たに CatBoost の結果が良好であった．

(2) 過学習の考慮

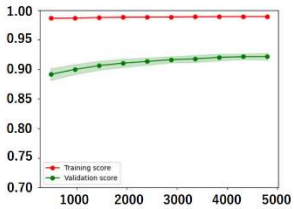

図-2 RandomForest の学習曲線(1 月)

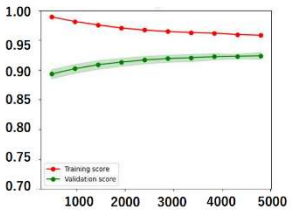

図-3 CatBoost の学習曲線(１月)

(1)の結果から RamdomForest と CatBoost の学習曲線を作成した結果が次の図-2，図-3 のようになった．学習曲線の状態が 2 月から 12 月まで同様の結果となったことから，RandomForest は過学習の傾向にあり，一方 CatBoost は比較的良好な学習状態であると判断した．よって本研究では，機械学習手法として CatBoost を用いることとした．

(3) 土地利用形態の考慮

各観測地点から半径 5 kmの土地利用割合を集計、植生，建築物，水域に統合した結果，それぞれの割合は表-2 のようになった．これらの数値を新たな説明変数として追加し，推定モデルの構築を行った．

表-2 観測地点から半径 5 km以内の土地利用割合

	植生	建築物	水域
高知	0.48	0.43	0.087
清水	0.23	0.0092	0.76
宿毛	0.57	0.069	0.36
室戸岬	0.15	0.035	0.82
徳島	0.20	0.43	0.37
高松	0.20	0.66	0.14
多度津	0.27	0.32	0.41
松山	0.37	0.59	0.038
宇和島	0.68	0.14	0.18

(4) 推定モデルの構築

全気象官署における気象観測データと可降水量データを用いて CatBoost 回帰により，相対湿度推定モデルの構築，精度検証を行った(表-3)．

表-3 RandomForest と CatBoost の相対湿度推定精度(RMSE)の比較

月	RandomForest(%)	CatBoost(%)
1月	7.96	8.03
2月	7.69	7.74
3月	8.85	8.95
4月	8.25	8.24
5月	8.61	8.70
6月	5.29	5.29
7月	4.25	4.07
8月	4.32	4.20
9月	4.97	4.90
10月	7.41	7.41
11月	6.00	5.98
12月	7.10	7.08

先行研究 [2]の RandomForest 回帰による推定モデル構築の結果との比較から，RandomForest と CatBoost の精度に大きな差は無いが，6~12 月は CatBoost の精度の方が同程度，またはわずかに良好であることが分かった．

さらに，土地利用割合を説明変数として追加する前の精度と追加後の精度を比較した結果を表-4 に示す．

表-4 土地利用割合を追加した前後の相対湿度推定精度(RMSE)

月	基準(%)	追加前(%)	追加後(%)
1月	9.84	8.03	7.35
2月	10.4	7.74	7.00
3月	10.9	8.95	7.53
4月	13.2	8.24	7.21
5月	10.8	8.70	7.94
6月	10.3	5.29	4.34
7月	9.73	4.07	3.39
8月	10.0	4.20	3.52
9月	9.38	4.90	4.33
10月	10.1	7.41	6.35
11月	10.4	5.98	5.66
12月	8.93	7.08	6.46

この結果から，四国全体では土地利用形態割合の追加前より追加後の方の精度が良好となり，基準値も満たしていることが分かる．よって，推定モデル構築において土地利用形態を説明変数として追加することで精度は向上すると考えられる．

この結果と推定モデルの構築に伴ってその説明変数を予測に用いた場合と用いなかった場合でどの程度モデルの予測値が変わるのかを指標化 [3]した特徴量重要度のグラフ（図-4，図-5）から特徴量重要度と相対湿度推定精度との関連性を調べた．

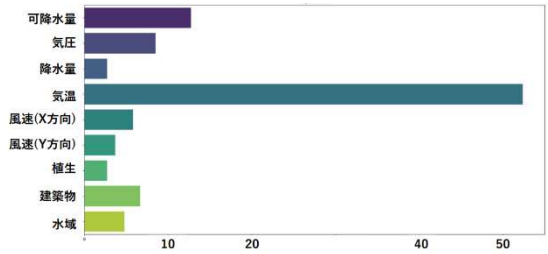

図-4 特徴量重要度グラフ（1 月）

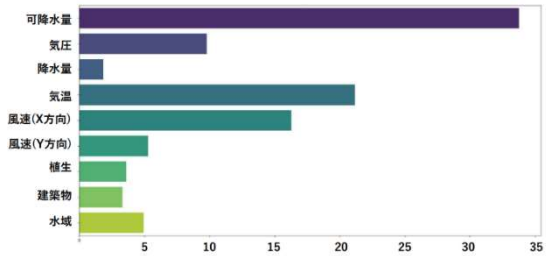

図-5 特徴量重要度グラフ（8 月）

各観測地点の精度が悪い月と良い月を調べた結果，精度の悪い月は 1 月，2 月，5 月，10 月，精度の良い月は 7 月，8 月，9 月が主に挙げられた．特徴量重要度のグラフでは 1~5 月，10~12 月は気温の特徴量重要度が低下しており，7 月，8 月は気温の特徴量重要度が高い数値を示していた．これらの結果から，夏季は，ほぼ気温のみで相対湿度が決まることが示唆された．

(5) AMeDAS 観測点での推定

表-5 土地利用割合を追加した前後の AMeDAS の気象観側点での相対湿度推定値の RMSE(2021 年 12 月)

	安芸	窪川	穴吹	引田	内海
基準(%)	10.7	16.4	8.82	8.89	11.2
追加前(%)	13.5	12.9	10.5	9.58	10.6
追加後(%)	13.8	12.9	10.5	9.38	10.2

3(4)にてモデルの構築に用いていない AMeDAS 観測点を対象に，気象観測データをモデルに入力して相対湿度を推定し，RMSE 値を計算し精度検証を行

った(表-5).

説明変数を追加する前の結果と先行研究[2)]のRandomForest を用いた場合の結果を比較した結果，窪川，穴吹では精度の向上は見られたが，その他の地点では精度は低下した．また 3(4)の説明変数追加前の結果(表-5)と比較すると，精度が悪いことが分かる．説明変数を追加した後の結果も精度向上は見られなかった．また，土地利用割合の追加前後で推定精度に大きな違いは見られなかった．よって構築した推定モデルを用いて，モデルの構築に用いていない場所における相対湿度を精度よく推定することは困難であると言える．

(6) MSM データを用いた推定

3(4)で構築した推定モデルに MSM データの値，1km メッシュ毎で集計した土地利用割合を入力し，相対湿度分布のマップ化，観測地点ごとに基準値との精度検証を行った結果が表-6，表-7 である．

表-6 土地利用割合を追加する前の MSM データを用いた相対湿度推定値の RMSE

月	説明変数追加前(%)								
	高知	清水	宿毛	室戸岬	徳島	高松	多度津	松山	宇和島
1月	19.6	15.7	18.0	15.4	18.2	19.7	17.7	17.9	18.7
2月	20.2	14.2	16.3	12.6	17.6	17.3	14.0	15.9	18.1
3月	18.9	14.9	15.3	14.8	17.4	17.5	19.5	17.0	14.6
4月	17.2	14.7	15.8	14.7	16.5	19.4	21.2	19.0	17.5
5月	15.5	15.7	13.5	16.7	18.8	19.1	18.7	16.9	15.1
6月	14.9	13.8	13.3	14.6	19.3	21.2	16.0	18.4	14.5
7月	10.2	10.1	8.17	15.4	11.7	14.8	13.9	12.0	10.4
8月	17.0	11.5	13.5	14.7	18.3	18.6	16.3	17.1	13.8
9月	12.1	8.53	10.0	10.1	12.4	12.0	15.7	10.9	10.7
10月	17.8	13.6	17.4	11.5	14.3	16.3	19.1	16.4	18.0
11月	18.6	18.0	17.2	15.7	16.7	19.4	17.3	18.4	17.9
12月	19.4	11.9	17.4	15.3	15.5	16.5	15.3	14.9	17.5

表-7 土地利用割合を追加した後の MSM データを用いた相対湿度推定値の RMSE

月	説明変数追加後(%)								
	高知	清水	宿毛	室戸岬	徳島	高松	多度津	松山	宇和島
1月	18.9	16.1	17.5	15.7	17.7	18.6	17.5	17.2	18.3
2月	20.1	14.4	16.4	12.8	17.1	16.7	14.0	16.1	18.9
3月	20.5	16.4	16.8	16.5	19.1	20.5	19.6	17.4	15.7
4月	16.7	16.7	15.3	14.1	15.6	18.7	21.1	18.7	16.3
5月	15.1	14.1	14.1	17.1	18.4	17.1	18.8	16.2	14.7
6月	15.0	13.1	12.7	14.1	18.2	20.4	14.3	17.6	14.4
7月	11.1	12.2	10.1	11.1	12.0	13.7	12.1	11.3	10.5
8月	16.8	12.4	13.6	15.5	17.4	17.6	16.2	16.3	13.7
9月	12.7	8.37	10.3	9.65	12.1	11.9	15.4	10.9	10.9
10月	16.4	13.7	16.8	11.7	13.6	15.2	18.4	15.5	17.2
11月	17.6	17.0	16.7	14.8	15.8	18.0	16.8	17.3	16.9
12月	20.1	12.6	17.6	14.3	16.6	17.2	16.5	16.1	17.6

説明変数を追加する前の結果(表-6)と先行研究[2)]の結果を比較すると精度の低下が見られたが，先行研究では過学習が起きていた可能性がある．

また 3(4)の結果と比較すると，土地利用割合を説明変数に追加してもしなくても，MSM データを入力データとすることで精度の低下が見られた．

従って，機械学習手法を CatBoost に変更しても，MSM データを用いて面的に相対湿度を高精度に推定することは困難であることが分かった．

4．まとめ

先行研究である機械学習の回帰により数値予報データを用いて湿度を推定する手法の再検討として，機械学習手法の再検討，過学習の考慮を行った結果，RandomForest は過学習の傾向があり，CatBoost を用いることで精度が向上することが分かった．また土地利用形態の割合を説明変数として追加すると精度が向上した．しかし，気象観測データを入力データとしても，相対湿度推定モデル構築に利用した観測点以外の場所では，精度よく推定することは困難であった．さらに，MSM データを用いて面的に相対湿度を高精度に推定することは困難であることが分かった．

参考文献

1)環境省：環境省熱中症予防情報サイト　暑さ指数（WBGT）について学ぼう
https://www.wbgt.env.go.jp/wbgt_lp.php

2)川上育海：数値予報データを用いた絶対湿度分布の再現 2022 年高知工科大学システム工学群卒業研究概要書

3)国土地理院：電子基準点データ提供サービス
https://terras.gsi.go.jp/

4)気象庁：気象庁｜過去の気象データ検索
https://www.data.jma.go.jp/obd/stats/etrn/index.php

5)京都大学：グローバル大気観測データ
http://database.rish.kyoto-u.ac.jp/arch/glob-atmos/

6)国土数値情報ダウンロードサイト
https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-L03-b.html

7)Sebastian Raschka, Vahid Mirjalili（株式会社クイープ訳，福島真太郎監訳）：Python 機械学習プログラミング 達人データサイエンスによる理論と実践 pp191, pp373