기상청

결측값 대체를 중심으로 한 효율적 지면/지상 온도 예측 모델 개발 (RMSE = 1.298032)

참 가 번 호	220242	팀 명	기대값

1차 대회- 2과제: 기상위성 자료를 활용한 지면/지상 온도 산출 기술 개발·개선

1. 분석 배경 및 목표

지면 온도는 기후, 수문 모형 및 농업의 기본적인 자료이며 기후변화, 물순환, 황사 탐지, 도심 열섬현상 분석, 화산활동 모니터링, 가뭄 감시 등의 분야에서 다양하게 활용될 수 있다. 또한 지상 온도는 매일 기상청에서 발표되는 일기예보에 반영되는 기온이며 이는 생활에 밀접하게 연관되어 있다. 이러한 지면/지상 온도는 보통 관측 지점에서 측정하나, 직접적인 측정으로 자료를 얻기 위해서는 인력과 경제적 자원의 소모와 시·공간적 제한 등의 문제점이 있다. 따라서 기상 빅데이터를 활용해 실생활과 현업에 널리 사용될 수 있는 지면·지상 온도 산출 기술의 개발이 필요하다.

2. 데이터 정의

연도	평균(지면 온도)	표준편차(지면 온도)	평균(지상 온도)	표준편차(지상 온도)
2020년	26.2519	6.0659	24.5201	3.5127
2021년	30.5335	8.0691	26.3872	3.5391

<표 1> 20/21년 통계량 차이

<표 1>은 2020년 7, 8월과 2021년 동기간의 지면/지상 온도 통계량을 비교한 것으로 2020년의 기상 자료는 2021년의 기상 상황을 잘 대변하지 못한다고 분석되었다. 따라서 모델의 검증 기간이 2021년 7, 8월인 점을 고려해 2020년과 2021년의 기상 데이터 중 2021년 데이터만을 사용하는 것이 합리적이라 판단하였다. 또한 검증 기간인 여름철과 기상 분포에 차이가 있는 봄, 초여름, 늦가을, 겨울 데이터 또한 학습에서 제외했다. 즉, 2021년의 여름 절기 중 망종(6월 5일)부터 2021년의 가을 절기 중 추분의 전날(9월 22일)까지의 데이터를 모델 학습과 테스트에 이용하였다. 추가로 예측 모델의 검증 기간은 2021년 7월 1~11일, 8월 21~31일이며 테스트 기간은 검증 기간에서 ±5일로 설정하였다. 정리하면 학습/테스트/검증 데이터의 구성은 <표 2>와 같으며 학습 데이터와 테스트 데이터의 비율은 약 77 : 23이다.

06.05 ~ 06.25	06.26 ~ 06.30	07.01 ~ 07.11	07.12 ~ 07.16	07.17 ~ 08.15	08.16 ~ 08.20	08.21 ~ 08.31	09.01 ~ 09.05	09.06 ~ 09.22
학습	테스트	검증	테스트	학습	테스트	검증	테스트	학습

<표 2> 데이터셋 구성

이때 주어진 데이터에서 종속변수인 지면 온도와 지상 온도에 –999와 같이 일반적인 범주에서 벗어나는 값들이 상당수 확인되었는데 평균, 분위 수 등의 통계량을 확인하였을 때 –999는 실제 값이 아닌 결측값이라고 판단하였다. 따라서 종속변수가 -999만으로 이루어져 있는 Station(STN) 들은 학습에서 제외했고 그 결과 총 706개의 STN 중 104개만이 남게 되었다. 즉, 104개의 STN에서 약 3개월가량 수집한 시공간 데이터를 분석하였다.

또한 온도 산출 모델 개발에 있어 모델의 성능을 높이기 위해 기상청 기상 자료 개발 포털에서 일조량 데이터를 받아 이를 새로운 독립변수로 추가하였다. 일조량은 '태양 광선이 구름이나 안개에 가려지지 않고 실제로 지면에 도달하는 양'으로 지면/지상 온도 변화에 영향을 미칠 것으로 판단하여 추가하였으며 변수 이름은 'Solar_amount'로 정의했다. 최종 데이터의 변

수는 총 45개로 <표 3>과 같다.

변수명	정의	변수명	정의	변수명	정의
insitu-TA	지상 온도	Band10	하층 수증기 밴드	insitu-HM	상대습도
isitu-LST	지면 온도	Band11	구름상 밴드	insitu-TD	이슬점온도
Year~Minute	년월일시분	Band12	오존 밴드	insitu-TG	초상 온도
STN	지점	Band13	대기창 밴드	insitu-TED0.05	5cm 지중온도
Lon	경도	Band14	깨끗한 대기창 밴드	insitu-TED0.1	10cm 지중온도
Lat	위도	Band15	오염된 대기창 밴드	insitu-TED0.2	20cm 지중온도
Band1	파랑 가시밴드	Band16	이산화탄소 밴드	insitu-TED0.3	30cm 지중온도
Band2	초록 가시밴드	30daysBand3	30일 가시 밴드	insitu-TED0.5	50cm 지중온도
Band3	빨강 가시밴드	30daysBand13	30일 대기창 밴드	insitu-TED1.0	1.0m 지중온도
Band4	식생 가시밴드	GK2A-LST	위성 관측 지표 온도	insitu-TED1.5	1.5m 지중온도
Band5	권운 밴드	SolarZA	태양 천정각	insitu-TED3.0	3.0m 지중온도
Band6	눈/얼음 채널	SateZA	위성 천정각	insitu-TED5.0	5.0m 지중온도
Band7	야간안개/하층운 밴드	ESR	대기 외 일사량	insitu-PA	현지기압
Band8	상층 수증기 밴드	Height	관측고도	insitu-PS	해면기압
Band9	중층 수증기 밴드	LandType	지면 타입	Solar_amount	일조량

<표 3> 변수 개요

3. 탐색적 자료 분석(EDA)을 통한 전처리 (Preprocessing)

산점도, 히트맵 등의 시각화를 통해 EDA를 구축하였으며 이를 바탕으로 독립변수와 종속변수의 전처리를 수행하였다.

3.1. 지중온도 변수 축소

지중온도(insitu-TED)는 지표면 밑 토양의 온도를 측정한 것으로 지하 5cm, 10cm, 20cm, 30cm, 50cm, 1.0m, 1.5m, 3.0m, 5.0m까지 측정되어 있다. TED 변수끼리는 모두 의존적이지만 종속변수와의 상관관계를 비교했을 때, 지표면에서 멀어질수록 상관관계가 지속해서 약해지는 것을 확인하였다. 따라서 지표면에서 가장 가까운 TED0.05 변수만을 분석에 이용하였다.

3.2. day, hour 변수 추가

연월일시분(YearMonthDayHourMinute) 변수로부터 날짜(day)와 시간(hour) 변수를 추출하고 원 변수는 삭제했다. 2021년 데이터만을 사용했기에 연도는 추출하지 않았으며 월이 바뀔 때 일에 30 또는 31을 더해주어 월의 변화를 반영하였기에 월도 추출하지 않았다. 마지막으로 1시간 간격 안에서는 온도 변화가 미세할 것이라는 가정하에 분 또한 추출하지 않았다.

3.3. 이상값(Outlier) 처리

이상값 처리를 위해 날짜별로 모든 독립변수의 산점도를 확인하였다. 그 결과 10개의 독립변수에서 데이터 분포의 양상과 명백히 동떨어진 이상값으로 추정되는 데이터가 존재하여 이러한 데이터는 값을 삭제하고 결측값으로 처리하였다.

3.4. 결측값(Missing Value) 처리

종속변수뿐 아니라 독립변수 또한 결측값이 다량 존재했다. 이러한 결측값은 예측 모델 생성에 악영향을 끼치므로 시간/공간적으로 결측값 주변에 위치한 데이터를 이용해 값을 보간 및 추정하였다. 먼저 시간 정보를 이용해 보간하고 이후 공간 정보를 이용해 결측값을 추정했다.

3.4.1. 시간 선형 보간(Time Linear Interpolation)

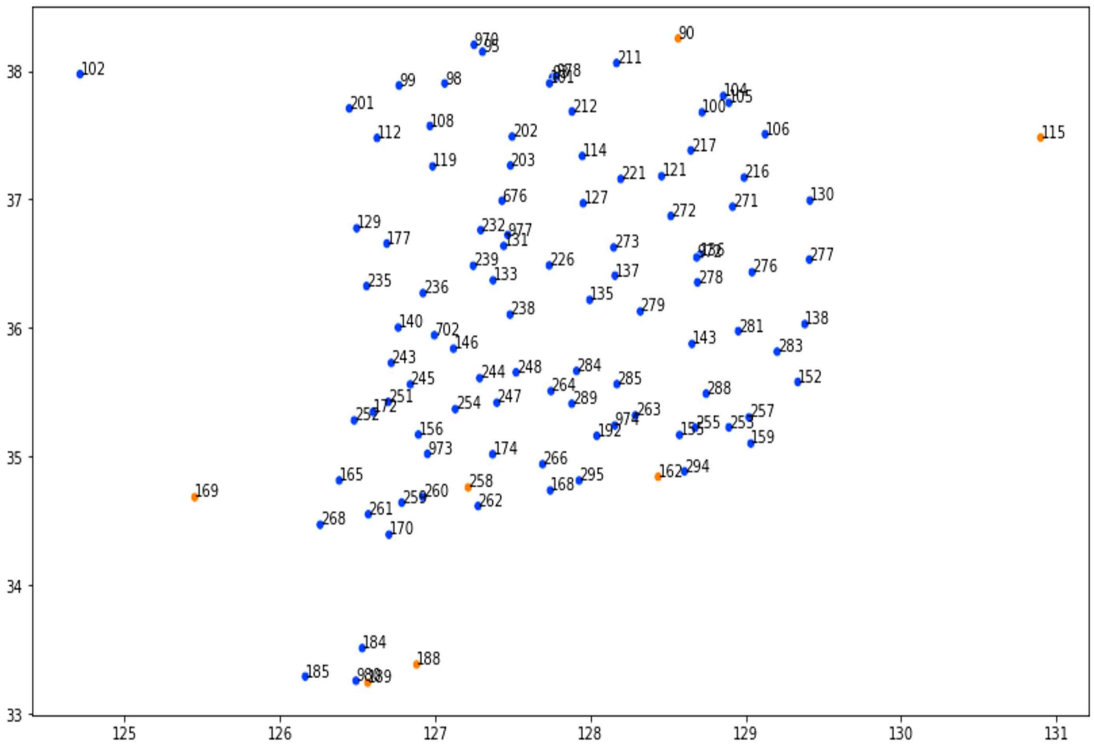
우선 결측값 양쪽의 실측값을 기준으로 시간 선형 보간을 해 주었다. 시간 선형 보간이란 x축을 시간, y축을 값으로 갖는 2차 공간에서 2개의 인접한 실측값을 직선으로 연결하고 그사이의 결측값들을 이 직선 위의 값으로 추측하는 방법이다. 이 방법은 두 실측값 사이의 시간 간격이 길어질수록 보간의 정확도가 떨어지게 된다. 따라서 최대 보간 정도를 양 끝 실측값으로부터 1시간씩 총 2시간으로 제한하여 연속된 결측값이 과도하게 보간되는 것을 방지했다. 시간 선형 보간법에 대한 자세한 예시는 <표 4>와 같다.

Time	Data		Time	Data		Time	Data	
	보간 전	보간 후		보간 전	보간 후		보간 전	보간 후
5시 50분	28	28	6시 50분	NaN	25.18	7시 50분	NaN	22.35
6시 00분	NaN	27.53	7시 00분	NaN	NaN	8시 00분	NaN	21.88
6시 10분	NaN	27.06	7시 10분	NaN	NaN	8시 10분	NaN	21.41
6시 20분	NaN	26.59	7시 20분	NaN	NaN	8시 20분	NaN	20.94
6시 30분	NaN	26.12	7시 30분	NaN	NaN	8시 30분	NaN	20.47
6시 40분	NaN	25.65	7시 40분	NaN	22.82	8시 40분	20	20

<표 4> 시간 선형 보간 예시

이렇게 보간한 결과 종속변수 외에 결측값이 존재하던 26개의 변수 중 18개의 Band 변수에서 결측값이 전부 제거되었다. 결측값이 남아있는 'insitu-TED0.05', 'Solar_amount', 'GK2A-LST', 'insitu-HM', 'insitu-TD', 'insitu-TG', 'insitu-PA', 'insitu-PS'는 공간 추정을 적용하였다.

3.4.2. 공간 추정(Spatial Estimation)



<그림 1> STN별 GK2A-LST

<그림 1>은 STN별 GK2A-LST(위성 관측 지표 온도) 데이터를 나타낸 것으로 주황색으로 표시된 점(Target STN)들은 해당 STN에서 LST 관측값이 전부 결측으로 추정이 필요한 점이며 파란색으로 표시된 점(Known STN)들은 그렇지 않은 점들이다. Target STN들을 공간 추정하기 위해 우선 LST와 가장 상관계수의 절댓값이 높은 변수(LST의 경우 TG)를 찾고, 모든 Known

STN들의 LST 값과 TG 값을 이용해 식 (1)과 같이 회귀식을 추정한다.

$$LST = \alpha + \beta * TG \cdots (1)$$

회귀식을 추정한 이후, Target STN을 중심으로 초기 원을 그려 원 안에 포함된 Known STN(Candidate STN)의 개수를 확인한다. 이때 초기 원의 반지름은 실거리 약 30km인 0.3으로 설정하였다. 만약 Target STN의 Candidate STN이 2개 미만이면 추정을 보류하고 다음 Target STN으로 넘어가서 초기 원을 그리고 Candidate STN의 개수를 확인한다. 원 안에 2개 이상의 Candidate STN이 포함되면 원의 중심에 있는 Target STN과 원 안에 있는 모든 Candidate STN에 대해서 식 (2)처럼 모든 시간에 대한 TG 값 차이 제곱의 합을 계산한다.

$$\sum_{t}(TG_{Target}(t) - TG_{Candidate}(t))^2 \cdots (2)$$

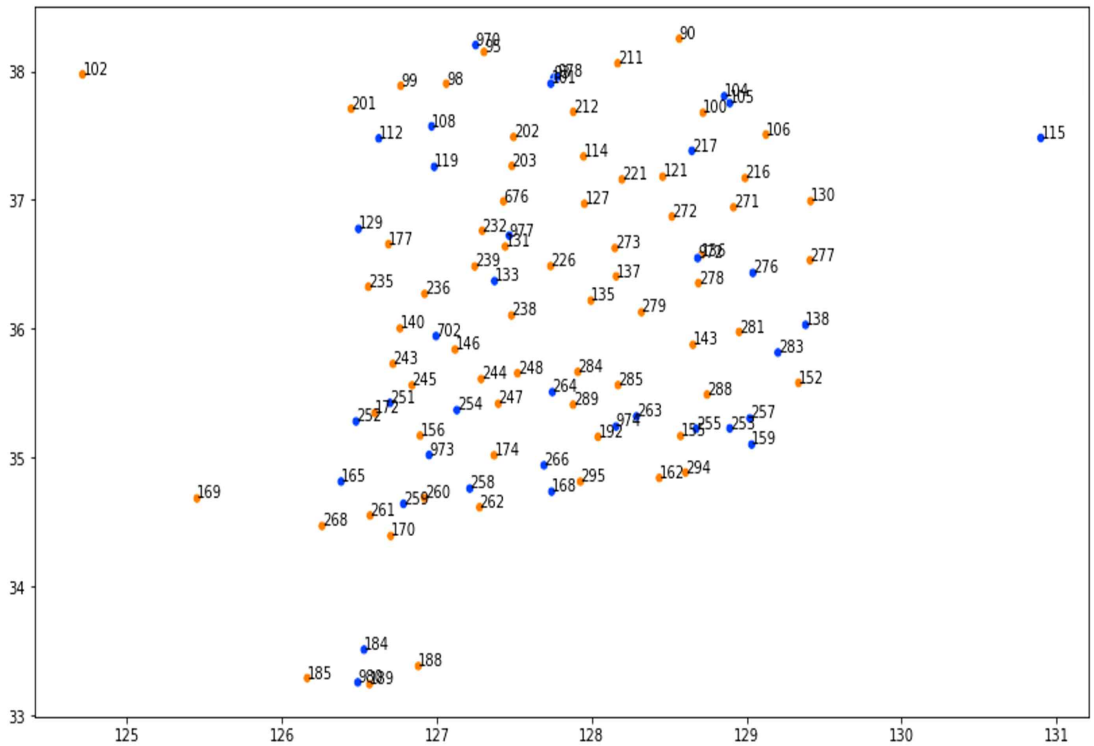
식 (2)의 차이 제곱의 합이 가장 작은 Candidate STN을 Elected STN으로 정의하고, 식 (1)에서 구한 회귀계수 β와 Target STN과 Elected STN의 TG 값 차이를 곱한 값과 Elected STN의 LST 값을 더해 Target STN의 LST 값으로 추정하였다.

$$LST_{Target} = \beta * (TG_{Target} - TG_{Elected}) + LST_{Elected} \cdots (3)$$

식 (2)와 (3)이 의미하는 바는 Target STN은 지리적으로 거리가 가까운 Candidate STN 중 Target STN과 TG 값의 분포가 가장 비슷한 Elected STN의 LST 값을 기저로 하되 미세한 오차를 TG 값의 차이와 회귀계수의 곱으로 추정하겠다는 것이다.

모든 Target STN에서 초기 원에 대한 추정을 진행한 이후 아직 추정되지 않은 Target STN들은 초기 원의 반지름을 10% 늘려서 위 과정을 반복하며, 원의 크기가 너무 커져 지리적으로 차이가 있는 Known STN을 참조하는 것을 방지하기 위해 최대 반복 횟수를 5회로 제한한다. 5회의 반복 동안 대체되지 않은 Target STN은 결측값으로 남겨두었다.

GK2A-LST를 추정한 것과 동일한 방식으로 insitu-HM, insitu-TD, insitu-TG, insitu-PA, Solar_amount의 Target STN의 값을 추정하였다. insitu-PS의 경우 높은 상관관계에 있는 변수가 없어 추정을 적용하지 않았다.



<그림 2> STN별 insitu-TED0.05

insitu-TED0.05(지중온도 5cm)의 경우 <그림 2>에서 보이듯이 전체 STN의 절반 이상이 추정이 필요함을 확인할 수 있다. 이를 추정하기 위해 앞과 동일한 방식을 이용하여 원의 반지름을 키워가며 반복하여 추정한다면 추정값을 받은 STN을 Known STN으로 삼아 다시 다른 Target STN을 추정하는 연쇄적인 추정으로 인해 부정확하게 추정될 수 있다고 판단하였다. 따라서 최대 반복 횟수를 2회로 제한하였으며 2회 반복 동안 대체되지 않은 Target STN들은 결측값으로 남겨두었다. <표 5>는 결측값이 존재하던 변수들의 초기 결측값 개수, 시간 보간 이후의 결측값 개수, 공간 추정 이후의 결측값 개수로 기존 총합 2,221,586개의 결측값을 1,264,129개로 줄여 전체 결측값의 약 43%를 대체하였다.

변수	초기	시간 보간	공간 추정	변수	초기	시간 보간	공간 추정
Band1	11,034	0	-	30days Band3	11,454	0	-
Band2	11,034	0	-				
Band3	11,027	0	-	30days Band13	11,452	0	-
Band4	11,034	0	-				
Band5	11,033	0	-	insitu-TED0.05	856,616	856,003	640,815
Band6	11,033	0	-				
Band7	11,038	0	-	Solar Amount	115,932	114,202	7,209
Band8	11,032	0	-				
Band9	11,033	0	-	GK2A-LST	853,499	544,650	515,857
Band10	11,031	0	-	insitu-HM	9,174	7,898	7,168
Band11	11,033	0	-	insitu-TD	10,259	8,951	8,221
Band12	11,031	0	-	insitu-TG	20,903	19,783	7,122
Band13	11,033	0	-	insitu-PA	77,345	76,336	348
Band14	11,033	0	-	insitu-PS	78,432	77,389	77,389
Band15	11,029	0	-	isitu-LST	8,170	-	
Band16	11,032	0	-	insitu-TA	9,322		

<표 5> 결측값 보간/추정 결과

3.5. Band 변수 병합

독립변수의 수를 줄여 모델의 학습 효율과 예측 성능을 높이기 위해 상관계수가 1에 근접한 값을 갖는 몇몇 무의미한 밴드 변수들을 병합했다. 그 결과 <표 6>과 같이 비슷한 의미를 갖는 13개의 Band 변수가 3개의 그룹으로 병합되었다.

병합 그룹 1	병합 그룹 2	병합 그룹 3
Band1(파랑 가시밴드)	Band8(상층 수증기 밴드)	Band11(구름상 밴드)
Band2(초록 가시밴드)		Band12(오존 밴드)
Band3(빨강 가시밴드)	Band9(중층 수증기 밴드)	Band13(대기창 밴드)
		Band14(깨끗한 대기창 밴드)
Band4(식생 가시밴드)	Band10(하층 수증기 밴드)	Band15(오염된 대기창 밴드)
		Bamd16(이산화탄소 밴드)

<표 6> Band 병합 결과

3.6. 종속변수 이상값 삭제

3.6.1. 지면 온도

지면 온도 분포의 99.99% 분위 수에 해당하는 값은 64.6이며 0.01% 분위 수에 해당하는 값은 8.8로 이 두 값을 벗어나는 데이터 약 0.02%는 이상값으로 규정하고 삭제하였다.

3.6.2. 지상 온도

지상 온도 분포의 99.99% 분위 수에 해당하는 값은 36.6이며 0.01% 분위 수에 해당하는 값은 9.3으로 지면 온도와 마찬가지로 두 값을 벗어나는 값을 삭제하였다.

3.7. LandType One Hot Encoding

LandType은 지면 타입을 표현한 변수로 0~4 사이의 정숫값을 갖는다. 이를 One Hot Encoding을 적용해 LandType 0~4까지 5개의 독립변수를 추가했으며 원 변수는 삭제했다.

4. 모델링(Modeling)

예측 모델로는 정형 데이터에 뛰어난 예측 능력을 보여주는 것으로 알려진 LightGBM (LGBM), CATBoost (CAT), Random Forest (RF), XGBoost (XGB) 네 가지를 사용하였다. 모델의 성능은 평균 제곱근 오차인 RMSE를 사용하여 평가하였다.

$$RMSE = \sqrt[2]{\sum_{i=1}^{N} \frac{(A_i - F_i)^2}{N}} \ (A_i = Actual, F_i = Prediction) \cdots (4)$$

우선 Hyper parameter가 Default 값으로 설정된 각 모델에 학습 데이터를 학습시키고 각 모델의 내장 함수(feature_importances_)를 통해 변수 중요도가 0.1% 미만인 변수를 모델별 확인 및 삭제하였다. 이후 Grid Search를 통해 네 가지 모델의 최적 Hyper parameter를 찾았으며 모델 조합 Grid Search를 바탕으로 가중치를 설정해 각 모델의 예측을 종합하는 최종 앙상블 모델을 만들어 모델별 취약점을 보완했다.

5. 모델을 이용한 예측(Prediction)

네 가지 모델 각각과 앙상블 모델의 테스트 데이터 예측 성능은 <표 7>, <표 8>과 같다.

5.1. 지면 온도

Model	RMSE	Ensemble Weight	Ensemble Prediction	Ensemble RMSE
LGBM	1.5999	0.4090	LGBM_pred*0.4090 + CAT_pred*0.3636 + RF_pred*0.0454 + XGB_pred*0.1818	1.5574
CAT	1.6298	0.3636		
RF	1.7430	0.0454		
XGB	1.6650	0.1818		

<표 7> 지면 온도 테스트 결과

5.2. 지상 온도

Model	RMSE	Ensemble Weight	Ensemble Prediction	Ensemble RMSE
LGBM	0.0454	0.0625	LGBM_pred*0.0625 + CAT_pred*0.1875 + RF_pred*0.6250 + XGB_pred*0.1250	0.0270
CAT	0.0428	0.1875		
RF	0.0290	0.6250		
XGB	0.0498	0.1250		

<표 8> 지상 온도 테스트 결과

각 모델의 Ensemble Weight은 식 (5)를 이용해 계산했다. 순서쌍 (L, C, R, X)는 각각 1부터 10까지 10,000(10*10*10*10)개의 조합 중 Grid Search를 통해 가장 높은 Ensemble RMSE를 산출하는 조합으로 결정했다.

$$Weight_A = \frac{\mathrm{A}}{L + C + R + X}, A \in \{L, C, R, X\} \cdots (5)$$

6. 활용 방안 및 기대효과

지면/지상 온도를 정확히 예측하는 것은 매우 중요하여 다양한 방법으로 기상 자료를 수집하고 있지만 결측값이 많이 존재하는 실정이다. 이러한 상황에서 이번 프로젝트에서 개발된 예측 모델을 통해 실제 지면/지상 온도에 근접한 추정값을 산출함으로써 가뭄, 태풍 등 자연재해의 사전적인 예방과 보수, 기후변화에 대한 환경 정책을 마련, 일상생활에서의 이용 등 다양한 분야에서 폭넓게 사용될 수 있을 것으로 기대된다.

References

1. 이용관.(2016).천리안 위성 자료를 활용한 한반도의 일별 지면 온도 산정을 위한적정 관측시간 설정 연구.한국농공학회논문집,58(4),37-46.
2. 백종진, 최민하.(2012).천리안 위성을 이용한 지표면 온도의 검증.한국수자원학회 학술발표회,(),99-102.
3. 최낙빈, 이명인, 탁선래, 이준리.(2021).2021년 여름철 폭염 특성.한국기상학회 학술대회 논문집,(),166-166.
4. 구민호(Min-Ho Koo);송윤호(Yoonho Song);이준학(Jun-Hak Lee). (2006). 국내 지면 온도의 시공간적 변화 분석. 자원환경지질, 39(3), 255-268.
5. 조창제. "시공간 정보를 고려한 온도 예측모형의 비교." 국내석사학위논문 대구대학교, 2020. 경상북도
6.신휴석, 장은미, 홍성욱.(2014).MODIS 지표면 온도 자료와 지구통계기법을 이용한 지상 기온 추정.한국공간정보학회지,22(1),55-63.
7.이민혁, 전인우, 전철민.(2017).개선된 DBSCAN 알고리즘을 이용한 대중교통 정류장 군집화 기법.대한공간정보학회지,25(4),97-106.
8. 윤희영, 구윤서, 최대련. (2017). 미세먼지 예보정확도 향상을 위한 군집분석에 의한 앙상블 기법 개발 : 기상자료 가중치를 중심으로. 한국도시환경학회지, 17(1), 33-42.
9. B. S. Panda and R. Kumar Adhikari, "A Method for Classification of Missing Values using Data Mining Techniques," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), 2020, pp. 1-5, doi: 10.1109/ICCSEA49143.2020.9132935.