환경위성 GEMS 데이터 기반 미세먼지 농도 추정 오류 개선을 위한 파생지표 개발

송유찬, 배정훈, 전경구*

인천대학교 임베디드시스템공학과

pndgm@inu.ac.kr, oiryh@inu.ac.kr, kjun@inu.ac.kr

Development of Derived Indicators for Improving Fine Dust Concentration Estimation Errors Based on GEMS Environmental Satellite Data

Yuchan Song, Jeonghun Bae, Kyungkoo Jun*

Dept. of Embedded Systems Engineering, Incheon National University

요 약

건강에 심각한 영향을 미치는 미세먼지에 대한 모니터링이 중요하지만, 개발도상국들은 지상 미세먼지 농도 측정 시설 구축에 소요되는 비용 문제로 어려움을 겪는다. 최근 인공위성 데이터를 이용하여 지상 미세먼지 농도를 추정하는 연구들이 진행되고 있다. 이러한 연구들은 인공위성 자료 외에도 지리, 기상, 경제 등의 자료를 활용하지만, 개발도상국들에 대한 이러한 자료들이 부족하여 연구 성과들을 적용하는데 한계가 있다. 본 논문에서는 부가 자료 없이 대한민국 환경위성 탑재체 Geostationary Environment Monitoring Spectrometer (GEMS) 데이터만을 이용하여 기계학습 알고리즘으로 지상 농도를 추정함에 있어, 정확도를 향상시킬 수 있는 파생지표들을 제안한다. 첫 번째는 관심영역의 구름 정도, 두 번째는 관측 지점과 주변의 구름 차이, 그리고 세 번째는 관심영역의 유효 관측 비율이다. 이 값들은 GEMS 데이터 구성 요소들 중 하나인 Cloud Fraction 기반으로 계산된다. 성능평가 실험은 GEMS 데이터를 이용하여 미세먼지 농도를 추정한 후, 이를 지상에서 실측한 값과 비교한다. 지상 미세먼지 농도 데이터가 공개된 대한민국을 대상으로 실험한 결과, 파생지표를 사용할 때, 기계학습 기반 알고리즘들의 경우 추정오류 RMSE가 평균 8.94% 감소하였다.

Ⅰ. 서 론

개발도상국가들은 지상 미세먼지 측정망이 부족하여, 대기 오염 상황을 지속적으로 관측하기 어렵다. 대한민국이 개발해서 운영 중인 GEMS[8]는 정지궤도에서 아시아 태평양 지역의 기후변화 및 대기오염물질을 관측하는 천리안위성 2B의 환경탑재체이다. 지상 관측을 위해 자외선 및 가시광선 초분광계를 이용하고, 에어로졸 및 대기화학 관련 데이터가 생산되며 그림 1과 같이 동남아를 포함한 지역들을 관측한다.

이러한 위성 산출물만으로 지상 Particulate Matter(PM) 농도를 추정할 수 있다면 개발도상국가들의 미세먼지 모니터링 수단으로 활용될 수 있다. 그러나 구름, 관측각, 지표 반사 등 관측 품질의 변동은 GEMS 데이터에 오차를 유발한다. 본 논문에서는 GEMS 데이터를 입력받아 지상 미세먼지 농도를 추정하는 기계학습 알고리즘들 RandomForest[7], XGBoost[5], LightGBM[6]의 오차를 개선하기 위한 파생지표들을 제안한다. 본 논문의 주요 기여는 다음과 같다.

- 기계학습 알고리즘에 활용 가능한 GEMS 데이터 전처리 방법 개발
- GEMS 데이터 중 구름 관련 산출물(Cloud Fraction)로부터 미세먼지 추정 정확도 향상에 기여할 수 있는 3가지 파생지표 개발
- 파생지표의 효과성 검증 실험으로 정확도 개선 확인

Ⅱ. 본론

GEMS 데이터 중 기계학습 입력으로 사용되는 요소들을 소개하고, 본 논문에서 제안하는 3가지 파생지표들을 설명한다. 또한, 이 지표들을 활용한 기계 학습 알고리즘의 지상 미세먼지 농도 추정 실험 결과를 제시한다.

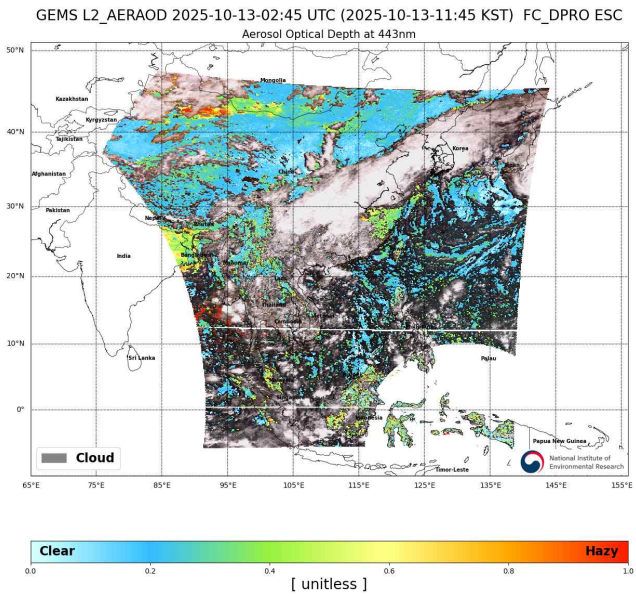

그림 1. 환경위성 GEMS 관측 영역과 에어로졸 광학두께 산출물 예시

GEMS 데이터

다양한 GEMS 산출물 들 중, 미세먼지 농도 추정을 위한 기계학습 알고리즘의 입력으로 사용되는 것들은 다음 15개 산출물, AOD, BSR, SSA, AEH(AGL), Cloud Fraction, NO_2 , O_3 T, SO_2 , HCHO, CHOCHO, UVI, NRAD이다. 이들은 미세먼지 농도와 각 산출물 간 상관관계를 표 1과 같이 분석한 후, 상위 15개로 선정한 것이다. Cloud fraction의 경우, 상대적으로 낮은 상관 관계를 보완하기 위해, 다음에 설명할 파생지표들로 변환하여 사용한다. 시간·위경도·외부 변수는 입력에서 제외하고, 좌표

와 관측 시각은 라벨 매칭 및 관심영역 산출에만 사용한다.

표 1. PM10·PM2.5와 GEMS 산출물의 상관계수 요약(상위 6개)

Feature	PM10 (r)	PM10 (p)	PM25 (r)	PM25 (p)
O_3T	0.3720	0.4459	0.3389	0.3604
AOD	0.2531	0.3095	0.2992	0.3204
NO_2	0.2398	0.3428	0.2674	0.2762
AEH	-0.1245	-0.1601	-0.1509	-0.1441
NRAD	-0.1561	-0.1945	-0.1492	-0.1359
Cloud Fraction	-0.1329	-0.1584	-0.1353	-0.1251

Cloud Fraction 관심영역 파생지표

위성 데이터를 활용한 지상 미세먼지 농도 추정에서 중요한 요소 중 하나는 구름이다. 따라서 본 논문에서는 GEMS 데이터 중 구름 관련한 데이터로부터 다음과 같은 3가지 파생지표를 만들어 낸다. CF_ROI_med는 추정 대상 영역의 구름 상태를 나타내고, CF_diff는 해당 영역의 상대적 날씨 쾌청도를 의미한다. 그리고. CF_ROI_valid_ratio는 구름이 없는 유효 관측 영역 비율을 의미한다.

이러한 파생지표들은 다음과 같이 계산된다. 우선 CF_ROI_med는 그림 2와 같이 추정 대상 영역의 중심에서 75km에 해당하는 Cloud Fraction의 중앙값으로 계산된다. CF_diff는 영역 내 위치별 cloud fraction 수치와 중앙값 CF_ROI_med의 차이이다. 마지막으로, CF_ROI_valid_ratio는 영역 내 cloud fraction 값이 임계값 이하를 가지는 유효 위치들의 비율이다. 이들은 모두 동일 시각의 Cloud Fraction으로부터 계산되며, 추정 대상 영역을 결정하는 반경 75km는 GEMS 데이터의 공간 해상도를 고려하여 결정하였다.

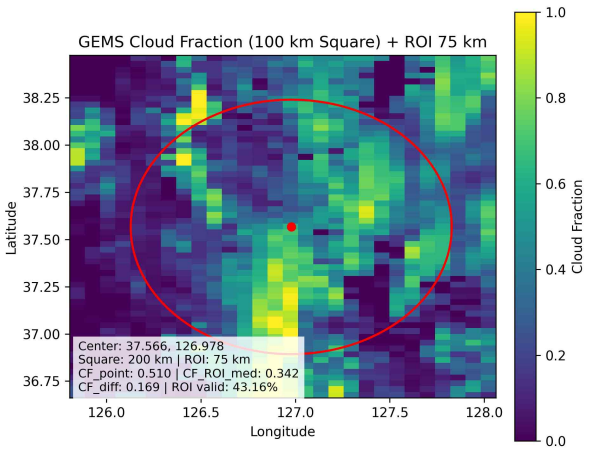

그림 2. Cloud Fraction 기반 ROI(반경 75 km)

성능 실험과 결과 분석

기계 학습 알고리즘 Random Forest, XGBoost, 그리고 LightGBM에서 GEMS 데이터만을 이용한 경우 (baseline)와 제안 지표들을 포함한 경우(CF-ROI)로 나누어 추정오류를 측정하였다. 성능 평가는 R^2, RMSE, MAE, Acc@±10, Acc@±5 를 기준으로 한다. Acc@±10/±5는 예측 오차가 각각 ±10/±5 $\mu g/m^3$ 이내인 표본의 비율(%)을 의미한다.

표 2의 실험 결과와 같이, Baseline 데이터에 대해, XGBoost는 $R^2 \approx$ 0.525를 보였고, 파생지표가 포함된 CF-ROI 데이터에 대해서는, 모든 기계 학습 알고리즘들에서 오류가 개선되었다. RandomForest가 0.548→0.625(Δ+0.077), XGBoost가 0.525→0.604(Δ+0.079), LightGBM이 0.550→0.630(Δ+0.080)으로 개선되었고, RMSE 또한 1.8 $\mu g/m^3$ 감소하였다. 이러한 오류 개선은 구름량과 미세먼지 농도 추정간 상관관계를 반영한 파생지표들의 활용으로 가능하였다. 파생지표의 한계로는 ROI 반경에 따른 민감도분석 미비와, 동남아 지역 적용시 검증을 위한 미세먼지 농도 데이터의 부재이다. 향후에는 반경(예: 25 - 100 km)에 대한 민감도 분석, 동남아 현지 자료 활용을 통한 보강 연구가 필요하다.

표 2. 미세먼지 농도 추정 실험 결과

Model	Data	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)	R^2	Acc@±10	Acc@±5
Random Forest	Baseline	20.429	11.726	0.548	60.8%	33.6%
XGBoost	Baseline	20.948	12.297	0.525	58.4%	32.0%
LightGBM	Baseline	20.385	12.037	0.550	58.8%	32.0%
Random Forest	CF-ROI 75×3	18.614	10.786	0.625	63.9%	36.0%
XGBoost	CF-ROI 75×3	19.140	11.476	0.604	60.9%	33.5%
LightGBM	CF-ROI 75×3	18.485	11.198	0.630	61.2%	33.9%

Ⅲ. 결론

기계학습 알고리즘에 GEMS 데이터만을 입력으로 활용하여 지상 미세먼지 농도를 추정할 때, 오류를 감소시키기 위한 구름 관련한 데이터 기반으로 파생지표들을 제안하였다. 성능 검증을 위해 대한민국 미세먼지 실측 데이터를 이용한 실험을 진행하여, 파생지표들이 오류를 유의미하게 개선할 수 있음을 확인하였다.

참 고 문 헌

[1] S. Park, M. Kim, and J. Im, “Estimation of ground-level PM10 and PM2.5 …,” Korean J. Remote Sens., vol. 37, no. 2, pp. 321 - 335, 2021.

[2] Y.-S. Choi, G.-Y. Kim, M.-J. Cho, B.-R. Kim, and M.-J. Kwon, GEMS ATBD: Cloud Retrieval Algorithm, Environmental Satellite Center, National Institute of Environmental Research (NIER), 2024.

[3] H. Choi, Y. Kang, and J. Im, “Estimation of TROPOMI-derived ground-level SO_2 …,” Korean J. Remote Sens., vol. 37, no. 2, pp. 275 - 290, 2021.

[4] Environmental Satellite Center, NIER (MOE), GEMS ATBD: Ground-Level PM Estimation Algorithm, 2024.

[5] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. KDD ’16, pp. 785 - 794, 2016.

[6] G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” in NeurIPS 2017, Long Beach, CA, USA, 2017.

[7] K. Fawagreh, M. M. Gaber, and E. Elyan, “Random forests: From early developments to recent advancements,” Syst. Sci. Control Eng., vol. 2, no. 1, pp. 602 - 609, 2014.

[8] 국립환경과학원 환경위성센터, “환경위성센터,” 2025. [온라인]. 이용 가능: https://nesc.nier.go.kr/ko/html/index.do