A Measurement Error Approach for Satellite-based Air Quality Modeling

Using GEMS Level-2 NO2 Products

Park, Rinseo

Yonsei University

October 15, 2021

Introduction

- Satellite-based air quality modeling
  - Satellite remote sensing has been widely used to conduct a broad range, real-time monitoring of air quality.
- Missingness and measurement errors
  - Satellite data have potential sources of error in measurement.
  - For example, *cloud cover* may increase uncertainties or even disable remote sensing as satellite sensors are negatively affected.
- The **threshold vs. measurement error** approach
  - It is a common practice to discard unreliable pixels by introducing certain thresholds to those sources of error.
  - From a statistical viewpoint, this can cause bias and efficiency loss.

Research Problem

1. We estimate the **mean fields** from satellite air quality data.
   - The data are *GEMS*[1] *Level-2* NO_2 *(nitrogen dioxide)* products.
   - Our mission is to calculate their cell means in re-gridded *Level-3*.

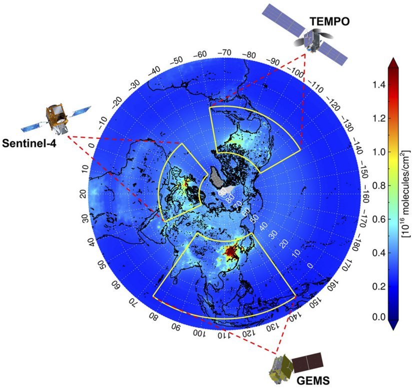


Figure 1: GEMS produces a variety of air quality data over East Asia.

[1] GEMS = Geostationary Environment Monitoring Spectrometer

Research Problem

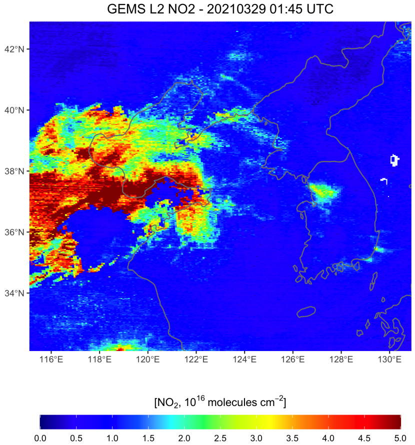


Figure 2: GEMS Level-2 NO2 products

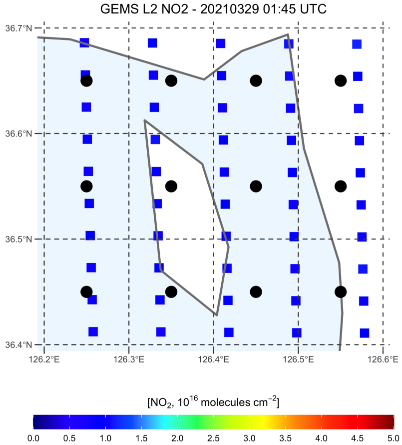


Figure 3: Level-2 and Level-3 products

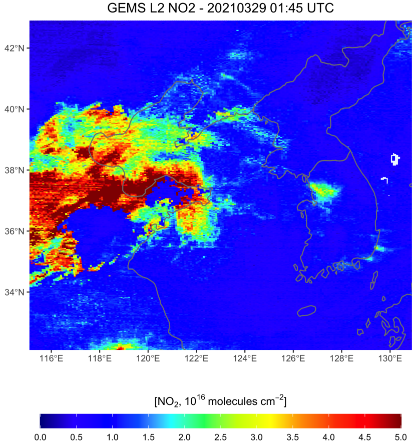
Research Problem

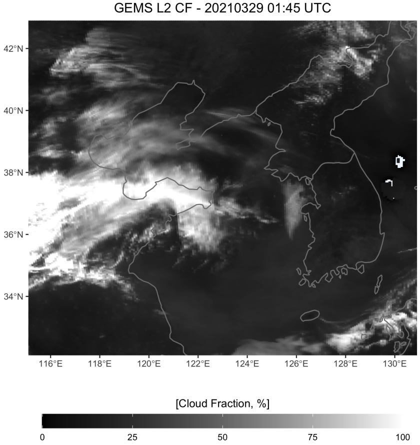
1. We estimate the **mean fields** from satellite air quality data.
   - The data are *GEMS Level-2 NO_2 (nitrogen dioxide)* products.
   - Our mission is to calculate their cell means in re-gridded *Level-3*.
2. Each observation is prone to **measurement error**.
   - GEMS also provides information about data quality, such as cloud fraction (i.e., CF) and viewing angles (i.e., SZA, VZA[2]).
   - Potential error sources should be considered to adjust possible error.
3. The data have **spatio-temporal** dependence.
   - GEMS observes around the Korean peninsula, hourly 8 times a day.

[2] SZA = Solar Zenith Angle, VZA = Viewing Zenith Angle.

Data and Related Studies

GEMS Level-2 NO2 products (1)

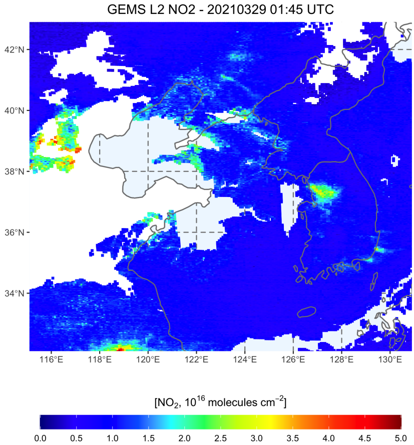


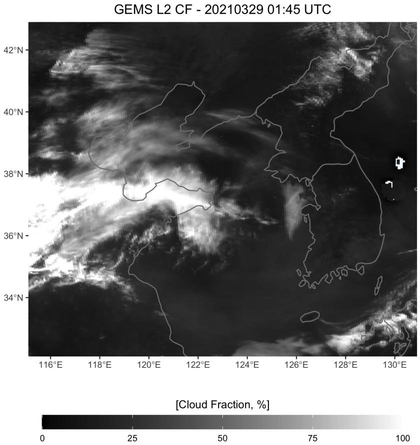


- 2021/03/29 South Korea hit by season's worst yellow dust

Data and Related Studies

GEMS Level-2 NO2 products (2)





- If unreliable data pixels are deleted as CF(cloud fraction) $\geq$ 30%, high concentration would not be included in the mean fields.

Data and Related Studies

GEMS Level-2 NO2 products (3)

Deleted pixels (left) NOT recovered after gridding (right).

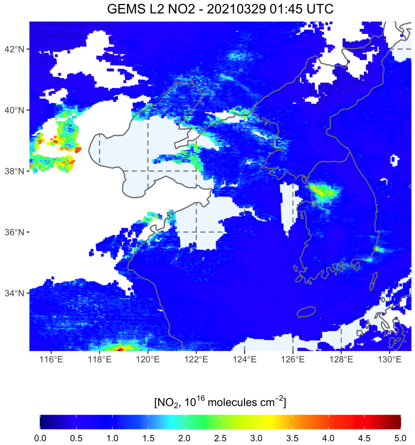


Figure 4: L2 products where CF<30%

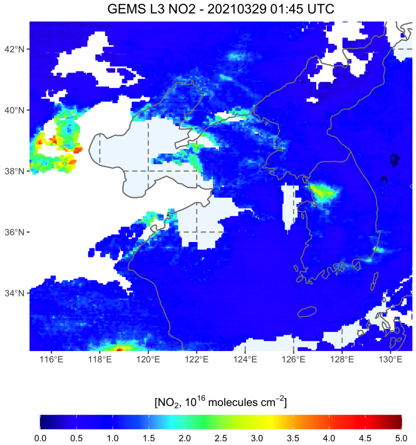


Figure 5: L3 products from Figure 4

Data and Related Studies

Previous methods in remote sensing

- Rely on the threshold approach
  - Set quality thresholds to create missing values and *fill in the blanks* using different geostatistical algorithms
  - For example, **the tessellation approach** calculates the cell means in Level-3 by area-weighted averages (Sun et al., 2018).[3]
- Focus on cloud-induced missingness (Kikuchi et al., 2018)
  1. Missing proportion is very high (almost 50%)
  2. Missing mechanism may not be random.
- Measurement error approach as a statistical alternative
  - Errors are in the reponse.
  - Errors can be explained by other sources (e.g., cloud fraction [CF]).
  - Errors are spatio-temporally correlated.

[3] The results are displayed in Figure 5 on the page 8.

Proposed Method

Basic setup

Error in the response (Buonaccorsi, 1996)

Given the structural error in the spatio-temporal response Y,

$$Y_{s,t} = \mu_{s,t} + \epsilon_{s,t} \quad \text{where} \quad \epsilon_{s,t} \sim (0, \sigma_\epsilon^2) \tag{1}$$

$$Y^*_{s,t} = Y_{s,t} + e_{s,t} \quad \text{where} \quad e_{s,t} \sim (0, \sigma_e^2) \tag{2}$$

the observed Y^* can be decomposed into three terms.

$$\begin{aligned} Y^*_{s,t} &= Y_{s,t} + e_{s,t} \\ &= (\mu_{s,t} + \epsilon_{s,t}) + e_{s,t} \;\; \text{where } Cov(\epsilon_{s,t}, e_{s,t}) = 0 \\ &= (\text{true mean} + \text{structural error}) + \text{measurement error} \end{aligned}$$

Proposed Method

Error assumption

- Now we expand the MEM framework to incorporate *two types of measurement error* (Carroll et al., 2006).
- First, **classical error** refers to random error.

$$Y^*_{s,t} = Y_{s,t} + e_{s,t} \tag{2}$$

- Second, **systematic error** can bias results if $\gamma_0 \neq 0$ and $\gamma_1 \neq 1$.

$$Y^*_{s,t} = \gamma_0 + \gamma_1 Y_{s,t} + e_{s,t} \tag{3}$$

- Error equations (2) and (3) combined, we have the following (4).

$$Y^*_{s,t} = \begin{cases} Y_{s,t} + e_{s,t} & \text{if} \quad \text{CF} < 30\% \\ \gamma_0 + \gamma_1 Y_{s,t} + e_{s,t} & \text{if} \quad \text{CF} \geq 30\% \end{cases} \tag{4}$$

Proposed Method

Model estimation

Model setup

$$Y_{s,t} = \mu_{s,t} + \epsilon_{s,t} \tag{1}$$

$$Y^*_{s,t} = \begin{cases} Y_{s,t} + e_{s,t} & \text{if} \quad \text{CF} < 30\% \\ \gamma_0 + \gamma_1 Y_{s,t} + e_{s,t} & \text{if} \quad \text{CF} \geq 30\% \end{cases} \tag{4}$$

Note. $\epsilon_{s,t} \sim (0, \sigma^2_\epsilon)$ and $e_{s,t} \sim (0, \sigma^2_e)$ are independent.

- Two common approaches for estimation
  1. The Buonaccorsi approach using the method of moments
  2. The maximum likelihood approach
- *The maximum likelihood approach* (Keogh et al., 2016)
  - Uses the joint distribution of both unbiased (i.e., CF $< 30\%$) and biased measurements (i.e., CF $\geq 30\%$)
  - Can incorporate the spatio-temporal covariance structure

Proposed Method

Data structure for estimation

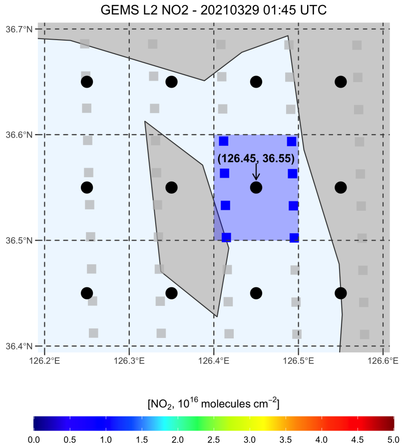


Figure 6: Estimation for Level-3 products

- Level-2 and Level-3 products
  - (Level-2) $Y^*_{s,t}$ = observed
  - (Level-3) $Y_{s,t}$ = true
  - $s = (126.45°\text{E}, 36.55°\text{N})$
  - t = 2021-03-29 01:45 UTC
- The true $Y_{s,t}$ is unknown.
  - Replicate measurements $Y^*_{s,t}$ are used to estimate $Y_{s,t}$.
- Error types may depend on CF.
  - Biased measurements (CF $\geq$ 30%) can be **calibrated** using unbiased ones (CF $<$ 30%).

Future Work

- Simulation study
  - For model evaluation, we plan a Monte Carlo simulation study.
  - Based on spatio-temporal kriging methods, we generate Level-2 NO2 data from which the true Level-3 means are known.
  - Then we will compare the Level-3 estimates of different methods using *MSE(mean squared error)* and *bias*.

Possible Contribution

1. Statistical applications in environmental research
   - The **threshold (before) vs. measurement error (after)** approach
   - Proposes a better way for satellite-based mean fields estimation
2. Extending models for measurement error in the response
   - The statistical literature on MEM has focused on explanatory variables.
   - Our method deals with *responses in the spatio-temporal framework*.

References

- Buonaccorsi, J. P. (1996). Measurement error in the response in the general linear model. *Journal of the American Statistical Association, 91*(434), 633. https://doi.org/10.2307/2291659
- Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. Chapman and Hall/CRC.
- Keogh, R. H., Carroll, R. J., Tooze, J. A., Kirkpatrick, S. I., & Freedman, L. S. (2016). Statistical issues related to dietary intake as the response variable in intervention trials. *Statistics in Medicine, 35*(25), 4493-4508. https://doi.org/10.1002/sim.7011
- Kikuchi, M., Murakami, H., Suzuki, K., Nagao, T. M., & Higurashi, A. (2018). Improved hourly estimates of aerosol optical thickness using spatiotemporal variability derived from Himawari-8 geostationary satellite. *IEEE Transactions on Geoscience and Remote Sensing, 56*(6), 3442–3455. https://doi.org/10.1109/tgrs.2018.2800060
- Sun, K., Zhu, L., Cady-Pereira, K., Chan Miller, C., Chance, K., Clarisse, L., Coheur, P., González Abad, G., Huang, G., Liu, X., Van Damme, M., Yang, K., & Zondlo, M. (2018). A physics-based approach to oversample multi-satellite, multispecies observations to a common grid. *Atmospheric Measurement Techniques, 11*(12), 6679-6701. https://doi.org/10.5194/amt-11-6679-2018