Spatial-temporal Modeling of Ambient Ozone

Lucy Lu, Sam Yin

DSS, Duke University

May 2, 2017

Outline

Introduction

- Ground level ozone (ppb)
- Collected at 1249 sites in the U.S. from 06/01/2015 00:00:00 to 06/30/2015 23:00:00 at hourly frequency.

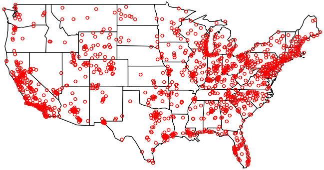

Figure 1: Map of ozone sites across the U.S.

Data Preprocessing

- Keep sites where data completeness is at least 95%
- Divide the U.S. into 25 subregions and model by region
- Impute missing values by averaging the two closest available observations along the temporal dimension

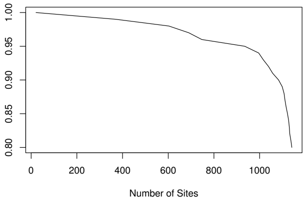

Figure 2: Cut-off percentage of completeness by number of sites

EDA

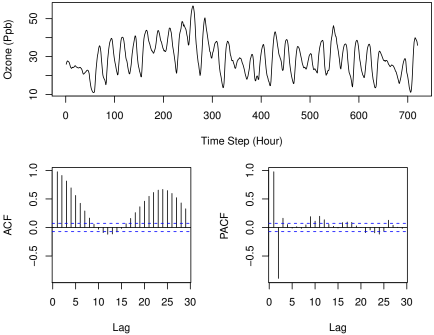

Figure 3: Time series display of average ozone (of all sites) in region 1

EDA (cont.)

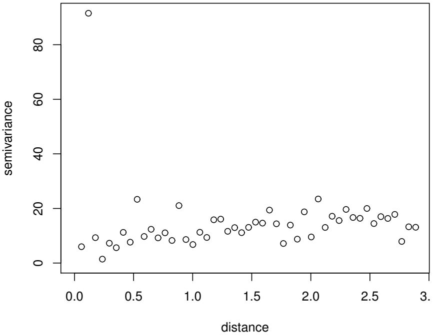

Figure 4: Variogram of average ozone (of all time steps) in region 1

Modeling Ideas

Assuming no dependence between spatial and temporal dimensions, we build models with the additive structure (2). In other words, $\alpha(t)$ is constant to all sites at a given time point, and $\omega(s)$ is constant at all time points given a specific site.

$$
\begin{aligned}
y(s,t) &= w(s,t) + e(s,t) && (1) \\
&= \alpha(t) + \omega(s) + \epsilon(s,t) && (2)
\end{aligned}
$$

Modeling Ideas (cont.)

Temporal specification:

1. AR(1): $\alpha(t) = \delta + \phi\alpha(t-1)$
2. MA(1): $\alpha(t) = \delta + w_t + \theta w_{t-1},\ w_{1:T} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_w^2)$

Spatial specification:

$$\omega(\tilde{s}) \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_\omega) \tag{3}$$

where $\{\Sigma_\omega\}_{ij} = \sigma_\omega^2 \exp(-\phi_\omega ||s_i - s_j||_2)$ for $i \neq j$
Error specification:

$$\epsilon(s, t) \sim N(0, \sigma_\epsilon^2) \tag{4}$$

Model Fitting

- Use a 120-hour learning window for every next-hour ozone forecast
- Use 47 sites for model training and 24 sites for kriging
- JAGS, 4000 iterations, 3000 burn-ins, 3 chains
- First make next-hour forecasts at observed sites, then krige to unobserved sites by sampling from conditional multivariate normal distribution with covariance specified by GP

Evaluation Criteria

$$\text{PMSE} = \frac{\sum_{t=1}^{m}\sum_{s=1}^{n}(\hat{y}(s,t) - y_{obs}(s,t))^2}{mn}$$

$$\text{Average Interval Length} = \frac{\sum_{t=1}^{m}\sum_{s=1}^{n}\text{Length of 95\% CI at Time } t \text{ Site } s}{mn}$$

$$\text{Coverage} = \frac{\sum_{t=1}^{m}\sum_{s=1}^{n}\mathbf{1}(y_{obs}(s,t) \in 95\% CI)}{mn}$$

$$\text{CRPS} = E_F|y(s,t) - y_{obs}(s,t)| - \frac{1}{2}E_F|y(s,t) - y'(s,t)|$$

Results

Model	PMSE	Interval	Coverage	CRPS
AR(1)	17.7	12.98	0.85	2.36
MA(1)	126.24	36.56	0.91	6.56

Table 1: Summary of predictive performance at fitting sites (47)

Model	PMSE	Interval	Coverage	CRPS
AR(1)	60.08	12.84	0.79	4.29
MA(1)	38.00	27.49	0.92	3.45

Table 2: Summary of predictive performance at kriging sites (24)

AR(1)	Mean	95% CI	MA(1)	Mean	95% CI
σ_ω^2	4.32	(3.59, 5.20)	σ_ω^2	49.97	(44.95, 55.55)
ϕ_ω	0.15	(0.10, 0.20)	ϕ_ω	0.57	(0.48, 0.67)
ϕ	0.88	(0.87, 0.89)	θ	1.06	(0.76, 1.44)
δ	3.83	(2.80, 5.14)	σ_w^2	13.26	(7.76, 20.21)
σ_ϵ^2	6.40	(6.07, 6.75)	δ	25.51	(24.11, 26.94)
			σ_ϵ^2	14.68	(13.00, 16.41)

Table 3: Summary of parameter posterior distributions

Results (cont.)

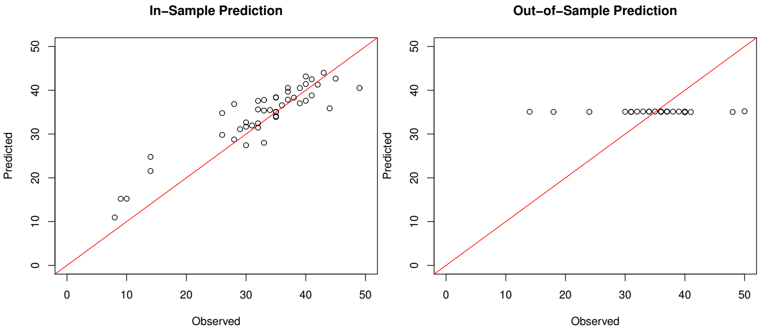

Figure 5: Pedicted vs observed ozones, AR(1)

Results (cont.)

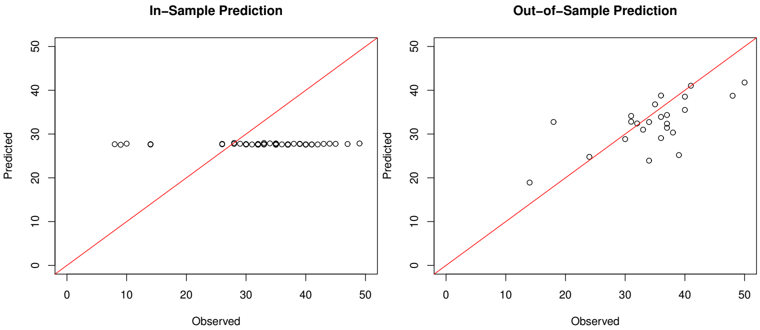

Figure 6: Pedicted vs observed ozones, MA(1)

Discussion

- In-sample vs. Out-of-sample (kriging) prediction
- Time series
- Additive assumption & Computational efficiency