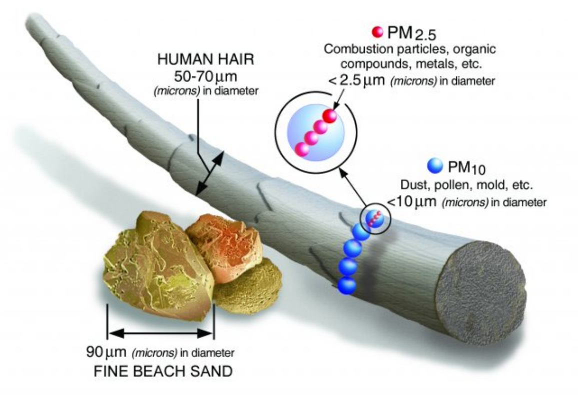
Spatio-Temporal Forecasting of PM2.5 in Texas

Taylor Grimm

April 26, 2023

What is PM2.5?

- particulate matter with diameters generally 2.5 micrometers or smaller
- a mixture of solid particles and liquid droplets found in the air
- inhalable



EPA Website source (click)

Why does PM2.5 Matter?

Health Risks

PM can contain harmful microscopic solids or liquids that may enter a person's lungs or bloodstream.

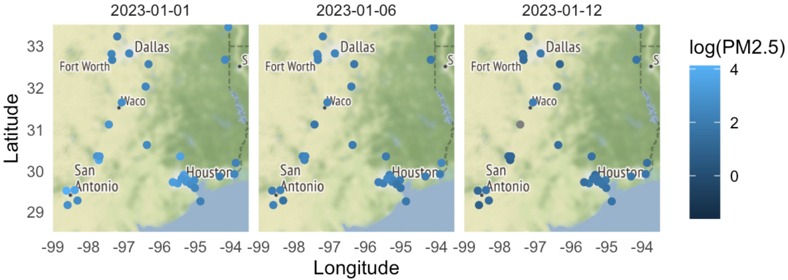
Examples of health effects include:

- premature death in people with heart or lung disease
- nonfatal heart attacks
- irregular heartbeat
- aggravated asthma
- increased respiratory symptoms
- decreased lung function

Data - PM2.5 in Texas

Daily average PM2.5 measurements across 50 sensors in Texas during January 2023

- Most sensors located around Houston, San Antonio, and Dallas
- Few sensors in rural areas



- Potential spatial & temporal components

Methods - Models Considered

Models

- independent error linear regression model (LM)
- Bayesian dynamic spatio-temporal model (spBayes)
- Bayesian autoregressive spatio-temporal model (spTimer)
- Bayesian autoregressive spatio-temporal model (inla)

Goal: which model is the best for predicting PM2.5?

Methods - Spatio-Temporal Model

A general spatio-temporal model is

$$y_t(s_i) = x_t'(s_i)\beta + w_t(s_i) + \epsilon_t(s_i)$$

$i = 1, \ldots, n$ is the number of spatial locations
$t = 1, \ldots, T$ is the number of observations (over time) at each location

Different distributions can be assigned to $w_t(s_i)$ and $\epsilon_t(s_i)$, which produces different models.

Methods - Dynamic Linear Model

Bayesian dynamic linear model implemented through the `spBayes` package via the `spDynLM` function:

$$
\begin{aligned}
y_t(s) &= x_t(s)'\beta_t + u_t(s) + \epsilon_t(s), t = 1, 2 \ldots, N_t \\
\epsilon_t(s) &\sim N(0, \tau_t^2) \\
\beta_t &= \beta_{t-1} + \eta_t \\
\eta_t &\sim N(0, \Sigma_\eta) \\
u_t(s) &= u_{t-1}(s) + w_t(s) \\
w_t(s) &\sim GP(0, C_t(\cdot, \theta_t)),
\end{aligned}
$$

$C_t(\cdot, \theta_t)$ here is an Exponential covariance function with parameters $\theta_t = \{\sigma_t^2, \phi_t\}$, where:
σ_t^2 is spatial variance
ϕ_t is the correlation decay parameter
We also let $\beta_0 \sim N(m_0, \Sigma_0)$ and $u_0(s) = 0$.

Methods - AR Model

Bayesian hierarchical autoregressive (AR) model used in the `spTimer` package:

$$\mathbf{y}_t = \mathbf{O}_t + \boldsymbol{\epsilon}_t$$
$$\mathbf{O}_t = \rho \mathbf{O}_{t-1} + \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\eta}_t$$

where $\rho \in (-1, 1)$ and

$$\boldsymbol{\epsilon}_t \sim N(0, \sigma_\epsilon^2 \mathbf{I}_n)$$
$$\boldsymbol{\eta}_t \sim N(0, \Sigma_\eta),$$

where $\Sigma_\eta = \sigma_\eta^2 S_\eta$, and S_η is the spatial correlation matrix obtained from the Matérn correlation function, which has parameters ϕ (range) and ν (order).
Priors: $\boldsymbol{\beta} \sim N(0, 10^{10})$, $1/\sigma^2 \sim \text{Gamma}(2, 1)$ (the default), and $\rho \sim N(0, 10^{10})$.

Methods - Spatial and Temporal Performance

Data: PM2.5 in Texas from January 1st - January 12th, 2023.
For each model, spatial and temporal predictive performance is evaluated:

Spatial Prediction

- Replace 10 random sites on a random day with NA's.
- Compute RMSE by comparing predictions with observed values.
- Repeat 10 times, take the mean as the overall spatial RMSE for each model.

Temporal Prediction

- One-step-ahead forecasts on an expanding window.
- Train Jan 1-5 and forecast Jan 6
  $\rightarrow$ Train Jan 2-6 and forecast Jan 7
  $\rightarrow \ldots$
  $\rightarrow$ Train Jan 7-11 and forecast Jan 12
- Compute RMSE for each model by comparing forecasted values to observed values on each day.

Results - Forecasting Coverage

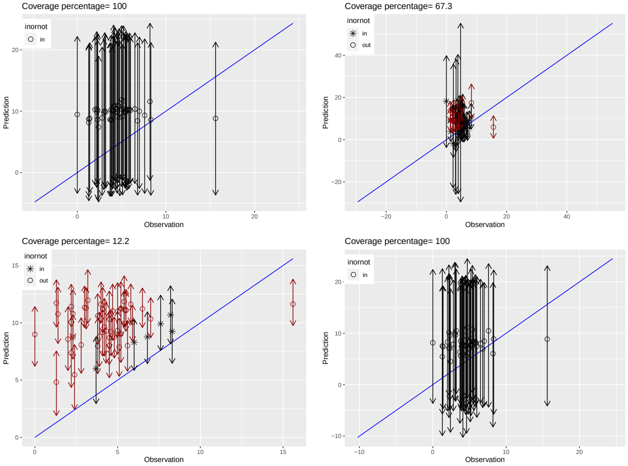


Figure: Prediction intervals and coverage for each model when forecasting Jan. 12th after training on Jan. 1st-11th, 2023. Plots are included for the linear model (top left), spBayes (top right), inla (bottom left), and spTimer (bottom right).

Results - Model RMSE

Spatial and temporal RMSPE from each model

Model	Spatial RMSPE	Temporal RMSPE
LM	5.18	5.04
inla	5.11	4.73
spBayes	7.96	9.83
spTimer	2.52	4.14

- the `spTimer` implementation of the AR model is the best
- the `spBayes` dynamic model needs more adjustment and tuning for this data (default settings are not adequate)

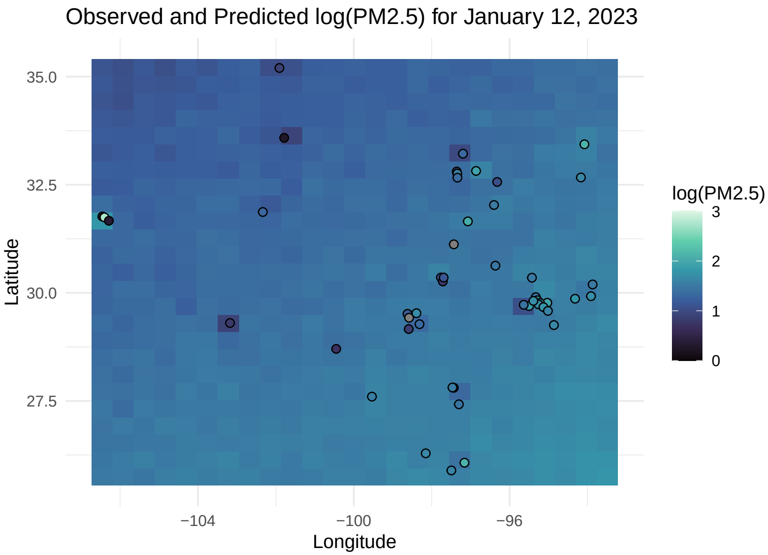
Results - Posterior Summary

Posterior estimates from the fitted `spTimer` Bayesian AR spatio-temporal model

Parameter	Mean	SD	2.5%	97.5%
Intercept	17.35	7.34	3.03	31.41
x (Longitude)	0.09	0.07	-0.05	0.22
y (Latitude)	-0.16	0.10	-0.38	0.04
ρ	0.48	0.02	0.43	0.52
σ_ϵ^2	2.55	0.63	1.42	3.93
σ_η^2	17.58	1.15	15.51	20.00
ϕ	0.08	0.01	0.06	0.12

- The temporal component is important $\rho = 0.48$
- Longitude and Latitude are not significant predictors of PM2.5

Results - Forecasted PM2.5 Map



Conclusion

Results

- Different models were compared for spatial and temporal prediction
- The `spTimer` implementation of an AR model was found to be the best
- An accurate (forecasted) prediction map of PM2.5 across Texas was produced

Analysis Limitations

- Only default settings and prior choices were used from each package
- No sensitivity analysis of prior choices
- Additional variables could be included for a multivariate model (such as CO, SO2, Pb, etc.)
- Most sensors are in cities, which have different PM2.5 than rural areas

Sources - PM2.5 and Data

Data: https://www.epa.gov/outdoor-air-quality-data/download-daily-data

Effects of PM on health: https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm

References - Packages

`spTimer`

Bakar KS, Sahu SK (2015). "spTimer: Spatio-Temporal Bayesian Modeling Using R." Journal of Statistical Software, 63(15), 1–32. https://doi.org/10.18637/jss.v063.i15.

Bakar KS, Sahu SK (2022). Spatio-Temporal Bayesian Modeling. R package version 3.3.2.

`bmstdr`

Sahu, S. (2022). Bayesian Modeling of Spatio-Temporal Data with R (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429318443

`spBayes`

Finley AO, Banerjee S, E.Gelfand A (2015). "spBayes for Large Univariate and Multivariate Point-Referenced Spatio-Temporal Data Models." Journal of Statistical Software, 63(13), 1–28. https://www.jstatsoft.org/article/view/v063i13.

`inla`

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society, Series B, 71(2):319392, 2009.