Spatio-Temporal Modeling of Pandemics

Nick Clark, United States Military Academy -
Jorge Mateu, Universitat Jaume I -
The Devil’s in the dependency!
Basic Statistical Model

- \( s_i \) - Spatial vector, usually location in \( \mathbb{R}^2 \)
- \( t \) - time
- \( Z(s_i, t) \) - Number of events observed at spatio-temporal location \( s_i \times t \)

\[
Z(s_i, t) \sim Po(\lambda(s_i, t))
\]

\[
\log(\lambda(s_i, t)) = \beta_0 + \sum_{j=1}^{n} x(s_i,t,j) \beta_j
\]

- All spatial-temporal correlation is captured in the large structure covariates
Castilla y Leon Confirmed COVID cases
Spatial Only Model with Large Scale Effects

\[
Z(s_i) \sim Po(\lambda(s_i))
\]

\[
\log(\lambda(s_i)) = \beta_0 + \log(Pop_{s_i}) + \beta_{Urban}x_{s_i}
\]

\[
\beta_0, \beta_{Urban} \sim N(0, 10)
\]

- \(E[\lambda|Z] \text{ vs } Z\)
Posterior Predictive Checks

- One goal of statistical modeling is to capture key elements of a scientific mechanism in small number of parameters.

- Predictive distribution $p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta$

- Posterior predictive checks compare key elements of original data with key elements from generated data.
Capturing small scale spatial effects

\[ Z(s_i) \sim Po(\lambda(s_i)) \]

\[ \log(\lambda(s_i)) = \beta_0 + \sum_{j=1}^{n} x_{(s_i,j)} \beta_j + \phi(s_i) \]

\[ \phi \sim MVN(0, \Sigma(\theta)) \]

- Structure of \( \Sigma(\theta)^{-1} \) yields conditional spatial dependence

\[ \phi(s_i) | \phi(s_j) \sim Ga u\left( \frac{1}{N|s_i|} \sum_{s_j \in N|s_i|} \phi(s_j), \sigma^2 \right) \]

- Precision Matrix only has entries where neighbors exist
Results

- In practice used BYM model (convolves spatial ICAR with heterogeneous RE)
- Even done efficiently 100 times slower, but appears to replicate data well
Alternatively INLA

- Compute marginals of hyper-parameters using Laplace approximation
- Compute conditional distribution of parameters given hyper-parameters and data
- Numerically integrate out hyperparameters
- Efficiently explore parameter space

```r
inla.formula <- y ~ 1 + pop + hzone + f(county, model = "bym2",
                                graph = q.mat, constr = TRUE)

model <- inla(inla.formula, family = "poisson", data = inla.df,
              control.compute = list(dic = TRUE, cpo = TRUE))
```

- 12 seconds to run vs 250 seconds per chain in stan
Is this overkill?

- Moran’s I fails to reject spatial randomness
- Model with only heterogeneous error
Need to keep in mind dynamic of virus

Hot Spots Emerge and Dissipate
Histogram of Moran’s I

- Not Uniformly distributed
Spatial-Temporal Models

- The spatial structure changes as time changes
- Challenge is how to structure model
  \[ \eta \equiv \log(\lambda(s_i, t)) = \mu(s_i, t) + \epsilon(s_i) + \gamma(t) + \kappa(s_i, t) + \delta(s_i, t) \]
- One option in separability
  \[ \Sigma_\kappa(\theta) = \Sigma_s \otimes \Sigma_t \]
- PDE Motivated Approach
  \[ \frac{\partial \eta}{\partial t} = \beta \frac{\partial^2 \eta}{\partial s} - \alpha \eta \]
  \[ \eta_t = M \eta_{t-1} + \psi_t \]
Fitting Spatio-Temporal Models

- INLA (to me) is more straightforward
- 223 time points, 247 spatial locations
- 177 minutes to fit with $\kappa(s_i, t) \sim MVN(0, I)$, 266 to fit with $\Sigma_\kappa = \Sigma_{BYM} \otimes \Sigma_{RW1}$
- DIC prefers simpler model
- Forecasting done as missing data
Data Driven Processes - Moving from Latent Gaussian

- Structure placed on $\lambda(s_i, t)$ instead of $\log(\lambda(s_i, t))$
- Convolution of latent spatial and explicit temporal
- Can no longer fit in INLA

$$\lambda(s_i, t) = \kappa Z(s_i, t - 1) + \exp(\mu(s_i, t) + \phi(s_i))$$
Data Driven Spatial Model with Hurdle

\[
Pr(Z(s_i, t) = 0) = \pi(s_i, t)
\]
\[
Pr(Z(s_i, t) > 0) = (1 - \pi(s_i, t))Po(Z(s_i, t | \lambda(s_i, t))1_{Z(s_i,t)>0}
\]
\[
\text{logit}(\pi(s_i, t)) = \beta_0 + \beta_1 x_{Pop(s_i)}
\]
\[
\text{log}(\lambda(s_i, t)) = \beta_0 + \beta_1 x_{Dow(t)} + \phi(s_i)
\]
Some preliminary results

ECDF, Maximum Value, Percent of Zeros
Interesting Lines of Research

- How do we capture longer temporal dynamics?
- What are practical differences between data driven processes and latent Gaussian driven processes?
- How should immune population be factored in? Cases divided by susceptible?
- How do we disentangle mobility from response variable?
Closing Thoughts

- Spatio-Temporal structure is primarily needed when large scale covariates fail to capture structure in data
- Things that are close together in time/space behave similarly
- Structure of covariance/precision matrix is necessary for computational reasons
- An appropriate statistical model should be able to replicate key characteristics of data
Some good resources

- “Statistics for Spatio-Temporal Data” - Cressie and Wikle
- “Spatial and Spatio-temporal Bayesian Models with R - INLA” - Blangiardo
- “Statistics for Spatial Data” - Cressie
- “Spatial and Spatio-Temporal Geostatistical Modeling and Kriging” - Montero, Fernandez-Aviles, Mateu