Final report

Shuyi

12/07/2020

Data(Covid-19)

Intro

Covid-19 dataset is a typical sptiotemporal data. We could only use the number of infected cases as our response y and we do statewise analysis, which means we record the the number of cases for each state and each day. If we write in formula it will be $y(x,t)$, where x is different state and t is date. The situation is getting more and more severe, which makes it necessary to analyse via the appropriate model and see if there is any way we could do to decrease its spread rate. Moreover, if we could obtain the correlation of different states we are likely to implement corresponding action to reduce the infected cases due to states dissemination. To summarize, we'd liek to first try to model this spatiotemporal process and study 2 issues:
Firstly, if we could capture the trend which means whether we could predict correctly w.r.t the response # of infected cases.
Secondly, if we could capture temporal evalution of spatio dependence(TESD), means explore the how spatial dependence changes with time goes by for the number of infected case of covid-19 dataset.

Preprocess

Let's take a look at the initial dataset(just a few first columns):

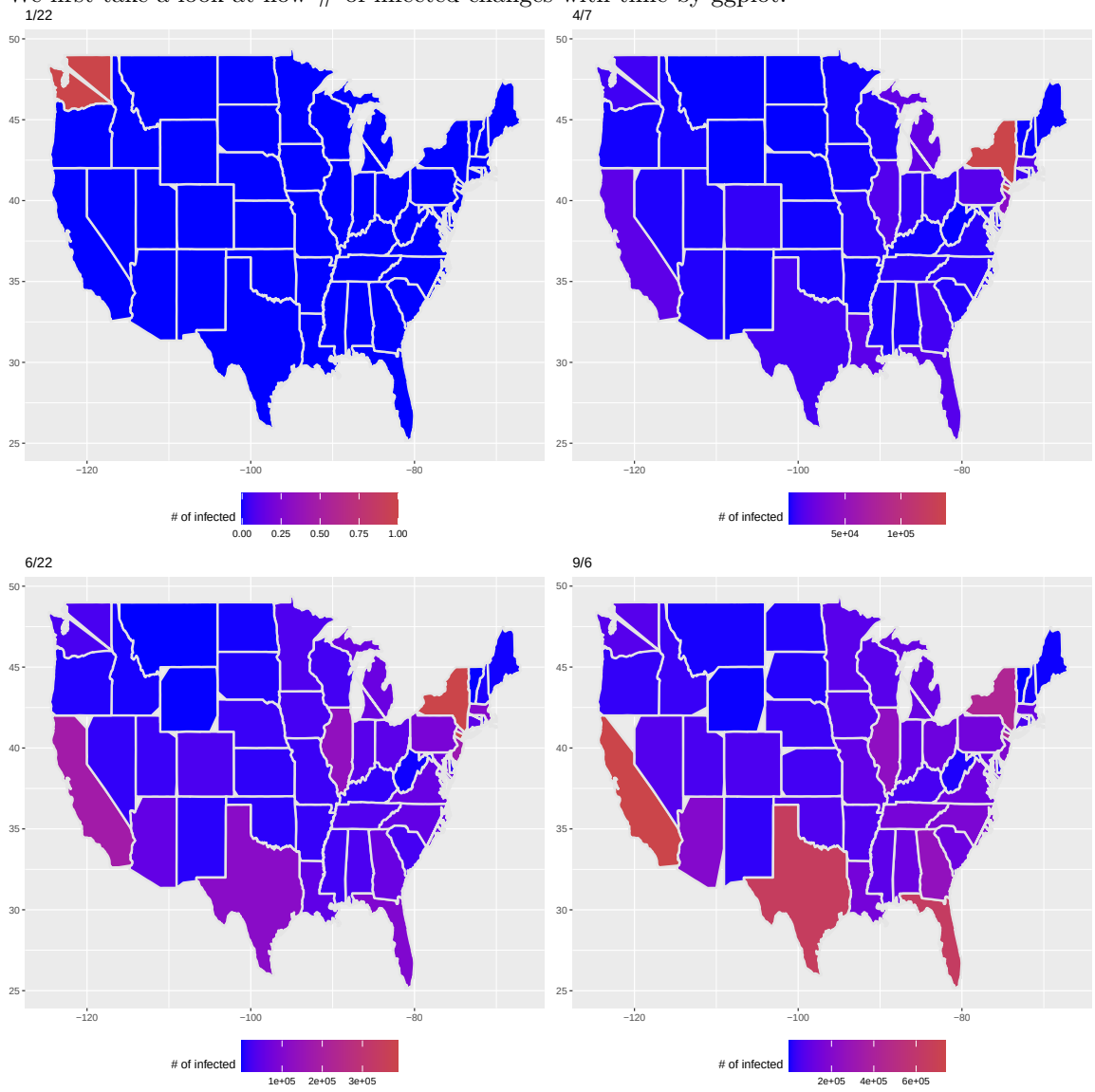
```
##    Admin2 Province_State Country_Region      Lat     Long_
## 1 Autauga        Alabama             US 32.53953 -86.64408
## 2 Baldwin        Alabama             US 30.72775 -87.72207
## 3 Barbour        Alabama             US 31.86826 -85.38713
## 4    Bibb        Alabama             US 32.99642 -87.12511
##           Combined_Key X1.22.20 X1.23.20
## 1 Autauga, Alabama, US        0        0
## 2 Baldwin, Alabama, US        0        0
## 3 Barbour, Alabama, US        0        0
## 4    Bibb, Alabama, US        0        0
```

We firstly preprocess our covid-19 dataset into state-wise, specifically we sum up infected case in each city on each day t within the same state k as the y_{kt}. Then we have y with dimension $49 * 229$, where 49 is the number of states and 229 is number of dates from 1/22 to 9/6. Also we extract the coordinate of each state as x like the table below, which will be used in second model(g-TGP).

```
##   Province_State       Lat      Long_
## 1        Alabama  32.88428  -86.71012
## 2         Alaska  60.24751 -148.03691
## 3 American Samoa -14.27100 -170.13200
## 4        Arizona  33.67590 -111.46323
```

EDA

We first take a look at how # of infected changes with time by ggplot:



Spatiotemporal model

Spatio-temporal structure with a multivariate first order autoregressive process(ST.CARar)

We fit a generalized linear mixed model to our data, whose general form is:

$$
\begin{aligned}
Y_{kt}|\mu_{kt} &\sim f(y_{kt}|\mu_{kt},\nu^2) \quad \text{for } k=1,\ldots,K, \quad t=1,\ldots,N,\\
g(\mu_{kt}) &= \mathbf{x}_{kt}^{\top}\boldsymbol{\beta} + O_{kt} + \psi_{kt},\\
\boldsymbol{\beta} &\sim \mathrm{N}(\boldsymbol{\mu}_{\beta}, \boldsymbol{\Sigma}_{\beta}).
\end{aligned}
$$

where $Y_{kt} \sim Poisson(\mu_{kt})$, covariate $\mathbf{x}_{kt}$ is a column of ones for the intercept term $\beta = \beta_0$ in this case since we don't have covariate, ψ_{kt} term is a latent component for areal unit k and time period t encompassing one or more sets of spatio-temporally autocorrelated random effects. We use the following setting for ψ_{kt}:

$$
\begin{aligned}
\psi_{kt} &= \phi_{kt},\\
\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1} &\sim \mathrm{N}\left(\rho_T\boldsymbol{\phi}_{t-1}, \tau^2\mathbf{Q}(\mathbf{W},\rho_S)^{-1}\right) \qquad t=2,\ldots,N,\\
\boldsymbol{\phi}_1 &\sim \mathrm{N}\left(\mathbf{0}, \tau^2\mathbf{Q}(\mathbf{W},\rho_S)^{-1}\right),\\
\tau^2 &\sim \text{Inverse-Gamma}(a,b),\\
\rho_S,\rho_T &\sim \text{Uniform}(0,1).
\end{aligned}
$$

This represents the spatio-temporal structure with a multivariate first order autoregressive process with a spatially correlated precision matrix W. Typically, W is assumed to be binary, where $w_{kj} = 1$ if areal units (S_k,S_j) share a common border (i.e., are spatially close) and is zero otherwise. From the model specification, we notice that we can't capture the TESD since y_{kt} is assumed to be independent and with identical observation variance v^2.
So let's compute the proximity (adjacency) matrix W of US states.

```
usa.state=map(database="state", fill=TRUE, plot=FALSE)
state.ID <- sapply(strsplit(usa.state$names, ":"), function(x) x[1])
usa.poly = map2SpatialPolygons(usa.state, IDs=state.ID)
usa.nb = poly2nb(usa.poly)
usa.adj.mat = nb2mat(usa.nb, style="B")
```

Generalised Spatiotemporal Gaussian Process(gSTGP)

The generalized STGP models can be summarized in the following unified form:

$$
\begin{aligned}
& y(\mathbf{z}) \sim \mathcal{GP}(m, \mathcal{C}_{y|m}), \quad m(\mathbf{z}) \sim \mathcal{GP}(0, \mathcal{C}_m)\\
\text{model } O: \quad & \mathcal{C}_{y|m} = \sigma_\epsilon^2 \mathcal{I}_{\mathbf{x}} \otimes \mathcal{I}_t, \quad \mathcal{C}_m = \mathcal{C}_{\mathbf{x}} \otimes \mathcal{C}_t\\
\text{model } I: \quad & \mathcal{C}_{y|m} = \sigma_\epsilon^2 \mathcal{I}_{\mathbf{x}} \otimes \mathcal{I}_t, \quad \mathcal{C}_m = \mathcal{C}_{\mathbf{x}|t} \otimes \mathcal{C}_t\\
\text{model } II: \quad & \mathcal{C}_{y|m} = \mathcal{C}_{\mathbf{x}|t} \otimes \mathcal{I}_t, \quad \mathcal{C}_m = \mathcal{I}_{\mathbf{x}} \otimes \mathcal{C}_t
\end{aligned}
\tag{1}
$$

We intend to use model II on our dataset since it not only requires less computation time but also captures TESD from the experiment.

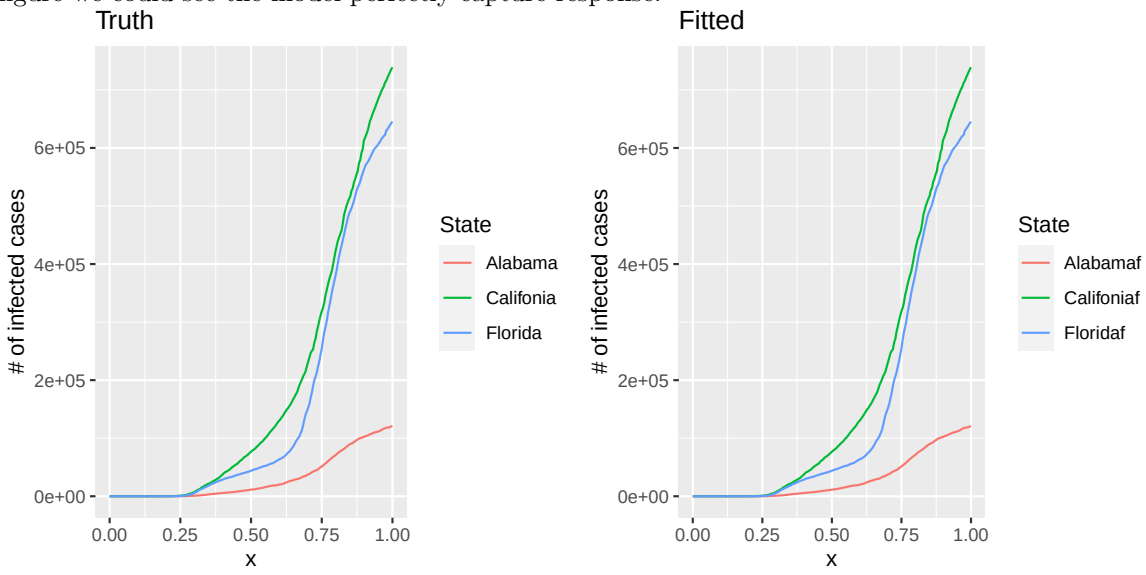
Result

ST.CARar model

We fit model as follows:

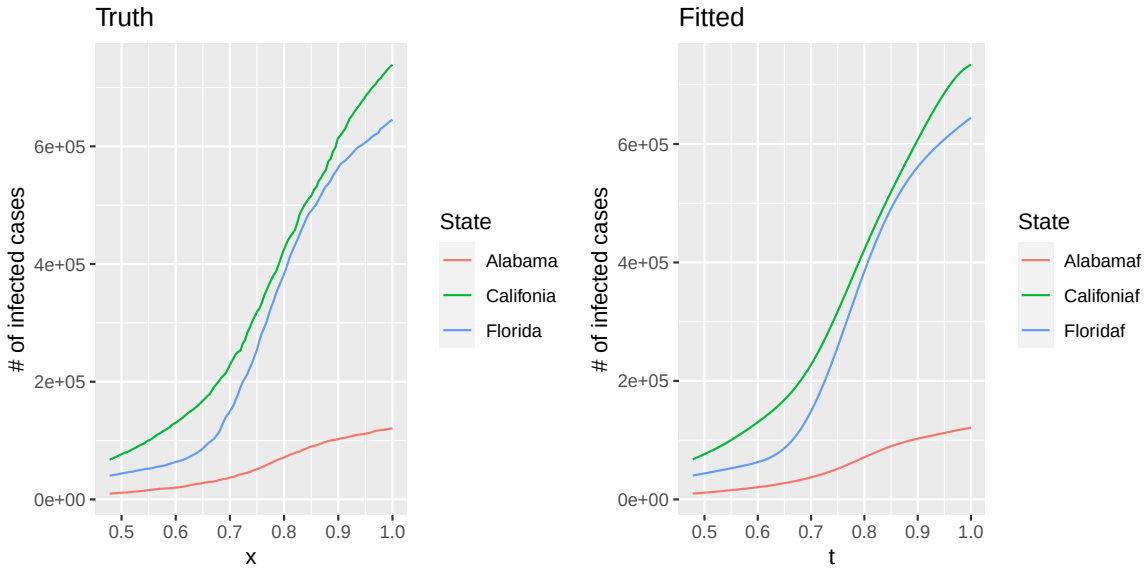
```
stcar = ST.CARar(formula=y[,3]~1, family="poisson",  trials=NULL, W=usa.adj.mat, burnin=10000, n.sam
```

Then we plot fitted case versus truth for selected 3 state: Alabama, Califonia, Florida. From the figure we could see the model perfectly capture response.

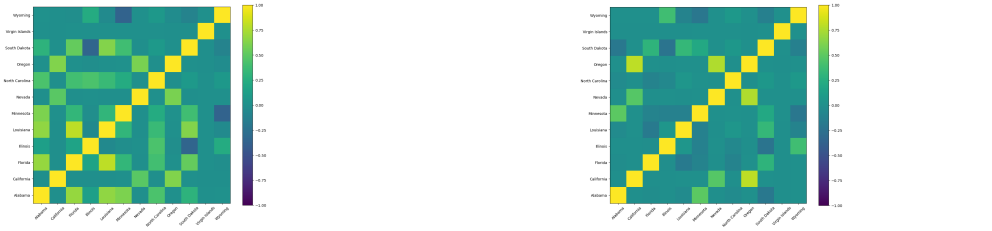


gSTGP

From the configuration of this model, we could capture both mean and covariance of the response. Limited by computation and some high-level issues, we run from May 10th to Sep. 6th. We first take a look at mean estimation:



The gSTGP also capture the mean very well. Another advantage for this model is it could also characterise covariance structure of our response, to be more specific, TESD we are interested. Below are the results:



(a) When t = May 10th

(b) When t = Sep. 6th

Figure 1: TESD of selected 12 states

We could see that the initial covariance is messy, but with time goes by and more data involved, it could capture the correlation like Califonia is highly correlated with Nevada and Oregon, considering they are physically close.

Future

For the covid-19 data, we aim to modify model a little bit by combining log-Gaussian Cox Processes(LGCP) with gSTGP due to non-negative discrete response Y. Cox process is inhomogeneous Poisson process with stochastic intensity λ written in the following formula:

$$Y(\mathbf{x}_i, t_j) \sim \text{Pois}(\lambda(\mathbf{x}_i, t_j)), \quad \lambda(\mathbf{z}) = \exp[f(\mathbf{z})] or \mu(x) * \beta(t) \exp[f(\mathbf{z})]$$
$$f(\mathbf{z}) \sim \text{gSTGP}(0, \mathcal{C}_\mathbf{z})$$

where f is a generalized STGP prior to capture the temporal change of the geographical dependence among 49 states .