Supplementary Materials for A Spatial Time-to-Event Approach for Estimating Associations between Air Pollution and Preterm Birth

Howard H. Chang

Department of Biostatistics and Bioinformatics
Emory University, Atlanta GA, USA
howard.chang@emory.edu

Brian J. Reich

Department of Statistics
North Carolina State University, Raleigh NC, USA

Marie Lynn Miranda

School of Natural Resources and Environment
University of Michigan, Ann Arbor MI, USA

1 Model for discrete-time spatially-referenced survival data

For subject $i = 1, ..., n$, we observe the follow-up time t_i, an indicator of whether the subject was censored c_i, and the spatial location s_i. We also observe potentially time-dependent covariates $\mathbf{X}_i(t) = (1, X_{1i}(t), ..., X_{pi}(t))'$, where $\mathbf{X}_i(t)$ is the $(p+1)$-vector of covariates for subject i at time t and the first element of $\mathbf{X}_i(t)$ is reserved for the intercept. We assume discrete domains for both the spatial locations and event times, that is, $s_i \in \{1, ..., n_s\}$ and $t_i \in \{1, ..., n_t\}$.

We model the survival probability as

$$P(t_i > m) = \prod_{t=1}^{m} [1 - \pi(\mathbf{X}_i(t), s_i, t)] \tag{1}$$

$$\pi(\mathbf{X}_i(t), s_i, t) = \Phi\left[\mathbf{X}_i(t)'\boldsymbol{\beta}(s_i, t)\right], \tag{2}$$

where Φ is the standard normal distribution function and $\boldsymbol{\beta}(s,t) = [\beta_0(s,t), ..., \beta_p(s,t)]'$ is a vector of regression coefficients for location s and time t.

Let $\beta_0(s,t) = \eta_0 + \mu_0(s) + \gamma_0(t) + \theta_0(s,t)$, where η_0 is the overall average; $\mu_0(s)$ is a spatial effect; $\gamma_0(t)$ is a temporal effect; and $\theta_0(s,t)$ is the space/time interaction. We temporarily drop the subscript j, and note that the covariate effects $\beta_j(s,t)$ for $j = 1, ..., p$ are modeled similarly below.

The spatial terms $\boldsymbol{\mu} = [\mu(1), ..., \mu(S)]'$ are modeled using the conditionally autoregressive model (CAR). Let $s \sim s'$ indicate that regions s and s' are spatial neighbors and m_s be the number of spatial neighbors of region s. The full conditional distribution is Gaussian with

$$E\left[\mu(s)|\mu(s'), s' \neq s\right] = \rho_\mu \sum_{s' \sim s} \mu(s')/m_s \tag{3}$$

$$V\left[\mu(s)|\mu(s'), s' \neq s\right] = \sigma_\mu^2/m_s. \tag{4}$$

The joint model for the vector $\boldsymbol{\mu}$ is multivariate normal with mean zero and covariance $\sigma_\mu^2 \left[M_S - \rho_\mu C_S\right]^{-1}$, where the (s, s') element of C_S is $C_S(s, s') = I(s \sim s')$ and M_S is diagonal with diagonal elements $\sum_{s' \neq s} C_S(s, s') = m_s$. We denote this model as $\boldsymbol{\mu} \sim \text{CAR}(\rho_\mu, \sigma_\mu^2, C_S)$.

The temporal effects $\boldsymbol{\gamma} = [\gamma(1), ..., \gamma(T)]'$ control the spatial average baseline hazard function. The vector $\boldsymbol{\gamma}$ has a lag-1 autoregressive model which can be written $\boldsymbol{\gamma} \sim \text{CAR}(\rho_\gamma, \sigma_\gamma^2, C_T)$, where C_T is the $n_t \times n_t$ temporal adjacency matrix with (t, t') element equal to $I(|t - t'| = 1)$. The spatio-temporal random effects have the dynamic spatial model (Banerjee *et al.*, 2003),

$$\theta(s, t) = \rho_\theta \theta(s, t-1) + \delta(s, t) \tag{5}$$

where $\rho_\theta \in (0, 1)$ and $\boldsymbol{\delta}_t = [\delta(1, t), ..., \delta(S, t)]' \sim \text{CAR}(\rho_\delta, \sigma_\delta^2, C_S)$. For identification purposes we fix $\theta(s, 1) = 0$ for all s.

1.1 Priors

To complete the Bayesian model, we specify priors for the remaining parameters. Let $\eta_j \sim \text{N}(0, c_j^2)$. To give vague priors for these overall averages we take $c_j = 100^2$. The variances $\sigma_{\mu j}^2$, $\sigma_{\gamma j}^2$, and $\theta_{\theta j}^2 \sim \text{Gamma}(a_1, b_1)$. Following Kelsall *et al.* (1999), we take $a_1 = 0.5$ and $b_1 = 0.005$. The CAR association parameters $\rho_{\mu j}$, $\rho_{\gamma j}$, $\rho_{\theta j}$, $\rho_{\delta j} \sim \text{Beta}(a_2, b_2)$. To facilitate MCMC sampling we discretize the prior to 1000 equally-spaced points spanning [0,1], and to give an uninformative prior we take $a_2 = b_2 = 1$.

1.2 Markov Chain Monte Carlo Algorithm

For notational convenience, in the description of the full conditionals let $\boldsymbol{\beta}^*(s_i, t)$ denote the coefficient vector $\boldsymbol{\beta}(s_i, t)$ with the element under consideration set to zero. For example, in the description of the full conditional for η_j, denote $\beta_j^*(s, t) = \mu_j(s) + \gamma_j(t) + \theta_j(s, t)$. Then $r_i(t) = Z_i(t) - \mathbf{X}_i(t)'\boldsymbol{\beta}^*(s_i, t)$ is the residual calculated without the variable under consideration. We initialize the MCMC algorithm by setting $\beta_j(s, t) = 0$ for all s and t and all CAR covariance parameters equal to one. Sampling then proceeds by repeatedly sampling each parameter conditioned on all others in the following steps.

1. $Z_i(t)|\text{rest} \sim \text{N}_A(\mathbf{X}_i(t)'\boldsymbol{\beta}(s_i, t), 1)$, where $N_A(\mu, \sigma^2)$ is the truncated normal distribution with domain A, location μ, and scale σ. For this probit model $A = (\infty, 0)$ if $Y_i(t) = 0$, and $A = (0, \infty)$ if $Y_i(t) = 1$.

2. $\eta_j|\text{ rest} \sim \text{N}\left(\frac{\sum_{i=1}^n \sum_{t=1}^{t_i} X_{ji} r_i(t)}{\sum_{i=1}^n \sum_{t=1}^{t_i} X_{ji}^2 + 1/c_j^2}, \frac{1}{\sum_{i=1}^n \sum_{t=1}^{t_i} X_{ji}^2 + 1/c_j^2}\right)$.

3. $\boldsymbol{\mu}_j|$ rest $\sim \mathrm{N}\Big([P+Q]^{-1}R, [P+Q]^{-1}\Big)$, where R is the vector with element s equal to $\sum_{i|s_i=s}\sum_{t=1}^{t_i} X_{ji}r_i(t)$, $Q = (M_S - \rho_{\mu j}C_S)/\sigma^2_{\mu j}$, and P is diagonal with s^{th} diagonal element equal to $\sum_{i|s_i=s}\sum_{t=1}^{t_i} X^2_{ji}$.

4. $\boldsymbol{\gamma}_j|$ rest $\sim \mathrm{N}\Big([P+Q]^{-1}R, [P+Q]^{-1}\Big)$, where R is the vector with element s equal to $\sum_{i|t_i\geq t} X_{ji}r_i(t)$, $Q = (M_T - \rho_{\gamma j}C_T)/\sigma^2_{\gamma j}$, and P is diagonal with s^{th} diagonal element equal to $\sum_{i|t_i\geq t} X^2_{ji}$.

5. $\theta_j(,t)|$ rest $\sim \mathrm{N}\Big(\left[P+(1+\rho^2_{\theta_j})Q\right]^{-1}\left[R+\rho_{\theta_j}\theta_j(,t-1)+\rho_{\theta_j}\theta_j(,t+1)\right], \left[P+(1+\rho^2_{\theta_j})Q\right]^{-1}\Big)$ where R is the vector with element s equal to $\sum_{i|s_i=s,t_i\geq t} X_{ji}r_i(t)$, $Q = (M_T - \rho_{\delta j}C_T)/\sigma^2_{\delta j}$, and P is diagonal with s^{th} diagonal element equal to $\sum_{i|s_i=s,t_i\geq t} X^2_{ji}$

6. $\sigma^2_{\mu j}|$ rest $\sim \mathrm{InvGamma}(n_s/2 + a_1, \boldsymbol{\mu}'_j(M_S - \rho_{\mu j}C_S)\boldsymbol{\mu}_j/2 + b_1)$

7. $\sigma^2_{\gamma j}|$ rest $\sim \mathrm{InvGamma}(n_t/2 + a_1, \boldsymbol{\gamma}'_j(M_T - \rho_{\gamma j}C_T)\boldsymbol{\gamma}_j/2 + b_1)$

8. $\sigma^2_{\delta j}|$ rest $\sim \mathrm{InvGamma}((nt-1)n_s/2+a_1, \sum_{t=2}^{n_t}(\theta_j(,t)-\rho_{\theta j}\theta_j(,t-1))'(M_S-\rho_{\delta j}C_S)(\theta_j(,t)-\rho_{\theta j}\theta_j(,t-1))/2 + b_1)$

Finally the association parameters $\rho_{\mu j}$, $\rho_{\gamma j}$, $\rho_{\theta j}$, and $\rho_{\delta j}$, have discrete priors, and thus discrete full conditionals. The full conditional probabilities are proportional to the product of the prior and the appropriate CAR density.

References

Banerjee, S., Carlin, B., Gelfand, A. (2003). Hierarchical modeling and analysis for spatial data. Chapman & Hall.

Kelsall, J. E., Wakefield, J. C., Bernado, J. M., Berger, J. O., Dawid, A. P., Smith, A. (ed.) (1999). Comment on Bayesian model for spatially correlated disease and exposure data in Bayesian Statistics 6 - Proceedings of the Sixth Valencia International Meeting Oxford *University Press*

2 Supplementary Application Results

Table 1: Deviance information criterion (DIC), effective degrees of freedom (p_D), and posterior predictive loss (PPD) for different preterm birth baseline hazard models and $PM_{2.5}$ exposure metrics.

	Cumulative			4-week lag		
Baseline Hazard	DIC	pD	PPD	DIC	pD	PPD
Non-spatial	44178	35.5	8429	44187	36.9	8423
Spatial frailty	44179	54.2	8429	44187	54.5	8429
Space-time interaction	44180	72.5	8424	44190	74.2	8433

Table 2: Posterior median and 95% posterior intervals (P.I.) of the baseline hazard parameters. Estimates are from a model that includes space-time interaction baseline hazards and average $PM_{2.5}$ levels over the entire pregnancy.

Parameter	Posterior median (95% P.I.)	Parameter	Posterior median (95% P.I.)
ρ_μ	0.96 (0.73, 1.00)	σ_μ	0.21 (0.13, 0.42)
ρ_γ	0.40 (0.02, 0.93)	σ_γ	0.06 (0.04, 0.10)
ρ_δ	0.39 (0.02, 0.94)	σ_δ	0.06 (0.03, 0.10)
ρ_θ	0.24 (0.01, 0.73)		

3 Supplementary Simulation Results

Table 3: Simulation study results: root mean-squared error (RMSE) and 95% confidence interval (C.I.) length based on 1000 simulated replicate datasets.

	RMSE ($\times$100)				95% C.I. length			
	Cumulative		4-week lag		Cumulative		4-week lag	
Relative Risk	Survival	Probit	Survival	Probit	Survival	Probit	Survival	Probit
1.00	0.58	0.83	0.34	0.57	0.96	0.94	0.96	0.90
1.01	0.62	0.94	0.35	0.59	0.95	0.91	0.95	0.87
1.02	0.61	0.98	0.35	0.61	0.95	0.90	0.95	0.85
1.03	0.61	1.02	0.34	0.65	0.94	0.89	0.95	0.84
1.04	0.61	1.15	0.34	0.67	0.95	0.85	0.96	0.81
1.05	0.62	1.24	0.35	0.69	0.95	0.80	0.95	0.79

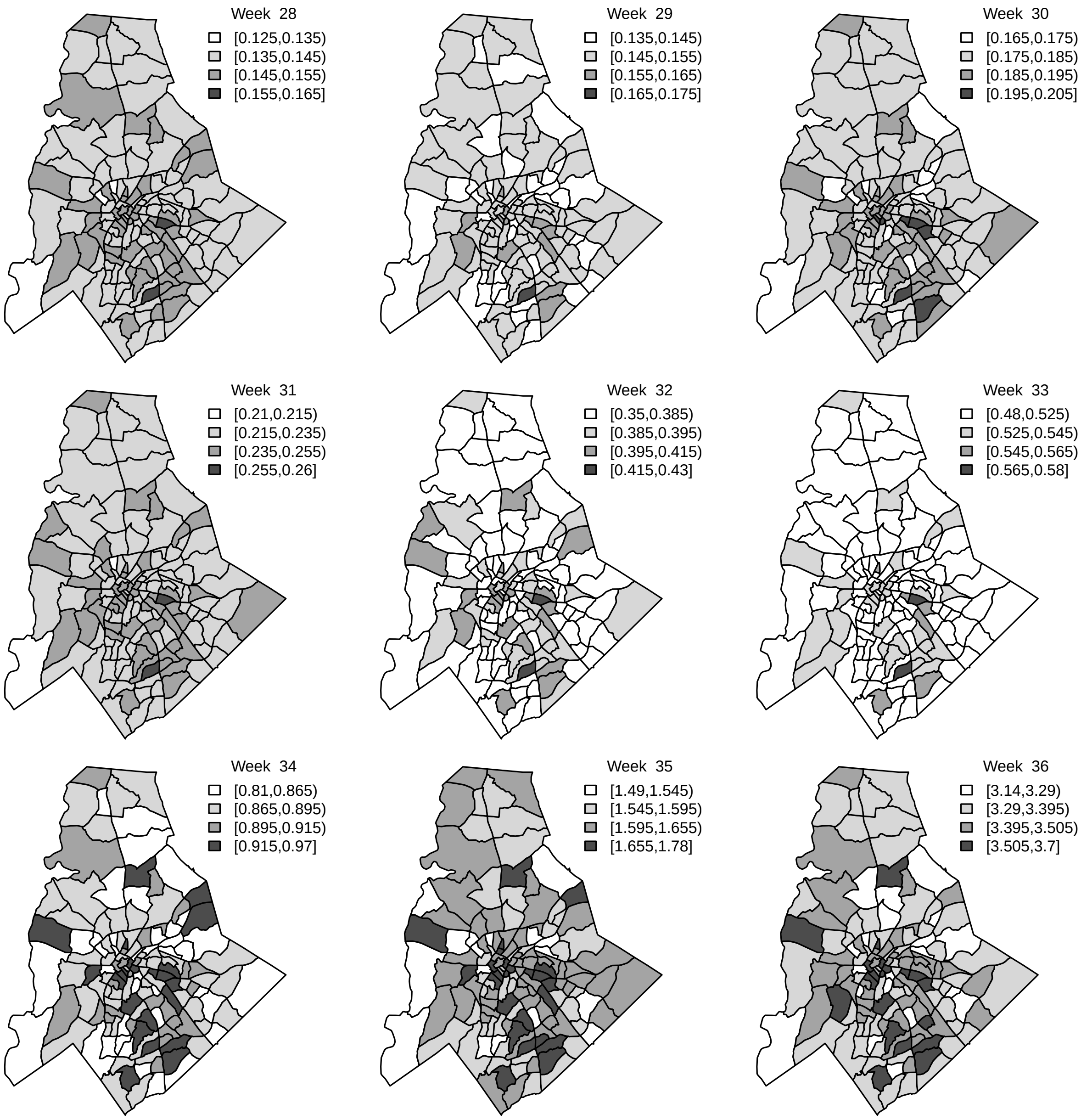

Figure 1: Baseline tract-specific hazard rates (%) of preterm birth. Baseline hazard rates are calculated at the average value of each covariate across all areas. Estimates are from a model that includes space-time interaction baseline hazards and average $PM_{2.5}$ levels over the entire pregnancy.

4 R Code

```
library(survival)
library(fields)

 #DATA:
    #Y[i,t]= 1 if sub i failed at time t,
    #      = 0 if sub i lived past time t
    #      = NA if sub i died before time t
    #x=covariates = n subs x ntimes x n covs
    #spatial location, s=1,...,ns (no regions without observations!)
    #The neighbor pairs are (np1[i],np2[i]),i=1,...,number of pairs

    #MODEL:
    #Y[i,t] = I(Z[i,t]>0)
    #Z[i,t] ~ N(x[i,t,]%*%(int+space[s[i],]+time[t,]+theta[t,s[i],]),1)
    #Space[,k]~CAR(tau1[k],rho1[k])
    #Time[,k]~CAR(tau2[k],rho2[k])
    #theta[,1,k]=0
    #theta[t,,k] ~ rho4*theta[t-1,,j]+CAR(tau3[k],rho3[k])

    #MODEL OPTIONS/HYPERPARAMETERS
    #spatial[k] = does beta_k have a spatial effect?
    #temporal[k] = does beta_k have a temporal effect?
    #spatiotemporal[k] = does beta_k have a spatiotemporal effect?
    #tau[k]~gamma(as,bs)
    #rho[k]~beta(ar,br)

probit.PH<-function(Y,x,s,np1,np2,
    spatial=rep(F,1000),temporal=rep(F,1000),spatiotemporal=rep(F,1000),
    sd.beta=100,
    as=0.5,bs=0.005,ar=1,br=1,
    runs=5000,burn=1000,update=10){

    #Set up data:
    n<-nrow(Y)
    nt<-ncol(Y)
    p<-dim(x)[3]
    ns<-max(c(np1,np2))
    O<-!is.na(Y)
    L<-ifelse(Y==0,-Inf,0)
    U<-ifelse(Y==0,0,Inf)
    for(k in 1:p){for(t in 1:nt){x[!O[,t],t,k]<-0}}

    #spatial adjacency matrix:
    ADJs<-matrix(0,ns,ns)
    for(j in 1:length(np1)){
      ADJs[np1[j],np2[j]]<-ADJs[np2[j],np1[j]]<-1
    }
    Ms<-diag(apply(ADJs,2,sum))

    #temporal adjacency matrix:
    ADJt<-matrix(0,nt,nt)
    for(j in 2:nt){
```

```
  ADJt[j,j-1]<-ADJt[j-1,j]<-1
}
Mt<-diag(apply(ADJt,2,sum))

#initial values
Z<-Y-0.5
Z[!O]<-0
int<-rep(0,p)
theta<-array(0,c(nt,ns,p))
space<-matrix(0,ns,p)
time<-matrix(0,nt,p)
tau1<-tau2<-tau3<-rep(1,p)
rho1<-rho2<-rho3<-rho4<-rep(0.9,p)

mn<-0*Z
for(t in 1:nt){
   mn[,t]<-x[,t,]%*%(int+time[t,])+
           apply(x[,t,]*(space[s,]+theta[t,s,]),1,sum)
}

#keep track of stuff:
keep.int<-matrix(0,runs,p)
beta.mn<-beta.var<-beta.pos<-0*theta
params<-matrix(0,runs,7*p)
#baseline.track <- array (0, c(nt, ns, runs) )
dev<-rep(0,runs)

dimnames(params)[[2]]<-c(paste("sd1[",1:p,"]",sep=""),
                         paste("sd2[",1:p,"]",sep=""),
                         paste("sd3[",1:p,"]",sep=""),
                         paste("rho1[",1:p,"]",sep=""),
                         paste("rho2[",1:p,"]",sep=""),
                         paste("rho3[",1:p,"]",sep=""),
                         paste("rho4[",1:p,"]",sep=""))

#set up prior for the CAR parameters
M2<-diag(1/sqrt(diag(Ms)))
ds<-eigen(M2%*%ADJs%*%M2)$values
M2<-diag(1/sqrt(diag(Mt)))
dt<-eigen(M2%*%ADJt%*%M2)$values
rm(M2)
nrho<-1000
canrho1<-detpart1<-qbeta(seq(0.001,0.999,length=nrho),ar,br)
canrho2<-detpart2<-qbeta(seq(0.001,0.999,length=nrho),ar,br)
canrho3<-detpart3<-qbeta(seq(0.001,0.999,length=nrho),ar,br)
canrho4<-qbeta(seq(0.001,0.999,length=nrho),ar,br)
for(j in 1:nrho){
  detpart1[j]<-0.5*sum(log(1-canrho1[j]*ds))
  detpart2[j]<-0.5*sum(log(1-canrho2[j]*dt))
  detpart3[j]<-0.5*(nt-1)*sum(log(1-canrho3[j]*ds))
}

#Save some loops:
x2<-x^2
```

```
x2forint<-apply(x2,3,sum)
x2forspace<-array(0,c(nt,ns,p))
for(k in 1:p){for(j in 1:nt){
  x2forspace[j,,k]<-tapply(x2[,j,k],s,sum)
}}
x2fortime<-apply(x2,2:3,sum)


#Start MCMC:
for(i in 1:runs){

  ####################################################
  ######         update the latent z's          #####
  ####################################################:
  for(j in 1:nt){
    Z[,j]<-rtnorm(mn[,j],L[,j],U[,j])
  }
  Z[!O]<-0

  ####################################################
  ######             update intercepts           #####
  ####################################################:
  for(k in 1:p){
     mn<-mn-x[,,k]*int[k]
     r<-Z-mn
     VVV<-x2forint[k]+1/sd.beta^2
     MMM<-sum(x[,,k]*r)
     int[k]<-rnorm(1,MMM/VVV,1/sqrt(VVV))
     mn<-mn+x[,,k]*int[k]
  }

  ####################################################
  ######     update spatial random effects      #####
  ####################################################:

  for(k in 1:p){if(spatial[k]){
    for(t in 1:nt){mn[,t]<-mn[,t]-x[,t,k]*space[s,k]}
    r<-Z-mn
    VVV<-tau1[k]*(Ms-rho1[k]*ADJs)+diag(apply(x2forspace[,,k],2,sum))
    MMM<-tapply(apply(x[,,k]*r,1,sum),s,sum)
    VVV<-solve(VVV)
    space[,k]<-VVV%*%MMM + t(chol(VVV))%*%rnorm(ns)
    for(t in 1:nt){mn[,t]<-mn[,t]+x[,t,k]*space[s,k]}

    #CAR covariance parameters:
    SS1<-t(space[,k])%*%Ms%*%space[,k]
    SS2<-t(space[,k])%*%ADJs%*%space[,k]
    tau1[k]<-rgamma(1,ns/2+as,(SS1-rho1[k]*SS2)/2+bs)
    R<-detpart1+0.5*tau1[k]*canrho1*SS2
    rho1[k]<-sample(canrho1,1,prob=exp(R-max(R)))
 }}

  ####################################################
  ######     update temporal random effects      #####
```

```
#####################################################:

for(k in 1:p){if(temporal[k]){
  for(t in 1:nt){mn[,t]<-mn[,t]-x[,t,k]*time[t,k]}
  r<-Z-mn
  VVV<-tau2[k]*(Mt-rho2[k]*ADJt)+diag(x2fortime[,k])
  MMM<-apply(x[,,k]*r,2,sum)
  VVV<-solve(VVV)
  time[,k]<-VVV%*%MMM + t(chol(VVV))%*%rnorm(nt)
  for(t in 1:nt){mn[,t]<-mn[,t]+x[,t,k]*time[t,k]}

  #CAR covariance parameters:
  SS1<-t(time[,k])%*%Mt%*%time[,k]
  SS2<-t(time[,k])%*%ADJt%*%time[,k]
  tau2[k]<-rgamma(1,nt/2+as,(SS1-rho2[k]*SS2)/2+bs)
  R<-detpart2+0.5*tau2[k]*canrho2*SS2
  rho2[k]<-sample(canrho2,1,prob=exp(R-max(R)))
}}

#####################################################
######   update spatiotemporal random effects  #####
#####################################################:
theta[1,,]<-0
for(k in 1:p){if(spatiotemporal[k]){
   for(t in 1:nt){mn[,t]<-mn[,t]-x[,t,k]*theta[t,s,k]}
   r<-Z-mn
   Q3<-tau3[k]*(Ms-rho3[k]*ADJs)
   for(j in 2:nt){
     #prior's contribution:
      VVV<-Q3
      MMM<-rho4[k]*Q3%*%theta[j-1,,k]
      if(j<nt){VVV<-VVV+rho4[k]*rho4[k]*Q3}
      if(j<nt){MMM<-MMM+rho4[k]*Q3%*%theta[j+1,,k]}

    #likelihood's contribution:
      diag(VVV)<-diag(VVV)+x2forspace[j,,k]
      MMM<-as.vector(MMM)+tapply(x[,j,k]*r[,j],s,sum)
      VVV<-solve(VVV)
      theta[j,,k]<-VVV%*%MMM + t(chol(VVV))%*%rnorm(ns)
  }
  for(t in 1:nt){mn[,t]<-mn[,t]+x[,t,k]*theta[t,s,k]}

  #CAR covariance parameters:
  SS1<-SS2<-rep(0,nt)
  for(j in 2:nt){
   ddd<-theta[j,,k]-rho4[k]*theta[j-1,,k]
   SS1[j]<-t(ddd)%*%Ms%*%ddd
   SS2[j]<-t(ddd)%*%ADJs%*%ddd
  }

  tau3[k]<-rgamma(1,ns*(nt-1)/2+as,sum(SS1-rho3[k]*SS2)/2+bs)
  R<-detpart3+0.5*tau3[k]*canrho3*sum(SS2)
  rho3[k]<-sample(canrho3,1,prob=exp(R-max(R)))
  Q3<-Ms-rho3[k]*ADJs
```

```
        SS3<-SS4<-rep(0,nt)
        for(j in 2:nt){
          SS3[j]<-t(theta[j,,k])%*%Q3%*%theta[j-1,,k]
          SS4[j]<-t(theta[j-1,,k])%*%Q3%*%theta[j-1,,k]
        }
        R<- -0.5*tau3[k]*(-2*canrho4*sum(SS3)+canrho4*canrho4*sum(SS4))
        rho4[k]<-sample(canrho4,1,prob=exp(R-max(R)))
      }}


     #keep track of stuff:
      beta<-theta
      for(k in 1:p){
        beta[,,k]<-beta[,,k]+int[k]
        for(j in 1:ns){beta[,j,k]<-beta[,j,k]+time[,k]}
        for(j in 1:nt){beta[j,,k]<-beta[j,,k]+space[,k]}
      }

        #baseline.track[,,i] <- beta[,,1]
      keep.int[i,]<-int
      dev[i]<-sum(-2*dbinom(Y[O],1,pnorm(mn[O]),log=T))
      params[i,]<-c(1/sqrt(tau1),1/sqrt(tau2),1/sqrt(tau3),rho1,rho2,rho3,rho4)

      if(i>burn){
          beta.mn<-beta.mn+beta/(runs-burn)
          beta.var<-beta.var+beta*beta/(runs-burn)
          beta.pos<-beta.pos+ifelse(beta>0,1,0)/(runs-burn)
      }


   }
   beta.var<-beta.var-beta.mn^2
   mn<-0*Z
   for(t in 1:nt){
     mn[,t]<-apply(x[,t,]*beta.mn[t,s,],1,sum)
   }
   dhat<-sum(-2*dbinom(Y[O],1,pnorm(mn[O]),log=T))
   dbar<-mean(dev[burn:runs])
   pD<-dbar-dhat
   DIC<-dbar+pD

list(beta.mn=beta.mn,beta.var=beta.var,beta.pos=beta.pos,
     CAR.params=params,int=keep.int, #baseline = baseline.track,
     dev=dev,dbar=dbar,pD=pD,DIC=DIC)}

#OUTPUT:
#beta[t,s,k] is the regression coefficient for time t, site s, covariate k
#beta.mn is the posterior mean of beta
#beta.var is the posterior variance of beta
#beta.pos is the posterior prob that beta is positive
#par.params has the draws of the CAR covariance parameters
#int has the draws of the overall average of each regression coefficient
#dev is the draws of the deviance
```

```
#DIC, dbar, and pD are DIC statistics



#generate truncated normals
rtnorm<-function(m,l,u){
   q1<-pnorm(l,m,1)
   q2<-pnorm(u,m,1)
   qnorm(runif(length(m),q1,q2),m,1)
}
```