**<<<< Type your name here >>>>**

**BST 6200**

**Final Exam**

**Spring 2020**

**Directions: Type your responses in this file. Start each problem (not each part to a problem) on a new page. Save your file as a PDF and upload to blackboard. Most of the questions require a short answer: two to three sentences. The only problem requiring computing is the last one. Budget your time carefully. There are six problems, totaling 54 points.**

Type your solutions using the TimesNewRoman font. You may copy and paste this line to get started on each problem.

**1. (8 Points: 2 each)** The map below shows the five Missouri counties in the St. Louis region. They are

1 = St. Louis City

2 = St. Louis

3 = St. Charles

4 = Franklin

5 = Jefferson

(a) Give the adjacency list. (That is, for each county, which counties are adjacent? Use the numbers 1, 2, 3, 4, 5, not the names of the counties.)

(b) Give the adjacency matrix. (This is a 5x5 matrix of 0s and 1s where 1 indicates adjacency.)

(c) and (d) When we send the adjacency information to nimble for the CAR model, we have to send it the adjacency information, which we called adj in our code, and the number of counties that each county is adjacent to, which we called num. What are adj and num for this map?

A close up of a logo

Description automatically generated

**2. (8 Points: 4,4)** A new breed of hornets (large bees) has migrated to the state of Washington from Asia. To learn about the habits of these insects, suppose that scientists selected a rectangular plot of land and mapped the location of the hornets’ nests. One question is whether the hornets tend to cluster together, space themselves out (social distancing for insects), or whether their nests are completely spatially random.

(a) The plot of the G function is below. Describe what the G function is and what it says about the spacing of nests.

A close up of a map

Description automatically generated

(b) Describe what the dclf test is. (For example, what is the null/alternative hypotheses, what is the test statistic, etc.) The output of the dclf test is below. What does this say about the spacing of nests?

Diggle-Cressie-Loosmore-Ford test of CSR

Monte Carlo test based on 99 simulations

Summary function: G(r)

Reference function: theoretical

Alternative: two.sided

Interval of distance values: [0, 0.182524668180823]

Test statistic: Integral of squared absolute deviation

Deviation = observed minus theoretical

u = 0.019401, rank = 1, p-value = 0.01

**3. (5 Points)** Suppose you would like to carry out an ecological analysis for the state of Texas where the response is the number of glaucoma (an eye disorder) cases in each county. You have the case count for each county. You also have demographic data from each of the 254 counties in Texas, including population, average age, proportion nonwhite, and others. You would like to use air pollution PM2.5 as a predictor, but you have data from just 31 stations spread across the state; you do not have PM2.5 measurements in each county. Explain how you might use kriging so that you are able to use PM2.5 as a predictor in a CAR model with glaucoma cases as the outcome.

**4. (5 Points)** Suppose that you sample people in a given county and 3 of the 10 test positive for COVID-19. Your prior distribution for the true proportion who have COVID-19 is given in the following table:

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 0.2 | 0.4 |
|  | 1/3 | 1/3 | 1/3 |

Find the posterior distribution of . (Hint: If # of positive cases out of trials, then has a binomial distribution with parameters and . We observed . In R you can get the binomial probability mass function with the dbinom(x,n,p) command.)

**5. (8 Points: 4,4)** Refer to the next problem on the COVID cases in Missouri. The code for Moran’s is as follows

moran.test( MO.sf$COVIDrate , MO.lw )

and gives output:

Moran I test under randomisation

data: MO.sf$COVIDrate

weights: MO.lw

Moran I statistic standard deviate = 4.2434, p-value = 1.101e-05

alternative hypothesis: greater

sample estimates:

Moran I statistic Expectation Variance

0.205149623 -0.008771930 0.002541389

(a) Interpret these results. Is there evidence of spatial autocorrelation? Why?

(b) Here, the method of randomization (randomisation with the UK spelling) was used to determine the significance. Explain how this randomization is done and how the P-value is obtained from it.

**6. (20 points: 4 each)** The file COVID19\_MO1.csv gives the number of COVID cases in Missouri by county as of May 4, 2020. Below, there is R code that you can copy and paste into RStudio. There are only a few lines missing, and these are clearly indicated. One is in the nimble code, where you have to add one line. Two others are later where you say how many simulations you want to run.

Run a CAR model with **just the correlated heterogeneity** (the s term only, not the v term) using the number of COVID cases as the outcome and the county’s population density as a predictor. You suspect that the rate of cases would be higher in more densely populated counties, like St. Louis City, Jackson County, etc., because of greater person-to-person contact.

(a) Give a choropleth map of the raw COVID rates by county.

(b) Run the CAR model with just the correlated heterogeneity. Give some trace plots (plots of the simulated values) for some of the important variables. You will first have to fill in the line in the nimble code, and you will have to give values for niter and nburn. What are you looking for in these plots? What do you conclude from these plots?

(c) Give a choropleth map of the COVID rates based on this spatial model by county.

(d) Bayesian spatial models like these tend to shrink the rates towards the overall rate. What county had the highest observed raw rate, and what was that county’s rate in the spatial model? How many counties had zero cases, and therefore a zero raw rate? Pick two of these counties with a zero COVID rate and determine their estimated rates in the spatial model. Based on your answers here, did this shrinking toward the overall rate occur as expected?

(e) Address from a Bayesian perspective the question of whether the population density has a predictive effect on the county’s COVID rate.

#### R Code

MO.COVID = read.csv("COVID19\_MO1.csv" , header=TRUE , stringsAsFactors=FALSE )

names(MO.COVID) = c("County","COVID","Population")

library( tigris )

library( sp )

library( sf )

library( spdep )

library( ggplot2 )

library( ggthemes )

library( tmap )

library( dplyr )

library( nimble )

#### Read in the CT shape file using the tigris package

MO = counties("Missouri" , cb=TRUE )

#### I used the following code to make the map for the previous problem

indx = ( MO$NAME == "St. Louis City" | MO$NAME == "St. Louis" |

MO$NAME == "St. Charles" | MO$NAME == "Franklin" | MO$NAME == "Jefferson" )

Region = MO[ indx , ]

plot( Region )

MOdata = MO@data

MOdata$NAME[ MOdata$COUNTYFP == 510 ] = "St. Louis City"

MOdata = left\_join( MOdata , MO.COVID , by = c("NAME"="County") )

#### Add the COVID19 rates and population densities to the MO file

MOdata$COVIDrate = MOdata$COVID / MOdata$Population

MOdata$PopDensity = MOdata$Population / as.numeric(MOdata$ALAND)

MO@data = MOdata

MO.sf = st\_as\_sf( MO )

windows( 9 , 7 )

tm\_shape( MO.sf ) +

tm\_fill( col="COVIDrate" , style="fixed" , breaks=seq(0,0.009,0.001) ) +

tm\_borders( col="black" ) +

tm\_text( "NAME" )

MO.nb = poly2nb( MO )

MO.net = nb2lines( MO.nb , coords=coordinates(MO) )

windows( 14 , 8 )

tm\_shape( MO ) +

tm\_borders( col="darkgray" ) +

tm\_shape( MO.net ) +

tm\_lines( col="darkgreen" , lwd=2 )

MO.lw = nb2listw( MO.nb )

k = length(MO@data$NAME) ## There are 115 counties

num = rep(0,k)

for (i in 1:k) num[i] = length( MO.lw$neighbours[[i]] )

adj = c()

for (i in 1:k) adj = c(adj,MO.lw$neighbours[[i]] )

L = length(adj)

MO.Code = nimbleCode({

alpha ~ dflat()

beta ~ dnorm( 0 , 0.01 )

tau ~ dgamma( 1 , 0.01 )

for (i in 1:L)

weights[i] <- 1

s[1:k] ~ dcar\_normal(adj[1:L],weights[1:L],num[1:k],tau,zero\_mean=1)

for (i in 1:k) {

log(theta[i]) <- **#### Finish this line**

y[i] ~ dpois( n[i]\*theta[i] )

}

})

n = MO@data$Population

y = MO@data$COVID

x = MO@data$PopDensity

MO.Consts = list( k=k , n=n , L=L , x=x , adj=adj , num=num )

MO.Data = list( y = y )

MO.Inits = list( alpha = 0 , beta = 0 , tau = 12 , s=rep(0,k) )

MO.Model = nimbleModel( MO.Code,

data = MO.Data,

constants = MO.Consts,

inits = MO.Inits )

compile.MO.Model = compileNimble( MO.Model )

MO.Conf = configureMCMC( MO.Model, print = TRUE )

MO.Conf$addMonitors(c("alpha","beta","tau","theta"))

MO.MCMC = buildMCMC( MO.Conf )

compile.MO.MCMC = compileNimble( MO.MCMC, project = MO.Model )

niter = **#### Finish this line**

nburn = **#### Finish this line**

set.seed(1)

start.time = proc.time()

samples = runMCMC( compile.MO.MCMC, niter = niter, nburnin = nburn,

inits = MO.Inits, nchains = 1, samplesAsCodaMCMC = TRUE )

stop.time = proc.time()

time.elapsed = stop.time - start.time

print( time.elapsed )

head( samples )

COVIDhat = apply( samples[,4:118] , 2 , mean )

MO.sf$COVIDhat = COVIDhat

windows( 9 , 7 )

tm\_shape( MO.sf ) +

tm\_fill( col="COVIDhat" , style="fixed" , breaks=seq(0,0.009,0.001) ) +

tm\_borders( col="black" ) +

tm\_text( "NAME" )

**#### Add whatever lines are needed to answer the questions.**