CSSS 554: Homework 4

Department of Biostatistics @ University of Washington

Alejandro Hernandez

Winter Quarter 2025

Question 1

(a) Obtain Fay-Herriot estimates at Admin1. To describe the model, define the logit of the weighted estimate, p_i^w, to be $\hat{\theta}_i = log[\hat{p_i^w}/(1-\hat{p_i^w})]$ and $\hat{V}_i$ to be the estimated design-based variance of $\hat{\theta}_i$. Fit the Fay-Herriot model, then plot $\widetilde{p_i}$ against $\widehat{p_i}^w$. Plot the posterior standard deviation from the Fay-Herriot model versus the standard errors of $\widehat{p_i}^w$. Comment on these two scatterplots.

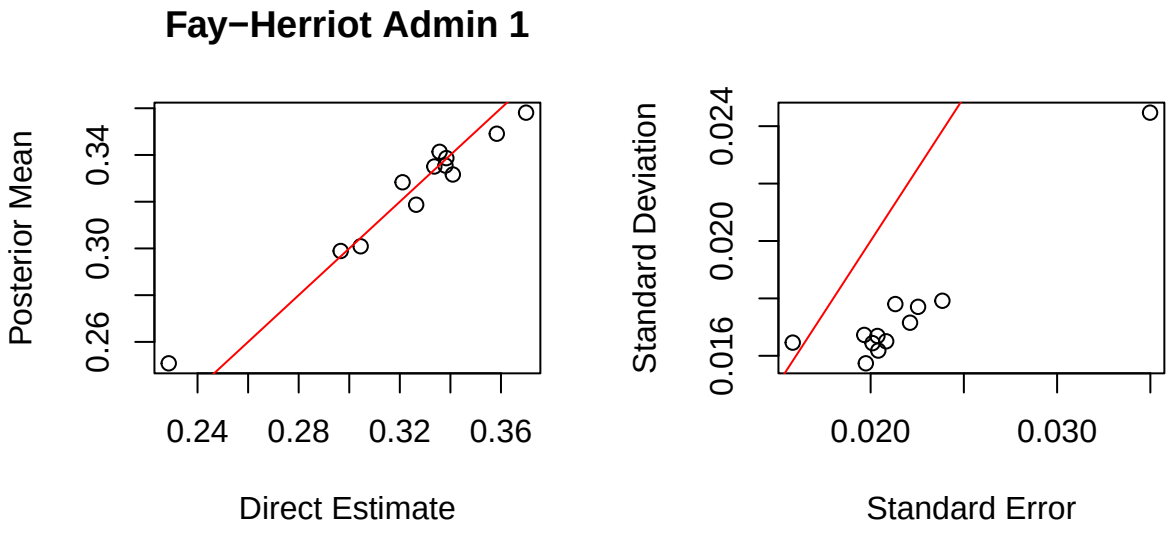

Figure 1: Comparison of Admin 1 prevalence estimates and uncertainty from the Faye-Herriot model.

The Admin 1 maps are shown in Figures 5 and 6. Comparing the direct weighted estimates of prevalence to those generated from the spatial Fay-Herriot model, we can see they agree with one another for most areas, save for a few outlying direct estimates, which F-H believes is more like the rest of the group (Figure 1). This is expected, because modeling spatial dependence grants the Fay-Herriot a reduced spread of area-level estimates (shrinking) compared to the direct estimates. The scatterplot comparing standard deviations of Fay-Herriot posteriors to standard errors of the direct estimates further illustrates the point that modeling spatial dependence through smoothing reduces uncertainty of estimates.

(b) Repeat the previous part but now for Admin2 areas. If there are too many areas with data difficulties, summarize the problem in terms of number of areas with each type of problem.

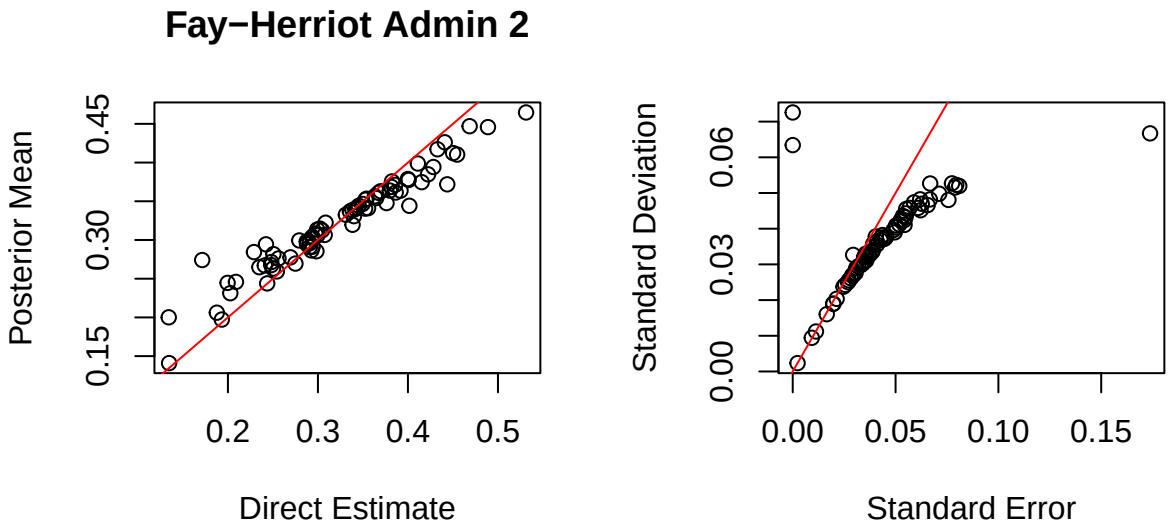

Figure 2: Comparison of Admin 2 prevalence estimates and uncertainty from the Faye-Herriot model.

The Admin 2 maps are shown in Figures 7 and 8. We see shrinking of prevalence estimates from the F-H model at the Admin 2 level as well (Figure 2). The spatial estimates of uncertainty are like those of their counterpart for low values, however they again produce less variation in their estimation as a result of smoothing.

(c) Fit a cluster-level beta-binomial model to your data and obtain Admin1 level estimates. Show maps of posterior mean estimates and posterior standard deviation maps of pi. Provide scatter plots comparing these estimates with the direct estimate analogs (so two scatterplots, one for point estimates and one for uncertainty measures).

The Admin 1 maps are shown in Figures 5 and 6. We see shrinking of prevalence estimates and reduced variance from the cluster-level model at the Admin 1 level (Figure 3).

(d) Fit a cluster-level beta-binomial model to your data and obtain Admin2 level estimates. Show maps of posterior mean estimates and posterior standard deviation maps of p_i.

The Admin 2 maps are shown in Figures 7 and 8. There is still evidence of shrinkage of prevalence estimates (compare axes of the left-hand-side plot) and again a large reduction of variance (Figure 4).

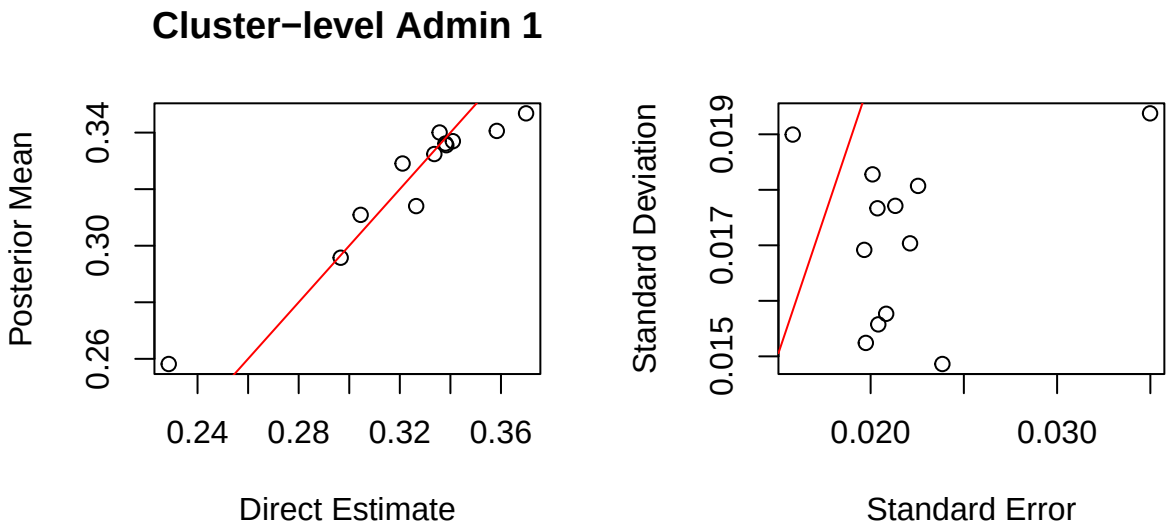

Figure 3: Comparison of Admin 1 prevalence estimates and uncertainty from the cluster-level model.

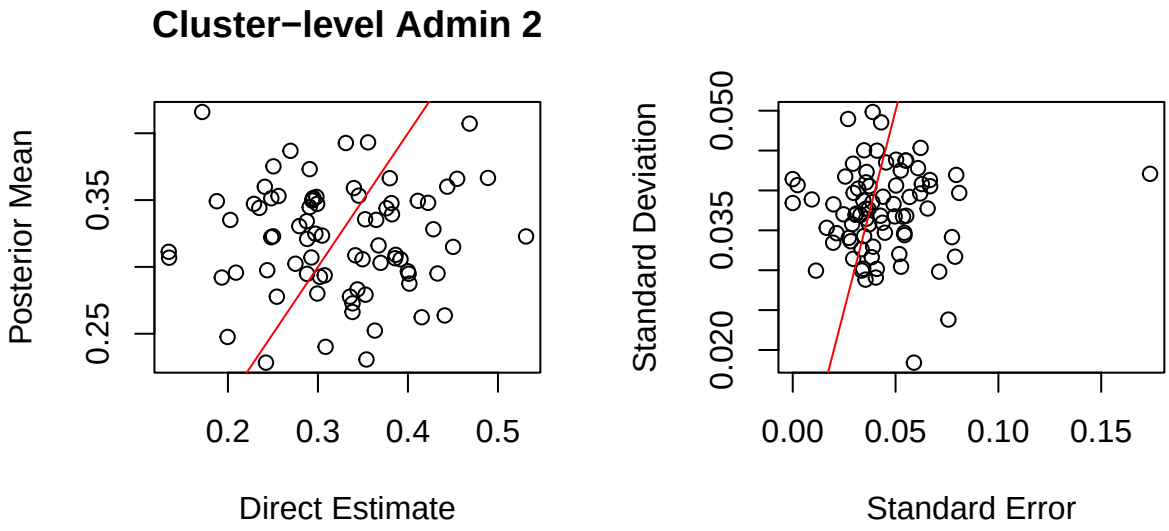

Figure 4: Comparison of Admin 2 prevalence estimates and uncertainty from the cluster-level model.

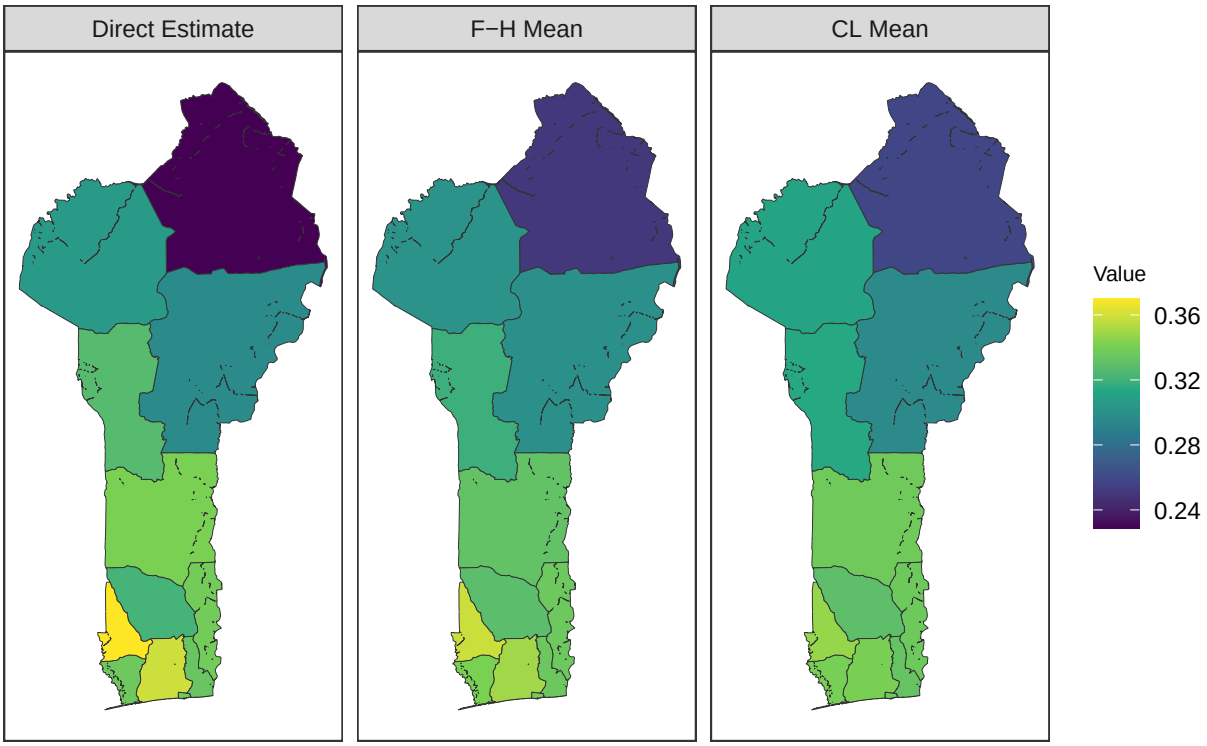

Figure 5: Maps of percentage with unmet family needs at the Admin 1 level of Benin, West Africa. Relative risk estimates are from direct (weighted), Fay-Heriot, and cluster-level models.

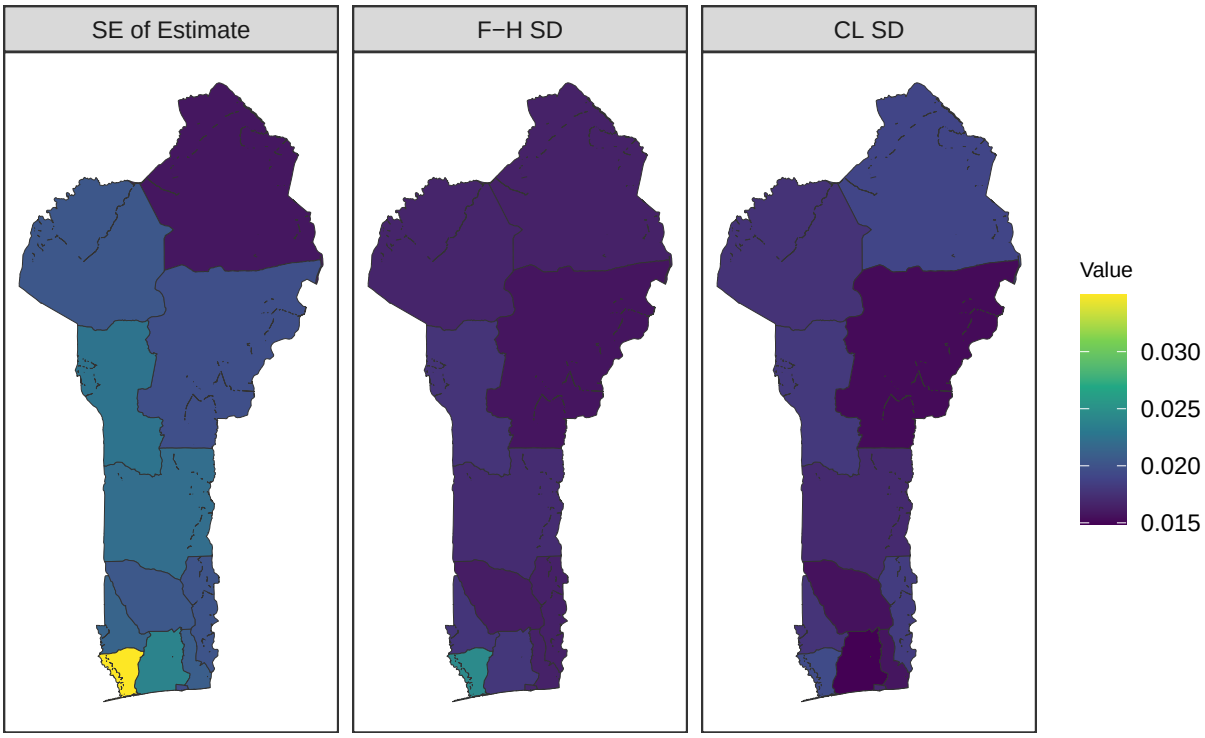

Figure 6: Map of uncertainty estimates from direct (weighted), Fay-Heriot, and cluster-level models at the Admin 1 level.

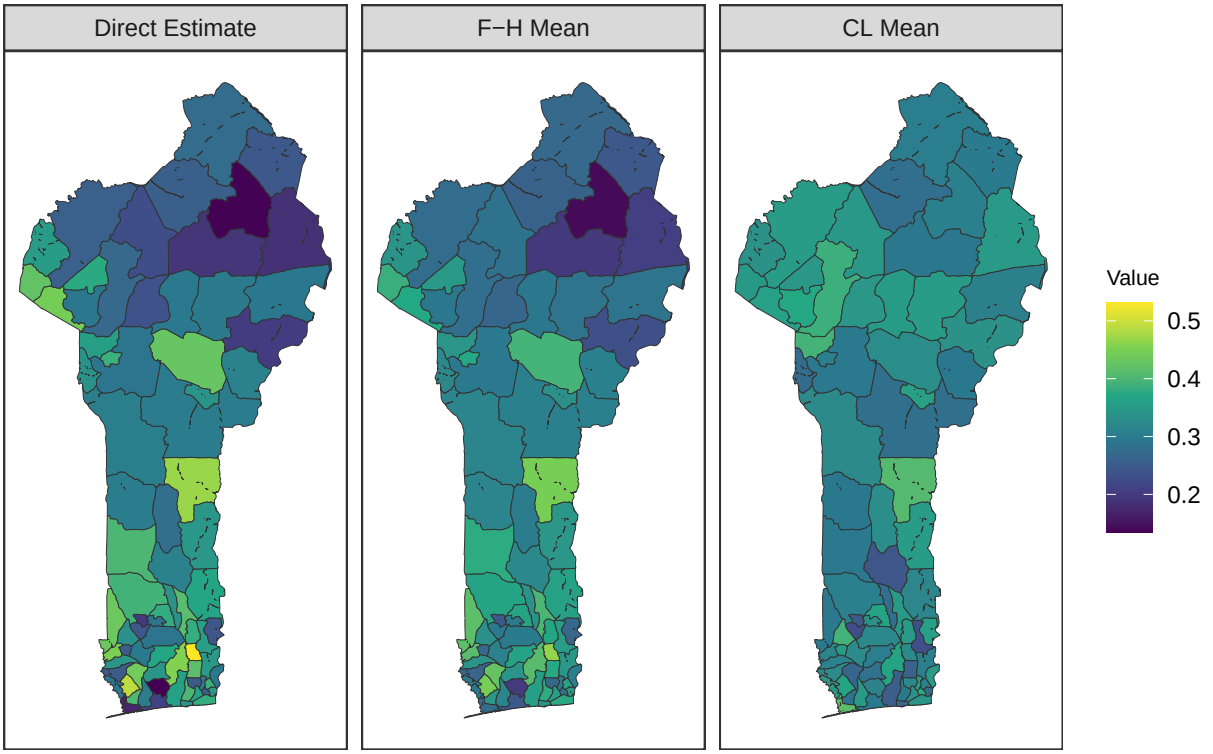

Figure 7: Maps of percentage with unmet family needs at the Admin 2 level of Benin, West Africa. Relative risk estimates are from direct (weighted), Fay-Heriot, and cluster-level models.

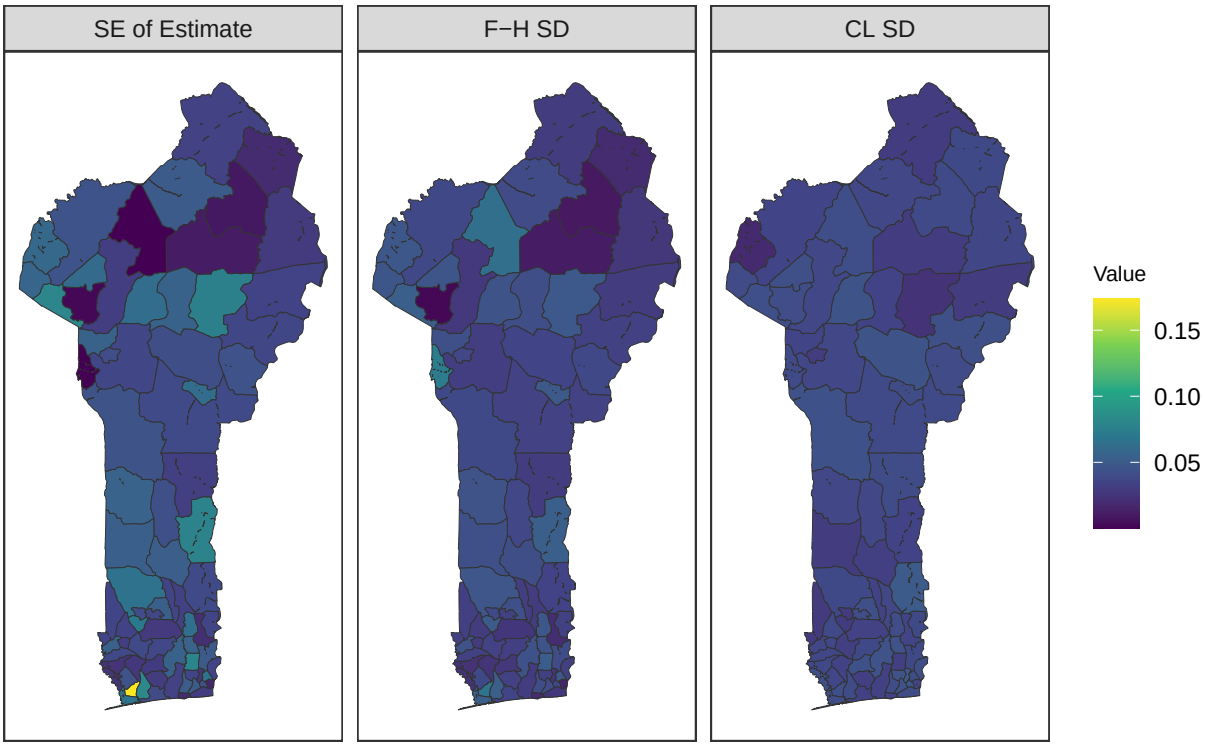

Figure 8: Map of uncertainty estimates from direct (weighted), Fay-Heriot, and cluster-level models at the Admin 2 level.

2. In this question we will carry out SAE for smoking prevalence in health reporting areas (HRAs) in King County, using the BRFSS data. Throughout this assignment, for the Bayesian analyses, use the default INLA hyperpriors.

In the following, we let Y_i and n_i denote the number of smokers and the number sampled respectively, and define p_i to be the proportion of smokers in HRA i, $i = 1, ..., n$.

(a) Weighted Estimation: Create a `svydesign` object and calculate weighted estimates $\hat{p}_i^w$, with associated standard errors and map each of these.

See Figures 9 and 10 for naive estimate and standard error maps.

(b) Smoothed Weighted (Fay-Herriot) Estimation: Define $\hat{\theta}_i = \text{logit}(\hat{p}_i^w)$ to be the transformed weighted estimates and $\hat{V}_i$ to be the estimated design-based variances of $\hat{\theta}_i$. Fit a spatial BYM2 model using the `SUMMER` package and extract posterior means and posterior standard deviations of p_i. Map these quantities.

See Figures 9 and 10 for smoothed weighted estimate and posterior standard deviation maps.

(c) Plot the weighted and smoothed weighted estimates of p_i against each other and comment. Plot the weighted and smoothed weighted standard errors of p_i against each other and comment. Which of the weighted or the smoothed weighted would you recommend using? Why?

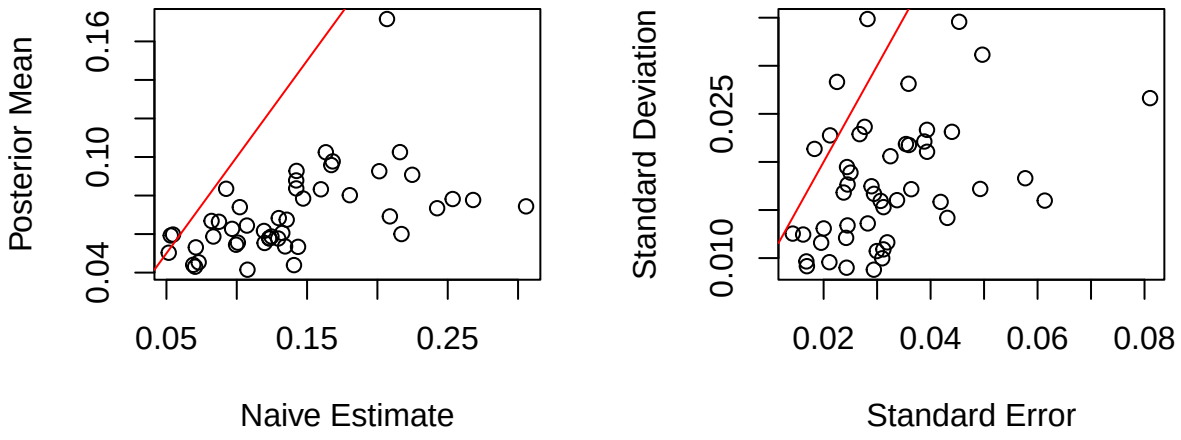

Figure 9: Comparison of smoking prevalence estimates and uncertainty from the weighted and smoothed weighted models.

Comparing the direct weighted estimates of smoking prevalence to those generated from the smoothed weighted estimates, we can see substantial shrinkage from the latter (Figure 9). The window of estimates is so narrow, and the smoothing in Figure 10 is so severe, I may caution the audience of over-shrinkage and recommend the naive estimator. On the right-hand-side, we can see the expected reduction in variation in estimates from the smoothed model, which is the result of smoothing to model spatial dependence.

(d) Summarize the HRA variation in smoking prevalence across King County.

From a non-spatial, weighted estimator of smoking prevalence, we estimate prevalence ranges from 5% to 30% across King County, with a median of 13.1% (IQR: 9.9-16.7 percent). Smoking is most prevalent around the south end of the Puget Sound, between Seattle and Tacoma. Mercer Island and regions East and Northeast of Lake Washington have the lowest prevalences.

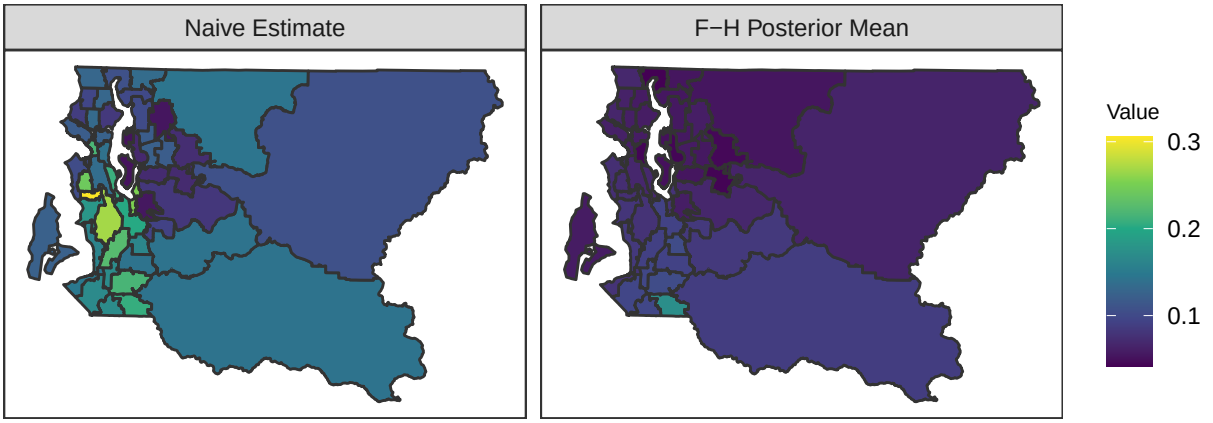

Figure 10: Maps of percentage smoking across King County, Washington USA. Relative risk estimates are from weighted models with (Fay-Herriot) and without (naive) spatial smoothing.

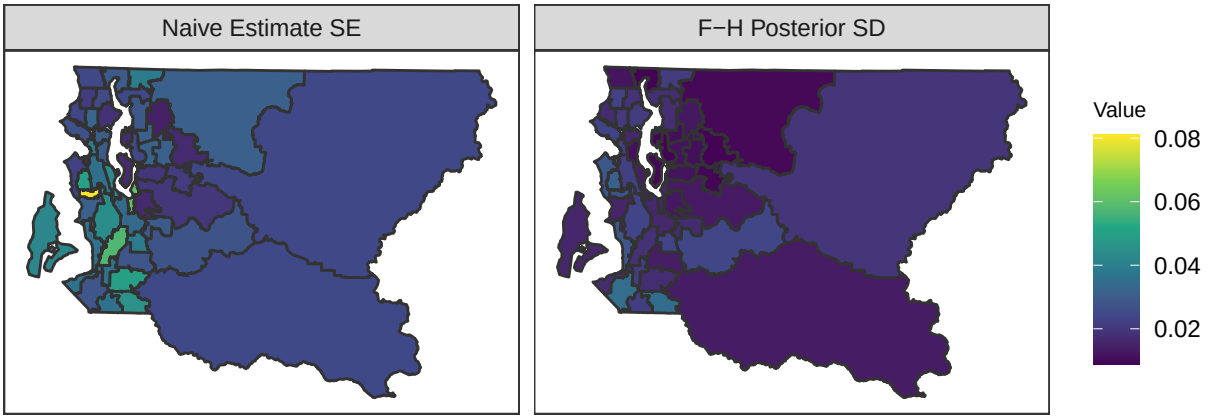

Figure 11: Map of uncertainty estimates from weighted models with (Fay-Herriot) and without (naive) spatial smoothing.

End of report. Code appendix begins on the next page.

Code Appendix

```
# Clear environment
rm(list=ls())

# Setup options
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE, results='hide')
options(knitr.kable.NA = '-', digits = 2)
options(survey.lonely.psu="adjust")
labs = knitr::all_labels()
labs = labs[!labs %in% c("setup", "allcode")]

# Load relevant packages
library(dplyr)
library(haven)
library(geodata)
library(INLA)
library(prioritizr)
library(sf)
library(sn)
library(SUMMER)
library(surveyPrev)

## Load data locally
# DHS data
indicator <- "unmet_family"
year <- 2018
country <- "Benin"
file_path <- "../data/Benin20172018/BJBR71DT/BJBR71FL.DTA"  # data file
dhsData <- as.data.frame(haven::read_dta(file_path))
indicator <- "unmet_family"
data <- surveyPrev::getDHSindicator(dhsData, indicator = indicator)
head(data)  # View head

# BRFSS data
data("BRFSS")
head(BRFSS)

# Handle missing data (there is none)
anyNA(data)

####################
#### QUESTION 1 ####
####################

# Get admin 1 info
poly.adm1 <- geodata::gadm(country, level=1, path=tempdir())
poly.adm1 <- sf::st_as_sf(poly.adm1)
admin1.info <- surveyPrev::adminInfo(poly.adm=poly.adm1, admin=1, by.adm="NAME_1")
head(admin1.info$data)  # View head

# Get admin 2 info
poly.adm2 <- gadm(country, level=2, path=tempdir())
```

```r
poly.adm2 <- sf::st_as_sf(poly.adm2)
admin2.info <- adminInfo(poly.adm=poly.adm2, admin=2,
                         by.adm="NAME_2", by.adm.upper="NAME_1")
head(admin2.info$data)  # View head

# Get geographic data locally
sf_use_s2(FALSE)  # IDK why this works but it fixes an error
file_path_geo <- "../data/Benin20172018/BJGE71FL/BJGE71FL.shp" # shapefile
geo <- sf::st_read(file_path_geo)

# Get cluster information at Admin 1 and 2
cluster.info <- clusterInfo(geo, poly.adm1, poly.adm2)

## Get direct estimates of Admin 1 prevalence
res_ad1 <- surveyPrev::directEST(data, cluster.info, admin = 1)
benin_ad1 <- res_ad1$res.admin1 %>%
  select("Admin 1 Name"=admin1.name, "Direct Estimate"=direct.est,
         "Variance of Estimate"=direct.var) %>%
  mutate("SE of Estimate" = sqrt(.$"Variance of Estimate"))

## Get direct estimates of Admin 2 prevalence
res_ad2 <- surveyPrev::directEST(data, cluster.info, admin = 2)
benin_ad2 <- res_ad2$res.admin2 %>%
  select("Admin 2 Name"=admin2.name, "Direct Estimate"=direct.est,
         "Variance of Estimate"=direct.var) %>%
  mutate("SE of Estimate" = sqrt(.$"Variance of Estimate"))

# Take a peek at estimates table
head(benin_ad1)
head(benin_ad2)

# Histogram of direct estimates
# hist(benin_ad1$`Direct Estimate`, main="Admin 1")
# hist(benin_ad2$`Direct Estimate`, main="Admin 2")

# Histogram of direct estimate variances
# plot(density(benin_ad1$`Variance of Estimate`, main="Admin 1"))
# plot(density(benin_ad2$`Variance of Estimate`, main="Admin 2"))

# Range of direct estimate variances
summary(benin_ad1$`Variance of Estimate`)
summary(benin_ad2$`Variance of Estimate`)

# Fit Fay-Herriot model of Admin 1 prevalence
benin_ad1_fh <- fhModel(data, cluster.info, admin.info=admin1.info, admin=1,
                        model="bym2", aggregation=FALSE)
head(benin_ad1_fh$res.admin1)

# Add estimates to Admin 1 table
benin_ad1 <- cbind(benin_ad1,
                   benin_ad1_fh$res.admin1 %>%
                     select("F-H Mean"=mean, "F-H Variance"=var, "F-H SD"=sd))
```

```r
# Assess whether posterior variances may be too small (like <1e-30)
summary(benin_ad1$`F-H Variance`)
benin_ad1_fh$res.admin1[c(1,6,2,7)]
# The variances are not extremely small nor the CIs extremely narrow

par(mfrow=c(1,2))

# Plot of direct estimates against posterior means
plot(x=benin_ad1$'Direct Estimate', y=benin_ad1$'F-H Mean',
     xlab="Direct Estimate", ylab="Posterior Mean")
abline(0,1,col="red")
title(main="Fay-Herriot Admin 1")

# Plot of direct estimate SE against posterior standard deviations
plot(x=benin_ad1$'SE of Estimate', y=benin_ad1$'F-H SD',
     xlab="Standard Error", ylab="Standard Deviation")
abline(0,1,col="red")

# Assess whether estimate SE may be too small (like <1e-30)
summary(benin_ad2$`Variance of Estimate`)
bad_admin2 <- subset(res_ad2$res.admin2, direct.var < 1e-30)$admin2.name.full
bad_clusters <- subset(cluster.info$data, admin2.name.full %in% bad_admin2)$cluster
# There are 2 areas with too small variance, do not model with these!

# Fit Fay-Herriot model of Admin 2 prevalence
benin_ad2_fh <- fhModel(subset(data, !cluster %in% bad_clusters),
                        cluster.info, admin.info=admin2.info, admin=2,
                        model="bym2", aggregation=FALSE)
head(benin_ad2_fh$res.admin2)

# Add estimates to Admin 2 table
benin_ad2 <- merge(
  x=benin_ad2,
  y=benin_ad2_fh$res.admin2 %>%
    select("Admin 2 Name"=admin2.name, "F-H Mean"=mean,
           "F-H Variance"=var, "F-H SD"=sd),
  by="Admin 2 Name")

par(mfrow=c(1,2))

# Plot of direct estimates against posterior means
plot(x=benin_ad2$'Direct Estimate', y=benin_ad2$'F-H Mean',
     xlab="Direct Estimate", ylab="Posterior Mean")
abline(0,1,col="red")
title(main="Fay-Herriot Admin 2")

# Plot of direct estimate SE against posterior standard deviations
plot(x=benin_ad2$'SE of Estimate', y=benin_ad2$'F-H SD',
     xlab="Standard Error", ylab="Standard Deviation")
abline(0,1,col="red")

# Fit cluster model of Admin 1 prevalence
benin_ad1_cl <- clusterModel(data=data,
```

```
                          cluster.info=cluster.info,
                          admin.info  = admin1.info,
                          stratification = FALSE,
                          model="bym2",
                          admin=1,
                          aggregation=FALSE,
                          CI=0.95)
# Add estimates to Admin 1 table
benin_ad1 <- cbind(benin_ad1,
                   benin_ad1_cl$res.admin1 %>%
                     select("CL Mean"=mean, "CL Variance"=var, "CL SD"=sd))

par(mfrow=c(1,2))

# Plot of direct estimates against posterior means
plot(x=benin_ad1$'Direct Estimate', y=benin_ad1$'CL Mean',
     xlab="Direct Estimate", ylab="Posterior Mean")
abline(0,1,col="red")
title(main="Cluster-level Admin 1")

# Plot of direct estimate SE against posterior standard deviations
plot(x=benin_ad1$'SE of Estimate', y=benin_ad1$'CL SD',
     xlab="Standard Error", ylab="Standard Deviation")
abline(0,1,col="red")

# Fit cluster model of Admin 2 prevalence
benin_ad2_cl <- clusterModel(data=data,
                             cluster.info=cluster.info,
                             admin.info=admin2.info,
                             model="bym2",
                             stratification=FALSE,
                             admin=2,
                             aggregation=FALSE,
                             CI=0.95)
# Add estimates to Admin 2 table
benin_ad2 <- cbind(benin_ad2,
                   benin_ad2_cl$res.admin2 %>%
                     select("CL Mean"=mean, "CL Variance"=var, "CL SD"=sd))

par(mfrow=c(1,2))

# Plot of direct estimates against posterior means
plot(x=benin_ad2$'Direct Estimate', y=benin_ad2$'CL Mean',
     xlab="Direct Estimate", ylab="Posterior Mean")
abline(0,1,col="red")
title(main="Cluster-level Admin 2")

# Plot of direct estimate SE against posterior standard deviations
plot(x=benin_ad2$'SE of Estimate', y=benin_ad2$'CL SD',
     xlab="Standard Error", ylab="Standard Deviation")
abline(0,1,col="red")

# Map Admin 1 estimates
```

```r
mapPlot(data=benin_ad1, geo=poly.adm1, by.data="Admin 1 Name", by.geo="NAME_1",
        variables=c("Direct Estimate", "F-H Mean", "CL Mean"),
        size = 0.01)

# Map Admin 1 uncertainty
mapPlot(data=benin_ad1, geo=poly.adm1, by.data="Admin 1 Name", by.geo="NAME_1",
        variables=c("SE of Estimate", "F-H SD", "CL SD"),
        size = 0.01)

# Map Admin 2 estimates
mapPlot(data=benin_ad2, geo=poly.adm2, by.data="Admin 2 Name", by.geo="NAME_2",
        variables=c("Direct Estimate", "F-H Mean", "CL Mean"),
        size = 0.01)

# Map Admin 2 uncertainty
mapPlot(data=benin_ad2, geo=poly.adm2, by.data="Admin 2 Name", by.geo="NAME_2",
        variables=c("SE of Estimate", "F-H SD", "CL SD"),
        size = 0.01)

####################
#### QUESTION 2 ####
####################

# Clean BRFSS
BRFSS <- subset(BRFSS, !is.na(BRFSS$smoker1))
BRFSS <- subset(BRFSS, !is.na(BRFSS$hracode))

data(KingCounty)
KingCounty <- st_as_sf(KingCounty)
mat <- adjacency_matrix(KingCounty)
colnames(mat) <- rownames(mat) <- KingCounty$HRA2010v2_
mat <- as.matrix(mat[1:dim(mat)[1], 1:dim(mat)[1]])

# Calculate the direct (weighted) estimates AKA  Horvitz-Thompson estimates
design <- survey::svydesign(
  ids=~1, weights=~rwt_llcp,
  strata=~strata, data=BRFSS)
direct <- survey::svyby(~smoker1, ~hracode, design, survey::svymean)
head(direct)

# Fit smoothed weighted model
svysmoothed <- SUMMER::smoothSurvey(
  data = BRFSS, geo = KingCounty, Amat = mat, response.type = "binary",
  responseVar = "diab2", strataVar = "strata", weightVar = "rwt_llcp",
  regionVar = "hracode", clusterVar = "~1", CI = 0.95)
svysmoothed$fit$summary.fixed[1:5]

# Combine prevalence and uncertainty estimates
kingcounty_prev <- svysmoothed$smooth
kingcounty_prev$sd <- sqrt(kingcounty_prev$var)
kingcounty_prev$naiveest <- direct$smoker1
kingcounty_prev$naivese <- direct$se
```

```r
par(mfrow=c(1,2))

# Plot of direct estimates against posterior means
plot(x=kingcounty_prev$naiveest, y=kingcounty_prev$mean,
     xlab="Naive Estimate", ylab="Posterior Mean")
abline(0, 1, col="red")

# Plot of direct estimate SE against posterior standard deviations
plot(x=kingcounty_prev$naivese, y=kingcounty_prev$sd,
     xlab="Standard Error", ylab="Standard Deviation")
abline(0, 1, col="red")
# Map prevalence estimates
mapPlot(data = kingcounty_prev, geo = KingCounty,
        variables = c("naiveest", "mean"),
        labels = c("Naive Estimate", "F-H Posterior Mean"),
        by.data = "region", by.geo = "HRA2010v2_")

# Map prevalence uncertainty
mapPlot(data = kingcounty_prev, geo = KingCounty,
        variables = c("naivese", "sd"),
        labels = c("Naive Estimate SE", "F-H Posterior SD"),
        by.data = "region", by.geo = "HRA2010v2_")
```

End of document.