Bayesian SAE using Complex Survey Data

Lecture 8B: Advanced SAE in R

Richard Li

Department of Statistics
University of Washington

U5MR demo

SUMMER package

- So far we haven't implemented any space-time smoothing models.
- We will show an example for small-area estimation of the under-5 child mortality rate using the `SUMMER` package (version 0.2.0)
- The model is based on Mercer *et al.* (2015) and later modifications of Li *et al.* (2018)

```
# library(devtools)
# install_github('bryandmartin/SUMMER')
library(SUMMER)
data(DemoData)
```

The model: an overview

- Discrete-Hazards model: for region i, time t, and survey s
  - Estimate ${}_5q_0^{its}$ and design-based variance using `survey` package.
  - Obtain estimates and asymptotic variance of $\text{logit}({}_5q_0^{its})$ using delta method.
- Meta-analysis estimator:
  - Combine estimators from multiple surveys
- Space-time smoothing

Demo data

- DemoData contains model survey data provided by DHS.
- DemoData is a list of 5 data frames where each row represent one person-month record and contains the 8 variables as shown below.
- Notice that ‘time‘ variable is turned into 5-year bins from ‘80-84‘ to ‘10-14‘.

```
summary(DemoData)

##      Length Class      Mode
## 1999 8      data.frame list
## 2003 8      data.frame list
## 2007 8      data.frame list
## 2011 8      data.frame list
## 2015 8      data.frame list

head(DemoData[[1]])

##   clustid id  region  time  age  weights        strata died
## 1       1  1 eastern 00-04    0 1.057703 eastern.rural    0
## 2       1  1 eastern 00-04 1-11 1.057703 eastern.rural    0
## 3       1  1 eastern 00-04 1-11 1.057703 eastern.rural    0
```

Demo data

- `DemoData` is obtained by processing the raw DHS birth data (in .dta format) in R.
- The raw file of birth recodes can be downloaded from the DHS website `https://dhsprogram.com/data/Download-Model-Datasets.cfm`.
- `DemoData` contains a small sample of the observations in this dataset randomly assigned to 5 example DHS surveys.

Demo data

- Here we demonstrate how to split the raw data into person-month format from.
- Notice that to read the file from early version of stata, the package ‘readstata13‘ is required.
- The following script is based on the example dataset ‘ZZBR62FL.DTA‘ available from the DHS website.
- We use the interaction of v024 and v025 as the strata indicator for the purpose of demonstration.

```
library(readstata13)
my_fp <- "data/ZZBR62DT/ZZBR62FL.DTA"
dat <- getBirths(filepath = my_fp, surveyyear = 2015,
    strata = c("v024", "v025"))
dat <- dat[, c("v001", "v002", "v024", "per5", "ageGrpD",
    "v005", "strata", "died")]
colnames(dat) <- c("clustid", "id", "region", "time",
    "age", "weights", "strata", "died")
```

Demo map

- DemoMap contains geographic data from the 1995 Uganda Admin 1 regions defined by DHS.
- As we have practiced so far, you can also use read_shape to read in maps and extract adjacency matrix.
- Here we use this built-in map for a quick illustration.

```
data(DemoMap)
geo <- DemoMap$geo
mat <- DemoMap$Amat
```

Direct estimates: Mercer et al. (2015)

- THe U5MR is calculated as ${}_5q_0 = 1 - \prod_j (1 - {}_{n_j} q_{x_j})$ over discrete time intervals of $[x_j, x_j + n_j)$.
- We adopt a discrete hazard model with age groups (in months)

$$[0,1), [1,12), [12,24), [24,36), [36,48), [48,60)$$

- We use logistic regression (`svyglm`) to obtain Horvitz-Thompson estimators for the monthly (conditional) probability of dying and then calculate ${}_{n_j}q_{x_j}$.
- Design-based variance and the asymptotic variance of $\text{logit}_5 q_0$ are calculated.

```
years <- levels(DemoData[[1]]$time)
data <- countrySummary_mult(births = DemoData, years = years,
        idVar = "id", regionVar = "region",
        timeVar = "time", clusterVar = "~clustid+id",
        ageVar = "age", weightsVar = "weights",
        geo.recode = NULL)
```

Combining multiple surveys

- Before fitting the model, we first aggregate estimators from different surveys by

$$
{}_5\widehat{q}_0^{\,it} = \text{expit}\left(\sum_{s=1}^{S_t} \underbrace{\left[\frac{\widehat{V}_{\text{DES},its}^{-1}}{\sum_{s=1}^{S_t} \widehat{V}_{\text{DES},its}^{-1}}\right]}_{\text{Weight for survey } s} \text{logit}({}_5\widehat{q}_0^{\,its})\right),
$$

and

$$
\widehat{V}_{\text{DES},it} = \frac{1}{\sum_{s=1}^{S_t} \widehat{V}_{\text{DES},its}^{-1}}.
$$

```
data <- aggregateSurvey(data)
```

National model: 5-year period

- Now we are ready to fit the models.
- First, we ignore the subnational estimates, and fit a model with temporal random effects only. In this part, we use the subset of data region variable being "All".
- We fit a second-order Random Walk model (RW2) on the scale of 5-year periods, i.e., 85-90, 90-94, ...
- We also project one interval into the future, i.e., 15-19.

```r
years.all <- c(years, "15-19")
priors <- simhyper(R = 2, nsamp = 1e+05, nsamp.check = 5000,
    Amat = mat, only.iid = TRUE)
fit1 <- fitINLA(data = data, geo = NULL, Amat = NULL,
    year_names = years.all, year_range = c(1985, 2019),
    priors = priors, rw = 2, is.yearly = FALSE, m = 5)
```

National model: 1-year period

- The temporal random effects in the previous slides is based on Random Walks on 5-year periods.
- To obtain yearly estimates, we need to interpolate, which is not ideal.
- It turns out we can parameterize random walks on the yearly scale as well.
- More details in Li *et al.* (2018).

```
fit2 <- fitINLA(data = data, geo = NULL, Amat = NULL,
    year_names = years.all, year_range = c(1985, 2019),
    priors = priors, rw = 2, is.yearly = TRUE, m = 5)
```

Extract output

- The `fit` fields contain the regular INLA fitted object. Codes we have practiced so far can be used to extract information from it.
- Alternatively, the `projINLA` function organizes the smoothed estimates more nicely.

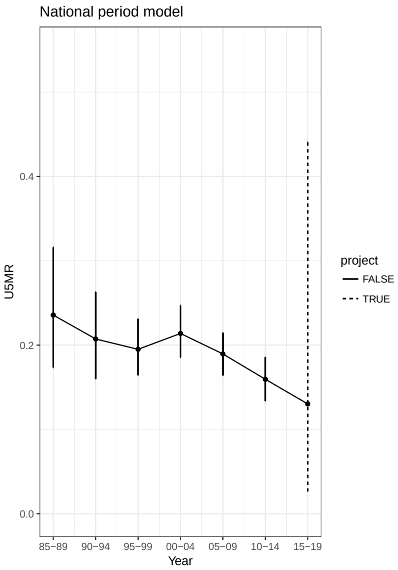
```
out1 <- projINLA(fit1, is.yearly = FALSE)
out2 <- projINLA(fit2, is.yearly = TRUE)
head(out2)

##   District Year logit.q975 logit.q025 logit.med      q975
## 1        0 1985 -0.3123454  -1.819250 -1.102872 0.4357433 0.1
## 2        0 1986 -0.5426282  -1.696759 -1.120144 0.3652167 0.1
## 3        0 1987 -0.6786606  -1.652128 -1.146574 0.3362382 0.1
## 4        0 1988 -0.7509157  -1.646109 -1.196325 0.3236092 0.1
## 5        0 1989 -0.7503514  -1.728796 -1.245580 0.3084886 0.1
## 6        0 1990 -0.8277108  -1.747986 -1.277273 0.3074410 0.1
##   Year.num
## 1     1985
## 2     1986
```

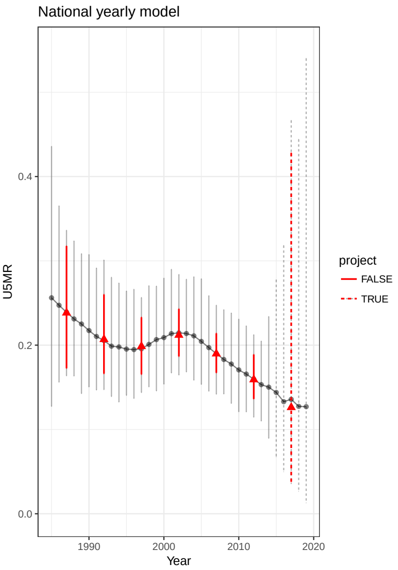
Visualization

- The default `plot` function plot the smoothed estimates over time
- It returns a `ggplot2` plot, which allows user to further edit the themes and elements.
- See `?plot.projINLA` for a list of arguments that help customize the plot

```
library(ggplot2)
library(gridExtra)
g1 <- plot(out1, is.yearly = FALSE, is.subnational = FALSE)
g1 <- g1 + ggtitle("National period model") + ylim(c(0,
    0.55))
g2 <- plot(out2, is.yearly = TRUE, is.subnational = FALSE)
g2 <- g2 + ggtitle("National yearly model") + ylim(c(0,
    0.55))
grid.arrange(grobs = list(g1, g2), ncol = 2)
```

Visualization





Subnational model: 5-year period

- Now we are ready to fit the subnational model with both spatial and temporal random effects.
- We also include a structured space-time interaction effect (type IV of Knorr-Held (2000), all 4 types of interactions are implemented)
- See Chapter 7 of Blangiardo and Cameletti (2015) for more details.
- Again we fit the period model and obtain the results first.

```
fit3 <- fitINLA(data = data, geo = geo, Amat = mat,
    year_names = years.all, year_range = c(1985, 2019),
    priors = priors, rw = 2, is.yearly = FALSE, m = 5)
out3 <- projINLA(fit3, Amat = mat, is.yearly = FALSE)
```

Subnational model: 1-year period

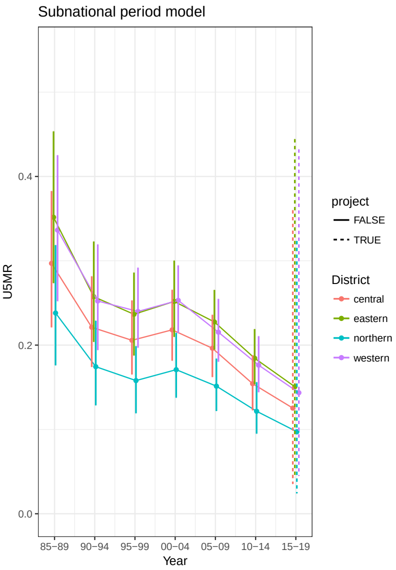
- We now fit the yearly RW2 model and obtain the results.

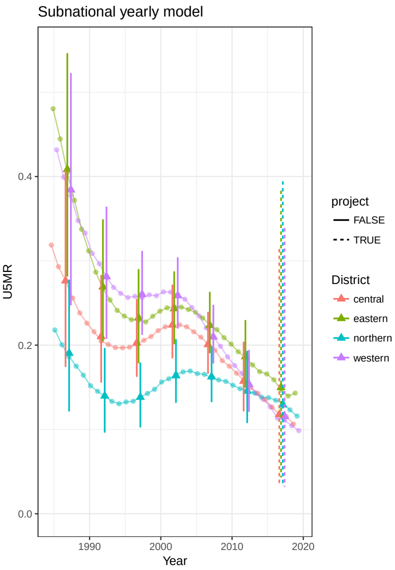
```
fit4 <- fitINLA(data = data, geo = geo, Amat = mat,
    year_names = years.all, year_range = c(1985, 2019),
    priors = priors, rw = 2, is.yearly = TRUE, m = 5,
    type.st = 4)
out4 <- projINLA(fit4, Amat = mat, is.yearly = TRUE)
```

Compare

```
g3 <- plot(out3, is.yearly = FALSE, is.subnational = TRUE) +
    ggtitle("Subnational period model") + ylim(c(0,
    0.55))
g4 <- plot(out4, is.yearly = TRUE, is.subnational = TRUE) +
    ggtitle("Subnational yearly model") + ylim(c(0,
    0.55))
grid.arrange(grobs = list(g3, g4), ncol = 2)
```

Compare

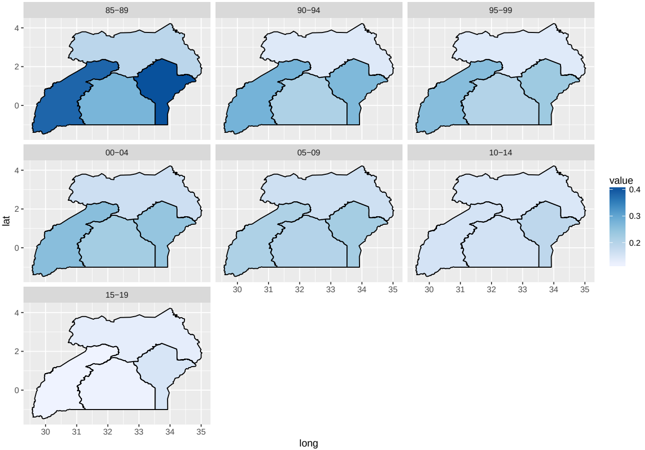

Subnational period model
U5MR
0.0
0.2
0.4
85–89
90–94
95–99
00–04
05–09
10–14
15–19
Year
project
FALSE
TRUE
District
central
eastern
northern
western


Subnational yearly model
U5MR
0.0
0.2
0.4
1990
2000
2010
2020
Year
project
FALSE
TRUE
District
central
eastern
northern
western

Visualization on maps

As we have seen, visualization of estimates on a map is straightforward with `mapPlot`.

```
mapPlot(data = subset(out4, is.yearly == F), geo = DemoMap$geo,
    variables = c("Year"), values = c("med"), by.data = "District",
    by.geo = "NAME_final", is.long = TRUE)
```



References

Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in medicine*, **19**(17-18), 2555–2567.

Li, Z. R., Godwin, J., Hsiao, Y., Martin, B., Wakefield, J., and Clark, S. J. (2018). Changes in the spatial distribution of the Under Five Mortality Rate: small-area analysis of 122 DHS Surveys in 262 subregions of 35 Countries in Africa.

Mercer, L. D., Wakefield, J., Pantazis, A., Lutambi, A. M., Masanja, H., and Clark, S. (2015). Space-time smoothing of complex survey data: small area estimation for child mortality. *The annals of applied statistics*, **9**(4), 1889.