Smoothing areal data

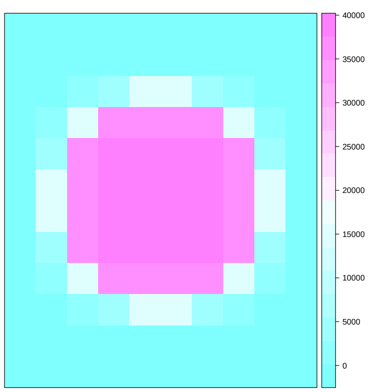
- observations are measured with error
- if obs. are spatially correlated, nearby obs. contain information relevant to the true value / prediction for this obs.
- ⇒ predict Y_i incorporating nearby obs.
- Nearby sometimes means "all", i.e. global mean
- Will talk about concepts with some details
  - Naturally leads to Bayesian inference, won't go that far in 406
- Bivand's example is disease mapping (Chapter 10)
- applies to count data for many outcomes
- We focus on normally distributed outcomes

Spatial smoothing

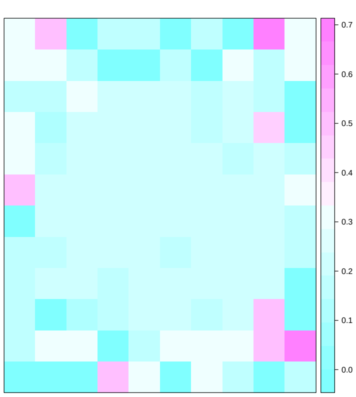
- Made up map of # cases of flu in central Iowa in January 2013
- Picture on next two slides
  - expressed as # cases / person
  - Overall rate in February expected to be similar to that observed in January
- Q: Where do you expect February rate to be the highest?

- it may help to know that # people large in middle, small at edges

"flu" data: sample size per area



"flu" data: empirical proportion of cases per area



Spatial smoothing

- Made up map of # cases of flu in central Iowa in January 2013
  - expressed as # cases / person
  - Rate in February expected to be similar to that observed in January
- Q: Where do you expect February rate to be the highest?
- A: Not one of the red areas on the edge !!

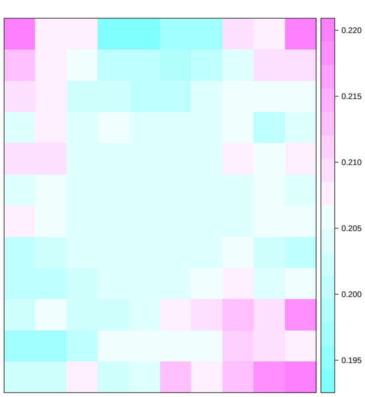
Spatial Smoothing

- Why an unusually large value on edge of map may be noise, not signal
  - often a consequence of sample size
  - each square in the middle is ca 10,000 people. Those on the edges are ca 100
  - e.g. big city in the middle of the mapped area
- sd of estimated rate 10x higher in edge squares.
- if assume rates are spatially correlated, nearby areas inform about poorly est. areas, especially if nearby areas have large N
- Another difference between areal and geostatistical data
  - Geostat: Often reasonable to assume constant variance
  - Areal: Often not.
  - Var, or sd, depends on something else about each region (e.g. N)

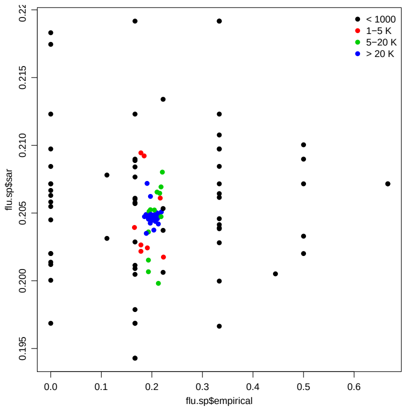
Smoothing using SAR or CAR models

- The CAR and SAR models in the previous section performed spatial smoothing
- Two types of predictions
  - ignoring neighbor values:
    prediction using only fixed effect part of the model
  - including neighbor values:
    fixed effect + sum of neighbor contributions
- Plot of 2nd on next two slides
- Predictions smoothed to overall mean because large values usually surrounded by smaller values
  - Average of neighboring residuals close to 0

"flu" data: SAR predictions



"flu" data: SAR predictions vs observed



Spatial smoothing

- Why something new?
  - Var Y_i not constant in a CAR or SAR model
    - depended on neighbor structure
    - not on anything else
    - the independent ν_i had constant variance
  - In many applications
    - Small area estimation of survey data
    - Disease mapping
  - Var Y_i depends on additional features of location i

Role of sample size

- Often the sample size
- When Y_i for a location is an average of N_i responses
- When modeling disease prevalence
  - $Y_i = D_i/N_i$, number of disease cases / population
  - Var $Y_i \propto 1/N_i$
- Larger $N_i \Rightarrow$ smaller variance
- amount of smoothing depends on sample size
  - when N_i large: little smoothing, $\hat{\mu}_i$ close to Y_i
  - when N_i small: want to smooth a lot
- Need to allow Var Y_i to depend on something we specify

Smoothing areal data

- Will keep things simple to emphasize concepts
- Observe Y_i in a region, have multiple regions
- Believe that each Y_i observed with some random error
- Want to predict "true" μ_i for each region
  - Normally distributed observations

$$Y_i \sim N(\mu_i,\ \sigma^2)$$

  - assume (to keep things simple) that σ^2 known
  - σ^2 may be different for each region
- Statistical problem:
  - Given Y_i predict μ_i
- Two common ways to solve:
  - mixed model: $Y_i = \mu + \alpha_i + \varepsilon_i,\ \alpha_i \sim N(0, \sigma_a^2), \varepsilon_i \sim N(0, \sigma_e^2)$
    Find BLUP of α_i
  - Bayes: $Y_i \sim N(\mu_i, \sigma^2)$, find posterior distribution of $\mu_i \mid Y_i$

Smoothing areal data using a mixed model

- model: $Y_i = \mu + \alpha_i + \varepsilon_i$, $\alpha_i \sim N(0, \sigma_a^2), \varepsilon_i \sim N(0, \sigma_e^2)$
- μ: fixed constant, α_i and ε_i are random effects
- Interpretation of the two random effects:
  - ε_i: measurement error, not repeatable, not part of "true" region-specific μ_i
  - α_i: repeatable characteristic of region i
- Goal: predict $\mu + \alpha_i$ for each region i
- Best predictor is E $\mu + \alpha_i \mid Y_i, \sigma_a^2, \sigma_e^2$
- Equation when both random variables have normal distributions, $\boldsymbol{y}$ has multivariate normal distribution

$$\text{BLUP } \mu + \alpha_i = \hat{\mu} + (Y_i - \hat{\mu}\frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

- When variances (σ_a^2 and σ_e^2) are estimated, more correctly called eBLUP (estimated BLUP)

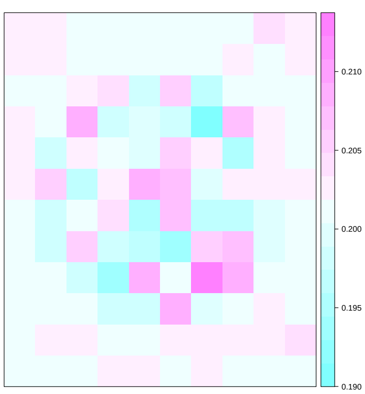
Smoothing areal data using a mixed model

- Often called Fay-Herriot model when both random variables have normal distributions

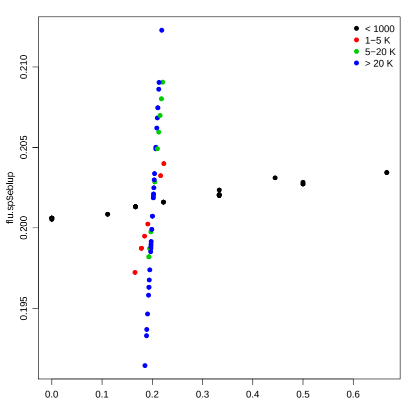
$$\text{BLUP } \mu + \alpha_i = \hat{\mu} + (Y_i - \hat{\mu})\left(\frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}\right)$$

- Behavior of the BLUP. Depends on σ_a^2 relative to σ_e^2
- Large "repeatable" variability, $\sigma_a^2 >> \sigma_e^2$: $\frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2} \approx 1$
  - BLUP $\mu + \alpha_i \approx Y_i$.
  - No (or little) smoothing
- Large measurement error, $\sigma_a^2 << \sigma_e^2$: $\frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2} \approx 0$
  - BLUP $\mu + \alpha_i \approx \hat{\mu}$.
  - extreme smoothing
  - Predictions are the estimated mean
- "Repeatable" variability usually assumed constant
- "Measurement error", σ_e^2 may vary between areas
  - Areas get different amounts of smoothing
  - how much depends on the size of measurement error

"Flu" data: Fay-Herriot smoothing



"Flu" data: Fay-Herriot vs empirical



Small area estimation: Intro

- Context: survey to estimate something in a large region
- e.g. Proportion of US adults who smoke more than 1 pack a month
  - Determine sample size based on desired precision of the estimated proportion
  - For the overall (all US, men + women, all ages) proportion
- Common to also report "small area" estimates
  - men-only, women-only, 20-24 year olds, 20-24 year old men, ...
  - Less precise because fewer responses for that subgroup
- But subgroups may be related
  - e.g. men/20-24 may be related to: women/20-24, men/25-30, ...

Small area estimation: Intro

- Variability between sub-group-specific estimates has two components
  - Measurement error: what happens when different people in men/20-24 sample?
  - Variability in "true" proportion: repeatable characteristic of group
- We have a Fay-Herriot model
- When groups are very different: $\sigma_a^2 >> \sigma_e^2$
  - No, or little, smoothing
  - sub-group estimate is the observed value
  - little (or no) improvement in precision
- When groups are quite similar (or very imprecise): $\sigma_a^2 << \sigma_e^2$
  - No repeatable variation between subgroups
  - Lots of smoothing
  - Sub-group estimate is close to the overall mean
  - Much more precise

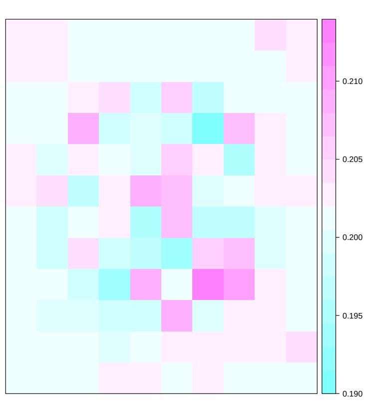
Local smoothing

- Previous smooth towards the global average
- location i ignored
- In the notation used below, the non-spatial FH model is $\boldsymbol{y} = \mu + \boldsymbol{u} + \varepsilon$
- What if you expect μ_i to vary spatially?
  - The "repeatable" variation ($\boldsymbol{u}$) is spatially correlated
  - The "measurement error" variation is independent

$$\boldsymbol{y} = \mu + (\boldsymbol{I} - \boldsymbol{W})^{-1}\boldsymbol{u} + \varepsilon$$

  - This is the spatial Fay-Herriot model
  - Uses a SAR model for the repeatable part (what's in software)
  - Requires specifying a spatial weight matrix
  - Same concept as measurement error kriging
    ε is the measurement error
- For flu data, very small spatial correlation
  so spatial and non-spatial FH give almost the same predictions

"Flu" data: Spatial Fay-Herriot smoothing



Local smoothing

- There are many other approaches - all similar concept but different details
  - Bivand describes Marshall's local Empirical Bayes (EB) estimator
- When responses are counts or yes/no:
  - Distribution of $\boldsymbol{y}$ is not multivariate normal
  - No explicit formulae for smoothed predictions
  - Various approximations
  - Best approach is Bayesian
- And you can combine regression approaches with smoothing
  - Bivand has many details
  - Active research area.

Summary of smoothing: Multivariate normal data

- one source of variability, spatially correlated
  - SAR (or CAR) on errors

$$\boldsymbol{y} = \mu + (\boldsymbol{I} - \boldsymbol{W})^{-1}\boldsymbol{\varepsilon}$$

  - Predictions averaged over neighboring residuals
  - Amount of smoothing depends on spatial correlation
- two sources of variability, no spatial correlation
  - Fay-Herriot model

$$\boldsymbol{y} = \mu + \boldsymbol{u} + \boldsymbol{\varepsilon}$$

  - Predictions are smoothed towards overall mean
  - Amount of smoothing depends on σ_a^2/σ_e^2
- two sources of variability, with spatial correlation
  - Spatial Fay-Herriot model

$$\boldsymbol{y} = \mu + (\boldsymbol{I} - \boldsymbol{W})^{-1}\boldsymbol{u} + \boldsymbol{\varepsilon}$$

  - Predictions are smoothed towards local mean
  - How local depends on spatial correlation
  - Amount of smoothing depends on σ_a^2/σ_e^2