Bayesian spatial smoothing for prevalence data from complex samples

Thomas Lumley
(Jon Wakefield, Cici Bauer)

2013–9–4

Motivating example

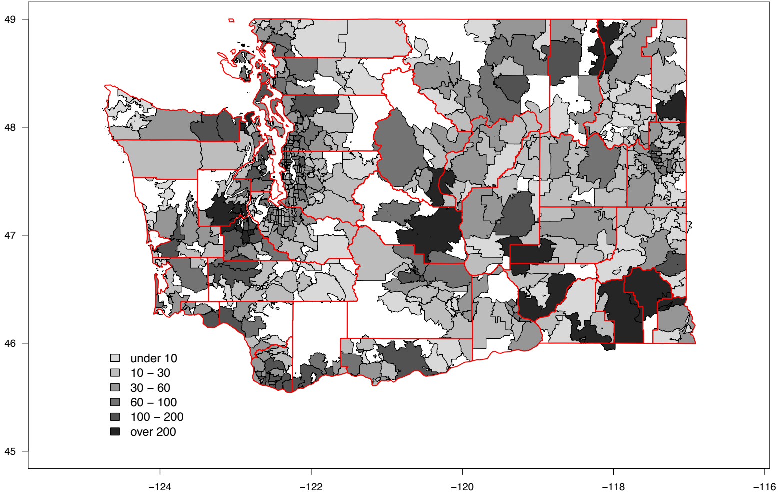
Postcode-level estimation of diabetes, obesity, etc, in Washington, USA, using data from Behavioral Risk Factor Surveillance System (BRFSS)

- Small area: 20% of zip codes have ≤ 9 observations
- Complex sampling: weights vary by a factor of 5000

Want to use Bayesian spatial model, as standard for spatial risk smoothing, but account for sampling design.

- Approximate (coarsened) likelihood for data

Sample size


49
48
47
46
45
−124
−122
−120
−118
−116
under 10
10 – 30
30 – 60
60 – 100
100 – 200
over 200

Approximate likelihood

$\hat{p}_i$ is the (Hajék) estimator of prevalence p_i in zip code i, based on m observations

- define an effective sample size m_i^* and model

$$m^* \hat{p}_i \sim Binomial(m_i^*, p_i)$$

- m_i^* chosen to match sampling and Binomial variances

$$m_i^* \widehat{var}[\hat{p}_i] = \hat{p}_i(1 - \hat{p}_i)$$

- Has correct mean, variance, approximately correct skewness and discreteness
- *cf* Raghunathan et al (2007, JASA) using

$$\sin^{-1}\sqrt{\hat{p}_i} \sim N\left(\sin^{-1}\sqrt{p_i}, \frac{1}{4m_i^*}\right)$$

But zeroes!

Problems if $\hat{p} = 0$ or $m_i < 2$.

For these areas only:

- Replace $\hat{p}_i$ by unweighted empirical-Bayes estimator $\tilde{p}_i$ in a Beta-Binomial model
- Use empirical-Bayes estimate based on Gamma model for residual sum of squares to get $\widehat{var}[\hat{p}]$
- Or add a single pseudo-observation with weight chosen to make $\hat{p}_i = \tilde{p}_i$

Areas with $m_i = 0$ can be treated as missing data and a posterior distribution will automatically be generated.

Shrinkage model

Random effects for each small area, but no explicit spatial structure

$$\begin{aligned} \mathrm{logit} p_i &\sim \alpha + \epsilon_i \\ \epsilon_i &\sim_{iid} N(0, \sigma_\epsilon^2) \\ \alpha &\sim \mathit{flat} \end{aligned}$$

Can easily add other area-level covariates

Spatial model

Spatial model: random effects plus conditional autoregressive spatial term linking area i to its neighbours $\mathcal{N}(i)$

$$\begin{aligned}
\mathrm{logit} p_i &\sim \alpha + \epsilon_i + U_i \\
\epsilon_i &\sim_{iid} N(0, \sigma^2_\epsilon) \\
U_i | U_{\mathcal{N}(i)} &\sim N\left(\bar{U}_{\mathcal{N}(i)}, \frac{\sigma^2}{|\mathcal{N}(i)|}\right) \\
\alpha &\sim \mathit{flat}
\end{aligned}$$

Again, easy to add more covariates

Computation: INLA

Accurate and faster ($\times 1000$) than MCMC, for models with latent Gaussian fields (η), small number of other parameters θ

- Gaussian approximation to $P(\eta|\theta, data)$ (optionally plus spline)
- Laplace approximation to $P(\theta|data)$
- Numerical quadrature for $P(\eta|data) = \int P(\eta|data, \theta)P(\theta|data)\,d\theta$

Heuristically, small-area data can't provide much information about shape of η distribution, so posterior is close to Gaussian.

[only gives marginal posteriors, so can't be used for ranking areas]

Simulations

Based on Washington BRFSS data, with varying spatial structure, calibration for non-response, calibration for age/sex.

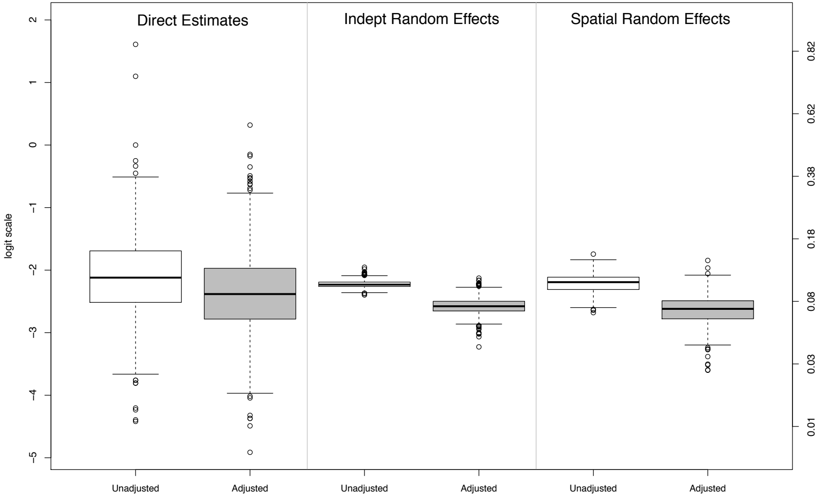
Shows bias reduction compared to unweighted spatial smoothing, variance reduction compared to direct estimates, MSE reduction compared to both

Fewer outlying estimates than arcsin-sqrt approach

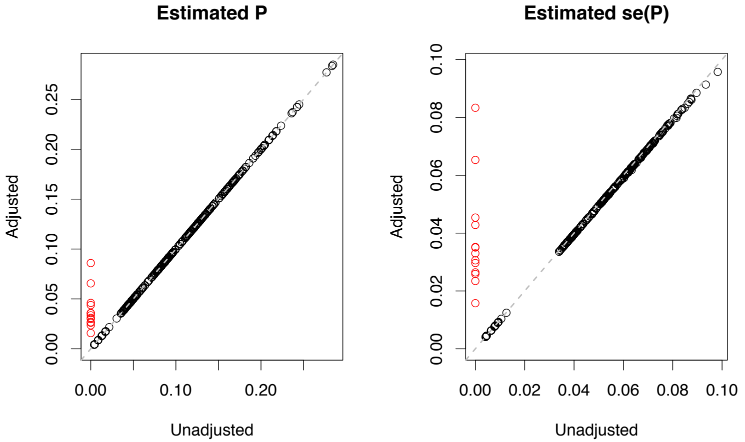
Adding covariates helps

Not as good as Bayesian smoothing adjusting for correctly-specified sampling model using design variables, if these are available.

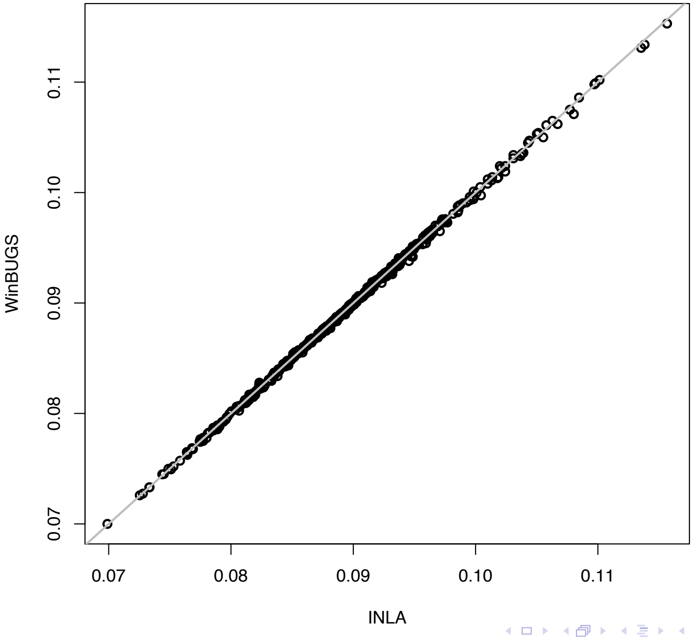
BRFSS example: shrinkage and bias



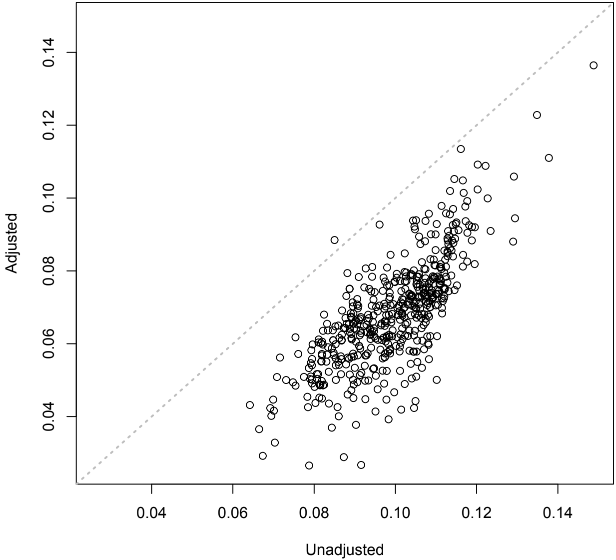
Zero correction



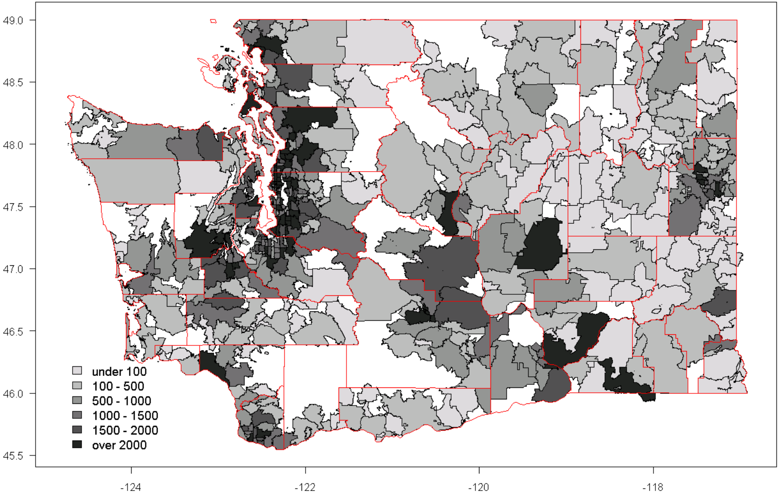
BRFSS example: INLA vs MCMC



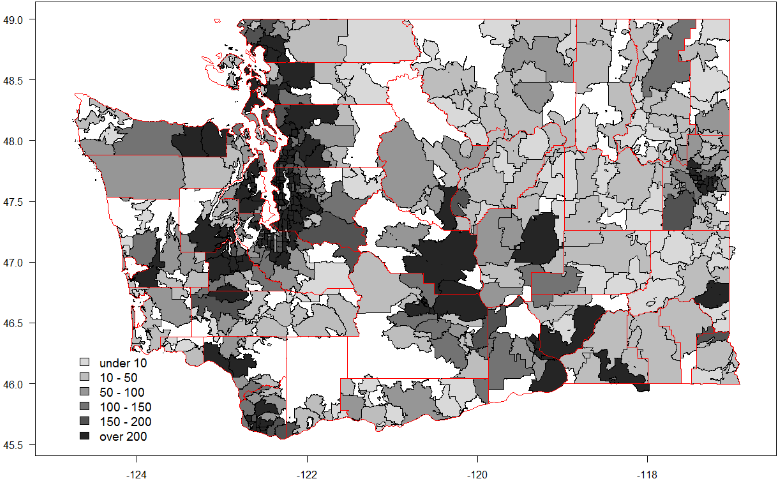
BRFSS example: impact of weights on spatial model



Posterior mean count of diabetes cases



Posterior SE of diabetes cases



Summary

- Approximate binomial likelihood allows simple use of standard Bayesian spatial models
- INLA fits the models well
- Reduces bias vs unweighted Bayesian model, variance vs unshrunk/nonspatial model
- Approximate binomial likelihood seems slightly better than approximate Normal likelihood
- Need to do *ad hoc* things to zeroes.