

Analysis of site and seasonal effects in Florida Panhandle road accident data

I will give a quick run-down of the Florida data, the statistical analysis performed and the results obtained. Note I have used only the accident data themselves - no covariates have been used, although we have been given what seems to be (simulated) flows. One idea could be to slot this method into our overall hotspot prediction model, where data on such a fine scale are available (we have monthly data over many years and for many sites); the main benefit would be much reduced uncertainty attached to model-based estimates of safety, and hopefully predictions.

1 The data

For a set of 52 ‘zones’ (all pretty small) in the Florida Panhandle, we have monthly average accident counts. Apparently, some zones had raw counts from as many as 22 ‘sites’, whilst others had as few as six. The data span the years 1960–2015 inclusive but, before you get excited, not a single zone has data for the entire period! In fact, many zones don’t have data for the majority of this period. The data now take the following format (after extensive spreadsheet manipulation):

Year	Month	Zone 1	Zone 2	...	Zone 52
1960	Sep	NA	4.235	...	8.452
1960	Oct	NA	5.670	...	9.234
⋮	⋮	⋮	⋮	⋮	⋮
2015	July	12.342	NA	...	NA

For each zone we have the longitude/latitude of what is believed to be the ‘zone centroid’, although these have been perturbed to retain actual zonal anonymity. As I said, I also have a file full of what seem to be flows, but that will take some time to bash into the required form for us to use.

2 The model

My idea is to account for zone and seasonal contributions to modelled accident rates across the entire geographic area. If this can be done, then:

- I think estimates of accident rates will have smaller standard errors/posterior standard deviations, and narrower confidence intervals: we can share information across all sites to estimate each seasonal contribution, and across all seasons to estimate each site contribution. In the context of our hotspot prediction model, I’m hoping this adds an extra layer of information to the model and could help reduce uncertainty on predictions;
- we can look for spatial dependence in the estimated site effects as a first step towards using a full-blown spatial model (perhaps allowing for interpolation for points on routes for which we have no data?) – a sort of exploratory analysis investigating the possible merits of a spatial analysis.

Let $y_{m,j}$ be accident rates in season (month) m and zone j . Then, as an initial stab at this, we can assume that

$$y_{i,j}|\lambda_{m,j} \sim \text{Poisson}(\lambda_{m,j}), \quad m = 1, \dots, 12; j = 1, \dots, 52.$$

Of course, we can replace this with a negative binomial model if we want, or do exactly what we do in the hotspot model where we have a Poisson for current/future time-points and negative binomial, with increasing variance, for past time-points. Now to allow for zonal and seasonal effects, we could have

$$\log(\lambda_{m,j}) = \gamma_\lambda^{(m)} + \epsilon_\lambda^{(j)},$$

where $\gamma_\lambda^{(m)}$ is our seasonal effect for (log) λ in month m , and $\epsilon_\lambda^{(j)}$ is our zonal effect for (log) λ in zone j . We might then assume the following priors for the seasonal/zonal effects (which I might now refer to as ‘random effects’):

$$\gamma_\lambda^{(m)} \sim N(0, \tau_\lambda^{-1}) \quad \text{and} \quad \epsilon_\lambda^{(j)} \sim N(c_\lambda, \zeta_\lambda^{-1}).$$

Note that the mean of one of the random effects needs to be fixed somewhere to avoid problems of identifiability (there are an infinite combination of zonal and seasonal effects which could sum to give the logged accident rate); we fix the mean of the prior for the seasonal effects at zero. To retain conjugacy, we then use

$$\tau_\lambda \sim \text{Gamma}(a_\lambda, b_\lambda)$$

and

$$c_\lambda \sim N(d_\lambda, e_\lambda) \quad \text{and} \quad \zeta_\lambda \sim \text{Gamma}(f_\lambda, g_\lambda).$$

Now let’s focus on the precision of the seasonal effects, τ_λ . We have the following density for the seasonal effects:

$$f(\gamma_\lambda^{(m)}) = \sqrt{\frac{\tau_\lambda}{2\pi}} \exp \left\{ -\frac{\tau_\lambda}{2} (\gamma_\lambda^{(m)})^2 \right\},$$

giving the following likelihood for τ_λ :

$$L(\tau_\lambda) = \left(\sqrt{\frac{\tau_\lambda}{2\pi}} \right)^{12} \exp \left\{ -\frac{\tau_\lambda}{2} \sum_{m=1}^{12} (\gamma_\lambda^{(m)})^2 \right\}.$$

The prior for τ_λ is:

$$\pi(\tau_\lambda) \propto \tau_\lambda^{a_\lambda-1} \exp \{ -b_\lambda \tau_\lambda \},$$

giving the following posterior:

$$\pi(\tau_\lambda | \gamma_\lambda^{(m)}) \propto \left(\frac{\tau_\lambda}{2\pi} \right)^6 \tau_\lambda^{(a_\lambda-1)} \exp \left\{ -\frac{1}{2} \left[2b_\lambda \tau_\lambda + \tau_\lambda \sum_{m=1}^{12} (\gamma_\lambda^{(m)})^2 \right] \right\}.$$

Thus, the full conditional for τ_λ is then

$$\pi(\tau_\lambda | \cdot) \propto \tau_\lambda^{(a_\lambda+6-1)} \exp \left\{ -\tau_\lambda \left[b_\lambda + \frac{1}{2} \sum_{m=1}^{12} (\gamma_\lambda^{(m)})^2 \right] \right\},$$

i.e.

$$\tau_\lambda | \cdot \sim \text{Gamma} \left(a_\lambda + 6, b_\lambda + \frac{1}{2} \sum_{m=1}^{12} (\gamma_\lambda^{(m)})^2 \right).$$

In a similar fashion, the full conditionals for the other parameters c_λ and ζ_λ can be derived as:

$$c_\lambda | \cdot \sim N \left(\frac{d_\lambda e_\lambda + \zeta_\lambda \sum_{j=1}^{52} \epsilon_\lambda^{(j)}}{e_\lambda + 52\zeta_\lambda}, (e_\lambda + 52\zeta_\lambda)^{-1} \right), \quad \text{and}$$

$$\zeta_\lambda | \cdot \sim \text{Gamma} \left(f_\lambda + 26, g_\lambda + \frac{1}{2} \sum_{j=1}^{52} (\epsilon_\lambda^{(j)} - c_\lambda)^2 \right).$$

(Joe: this is left as an exercise for you!)

3 Inference

We use a non-informative prior specification; that is,

$$a_\lambda = b_\lambda = f_\lambda = g_\lambda = 0.01; \quad d_\lambda = 0; \quad e_\lambda = 1000;$$

the random effects parameters $\gamma_\lambda^{(m)}$ and $\epsilon_\lambda^{(j)}$ are initialised at their prior means; and τ_λ , c_λ and ζ_λ are drawn from their full conditionals using the initial starting values for $\gamma_\lambda^{(m)}$, $\epsilon_\lambda^{(j)}$, τ_λ , c_λ and ζ_λ . Then random walk Metropolis steps are used to update each random effects parameter singly, before once again drawing τ_λ , c_λ and ζ_λ from their full conditional distributions. The random walk innovation variances are tuned to optimise the scheme; for all parameters, acceptance probabilities in the range 25%-35% were achieved, with thinning to every 10th iteration to reduced autocorrelation. The chains were allowed to run for 100,000 iterations, with the first 10,000 being discarded as burn-in.

4 Results and comparisons

Figure 1 below shows posterior summaries for the random effects, as well as a map of the zones (the longitude/latitude of the zone centroids) and a plot of the spatial effects against longitude. I think this shows that seasonal effects are certainly worthwhile modelling here, and there seems to be a correlation between site effects and longitude (but not latitude, so I skipped this plot). I think this is pretty interesting!

To recover estimates of λ for a particular month in a particular zone, we simply add the seasonal effect for that month to the zonal effect for the zone of interest. For example, the estimated accident rate for June in zone 23 (in effect our APM value here), we find

$$\hat{\lambda}_{6,23} = \hat{\gamma}_\lambda^{(6)} + \hat{\epsilon}_\lambda^{(23)}.$$

To get an idea of the degree of precision gained over an analysis which doesn't share information across sites and seasons, see table 1; here, a simple Poisson-Gamma model has been used to make posterior inferences on $\lambda_{m,j}$ in each month and for each zone separately.

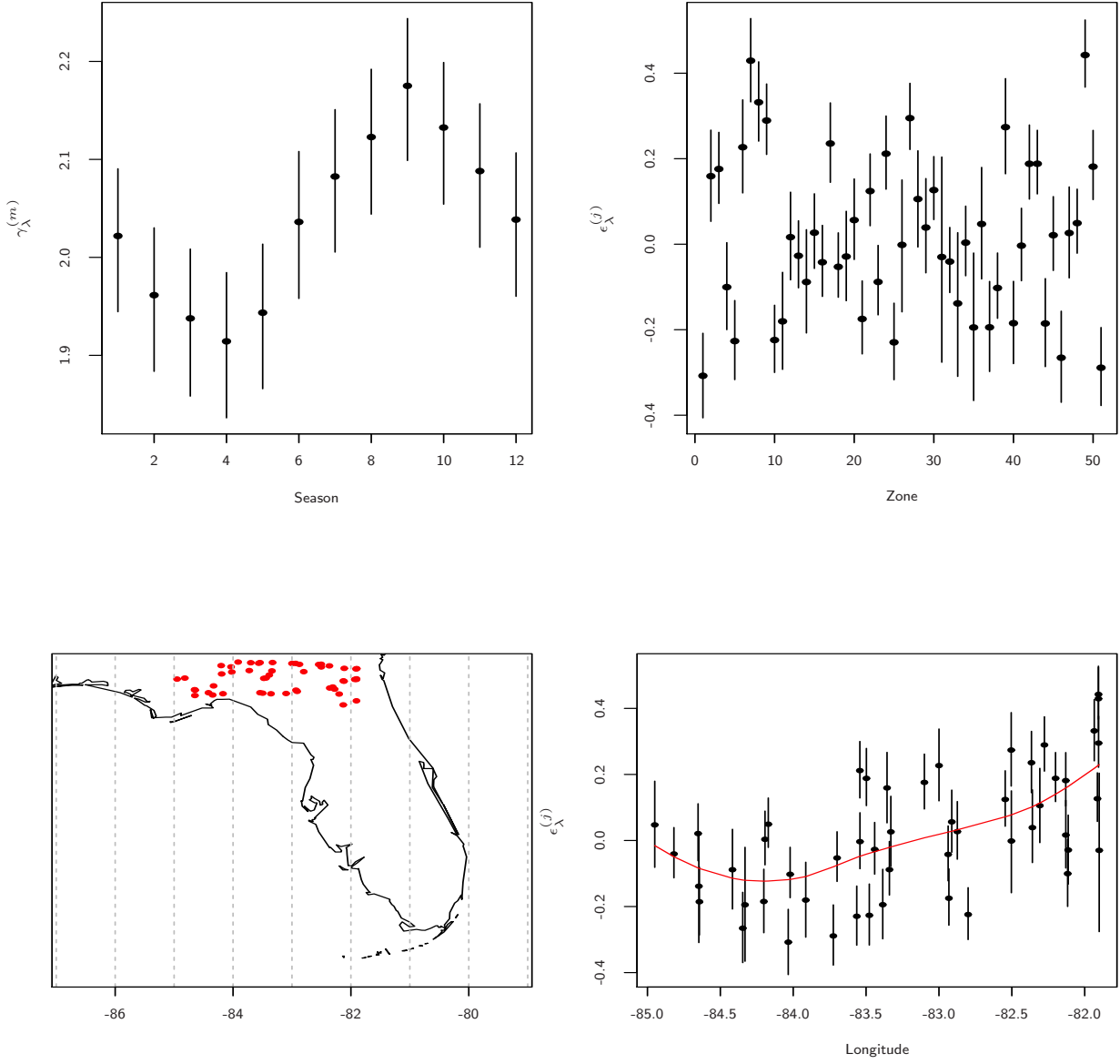


Figure 1: Clockwise, from top-left: Estimated seasonal effects $\gamma_{\lambda}^{(m)}$; estimated site effects $\epsilon_{\lambda}^{(j)}$; estimated site effects versus latitude; and a map showing the ‘location’ of the 52 zones. Points show the posterior mean, and vertical lines show the posterior 95% credible intervals. The red curve in the bottom-right plot shows the fit of a loess smoother.

		Site 23, June	Site 4, December
Random effects model	Posterior mean (St. dev.)	7.017 (0.178)	13.534 (0.635)
	95% CI	(6.677, 7.370)	(12.580, 14.523)
Separate analysis	Posterior mean (St. dev.)	6.946 (0.538)	13.541 (0.971)
	95% CI	(5.812, 8.130)	(11.211, 16.047)

Table 1: Posterior means (standard errors) and 95% credible intervals, for $\lambda_{m,j}$, for $(m = 6, j = 23)$ and $(m = 12, j = 4)$. For comparison, results are also shown for a separate seasons, separate zones approach, where sharing of information through seasonal and zonal effects is not permitted.