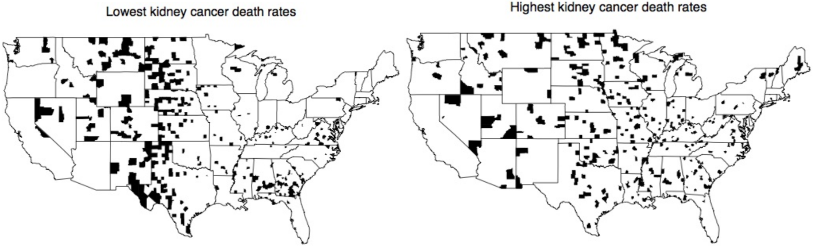
Spatial Models in Stan: Intrinsic Auto-Regressive Models for Areal Data

Mitzi Morris
Stan Developer, Columbia University

Motivating example - epidemiology

Areal data consists of a single aggregated measure per areal unit.

Example: number of cases of kidney cancer in the US, aggregated by county



Event counts for low-population counties display greater variance.

Background: areal data and how to model it.

Areal data consists of a single aggregated measure per areal unit, which may be a binary, count, or continuous value.

Areal units:

- partition a multi-dimensional volume D into a finite number of sub-volumes with well-defined boundaries
- the set of areal units is fixed

Geospatial data in R:

- Shapefile format: contains geometric locations (points, lines, and polygons) and associated attributes.
- R package `spdep` provides function `poly2nb` to extract neighbor relationships.

Example: pedestrian traffic fatalities in NYC

Areal unit is NYC census tract, data consists of event count and tract population

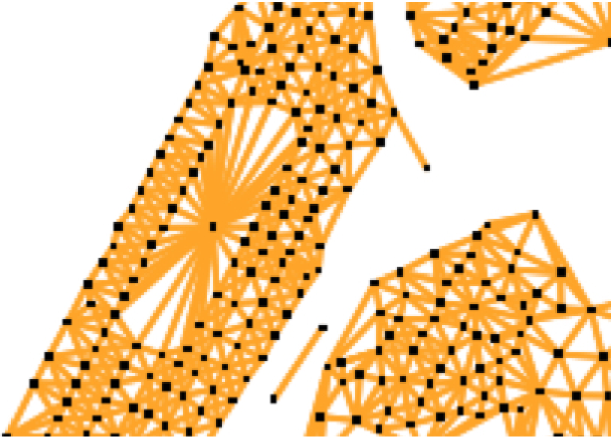
id	pop	cts	geo_id	geo_name
36005000200	4334	7	1400000US36005000200	Census Tract 2, Bronx County,New York
36005000400	5503	12	1400000US36005000400	Census Tract 4, Bronx County, New York
36005001900	1917	16	1400000US36005001900	Census Tract 19, Bronx County, New York
36005002000	8731	17	1400000US36005002000	Census Tract 20, Bronx County, New York
36005002701	3113	6	1400000US36005002701	Census Tract 27.01, Bronx County, New York
36005002702	4475	12	1400000US36005002702	Census Tract 27.02, Bronx County, New York

Columns `geo_id` and `geo_name` come from *shapefiles*

Plot of neighbor relationships between New York City census tract regions

Dots represent geographic center of each tract, lines connect neighbors.

Close-up of Central Park and neighboring census tract regions



Data structures for encoding adjacency

N times N Adjacency matrix:

- entries {i,j} and {j,i} are 1 when regions n_i and n_j are neighbors, zero otherwise

Undirected graph:

- regions are vertices
- pairs of neighbors are edges

For sparse matrices, graph representation is more efficient (memory, I/O, processing)

(Another option: list of lists: `nb_object`: R package `spdep`)

Intrinsic Conditional Auto-Regressive (ICAR) Models, Besag, 1974

Spatial interactions for fixed set of N areal units is N-length vector

$$\phi = (\phi_1, \ldots, \phi_n)^T$$

Neighborhood structure specified by $N \times N$ adjacency matrix W:

- W entries $w_{i,j}$ and $w_{j,i}$ are 1 when regions n_i and n_j are neighbors, zero otherwise

Number of neighbors for each region specified using $N \times N$ diagonal matrix D:

- element $d_{i,i}$ is number of neighbors for region n_i, all other elements zero

Conditional specification:

$$p\left(\phi_i \,|\, \phi_j \, j \neq i\right) = N\left(\frac{\sum_{i \sim j} \phi_i}{d_{i,i}}, \frac{1}{d_{i,i}\tau_i}\right)$$

Joint specification:

$$\phi \sim N(0, [\tau\,(D - W)]^{-1}).$$

Unit mulitivariate Gaussian: $\tau = 1$, joint distribution rewrites to:

$$p(\phi) \propto \exp\left\{-\frac{1}{2}\sum_{i \sim j}\left(\phi_i - \phi_j\right)^2\right\}$$

Pairwise difference formulation

But: ICAR model is non-identifiable, must add the constraint $\sum_i \phi_i = 0$.

Stan program: ICAR prior, hard sum-to-zero constraint

```
data {
  int<lower=0> N;
  int<lower=0> N_edges;
  int<lower=1, upper=N> node1[N_edges];  // node1[i] adj to node2[i]
  int<lower=1, upper=N> node2[N_edges];  // and node1[i] < node2[i]
}
parameters {
  vector[N - 1] phi_raw;
}
transformed parameters {
  vector[N] phi;
  phi[1:(N - 1)] = phi_raw;
  phi[N] = -sum(phi_raw);
}
model {
  target += -0.5 * dot_self(phi[node1] - phi[node2]);
}
```

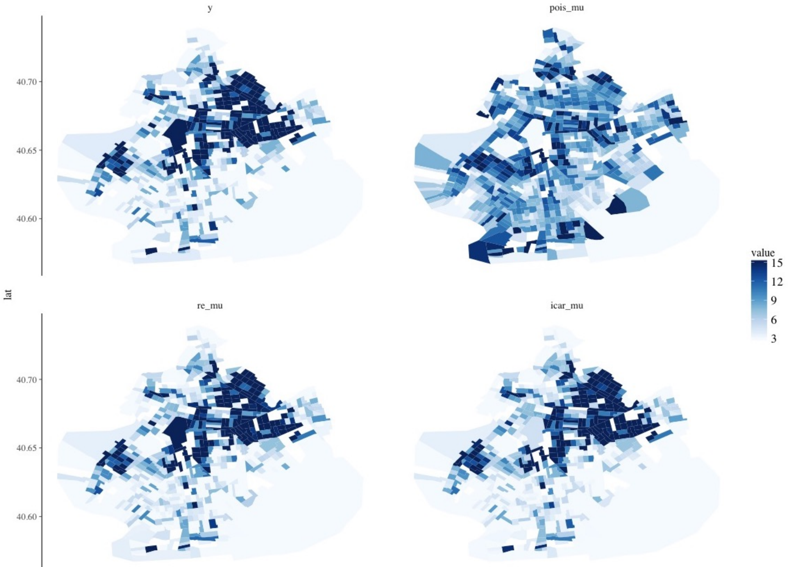
Stan program: ICAR prior, soft sum-to-zero constraint

An alternative sum-to-zero constraint can be implemented by putting a prior on phi as follows:

```
data {
  int<lower=0> N;
  int<lower=0> N_edges;
  int<lower=1, upper=N> node1[N_edges];  // node1[i] adj to node2[i]
  int<lower=1, upper=N> node2[N_edges];  // and node1[i] < node2[i]
}
parameters {
  vector[N] phi;
}
model {
  target += -0.5 * dot_self(phi[node1] - phi[node2]);

  // soft sum-to-zero constraint on phi,
  // equivalent to mean(phi) ~ normal(0,0.01)
  sum(phi) ~ normal(0, 0.01 * N);
}
```

Plot of traffic accident data for Brooklyn, raw counts vs. fitted ICAR model



Highlighting regions where ICAR smoothing has a noticable difference

Conclusion

- Simplified encoding used to construct ICAR model.
- ICAR is improper, can only be used as a prior.
- ICAR is computationally tractable for large-ish numbers of areal units.
- Coding up pairwise difference formula is straightforward in Stan.
- More information, more complex models covered in Stan Case Study: Spatial Models in Stan: Intrinsic Auto-Regressive Models for Areal Data