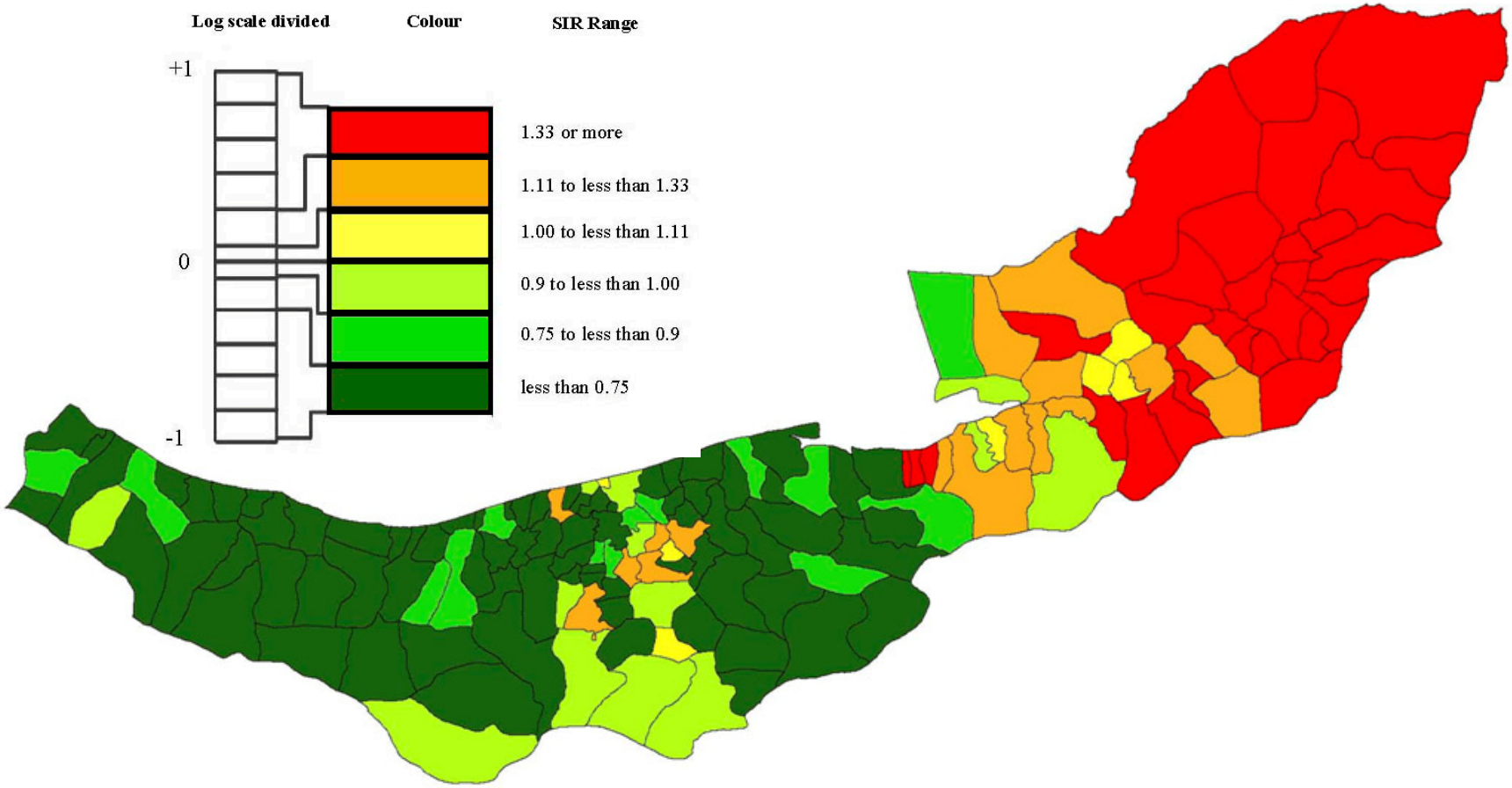
Application of a Bayesian Approach for Analysing Disease Mapping Data: Modelling Spatially Correlated Small Area Counts

Mohammadreza Mohebbi
Rory Wolfe
Department of Epidemiology and Preventive Medicine,
Faculty of Medicine, Nursing and Health Sciences,
Monash University, Melbourne

Mapping Relative Risk

- Relative risk measures how much a particular risk factor influences the risk of a specified outcome (e.g., cancer mortality)
- Classical approach is mapping SMRs (standardized mortality/morbidity rates) for subregions based on Poisson model

Standardised incidence rate (SIR) of esophageal cancer; both sexes combined



Poisson Model

The raw data are in the form of disease counts, Y_j , and population counts, N_j , where j=1,...,n, indexes geographical areas.

For rare and non-infectious diseases we may then assume

$$Y_j|E_j,\Psi_j \sim \text{Poisson}(E_j\Psi_j)$$

Where E_j denote the expected number and Ψ_j represents the relative risk of cases

in area j.

Bayesian approach: Hierarchical model

Enable us to incorporate multiple sources of data and knowledge (e.g., covariates, nonspatial random effect, and spatial autocorrelation)

Prior specification

- Nonspatial random effect to describe unstructured heterogeneity.
- Spatial random effect can be expressed via two approaches:
  - Distance-based V-C structure
  - Neighbourhood-based V-C structure

The Poisson regression

$$\log\Psi_j = X_j \beta^T_j + \theta_j + \Phi_j$$

- where $X^T_j = (1, X_{j1}, \ldots, X_{jk})^T$ is vector of area-level risk factors
- $\beta_j = (0, 1, \ldots, k)^T$ is vector of regression parameters
- θ_j , $j=1,\ldots,n$ represents a residual with no spatial structure
- Φ_j , $j=1,\ldots,n$ represents a residual with spatial structure

Elements of Distance-based Modelling

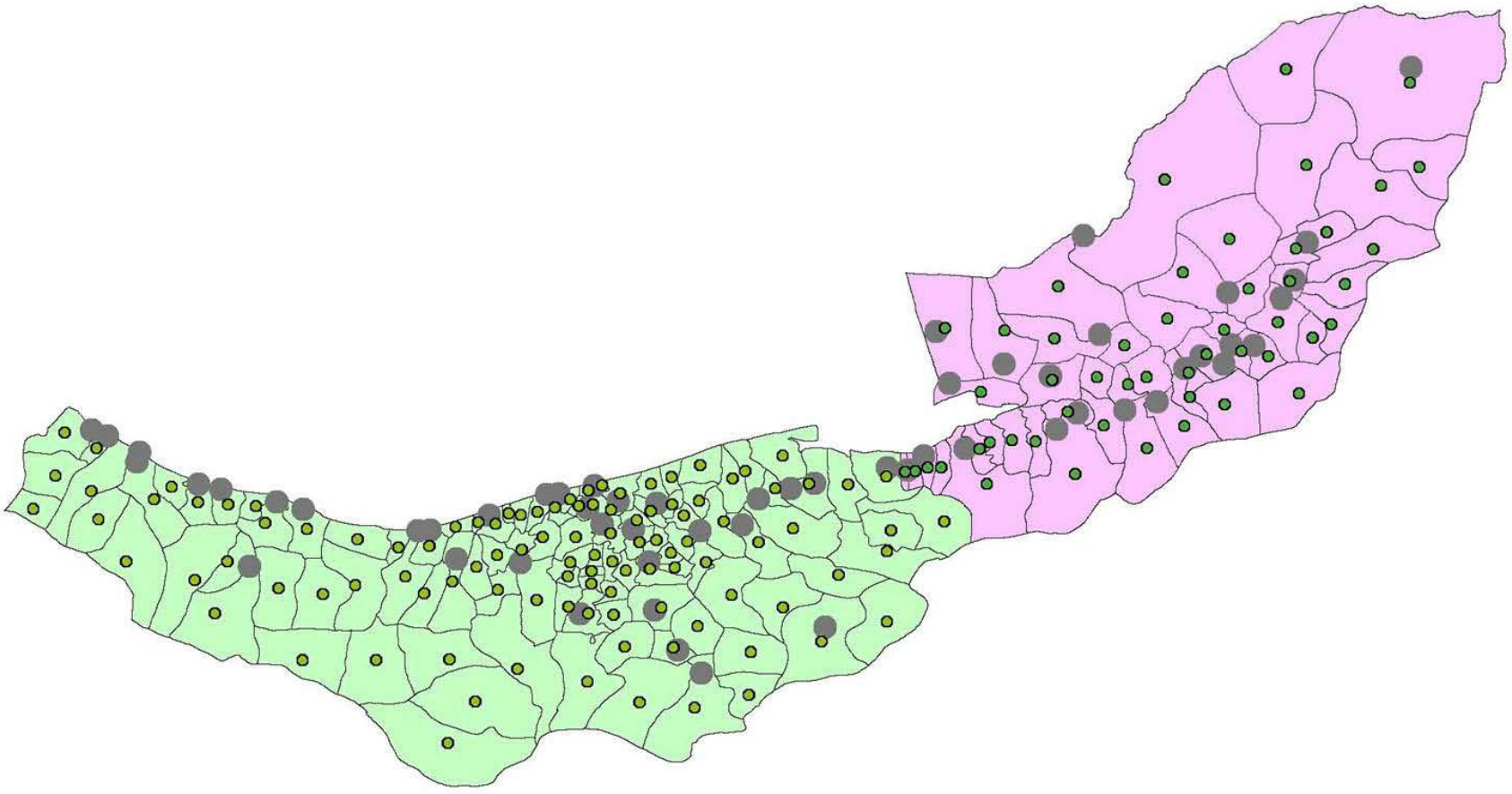
- Distance-based modelling refers to modelling of spatial data collected at locations referenced by coordinates
- Fundamental concept: Data from a spatial process

$$\{\log\Psi_j(s): s \in D\}$$

where D is a fixed subset in Euclidean space.

- Practically: Data will be a partial realization of a spatial process – observed at $\{s_1, \ldots, s_n\}$

Spatial Domain



Statistical Modelling

- Spatial model

$$\log \Psi_j(s) = \mu(s) + \Phi(s) + \theta(s)$$

- $\Phi(s) : s \in D \subset R_d$: Gaussian spatial process
- The covariance function:

$C(s, s') = K(s - s') \sim K(\|s - s'\|)$ (isotropic)

- and θ_i and θ_j are independent for $i \neq j$

The Gaussian process

- We assume $\Phi(s)$ has zero mean multivariate normal distribution $N(0,\Sigma)$
- For a model having a nugget effect, we set

$$\Sigma = \sigma^2 H(\varphi) + \tau^2 I$$

where $(H(\varphi))_{ij} = \rho(\varphi; \tau; d_{ij})$

  - $d_{ij} = ||si - sj||$, the distance between si and sj
  - ρ is a valid correlation function on R_r

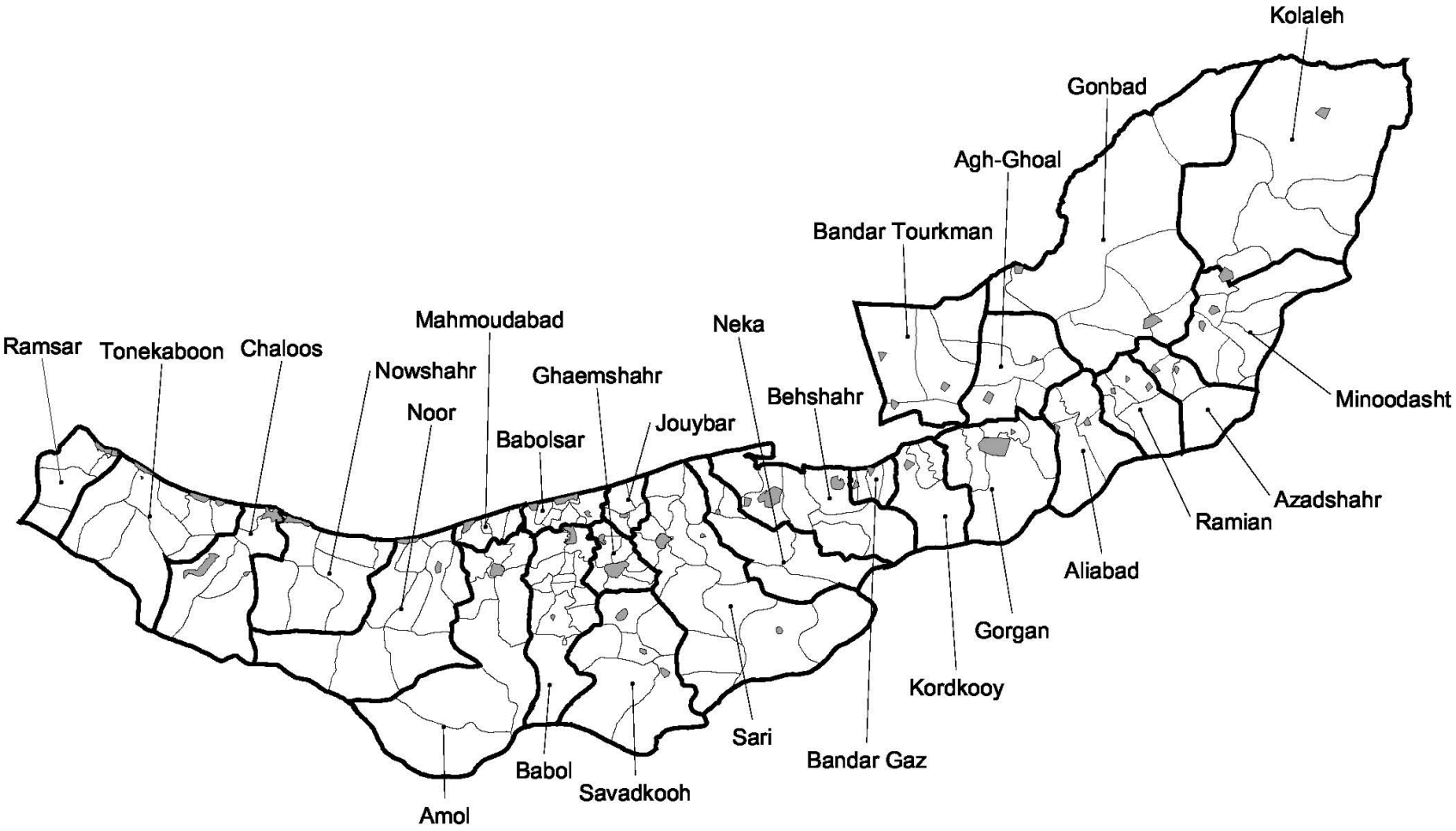
Some common V-C functions

Model	Covariance function, $C(t)$
Linear	$C(t)$ does not exist
Spherical	$C(t) = \begin{cases} 0 & \text{if } t \geq 1/\phi \\ \sigma^2\left[1 - \frac{3}{2}\phi t + \frac{1}{2}(\phi t)^3\right] & \text{if } 0 < t \leq 1/\phi \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Exponential	$C(t) = \begin{cases} \sigma^2 \exp(-\phi t) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Powered exponential	$C(t) = \begin{cases} \sigma^2 \exp(-\lvert\phi t\rvert^p) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Matérn at $\nu = 3/2$	$C(t) = \begin{cases} \sigma^2 (1 + \phi t)\exp(-\phi t) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$

Elements of Neighbourhood-based Modelling: Proximity matrices

- W entries w_{ij} (with $w_{ii} = 0$)
- Choices for w_{ij}:
  - $w_{ij} = 1$ if i, j share a common boundary w_{ij} is an inverse distance between units
  - $w_{ij} = 1$ if distance between units is $\leq K$
  - $w_{ij} = 1$ for m nearest neighbours.
- W is typically symmetric, but need not be

Geographic boundaries of wards (bold polygons), and cities (gray polygons) and rural agglomerations within wards, in the Caspian region



Conditional autoregressive (CAR) structure

- For spatial model

$$\log\Psi_j(s) = \mu(\omega) + \eta(\omega) + \theta(\omega)$$

we assume

$$P(\eta_i | \eta_j, j \neq i) = N(b_{ij} y_j, \sigma_i^2)$$

- Using Brook's Lemma we can obtain

$$p(\eta_1, \eta_2, \ldots \eta_n) \propto \exp\{-\tfrac{1}{2} \eta^T (I-B)\eta\}$$

where $B = \{b_{ij}\}$ and D is diagonal with $D_{ii} = \sigma_i^2$

- suggests a multivariate normal distribution with

$$\mu_\eta = 0 \text{ and } \Sigma_\eta = (I - B)^{-1} D$$

Intrinsic autoregressive (IAR) model!

Fully Bayesian estimation

the Bayesian approach that we follow requires

specification of prior distributions for the second-stage parameters θ_j and Φ_j.

This prior distribution usually depends on hyperparameters γ so that the marginal posterior of $\Psi$is given by

$$P(\Psi|y)= \int p(\Psi, \Upsilon|y)dy$$

- Markov chain Monte Carlo methods employed to obtain a sample from the joint posterior distribution of (Ψ, γ)
- The joint posterior distribution of all parameters is expressed as

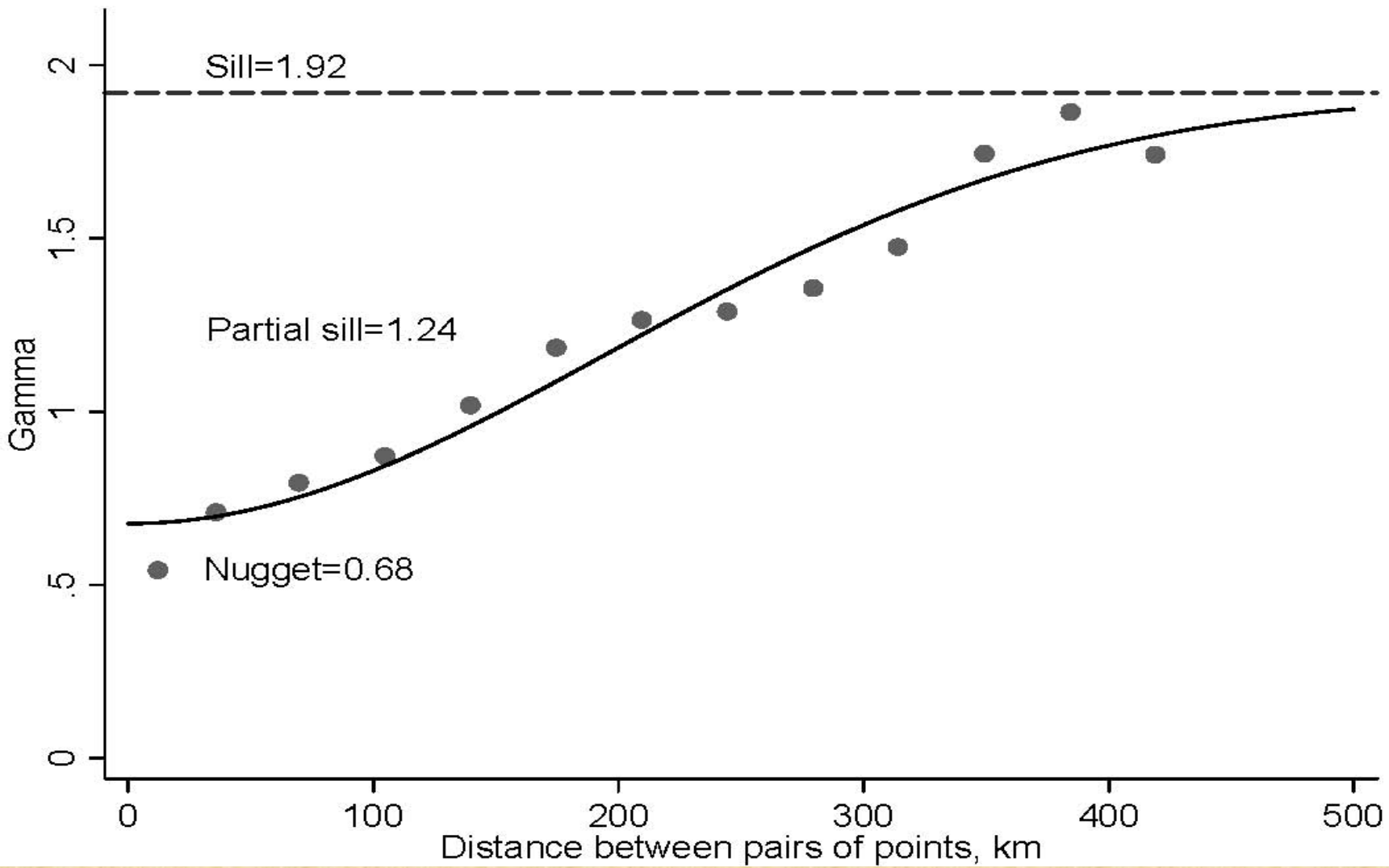
$$P(\theta,\Phi,\beta,\sigma_\theta,\sigma_\Phi,\sigma_\beta) \sim p(y|\theta,\Phi,\beta)\, p(\theta,\sigma_\theta)\, p(\Phi,\sigma_\Phi)\, p(\beta|\sigma_\beta)\, p(\sigma_\theta)\, p(\sigma_\Phi)\, p(\sigma_\beta)$$

Application: Mapping esophageal cancer SIR in the Caspian region of Iran

Sex	No. of Cases	Incidence Rate	1970 world population	2000 world population	Moran's I#
Male	891	8.10	12.16	14.61	0.28
Female	810	7.23	11.27	12.73	0.30
Both sexes	1693	7.67	11.72	13.71	0.22

E(I) for all tests are -0.0066, and p-values for Moran's I were less than 0.001 for analyses

Gaussian semivariograms fit to the empirical semivariograms points



Model fitting

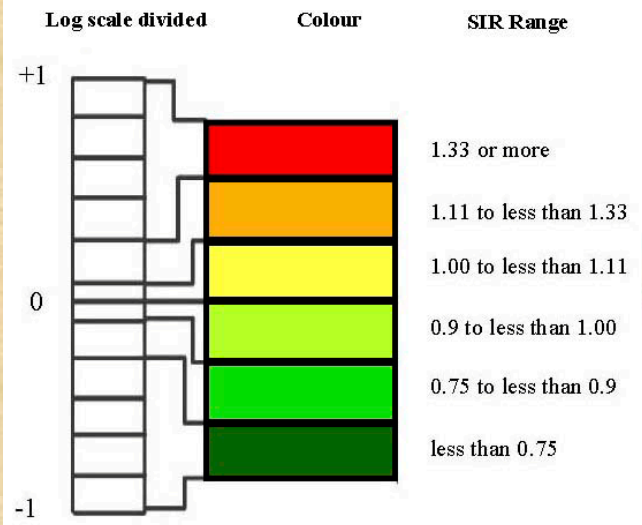
- WinBUGS was used to perform 200,000 simulations from the full conditional posterior distributions.
- Three parallel sampling chains were run with different initial values.
- The first 50,000 were discarded as burn-in.
- The three models described above had different burn-in periods, with slower convergence for the more complex models.

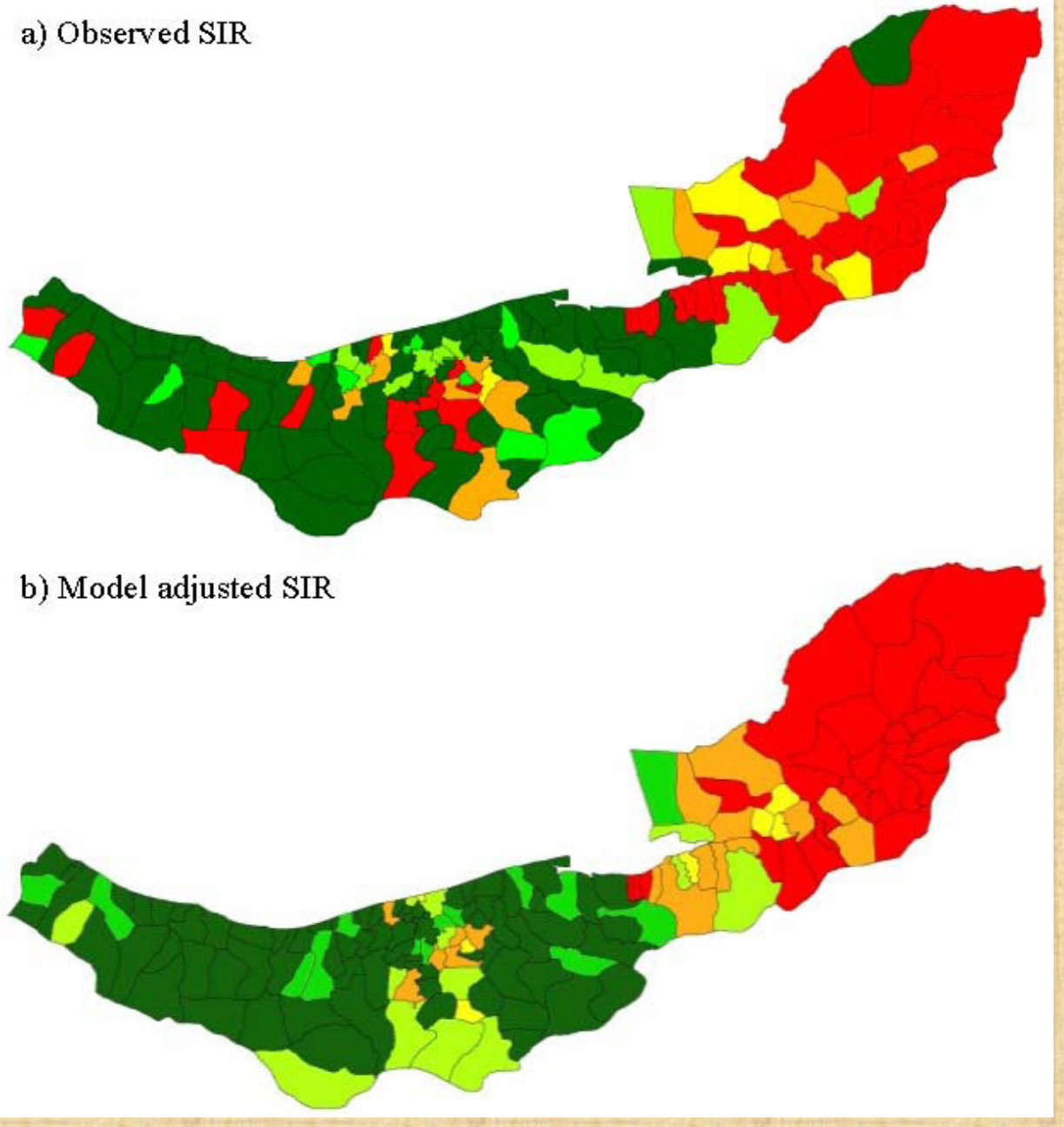
Goodness of fit comparison for three selected models: non spatial structure, joint model with nonspatial and distance-based spatial structure, and joint model with nonspatial and neighbourhood-based spatial structure

Model	ρ_D[1]	DIC[2]	MAPE[3]	MSPE[4]
Heterogeneity	78.3	661.4	2.4	15.5
Distance-based	124.1	658.7	2.0	10.4
Neighbourhood-based	61.9	649.2	2.1	10.2

1. the effective number of parameters
2. Deviance Information Criterion
3. Mean absolute prediction error
4. Mean squared prediction error

Observed spatial pattern (a), and adjusted spatial pattern of esophageal cancer's SIR from a joint model with nonspatial and neighbourhood-based spatial structure (b)





Monitoring MCMC convergence

- i)Simple graphical methods (working on single/multiple chains)
- ii) Methods using ratio of dispersions (multiple chains)
  - Gelman-Rubin Potential Scale Reduction Factor