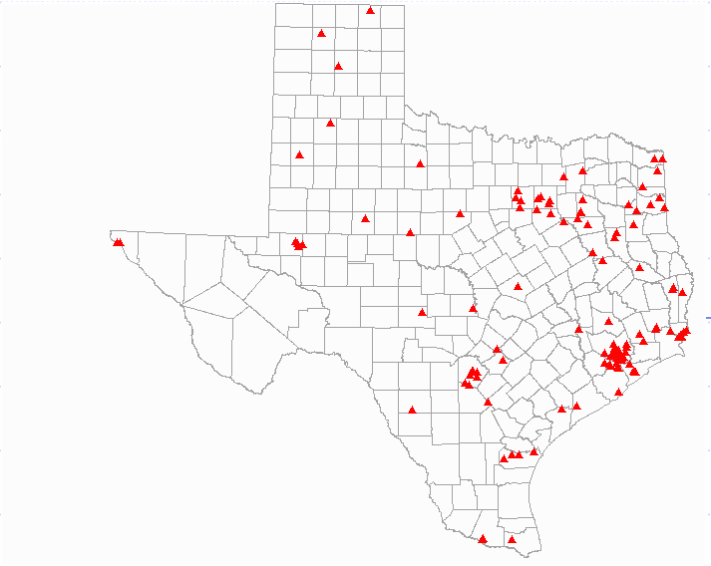
Bayesian Risk Mapping of Childhood Cancer Around Texas Superfund Sites

James A. Thompson, DVM, DVSc



“Bayesian” refers to a method of inference

Reverend Thomas Bayes

- The Bayesian result (posterior distribution) incorporates prior knowledge
  - Especially well practiced in clinical science
- At present, Bayesian analysis of research, usually uses uninformative priors
- Why?

“Bayesian” refers to a method of inference

Reverend Thomas Bayes

- Results have a “friendly” interpretation
  - Especially in clinical science
- The ability to implement complex models
  - Hierarchal models
  - Non-Gaussian measurements
  - Multivariate outcomes
    - “Multi-variable” means the modeling of multiple factors
    - “Multivariate” means the modeling of multiple outcome

Results have a “friendly” interpretation (a patient interpretation vs. a population interpretation)

- Cow is recumbent / unable to rise
- Odds of hypocalcemia (milk fever) depends on the amount of times since calving
  - In the week following calving the odds are 4:1
  - At 3 months post-calving the odds are 1:4
- Absence of a pupillary light reflex has an OR of 4

Bayesian Advantage: Results have a "friendly" interpretation (a patient interpretation vs. a population interpretation)

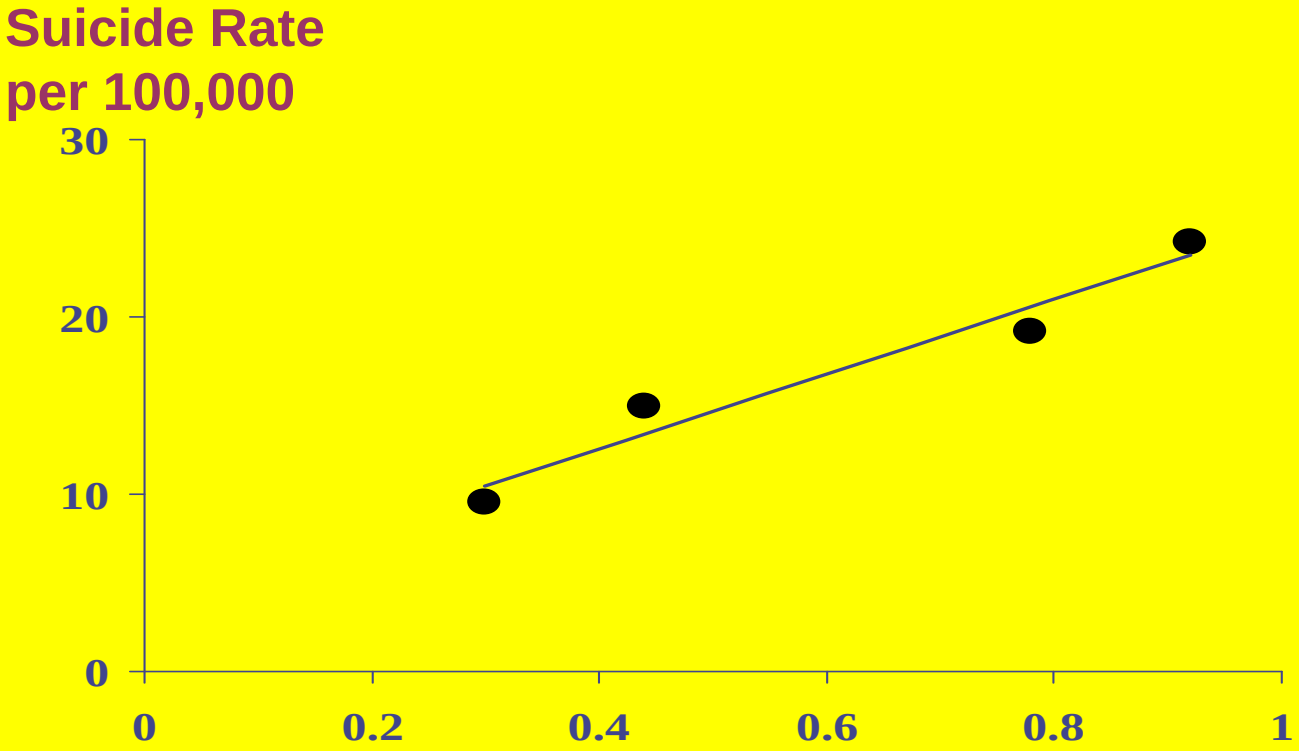
- Cow is recumbent / unable to rise and has an absent pupillary light reflex
  - In the week following calving the posterior likelihood of hypocalcemia is 94%
  - At 3 months post-calving the posterior likelihood o hypocalcemia is 50%
- Knowledge of the prior likelihoods and Bayesian updating are the foundations of clinical science

Bayesian Advantage: Hierarchal modeling for simultaneous modeling of cluster and individual effects

- In veterinary clinical science priors are based on two levels of factors
  - Herd
    - Tie stall
    - Free stall
  - Individual
    - Age
    - Milk production

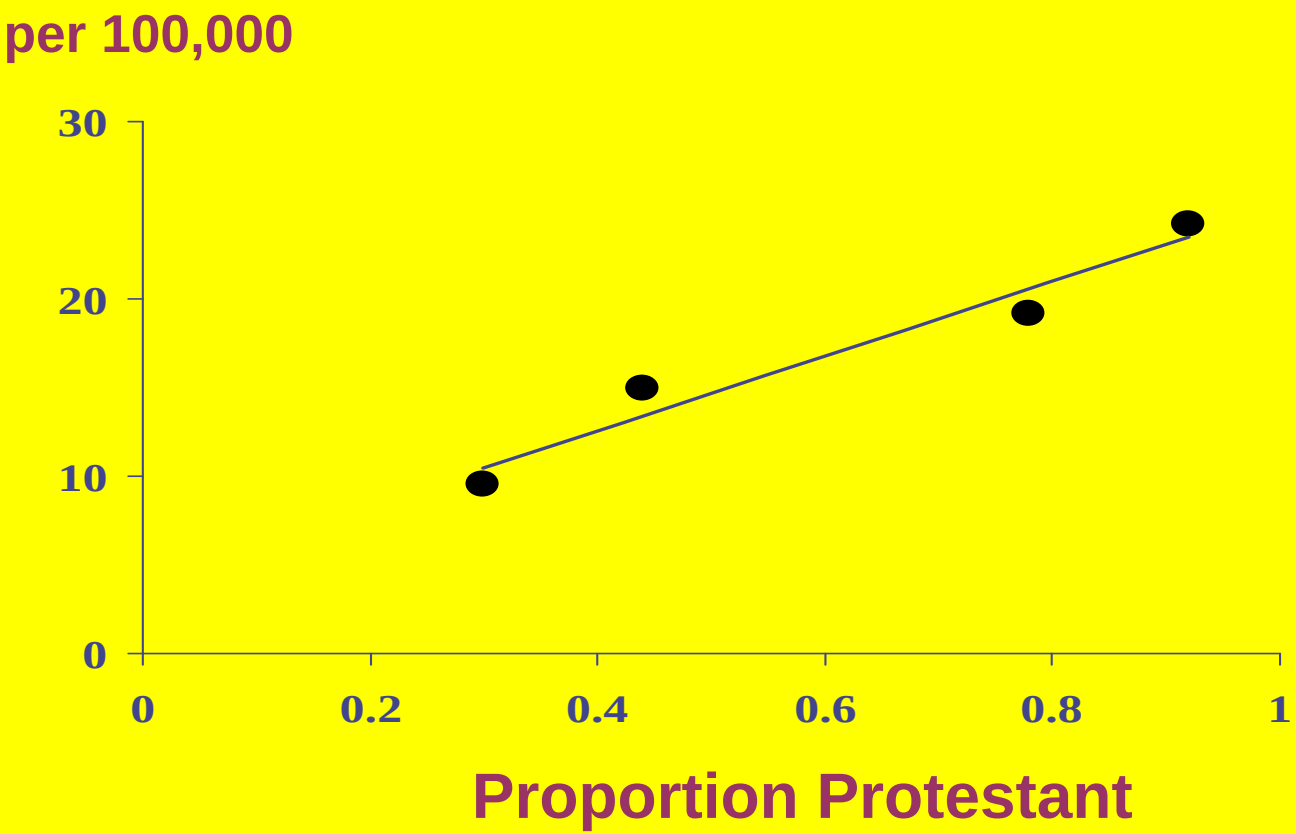
Bayesian Advantage: Hierarchal modeling for simultaneous modeling of cluster and individual effects

- In my graduate training experience it was suggested
  - Never measure herd variables if cow risk factors were being investigated
    - That would be an “ecologic fallacy”
  - Never measure cow variables if herd risk factors were being investigated
    - That would be an “atomistic fallacy”



In 1951, Durkheim, reported an association between the rate of suicide and proportion of four Prussian provinces of Protestant religion.

How might the ecologic fallacy be responsible for this association?



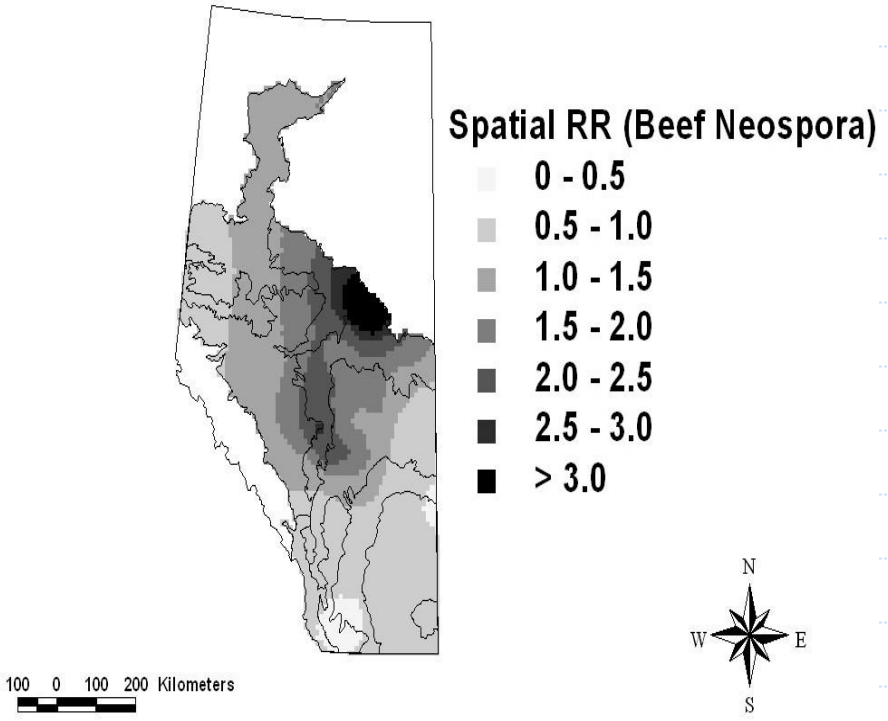
As non-Protestants in each province became more in the minority, these individuals, rather than Protestants, may have been more likely to commit suicide.

Bayesian Advantage: Hierarchal modeling for simultaneous modeling of cluster and individual effects

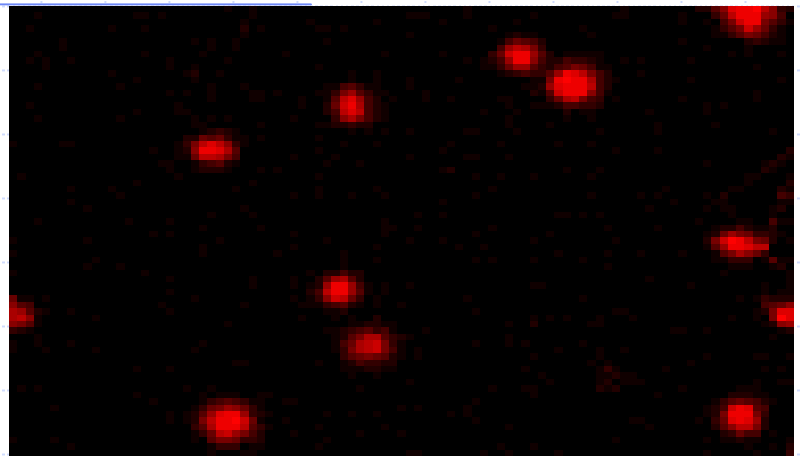
- In veterinary medicine, it has been impossible to avoid hierarchal models
  - These models have predominated in farm animal epidemiology
    - Adjustment for over-dispersion
    - Random effects models (generalized linear models)
    - Mixed effects models (generalized linear mixed models)
      - Penalized quasi-likelihood (also called Empirical Bayes)
- What if the “cluster/herd” effects are not “random” but in fact have a spatial distribution?

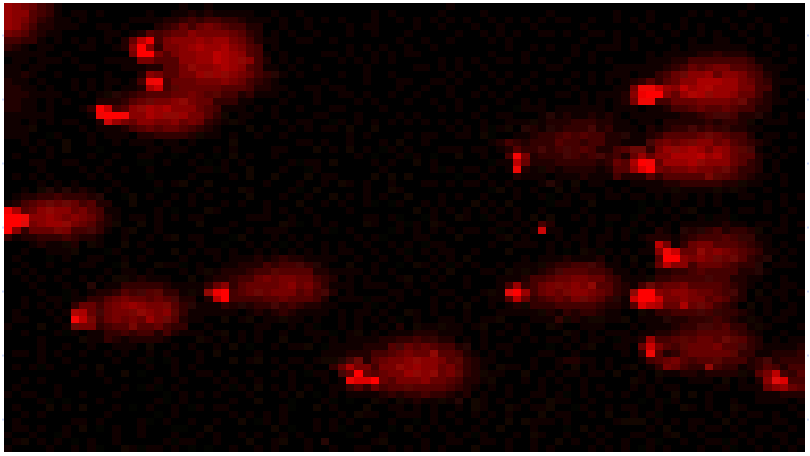
Bayesian Advantage: Hierarchal modeling for simultaneous modeling of cluster and individual effects

- It is possible with the Bayesian convolution geo-statistical model to parse herd effects into locational and management effects while controlling for individual effects



Sentinels in Environmental Investigations

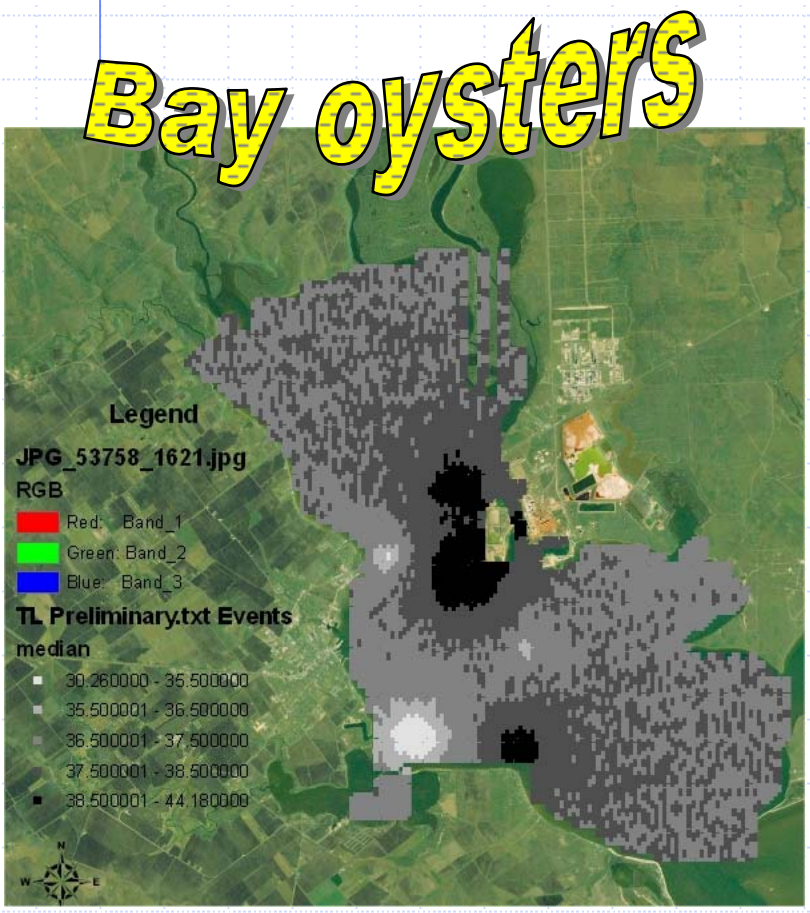


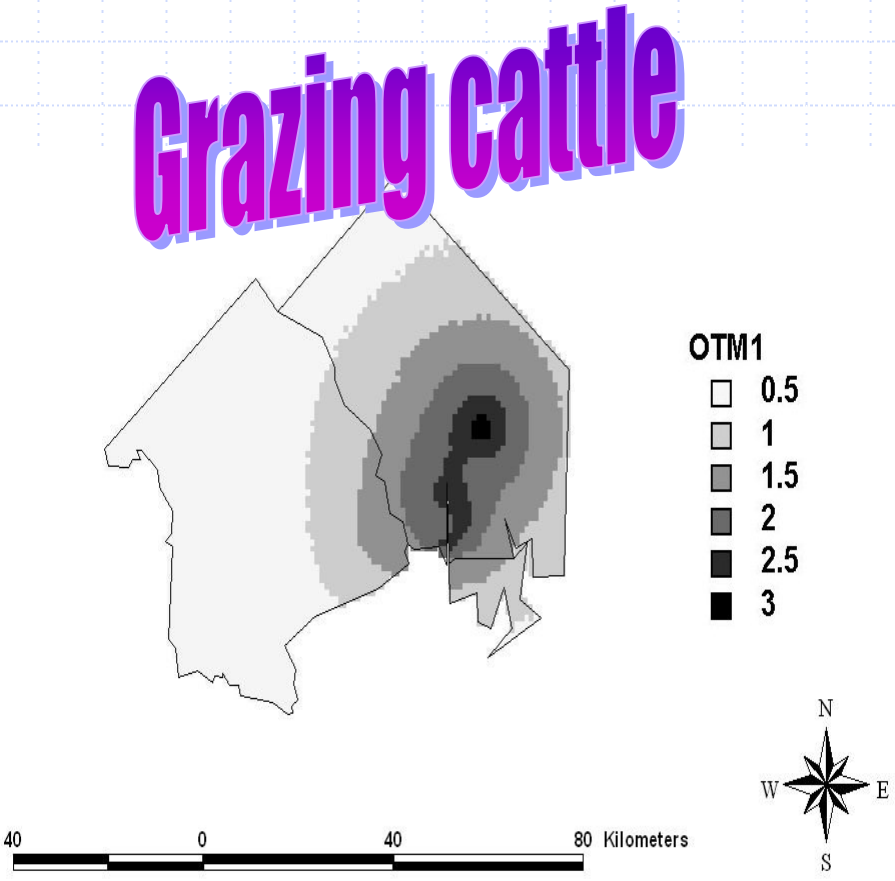


DNA fragments: comet assay

- DNA is released from its packing
- An electric field (electrophoresis) is applied and the DNA can migrate
- Small pieces migrate further than normal DNA
- DNA is fluorescent-stained

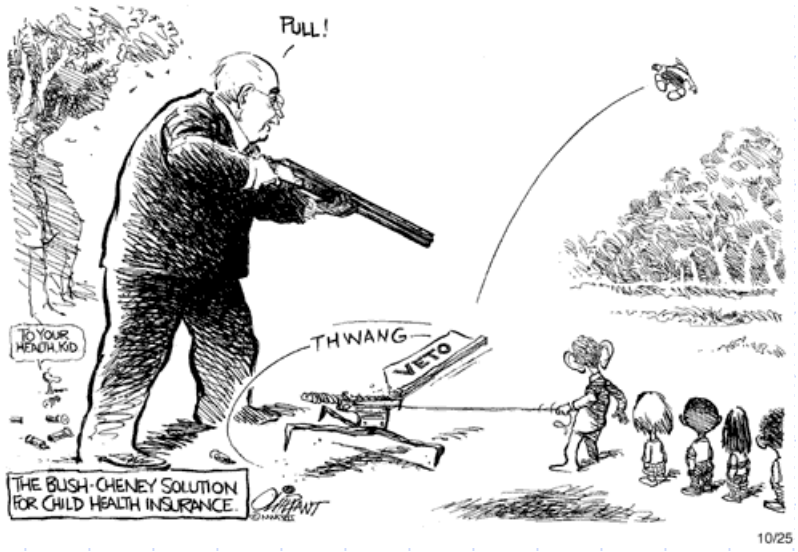
Bayesian Geo-statistical modeling of Comet Results





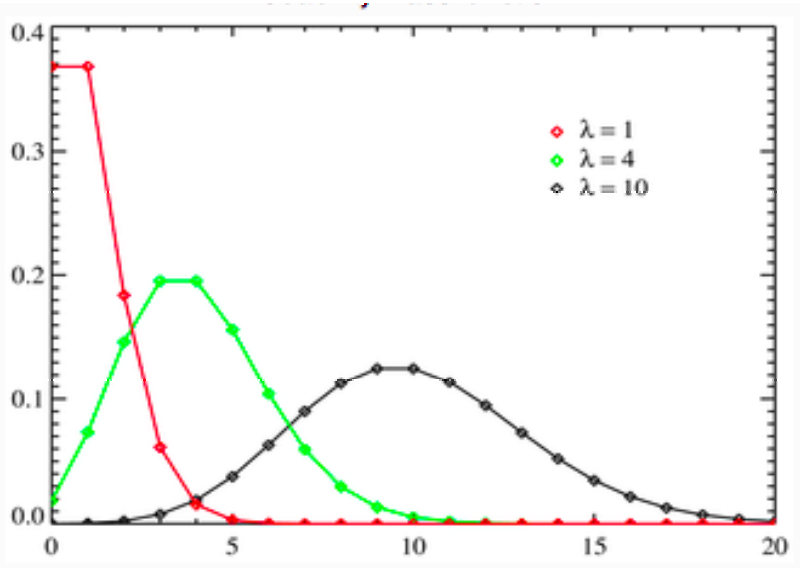
Complex modeling: Implementation by Markov Chain Monte Carlo (MCMC)

Example: skeet shooting



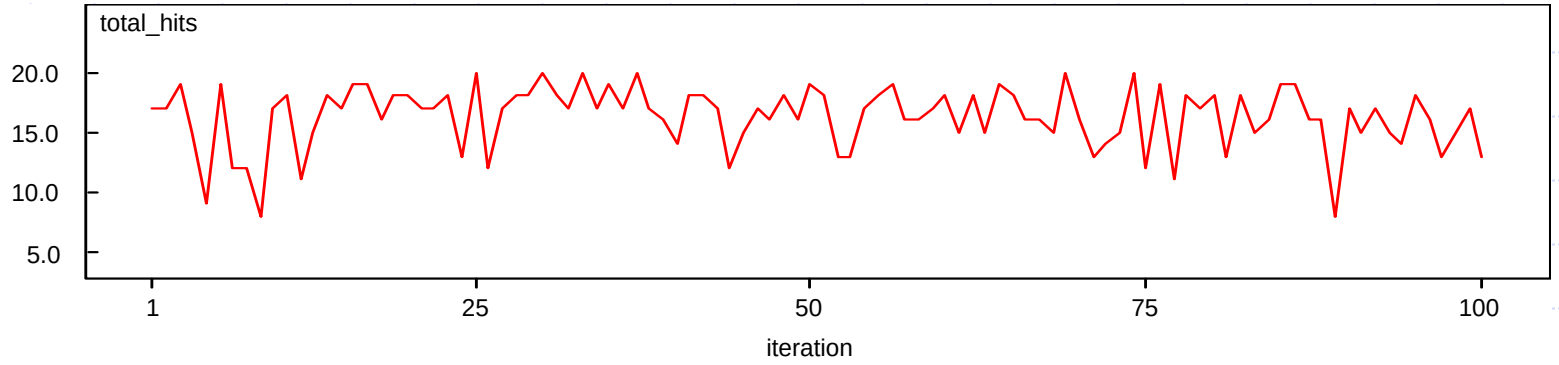
- For the first clay pigeon the likelihood of a hit is 50%
- After a miss, the next shot has a likelihood of a hit of 50%
- After a hit, the next shot has a likelihood of a hit of 90%
- What is the most likely number of hits for 20 tries?

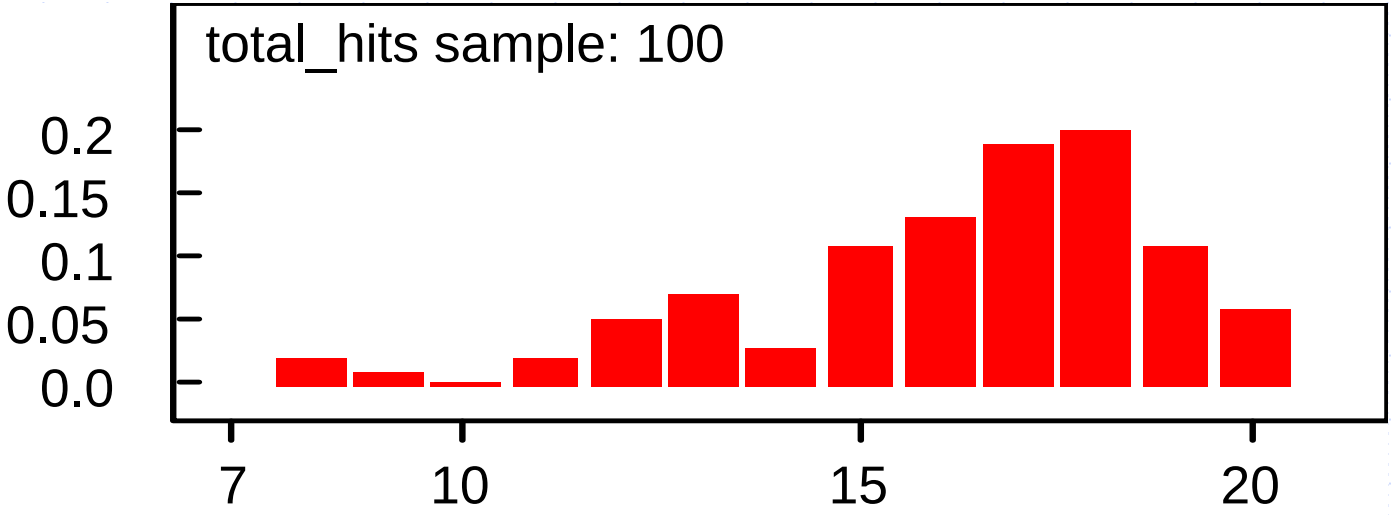
Traditional Maximum Likelihood Estimation (MLE)



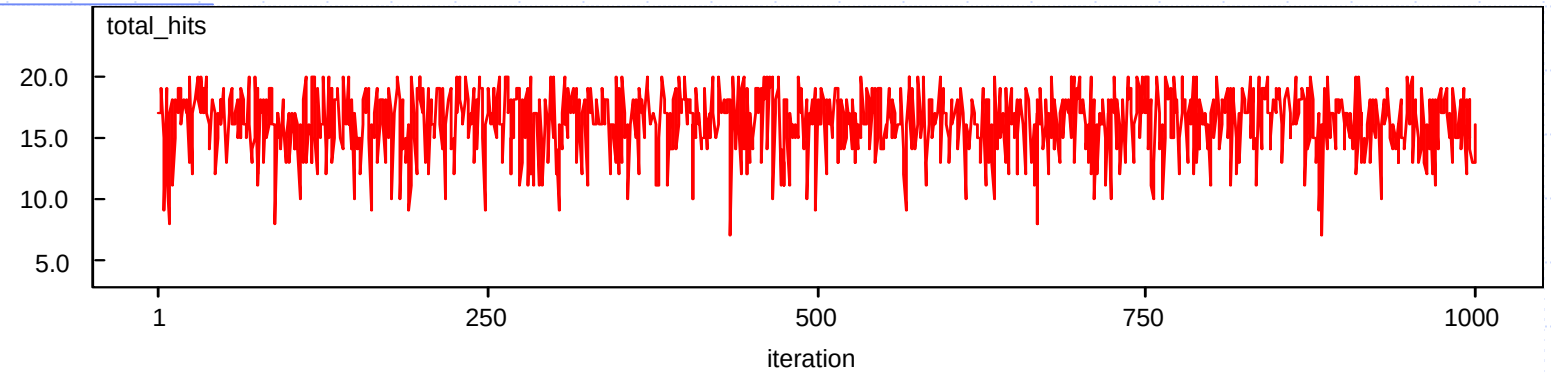
- Determine the likelihood function (maybe the log likelihood function)
- Take its derivative
- Set the derivative to zero
- Determine if the inflection point is a minimum or maximum

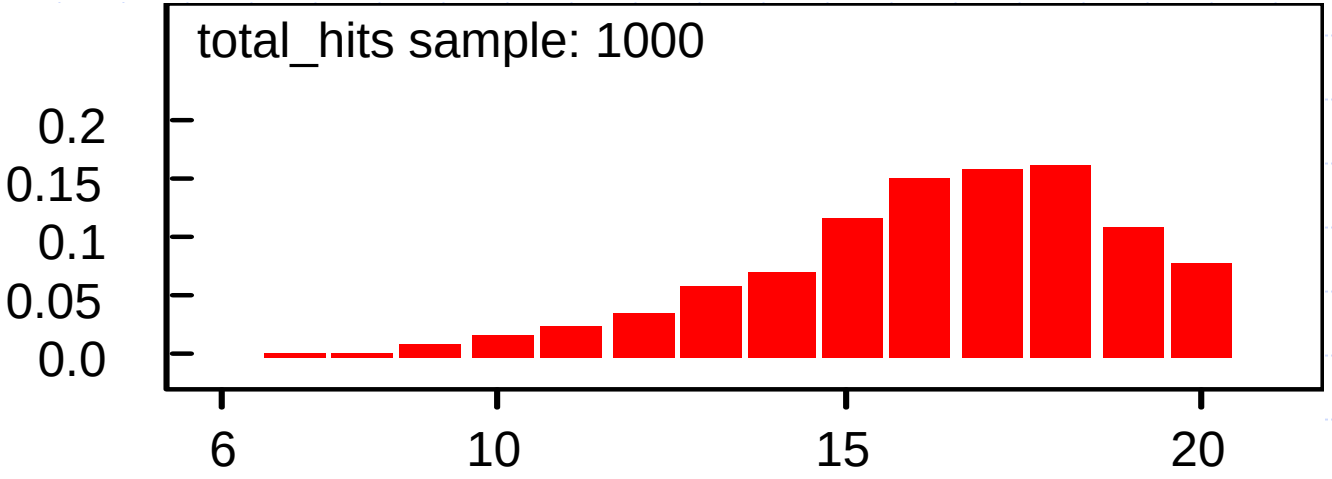
MCMC: set up a computer to “simulate” the chain of probability – do it 100 times



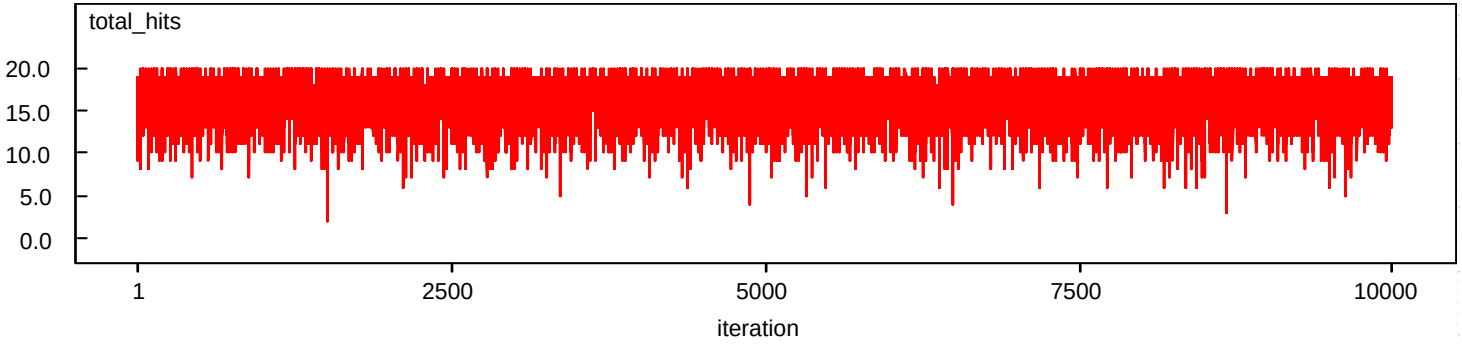


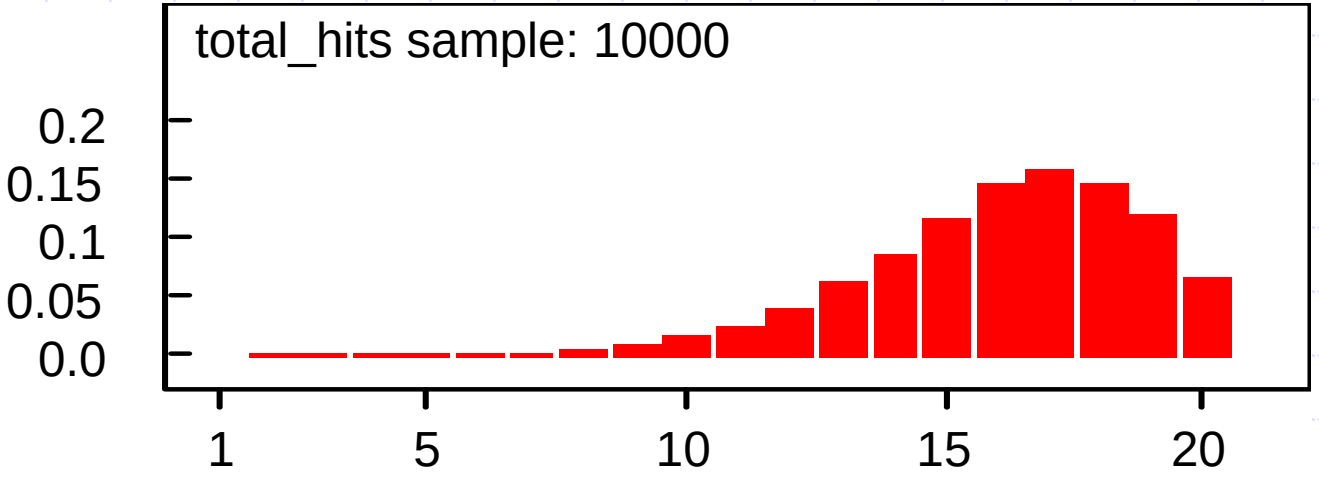
MCMC: set up a computer to “simulate” the chain of probability – do it 1,000 times



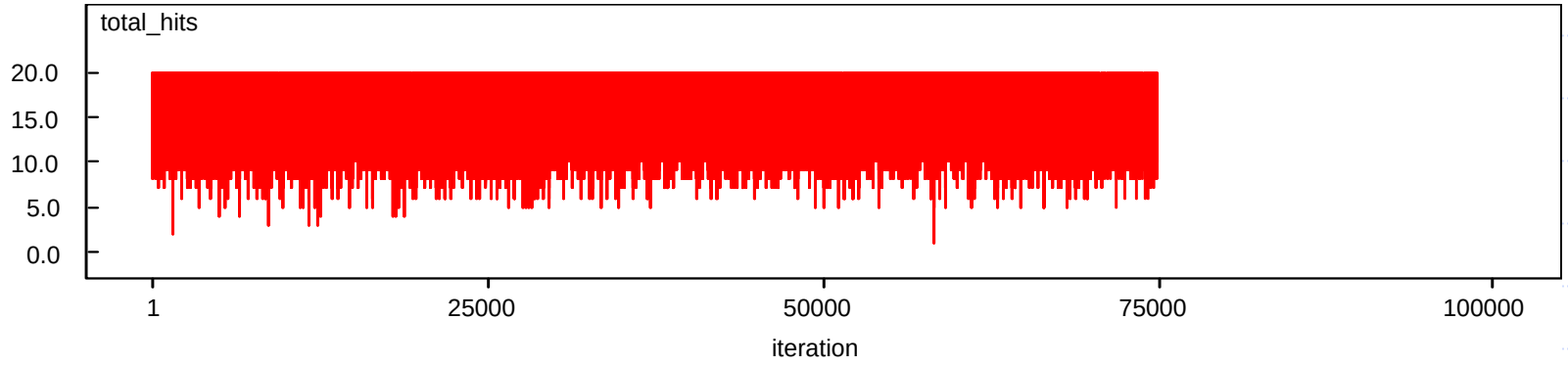


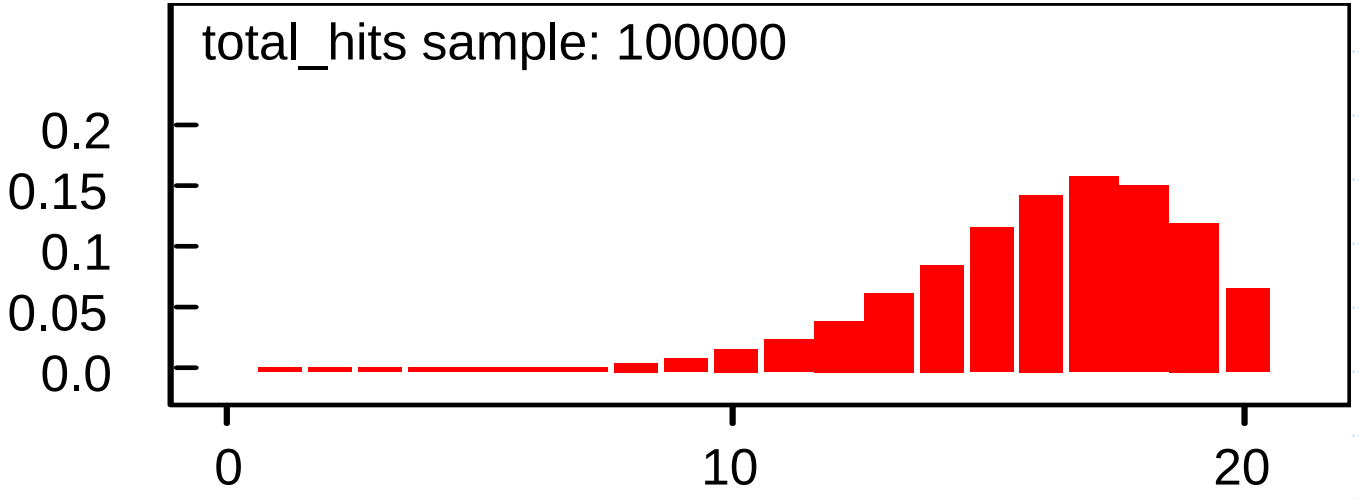
MCMC: set up a computer to “simulate” the chain of probability – do it 10,000 times



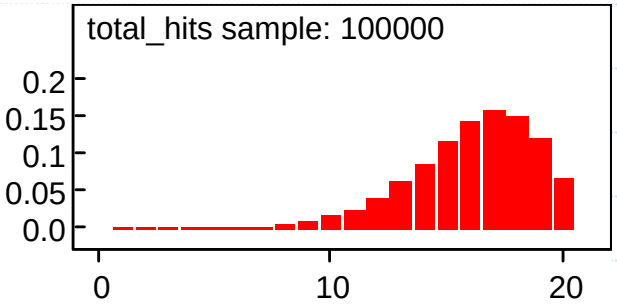


MCMC: set up a computer to “simulate” the chain of probability – do it 100,000 times





MCMC: set up a computer to “simulate” the chain of probability – do it 100,000 times



- We have found the most likely result (17 hits)
- We know the entire distribution of outcomes (the full posterior)
- More statistics are readily available without distributional assumptions
  - Mean
  - Mode
  - 95% confidence interval

Childhood Cancer

- Leading cause of fatal disease in young children
  - Children are especially sensitive to mutagenesis
  - Some cancers are increasing in incidence in developed countries and occurring at younger ages
    - Brain cancers
    - Leukemia
- Compared to adult cancers each cancer type is rare
- Causes are poorly known and believed to be complex
  - Hierarchal
    - Individual risks
    - Locational risks
  - Multivariate

Childhood Cancer is Multivariate

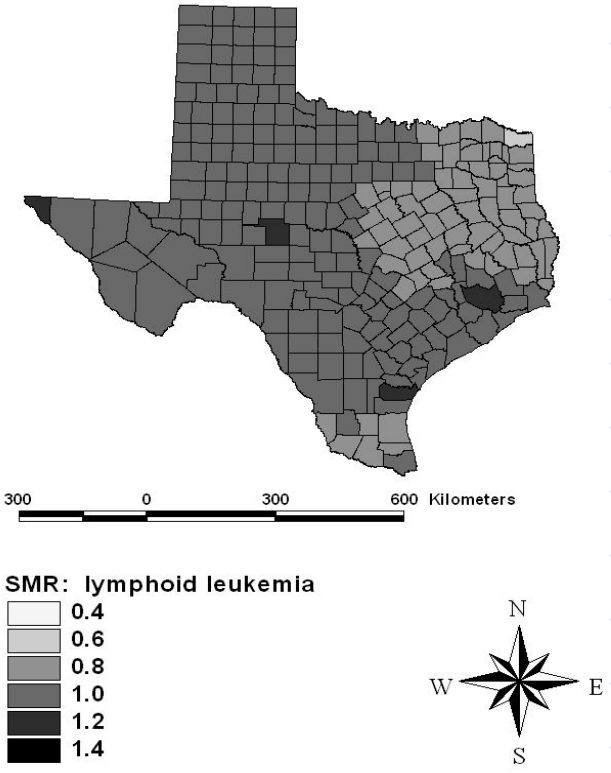
ICCC-3

I. Leukemia
II. Lymphoma
III. Central Nervous System (CNS)
IV. Peripheral Nervous System
V. Retinoblastoma
VI. Renal
VII. Hepatic
VIII. Bone
IX. Soft tissue
X. Gonad
XI. Epithelial
XII. Other

- Leukemia
  - Acute lymphocytic leukemia
  - Acute myeloid leukemia
  - Other
- CNS
  - Ependymoma
  - Astrocytoma
  - Embryonal
  - Gliomas
  - Other
- Other diseases
  - Birth defects
  - LBW/VLBW

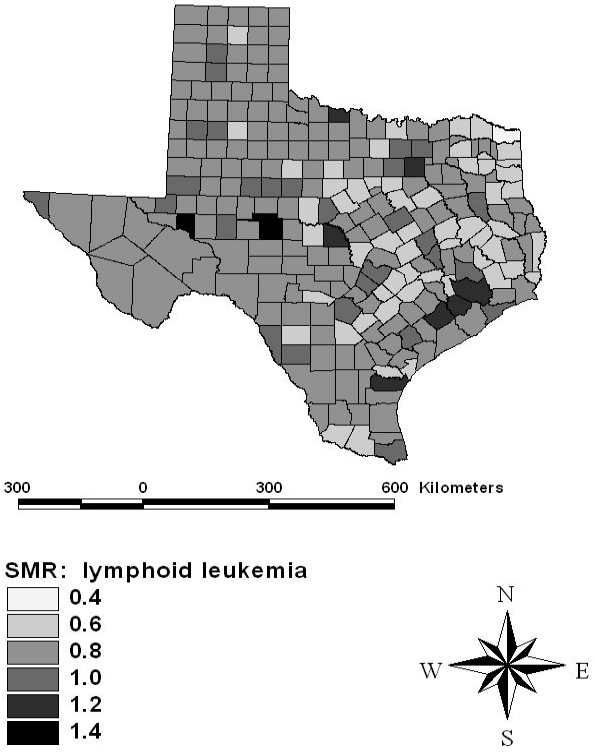
Bayesian spatial modeling (ICAR)

- Intrinsic Conditional Autoregressive modeling adjusts an area's risk toward the risk in neighboring counties
- The extent of smoothing is determined by the data (not by the prior) when uninformative priors are used
- Even the common cancers are so rare that spatial smoothing toward the overall mean is marked



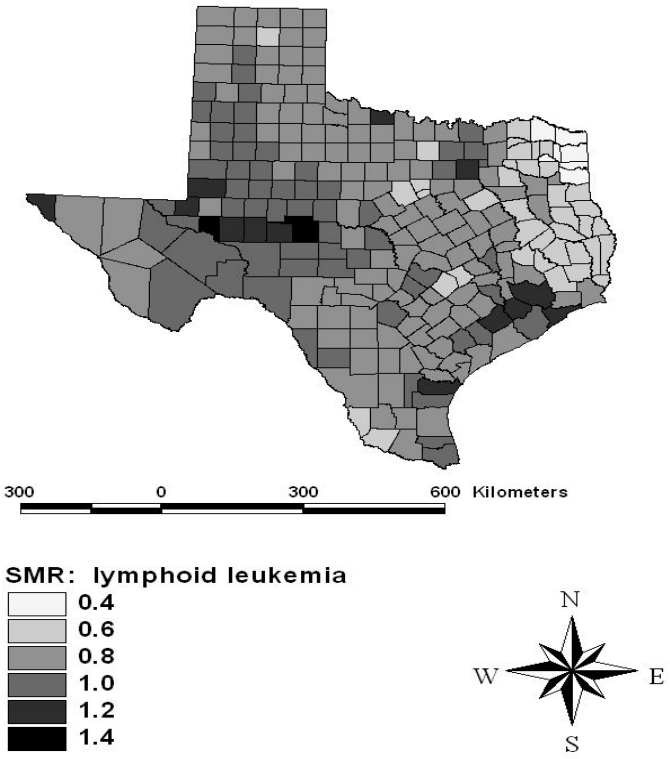
Bayesian Multivariate Modeling

- Multivariate modeling accounts for the correlation in log-SMR for multiple cancer types within each “area”
- The amount of smoothing is determined by the extent of correlation in the data when non-informative priors are used
- Preserves more extreme SMR



Bayesian multivariate and spatial (MCAR)

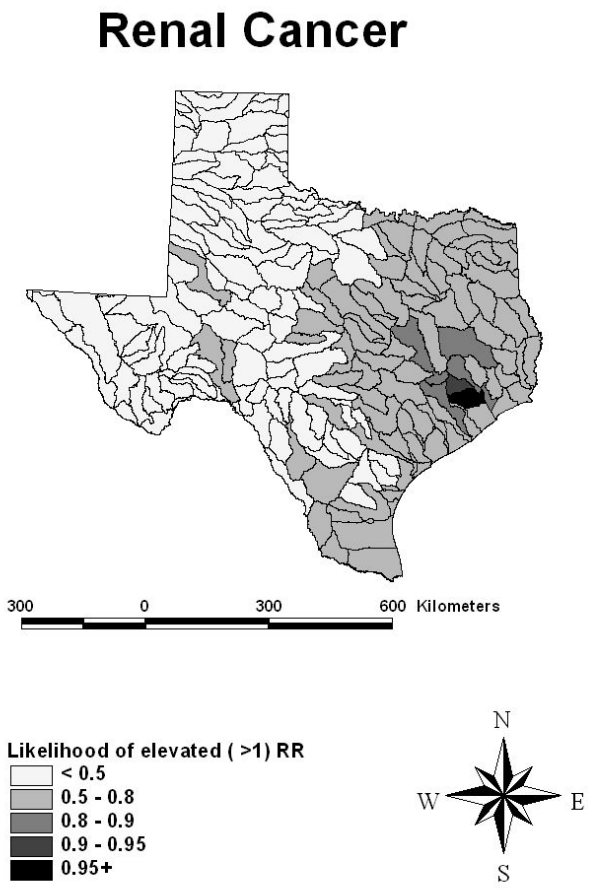
- Combining multivariate and spatial modeling preserves more extreme SMR and spatial patterns



Model Comparison

Model	Deviance Information Criterion
Spatial	3972.2
Multivariate	3888.0
Spatial and Multivariate (MCAR)	3873.9

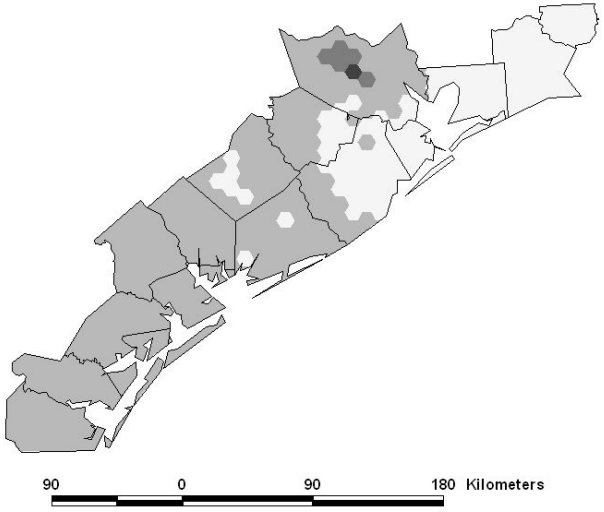
MCAR modeling is not restricted to administrative units (counties)



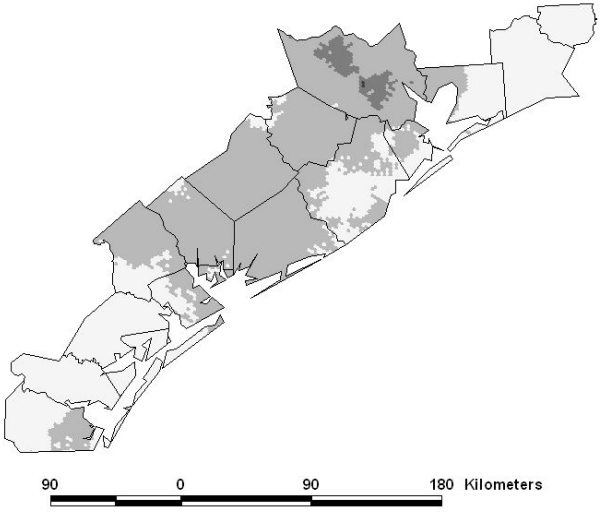
- All births were plotted based on the geocoordinates of mother's living location at the time of birth
- Birth data were merged with cancer data to identify patients/non-patients
- MCAR modeling was performed for watersheds
- Parameter of interest was the posterior probability of an increased RR

Risks can be plotted with increased detail to evaluate risk gradients

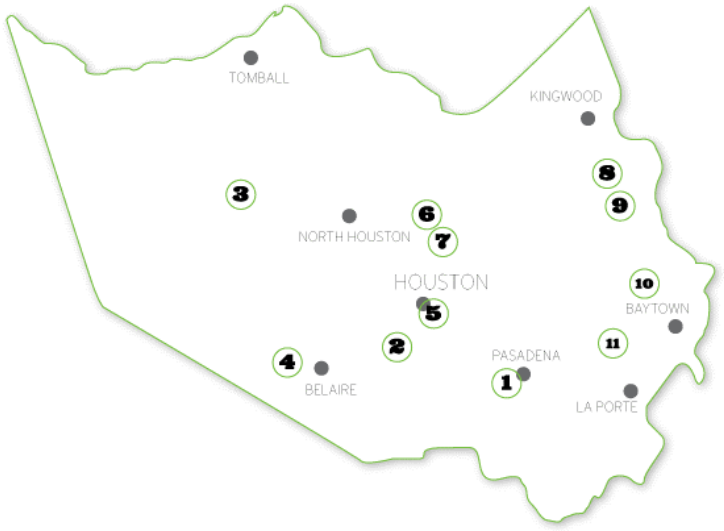
Childhood Renal Cancer
100 km Pixels

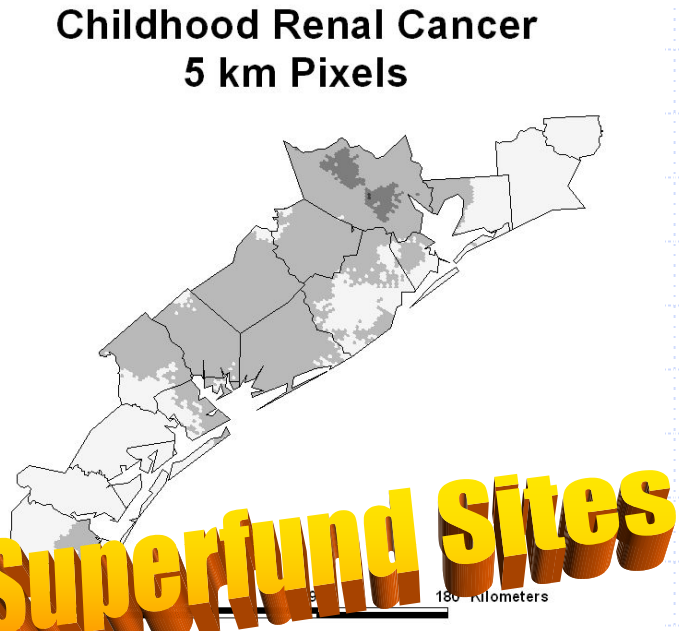


Childhood Renal Cancer
5 km Pixels



Risks can be plotted with increased detail to evaluate risk gradients





Harris County: 13 Federal Superfund Sites

Pixel-specific MCAR Modeling

- Y_{ki}, the counts of childhood cancer histotype k in pixel i are modeled as independent Poisson distributions
  - $Y_{ki} \sim Poisson(E_{ki} * RR_{ki})$
- $Log(RR_{ki})$ is modeled linearly as a function of a flat, improper intercept and the spatial effect S_{ki}
  - $Log(RR_{ki}) = alpha_k + S_{ki}$
- S_{ki} is a matrix of histotype- and pixel-specific risks and is the Multivariate Intrinsic Conditionally Autoregressive (MCAR) prior

The MCAR Prior

- The columns of dimensional matrix ***S*** represent the spatial units (pixels) and rows represent the variables (cancer histotypes).
- Within each pixel the log(RR) for cancer histotypes are multivariate normal represented by a k x k matrix. The diagonal of the matrix representing the variance of the risk estimate and the off-diagonals represent the covariances between pairs of cancer histotypes
  - The k x k matrix is given an Inverse Wishart (R, h). The Inverse Wishart distribution has h degrees of freedom, equal to the number of cancer histotypes and **R** is a k x k identity matrix.
- Among pixels, the log(RR) for each cancer histotype is conditionally autoregressive (correlated to the log(RR) for the same cancer in pixels that are first order neighbors
  - If we let 1 pixel be anything (a flat, improper prior), the rest of the risks can be estimated as conditional to the first using our Markov Chain Monte Carlo. When all the values are estimated, the intercept can be subtracted from all values so that they sum to zero.

Risk Gradients for Childhood Cancer Around Superfund Sites

1. Will provide useful causal evidence
2. Will identify locations for further investigation