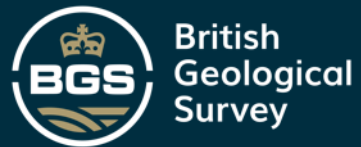




MATT ARRAN

# FOSS for large- dataset geostatistics



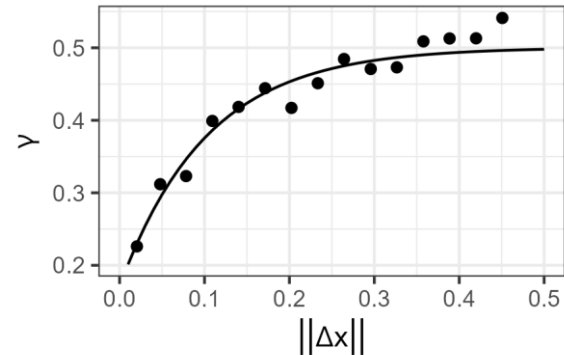
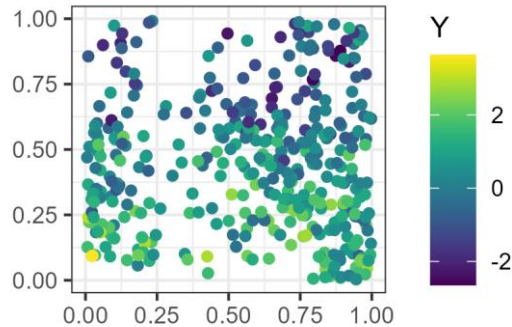
# Outline

- Motivation
- New software for variogram calculation
- Existing software for model fitting
- Conclusions

# Motivation

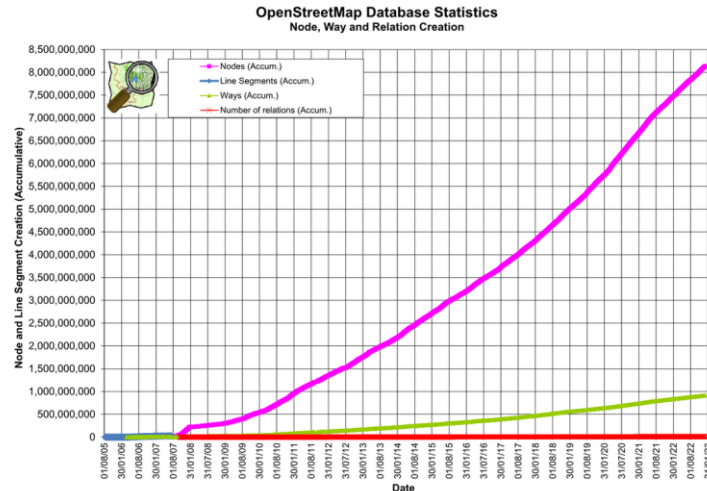
# Geostatistics

- Tobler's first law of geography:  
“everything is related to everything else, but near things are more related than distant things”
- Standard approach:
  1. Examine how correlation varies with distance, via an empirical variogram
  2. Fit a model, via an inferred spatial covariance matrix
  3. Predict values at new locations, via kriging



# Large datasets

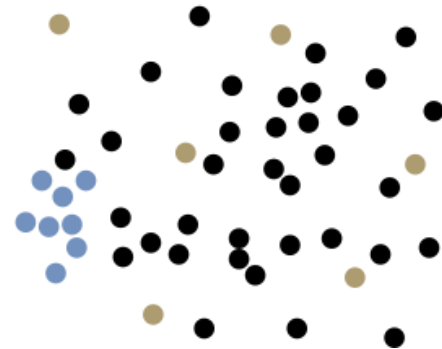
- Datasets are growing with increasing computing power
- Geostatistics' computational costs grow faster than dataset size  $n$ 
  - # of inter-datapoint distances/covariance matrix entries:  $O(n^2)$
  - # of operations for covariance matrix decomposition:  $O(n^3)$



# Variogram calculation

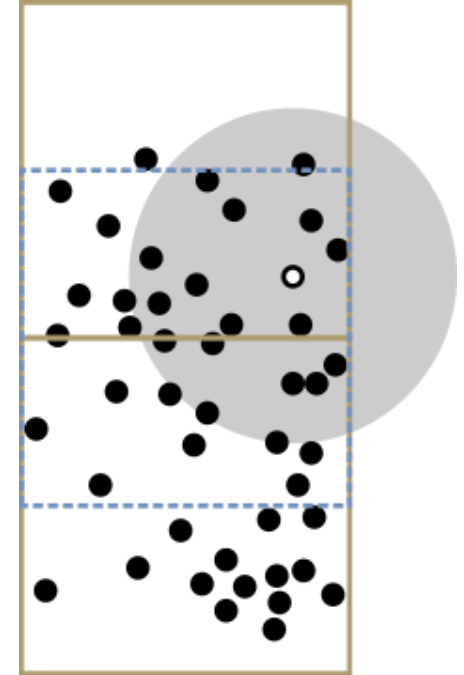
# Sampling risks

- Excessive spatial span:
  - Covariate variation dominates variability
  - Spatially correlated effect unconstrained
  - Inefficiency from low intercorrelation
- Insufficient spatial span:
  - Correlation dominates variability
  - Covariate dependence unconstrained
  - Inefficiency from high intercorrelation



# Efficient sampling

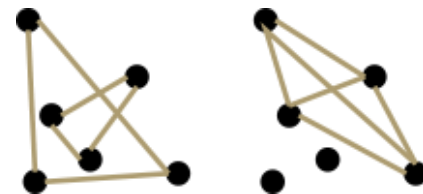
- Estimate maximum significant correlation length  $\lambda$
- Divide datapoints into  $2\lambda \times 2\lambda$  boxes with overlaps  $\lambda$ 
  - All significant pair-interactions within some box
- From each box, sample # pairs  $\propto$  # datapoints
  - First approx. to Reilly and Gelman (2012)
- # of comparisons  $O[A\lambda^2\rho\rho_{min}]$  rather than  $O[(A\rho)^2]$   
(for Area  $A$ , typical and low datapoint density  $\rho, \rho_{min}$ )



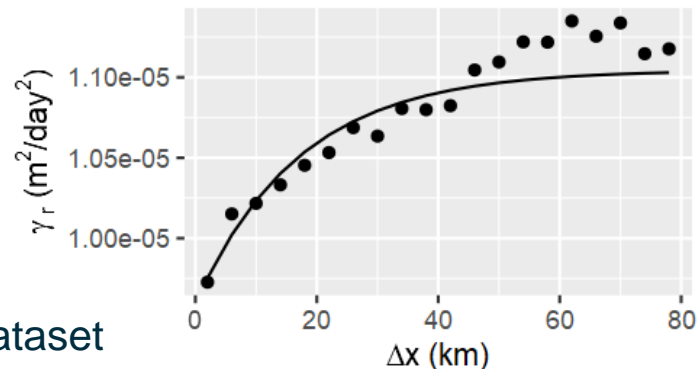


# Variogram implementation

- Development version implemented in R
- Euclidean and WGS84 distances supported
- Options for subsampling:
  - Complete coverage
  - Complete network
- Options for estimator:
  - Matheron (1962)'s
  - Cressie & Hawkins (1980)'s
  - Genton (1998)'s
- Applied to 78,878-datapoint Punjab groundwater dataset
  - ~ 1 hour to identify ~ 1.6 million interacting pairs



$\Delta t < 30$  days



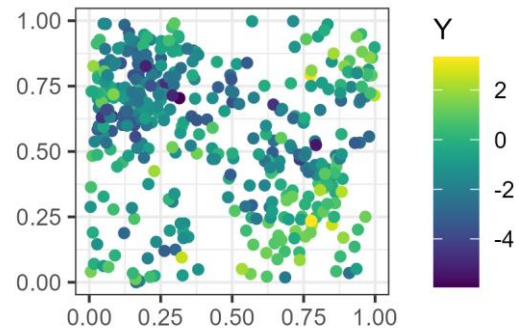
# Model fitting

# Published approaches

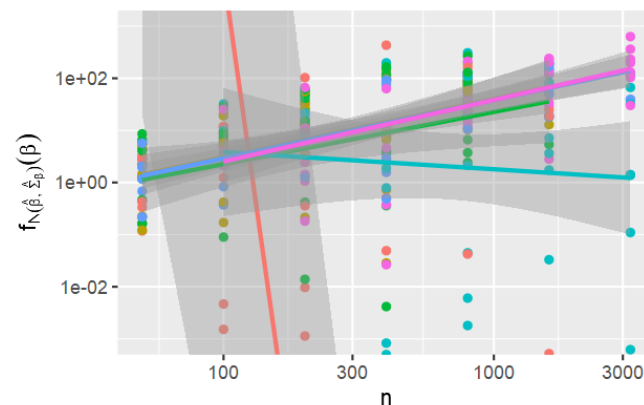
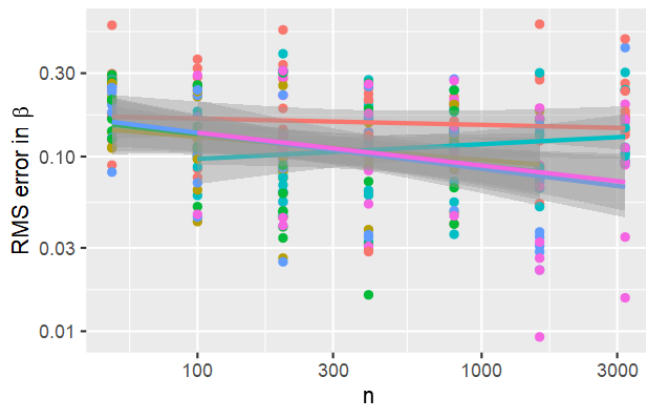
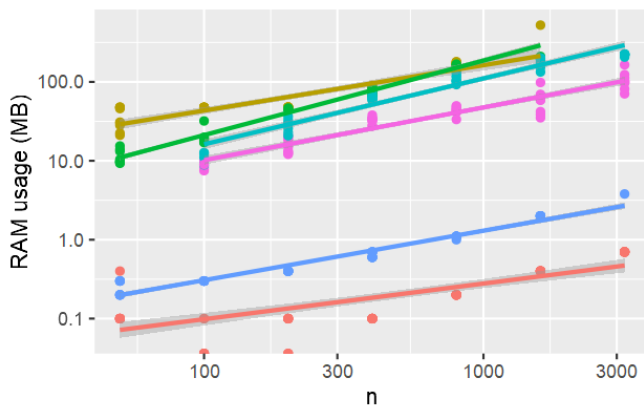
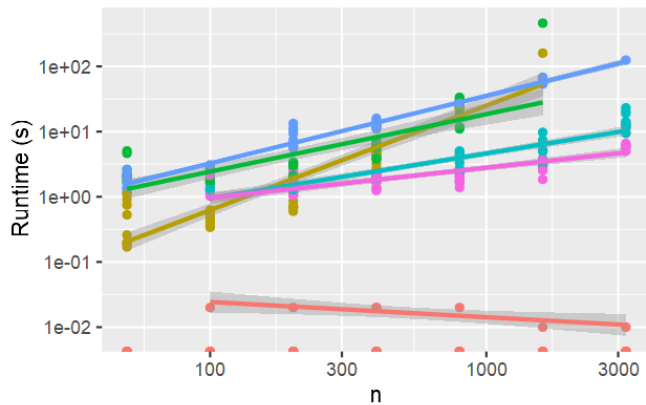
- Take advantage of low correlation at distance:
  - Tapered covariance matrices (e.g. with R package spaMM)
  - Nearest-Neighbour Gaussian Process (R package spNNGP)
- Take advantage of high correlation in proximity:
  - Fixed Rank Kriging (R package FRK)
  - Predictive Process (R package spBayes)
- Both: Multi-Resolution Approximation (Julia package MRA\_JASA, Python pyMRA)
- Solve equivalent problem with more sparsity:
  - Lattice Kriging (R package LatticeKrig)
  - INLA for equivalent SPDE (R package R-INLA)
- Few direct comparisons between different approaches

# Comparison method

- Simulate datapoints at  $n = 50, 100, \dots, 3200$  locations in the unit square
  - *Locations*: mixture of three Gaussian clusters and a uniform distribution
  - *Covariates*: one uncorrelated, one spatially correlated
  - *Errors*: exponential correlation structure with nugget
- Estimate coefficients with linear regression `lm`, as control, &:
  - Dense covariance matrix, using `spaMM`
  - Sparse, spherically tapered covariance matrix, using `spaMM`
  - Nearest-Neighbours Gaussian Process, using `spNNGP`
- Specify correlation structure:
  - As part of the model fit
  - Using the empirical semivariogram

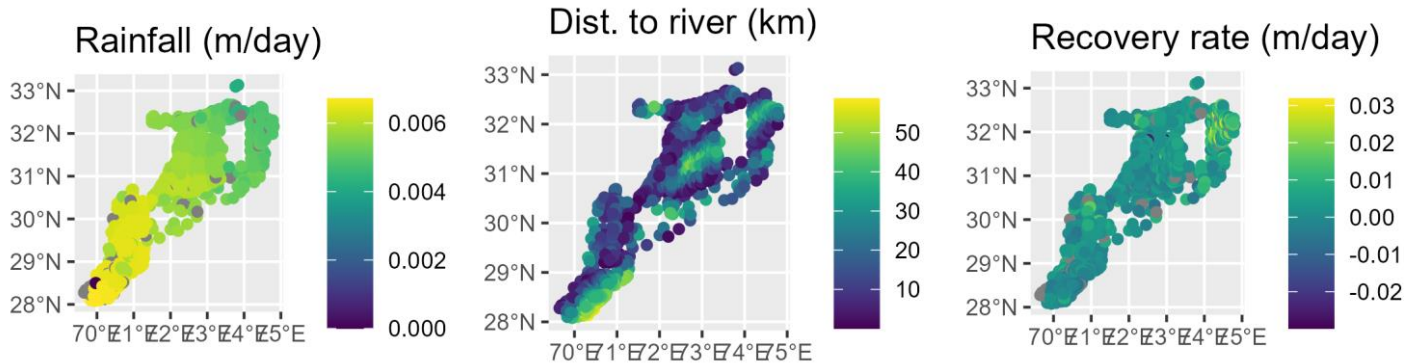


# Comparison results



# Application

- Consider monsoonal groundwater change rates in Punjab, Pakistan, 1979 – 2009
  - 41,852 records at 2,967 sites
  - Low residual inter-year temporal correlation
  - Spatially correlated covariates and errors
- Fit a linear model with a spatially correlated error term
  - Using spaMM, with a tapered covariance matrix specified from the variogram



```
----- Fixed effects (beta) -----  
[one-time computation of covariance matrix, which may be slow]  
                Estimate Cond. SE t-value  
(Intercept)    -0.001441 5.840e-04 -2.4670  
prev_prec      1.307351 5.200e-02 25.1431  
prev_pet       0.110728 9.814e-02 1.1282  
tubewell_density -58.908726 1.751e+01 -3.3642  
prev_dwl       -0.001137 5.108e-05 -22.2523  
prev_prec_anom  0.933084 1.780e-01 5.2423  
prev_irrig      0.396764 6.611e-02 6.0020  
prev_prec_anom:riv -0.005685 6.968e-03 -0.8159  
prev_irrig:canal -0.004589 3.522e-03 -1.3031  
----- Random effects -----  
Family: gaussian( link = identity )  
--- Variance parameters ('lambda'):  
lambda = var(u) for u ~ Gaussian;  
id : 1.982e-05  
# of obs: 9039; # of groups: id, 9039  
----- Residual variance -----  
phi estimate was 1.72108e-06  
----- Likelihood values -----  
logLik  
logL      (p_v(h)): 35846.38
```

# Conclusions

# Conclusions

- Computationally efficient methods key to large-dataset geostatistics
- Variety of free, open-source software available, especially in R
- New tool promising for empirical variogram calculation
- For model fitting and prediction, spNNGP's the best option when applicable



# Future work

- Rewrite variogram calculation as production code
  - Which language would be most useful?
- Test more existing methods
  - Which dataset complications are most important to model?
- Extend tests to prediction
  - Against which non-statistical methods would tests be most useful?