Survey unit selection for sample representativeness in site contamination studies

Narmadha Meenu Mohankumar
Moses Obiri
Deb Fagan
Jen Huckett

Pacific Northwest National Laboratory

PNNL is operated by Battelle for the U.S. Department of Energy

Outline

- Research question
- A scenario example (1-sample hypothesis test)
- Approach
- Results
- Conclusions, recommendations, and future work

Research question

- Investigating a site or a facility for possible contamination
- Collect samples and determine if the average contamination level of the site exceed a threshold value?
- What is the optimal sample size and sample placement to get a representative sample of the site?
- One-sample hypothesis test

Spatial autocorrelation

A site further from the river: nearby points are less correlated

A site surrounding the river: nearby points can be highly correlated

Contamination from a river

Spatial autocorrelation

Contamination from a river

A site further from the river: nearby points are less correlated

A site surrounding the river: nearby points can be highly correlated

Duplicate information

One sample hypothesis test

- To determine if the average contamination level of the site exceed a threshold value

$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

- μ: Mean contamination at the site
- μ_0: Threshold value

- Assumptions
  - Data are distributed independently and have homogeneous variance

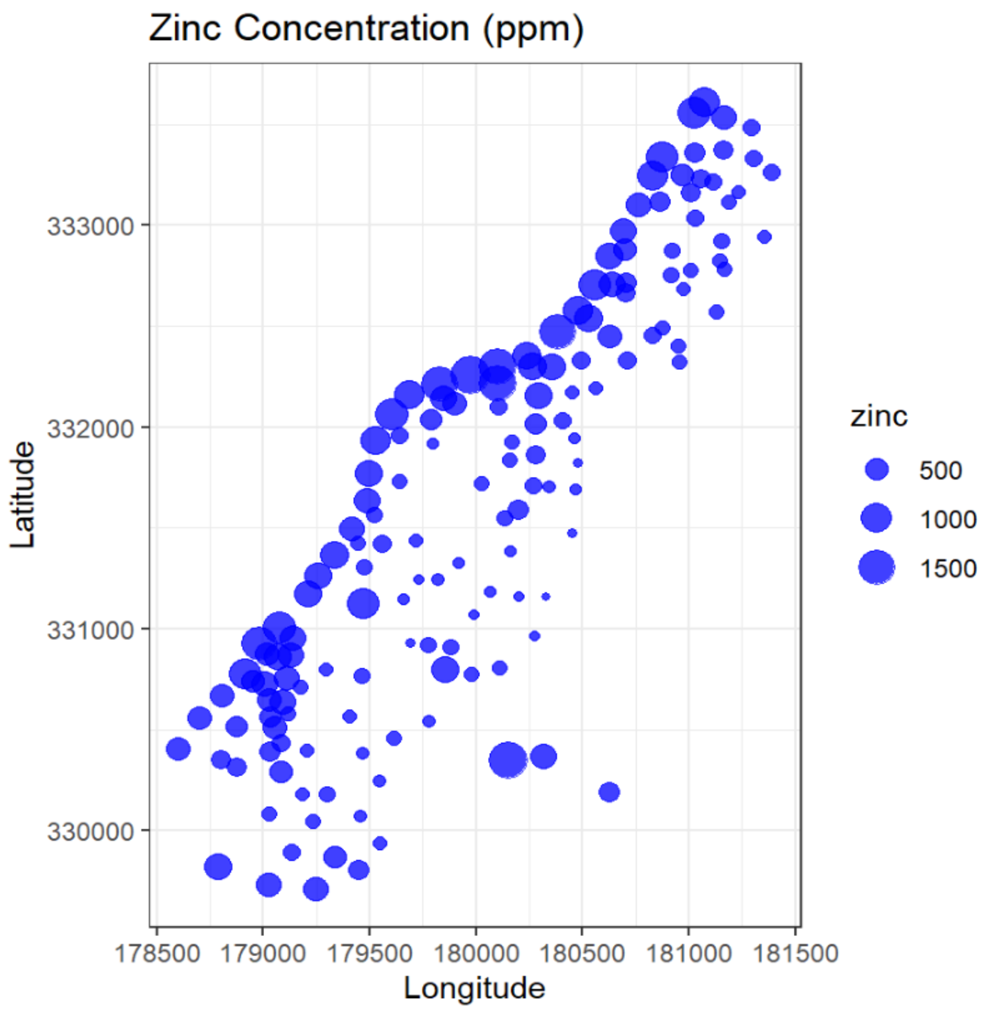
Influence of spatial autocorrelation on one-sample hypothesis test

E.g., Zinc concentration (ppm) in a flood plain of the Meuse River near the village of Stein, Netherlands

$H_o: \mu \geq 5.75$ (the mean concentration of zinc exceeds the historical average)

$H_a: \mu < 5.75$ (the mean concentration is less than the historical average).

Model	Estimate	Standard Error	t-value	p-value
Non-spatial model	0.1358	0.0580	2.3417	0.0205
Spatial model	0.7540	0.6296	1.1977	0.2329



Influence of spatial autocorrelation on one-sample hypothesis test

- Duplicate information from correlated sampled data (pseudo-replicates) violate the independence assumption

- Misleading conclusions
  - E.g., A site being classified as not contaminated when it's contaminated.

- Solutions
  - Use generalized least squares (GLS) rather than the traditional one-sample hypothesis test using ordinary least squares (OLS)
  - Collect samples in a way that do not violate the assumptions (may also reduce sampling effort).

Approach: Using Moran's I

A correlation coefficient that measures the spatial autocorrelation within a data set

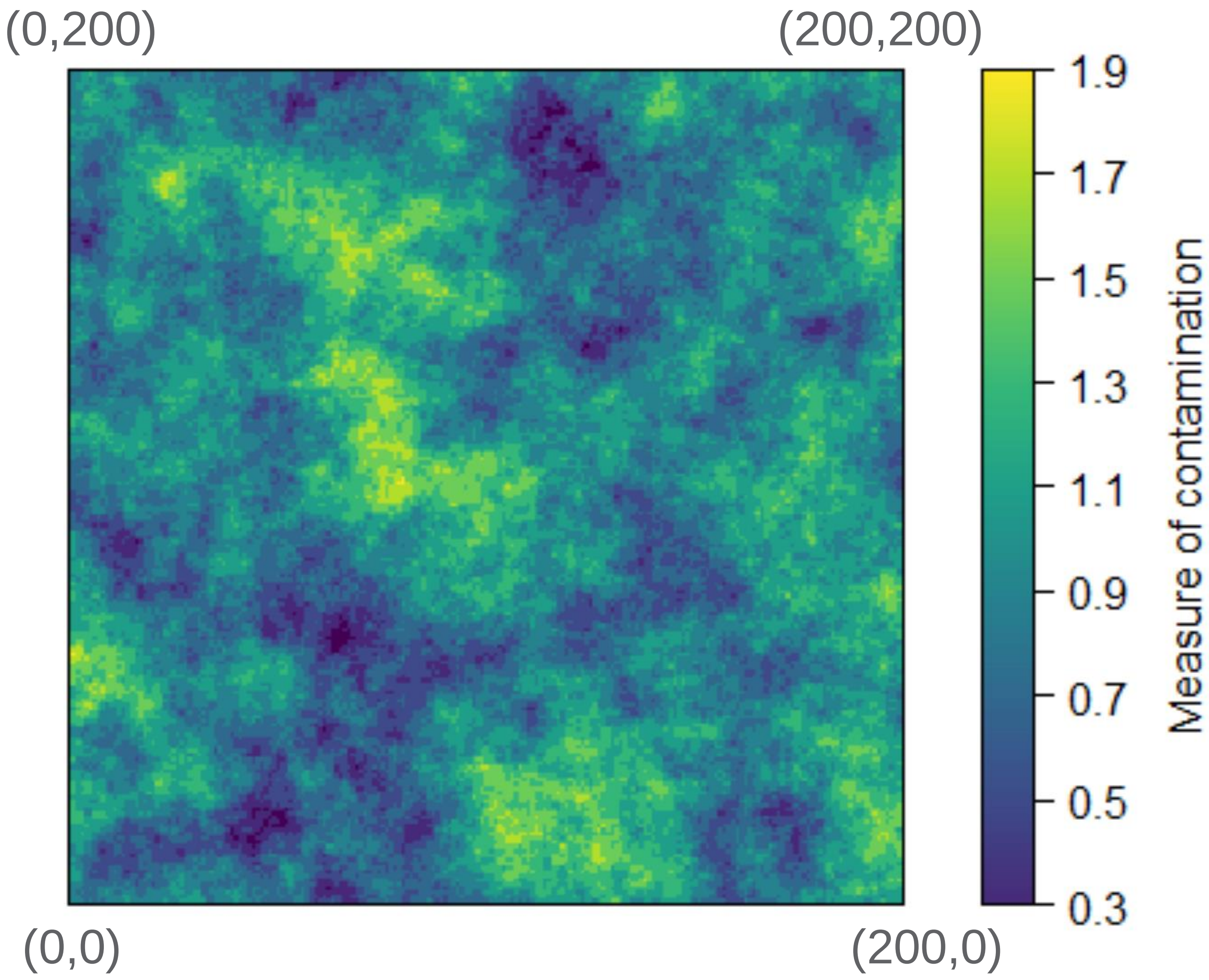
$$I_i = \frac{x_i - \bar{x}}{s_i^2} \sum_{j=1\, i\neq j}^{n} w_{ij}(x_j - \bar{x})$$

- x_i and x_j are the concentrations at location i and j, respectively,
- $\bar{x}$ is the mean concentration of the site,
- w_{ij} is the spatial weight between locations i and j, and
- s_i^2 is the sample standard deviation at location i.

Moran's I test: the null hypothesis states that the spatial process observed by sampled points is random chance (not enough evidence of a significant spatial autocorrelation).

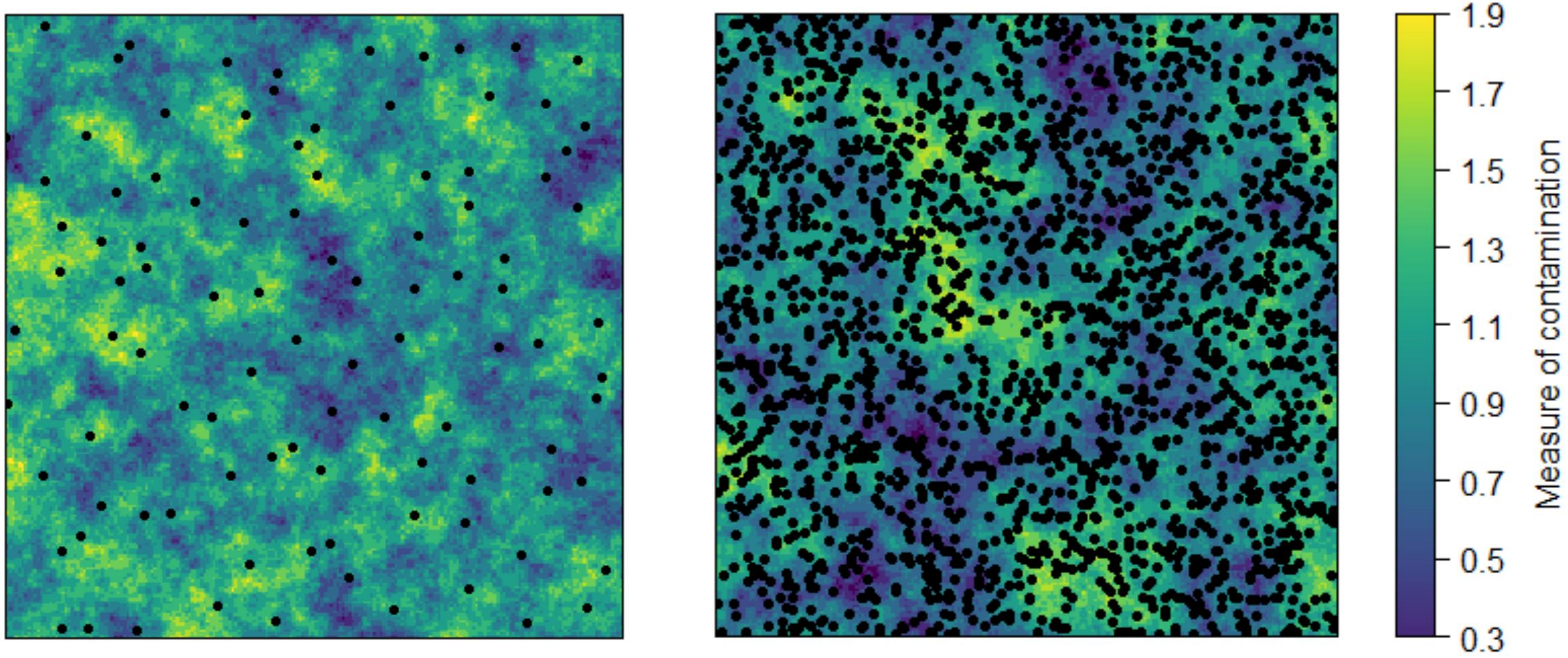
Cliff, A. D., Ord, J. K. 1981 Spatial processes, Pion, p. 21; Bivand RS, Wong DWS 2018 Comparing implementations of global and local indicators of spatial association. TEST, 27(3), 716--748

Approach: Simulation experiment



Exponential variogram model with a range = 20.

Approach: Simulation experiment

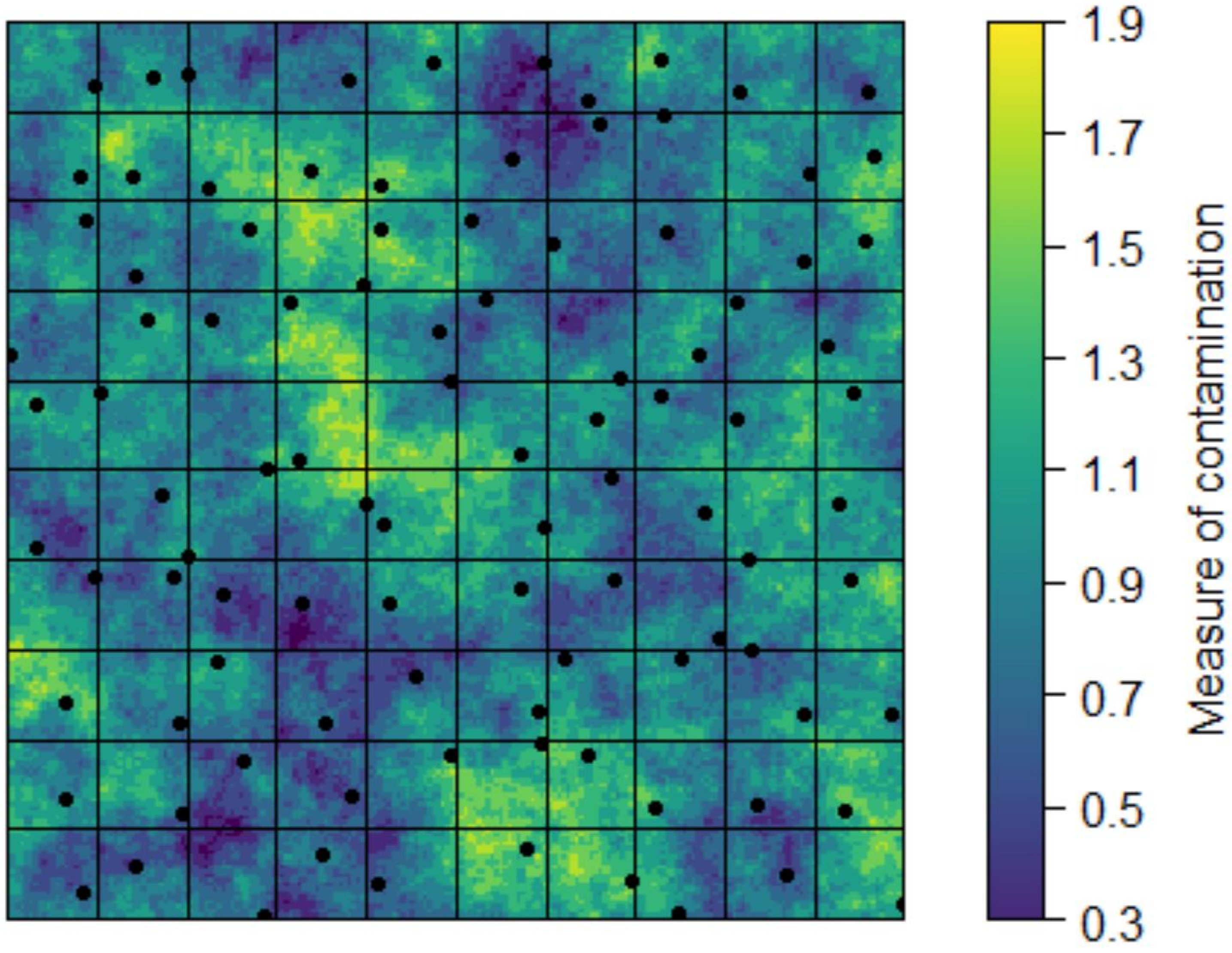


Uniformly placed 100 samples

Randomly placed 2000 samples

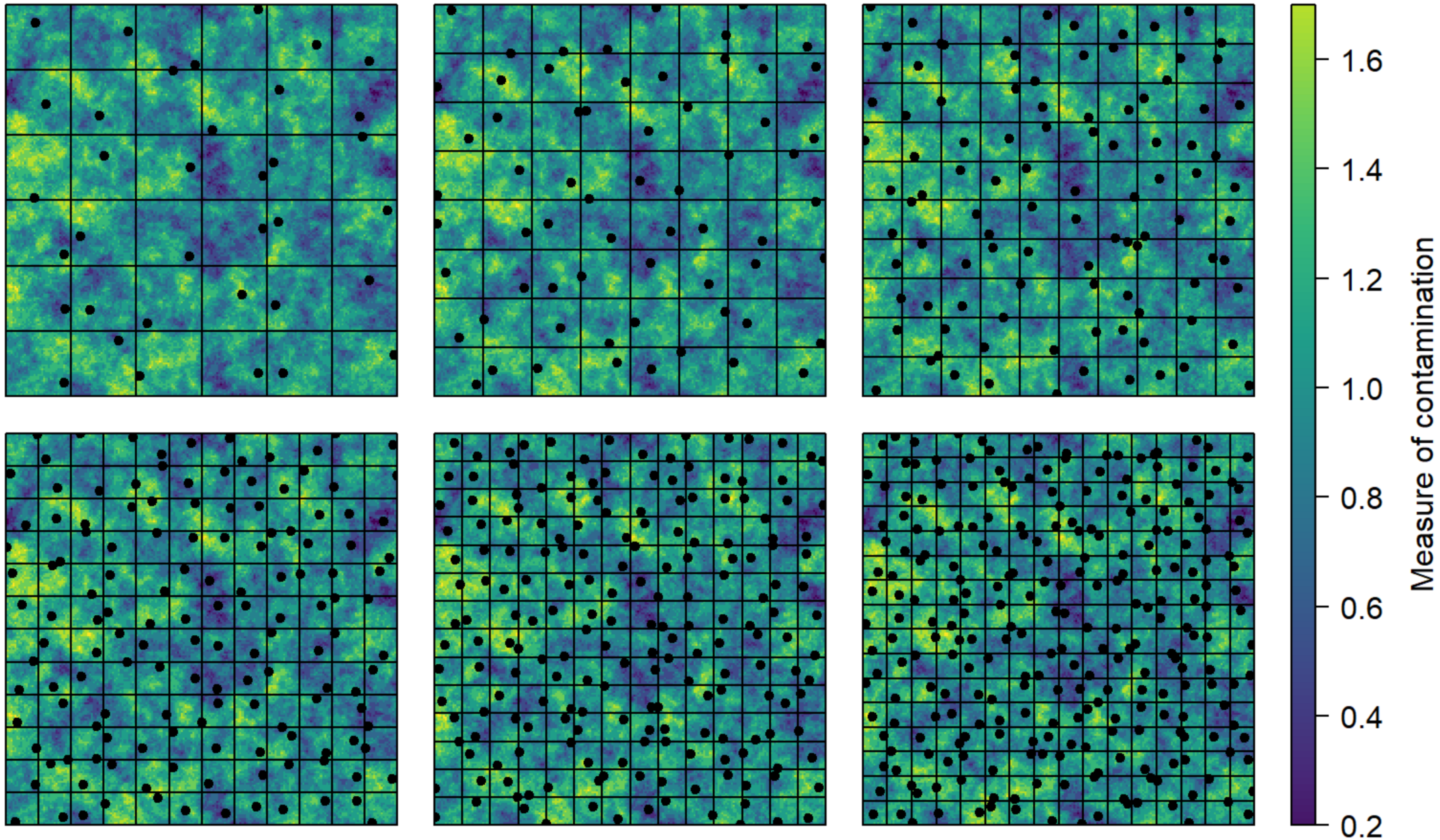
Number of samples?
Sample placement?

Approach: Simulation experiment



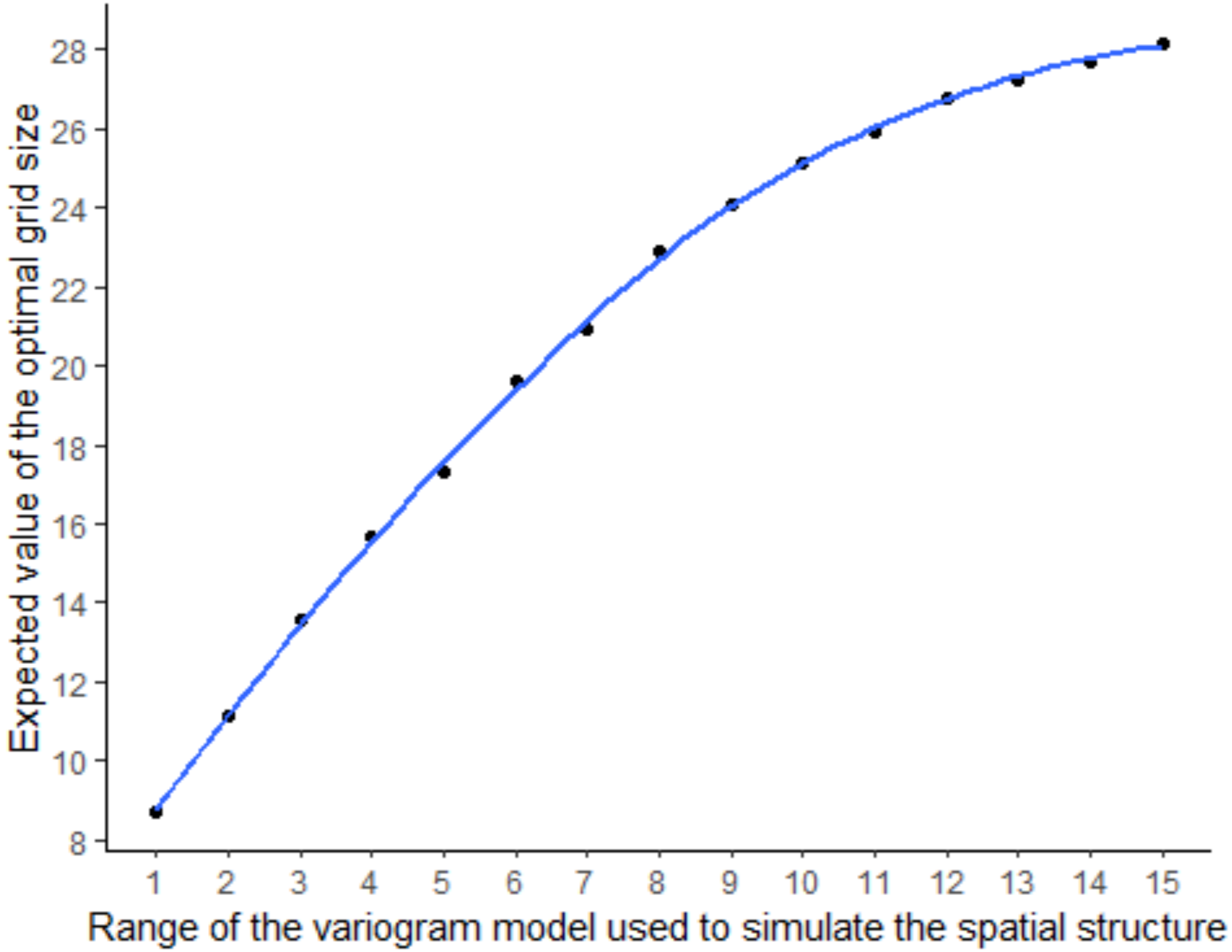
- The study area is partitioned into non-overlapping partitions
- A random sample is taken from each of the partitions as a representative sample of the partition
- Moran's I statistical test is performed to test whether the sampled points shows evidence of spatial autocorrelation (i.e., sampled points are correlated)

Approach: Simulation experiment



- The optimal grid size is obtained by iterating the number of partitions and determine when the Moran's I reject the null hypothesis and conclude there is spatial autocorrelation

Results: Simulation experiment



Range of the variogram model ↓

↳ Optimal grid size ↓

Range of the variogram model ↑

↳ Optimal grid size ↑

Conclusions, recommendations and future work

- To conduct a hypothesis test, need to have an idea of the spatial model, action level, the type I error rate, type II error rate, and the lower bound of the region that represent the probability of failing to reject null hypothesis.

- To account for spatial autocorrelation
  - Can use generalized least squares (GLS) modeling the spatial autocorrelation.
  - Use Moran’s I to determine the optimal grid size based on the spatial model.

- Future work:
  - Determining the minimum sample size to achieve a required statistical power

  - Determine how this approach works for different sampling goals (1-sample hypothesis test, presence/absence, etc.)

  - Investigate the effective sample size formula for different sampling goals using generalized least squares (GLS) model

Griffith, D. A. (2005). Effective geographic sample size in the presence of spatial autocorrelation. Annals of the Association of American Geographers, 95(4), 740-760.
Acosta, J., & Vallejos, R. (2018). Effective sample size for spatial regression models. Electronic Journal of Statistics, 12(2), 3147-3180.
Vallejos, R., & Acosta, J. (2021). The effective sample size for multivariate spatial processes with an application to soil contamination. Natural Resource Modeling, 34(4), e12322.

Contact:

Narmadha Meenu Mohankumar

Data Scientist

Applied Statistics & Computational Modeling

National Security Directorate

Pacific Northwest National Laboratory

narmadha.mohankumar@pnnl.gov

Moses Obiri

Data Scientist

Applied Statistics & Computational Modeling

National Security Directorate

Pacific Northwest National Laboratory

moses.obiri@pnnl.gov

Thank you