



K-d Tree Search Algorithms for a Nearest Neighbor Gaussian Process Model

Student: Alexander McKim

Advisor: Andrew Finley

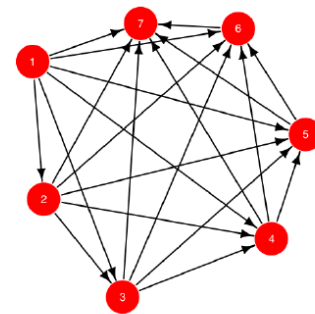


Problem Introduction

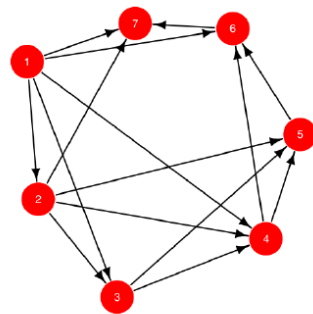
- ▶ What is a nearest neighbor search?
- ▶ Why is it important with spatial or environmental data?
- ▶ Why is it difficult?
 - ▶ A brute force solution is exponential time, thus being extremely costly with real world data sets

Nearest Neighbor Gaussian Process Model (NNGP)

- The NNGP model is an approximation to a Gaussian Process
- The NNGP model is a sparse process, allowing parameters to be estimated much faster.
- The algorithms implemented during this project are made to fit this model and utilize its assumptions.



(a) Full graph



(b) Sparse graph

Nearest Neighbor Gaussian Process Model (NNGP)

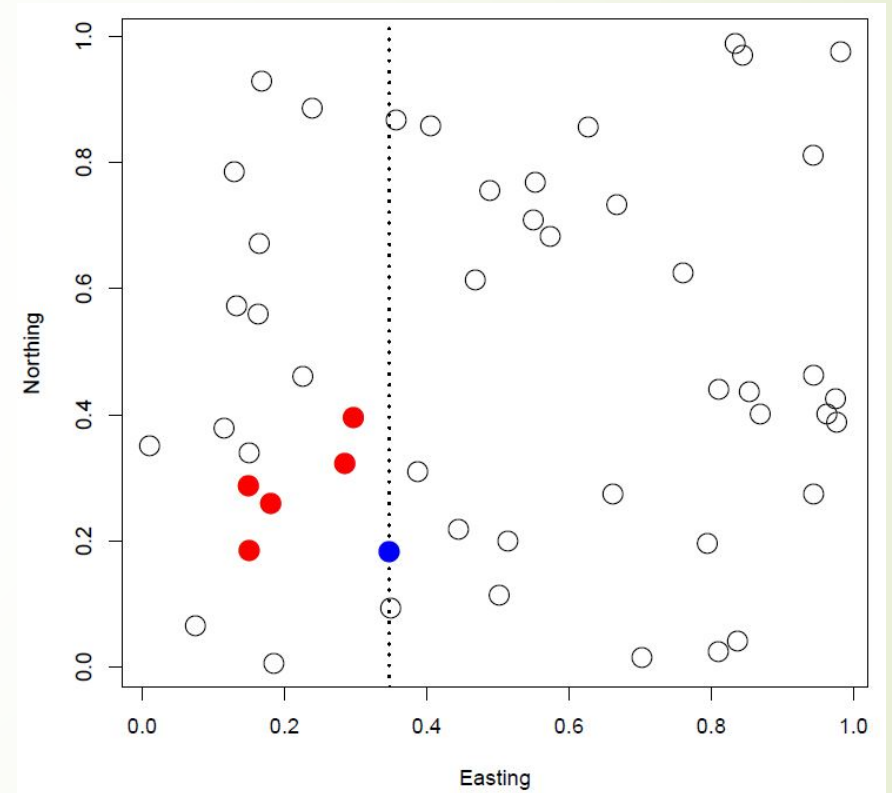
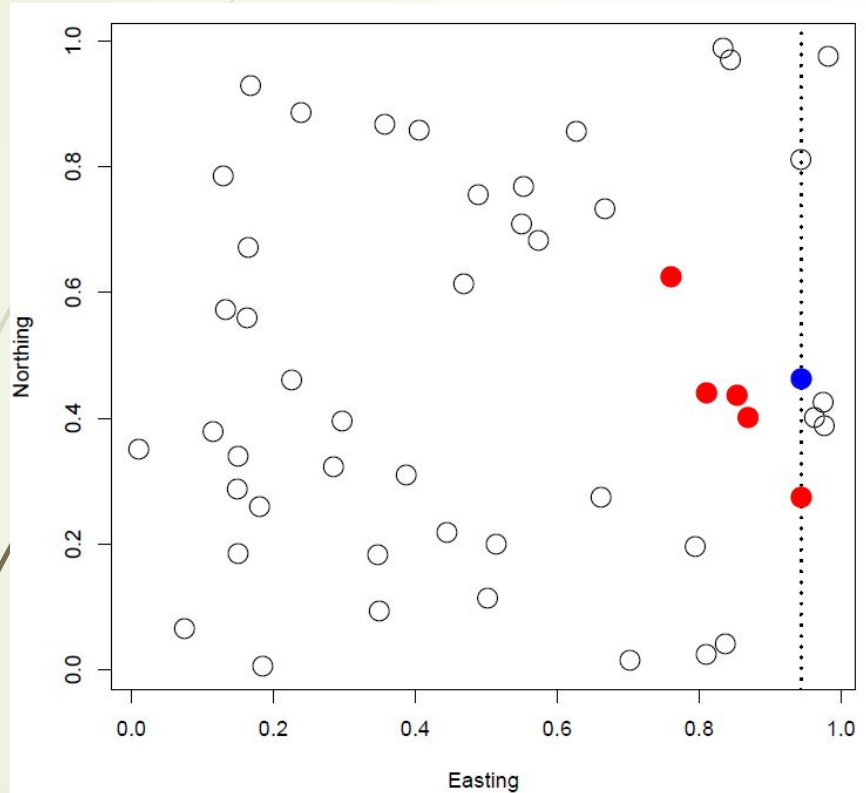



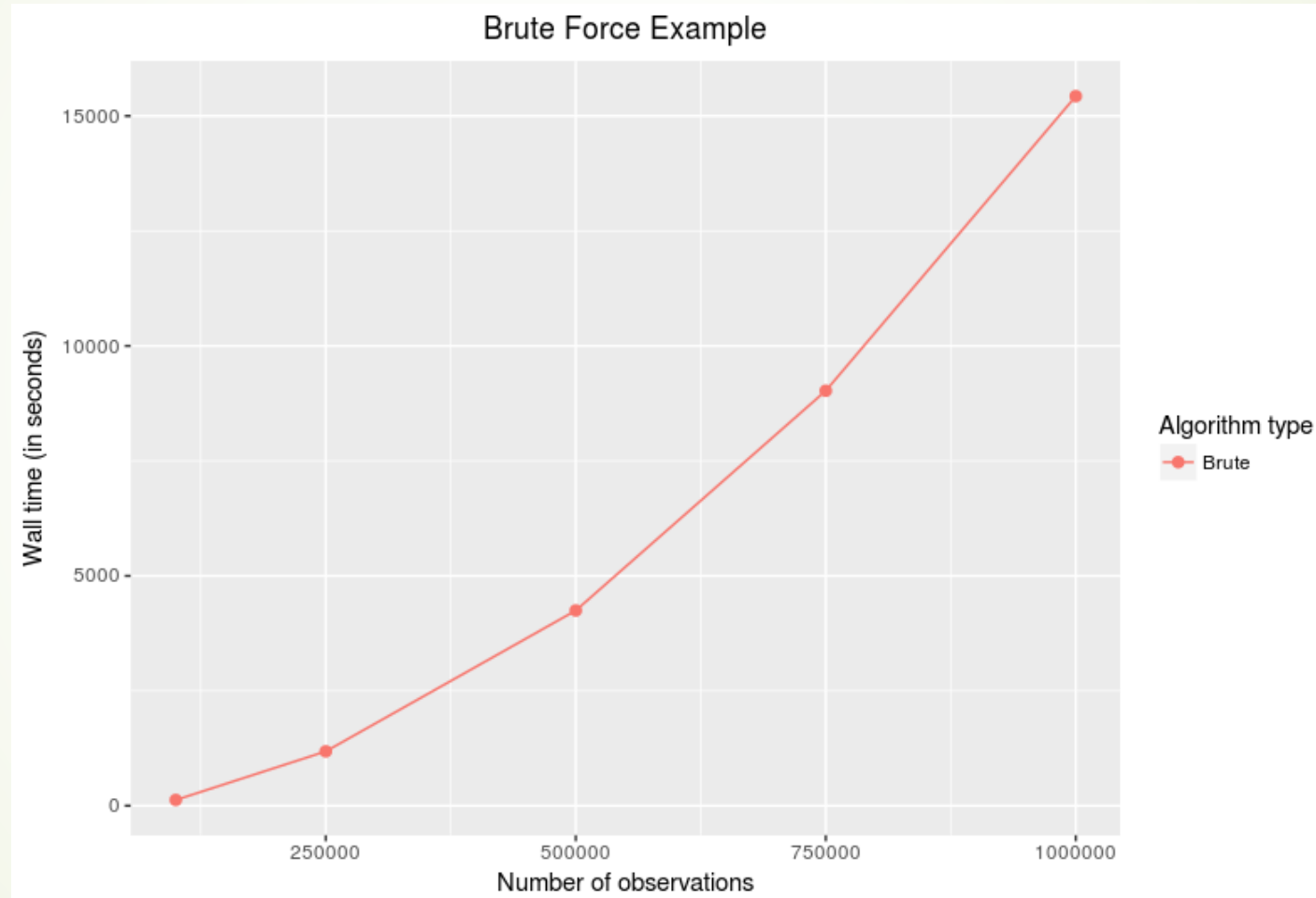
Figure 1: Nearest neighbors under x-axis ordering constraint. Only locations to the left of the given location are candidate neighbors. Left figure shows five neighbors (red point) for location 18 (blue point) among all observations (open circles). Right figure same set up but for location 45 (blue point).



Algorithms Created

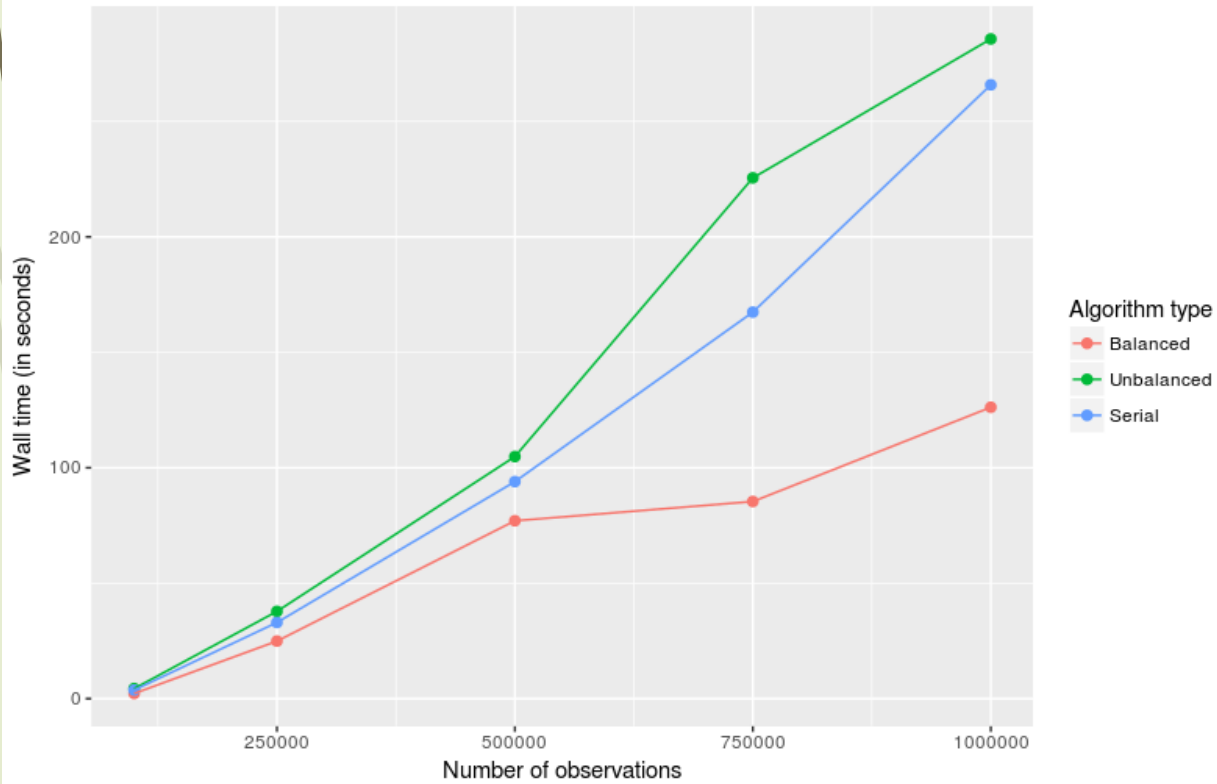
- What is a k-d tree?
 - Four algorithms/structures were implemented
 - Serial
 - Parallel using an unbalanced tree
 - Parallel using a balanced tree
 - Parallel using a balanced tree and well clustered observations
 - Tested algorithms using randomly generated data sets of various sizes and constraints
 - Algorithms implemented using C++, parallelism implemented using OpenMP
- 

Brute force time

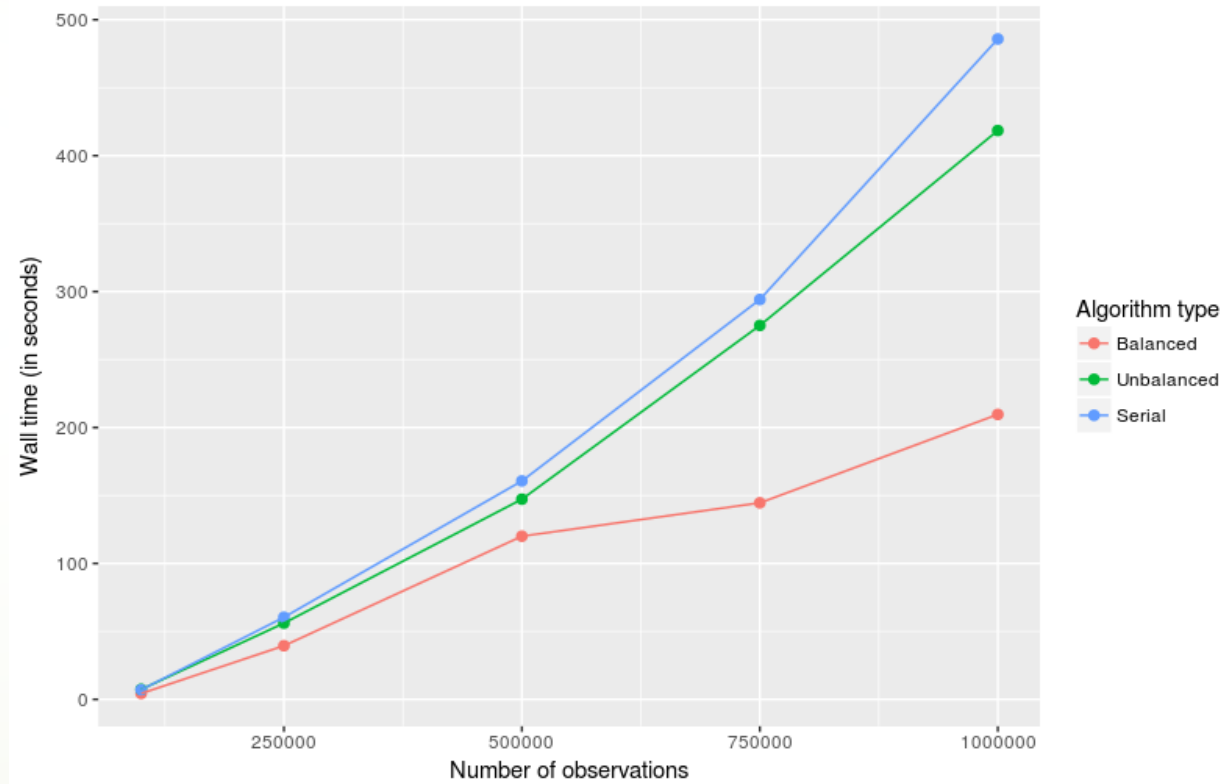


Sample of results

15 neighbors, 1 cpu

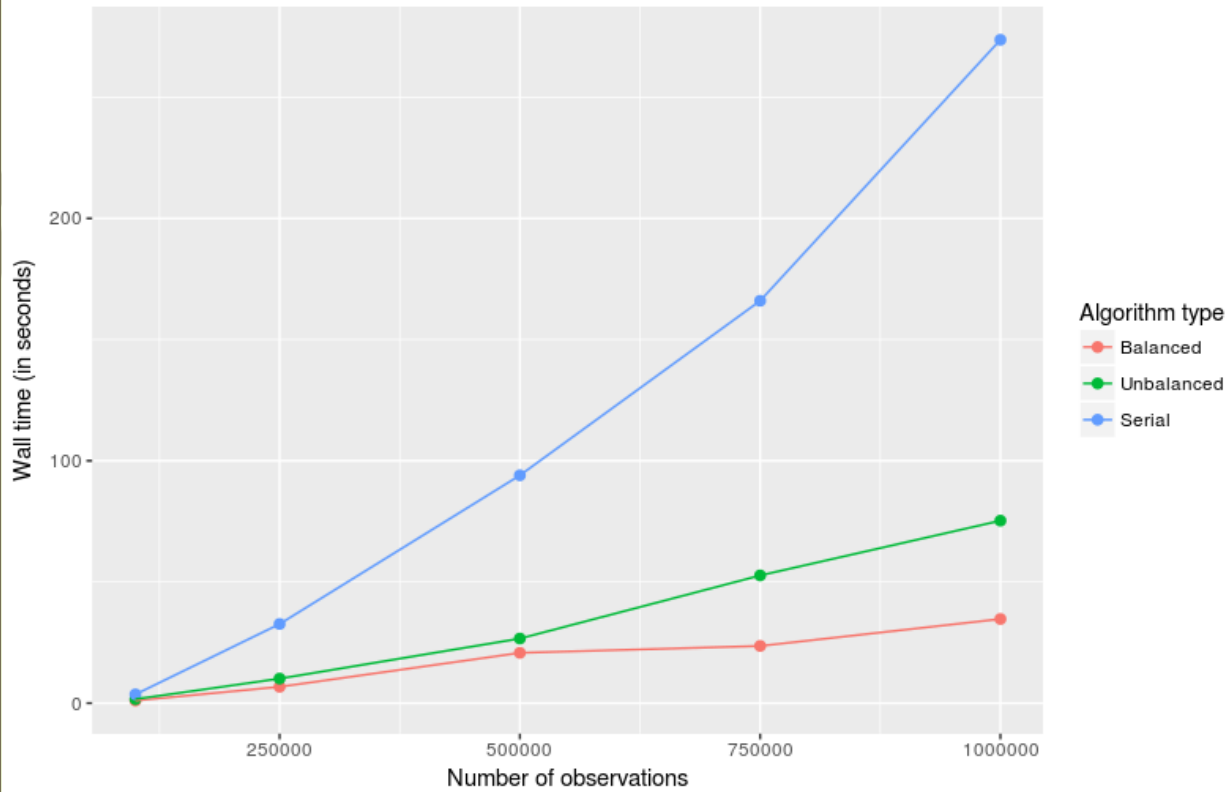


30 neighbors, 1 cpu

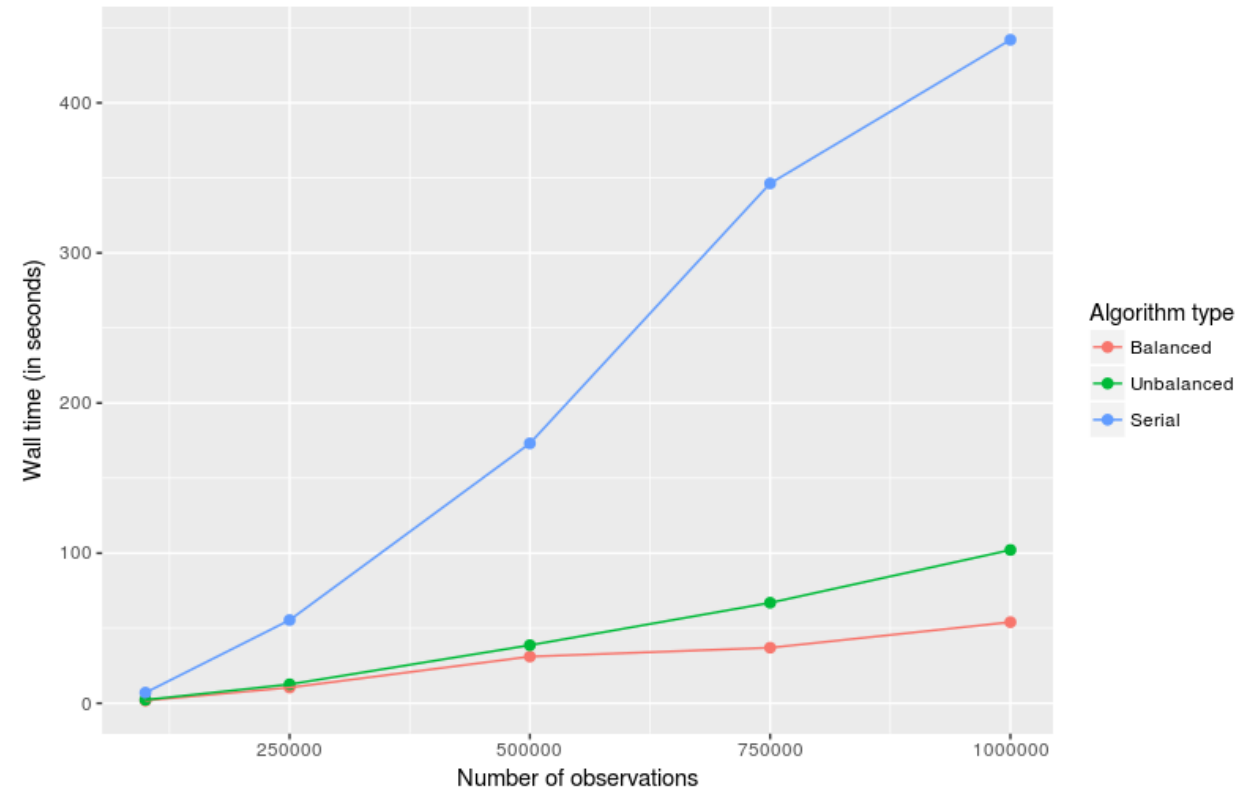


More results

15 neighbors, 6 cpu



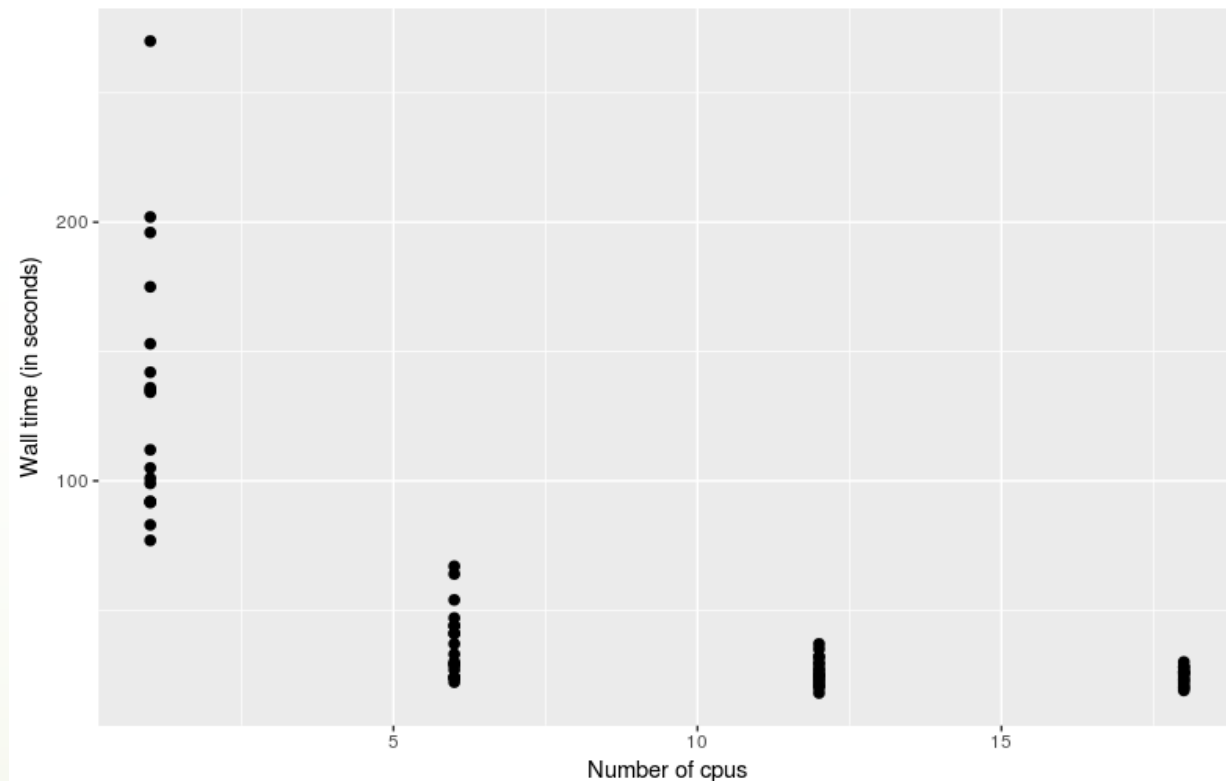
30 neighbors, 6 cpu



Discussion

- ▶ The implemented algorithms are clearly more faster than brute force speed
- ▶ Randomization in the balancing algorithm causes variety
 - ▶ For example, can get as low as between 10-20 seconds for 1 million observations and 15 neighbors, typically between 20-30, bad cases will be near speed of tree that is not balanced.
- ▶ Unbalanced algorithm has been released in spNNGP R package on CRAN
 - ▶ Shows context of how this algorithm is being used, and the purpose of this project
- ▶ <https://cran.rproject.org/web/packages/spNNGP/index.html>

Random Seed Variation with Balancing Tree Algorithm





Future Work



- ▶ Removing randomness on unbalanced algorithm is a priority for more consistency
- ▶ Implement this to work in three dimensions
- ▶ More pruning possibilities on the search itself using the assumptions of the NNGP model.



Works Cited/Acknowledgements

- ▶ Datta, A., S. Banerjee, A.O. Finley, and A.E. Gelfand. (2016) Hierarchical Nearest-Neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111:800-812.

Finley, A.O., A. Datta, B.C. Cook, D.C. Morton, H.E. Andersen, and S. Banerjee (2017) Computing Bayesian Nearest-Neighbor Gaussian Process Models for Massive Spatial Data Sets