A Designed Experiment on Effects of Dataset, Hadoop, and Hardware Factors on D&R Computational Performance

William S. Cleveland\textsuperscript{1,*}, Douglas G. Crabill\textsuperscript{1}

1. Purdue University, Department of Statistics
*Contact author: wsc@purdue.edu

Keywords: Tessera, D&R, Hadoop, computational performance, multi-factor experiment

D&R is a statistical approach to large complex data. The data are divided into subsets. Analytic methods are applied to each subset, and the outputs of each method recombined. Statistical thinking, theory, and methods are used to find division and recombination methods with good statistical performance. The Tessera D&R software has R at the front end, the Hadoop distributed database and parallel compute engine at the back end carrying out D&R computations, and the Tessera R packages in between that make D&R programing easy. Computational performance depends substantially on a number of controllable factors. There are dataset, Hadoop, and hardware factors. Past Hadoop performance measurement has been benchmarks, but they do not characterize D&R computation, and they almost always change one parameter one at a time despite substantial interactions. We ran a multi-factor designed experiment that reveals the structure of the dependence, including the interactions.