Data Extraction Evaluation

Matthew Leonawicz

December 24, 2014

1 Statistical sampling for spatial data extraction

1.1 Motivation: Data processing efficiency

We've gotten faster at SNAP, but so has our need for speed. What was once never bothered with (outside some of my own work), using statistical sampling to obtain results at no cost to validity, accuracy or precision compared to a census of our data, is now more relevant than ever. Before, we were content to let a process run in the background for hours and look at the results when done. There was little incentive to incorporate techniques like those laid out here. Now we have more types of data delivery and presentation, e.g., web applications, where it is intended for there to be a human watching and waiting for data processing to occur.

1.1.1 Assumptions, bad ones

An assumption I often encounter from those outside statistics, but involved in "big data" is that with today's processing power there is no reason not to use all the data. A corollary of this is that many statistical methods can be dispensed with, which is based on another assumption that this is what statistics basically exists for; to help us hobble along when we were in the stone age. However, both of these views are flawed. The latter suggests little knowledge of the broad uses of statistics. The former suggests little knowledge of statistics period, or the myriad ways data can be improperly analyzed and results interpreted.

1.1.2 Speed not for speed's sake

Making things go faster is perhaps the last area of application I would ever find for statistical methods, and since not a lot of traditional statistical analysis occurs at SNAP I do not want those untrained in statistics to get the wrong impression that speed improvements are all statistics is really good for. But it is relevant and beneficial in the context of some of our workflows, particularly my own. But I am also not the only one extracting and processing large amounts of data at SNAP. One use of statistics is data reduction. This is what I aim for when needing to "get things done faster," not really the speed itself. I'd rather see a decrease in computational time required for large data processing occur as a latent consequence of smart application of statistical methods than as something forced for its own sake. I will outline a simple and extremely common case.