Powerful statistical tool could significantly reduce the burden of analyzing very large datasets

October 30 2017

The KAUST supercomputer Shaheen II underpins the collaboration by providing high-performance computing applications and strategic advice and support. Credit: 2017 KAUST

By exploiting the power of high-performance computing, a new statistical tool has been developed by KAUST researchers that could

reduce the cost and improve the accuracy of analyzing large environmental and climate datasets.

Datasets containing environmental and climate observations, such as temperature, wind speeds and soil moisture, are often very large because of the high spatial resolution of the data. The cost of analyzing such datasets increases steeply as the size of the dataset increases: for instance, increasing the size of a dataset by a factor of 10 drives up the cost of the computation by a factor of a 1000, and the memory requirements by a factor of 100, creating a computational strain on standard statistical software.

This spurred postdoctoral fellow Sameh Abdulah to develop a standalone software framework through a collaboration between KAUST's Extreme Computing Research Center (ECRC) and statisticians specializing in spatio-temporal dynamics and the environment.

The new framework, called Exascale GeoStatistics or ExaGeoStat, is able to process large geospatial environmental and climate data by employing high-performance computing architectures with a high degree of concurrency not available through universally used statistical software.

"Existing statistical software frameworks are not able to fully exploit large datasets," says Abdulah. "For example, a computation that would normally require one minute to complete would take nearly 17h if the dataset were just 10 times larger. This leads to compromises due to the limitations in computing power, forcing researchers to turn to approximation methods that cloud their interpretation of results."

Leveraging linear algebra software developed by the ECRC, ExaGeoStat provides a framework for computing the maximum likelihood function for large geospatial environmental and climate datasets. It is able to

predict unknown or missing data as well as reduce the effect of individual measurement errors, allowing the data to be easily analyzed and represented in a statistical model used for making predictions.

The researchers successfully applied ExaGeoStat to a large, real-world dataset of soil moisture measurements from the Mississippi basin in the United States. This could lead to the routine analysis of the larger datasets that are becoming available to geospatial statisticians, and could be used in a wide range of applications from weather forecasting, crop-yield prediction, and early-warning systems for flood and drought.

David Keyes, Director of the ECRC, which hosts the project, plans significant further improvements, tracking a rapidly developing technique in linear algebra: "We are now working on taking ExaGeoStat a step further on the algorithmic side by introducing a new type of approximation, called hierarchical tile low-rank approximation, which reduces memory requirements and operations by allowing for small errors that can easily be understood and controlled."

More information: ExaGeoStat: A High Performance Unified Framework for Geostatistics on Manycore Systems arXiv:1708.02835 [cs.DC] arxiv.org/abs/1708.02835

Provided by King Abdullah University of Science and Technology