CCSU
DEPARTMENT OF MATHEMATICAL SCIENCES

COLLOQUIUM

Friday, April 24
9:30 – 10:00 AM
Davidson Hall, Room 203

USE OF PARALLEL COMPUTING TO FIT OLS REGRESSION MODELS USING SAS

MARCOS MARTINS SOUZA

(Data Mining MS Thesis Presentation)

CENTRAL CONNECTICUT STATE UNIVERSITY

Abstract: The physical universe is large – by 2020 containing as nearly as many digital bits as there are stars in the universe. It's doubling in size every two years, and by 2020 the digital universe – the data we create and copy annually – will reach 44 zettabytes, or 44 trillion gigabytes. The architecture used in current computer chips is reaching physical limits, mainly because of cooling requirements. In the last years, CPUs have not improved much in clock speed, but more cores have been added to chips. As it won't be possible to compute with faster CPUs (not using the actual technology), it's necessary to break processes in multiple parts, to achieve a reasonable total computational run time, while data grows and business requires additional analysis. This research presents a method of parallelizing the computation of the regression analysis across multiple CPUs and computers, to allow reducing total runtime and processing very large databases. Not every problem that is parallelized runs faster than its non-parallel version. Although, as computers become such a commodity asset, building large computing clusters and distributing the computation to fit models across all of them is a feature that analysts should have at their hands.