Parallel Additive Models for Big Data

Generalized additive models are generalized linear models in which the linear predictor is specified in terms of unknown smooth functions of covariates, and the functions are the target of inference. See e.g. Wood (2006) and the mgcv package in R.

In fields ranging from the psychology of language recognition to the prediction of electricity demand, these models are being used with increasingly large datasets containing hundreds of thousands or millions of observations. In part this is facilitated by the big data methods covered in Wood, Goude and Shaw (2014) and the associated bam function in R package mgcv, but most users would like to be able to use even more complex models with even large datasets, and here existing methods start to become impractical. This project will explore the development of new scale-able methods for GAM fitting that are susceptible to being made massively parallel. The project will involve statistics, applied numerical analysis, and programming in R and C as well as collaboration with the electricity company EDF.


Segments of the load, in Gigawatts, on the French national grid, for which accurate one day ahead forecasting is made possible by recent GAM methods for big data. Electricity company EDF would like to be able to use similar models with datasets very much larger that this.