Divide and conquer?
k-means clustering of demand data allows rapid and accurate simulations of the British electricity system.

Richard Green[†], Iain Staffell[†] and Nicholas Vasilakos[‡*]

[†] Imperial College
London, United Kingdom.

[‡] Centre for Competition Policy
and Norwich Business School,
University of East Anglia,
Norwich, NR4 7TJ,
United Kingdom.

Corresponding author: n.vasilakos@uea.ac.uk

(1) Overview

We use a k-means clustering algorithm to partition national electricity demand data for Great Britain and apply a novel profiling method to obtain a set of representative demand profiles for each year over the period 1994-2005. We then use a simulated dispatch model to assess the accuracy of these daily profiles against the complete dataset on a year to year basis. We find that the use of data partitioning does not compromise the accuracy of the simulations for most of the main variables considered, even when simulating significant intermittent wind generation. This technique yields fifty-fold gains in terms of computational speed, allowing complex Monte Carlo simulations and sensitivity analyses to be performed with modest computing resource.

(2) Methods

This paper presents a way of creating a small number of "typical" daily profiles that can represent a year of operation, allowing a dispatch model to be run very fast and thus many

times for stochastic or other simulations. To do this we apply a k-means clustering algorithm, a method that has been used widely in various disciplines to separate (often large) agglomerated datasets into a number of smaller groups. Membership for each cluster is decided using a similarity measure – in this case the Euclidean distance of each data point from the group mean. The objective of the process is to allocate every observation into a group in such a way that this distance is minimised. We then investigated several ways of using the demand levels and hour-to-hour changes from every day within each cluster to create a representative profile for that cluster. The most accurate simulations (compared to those for a full year of data) were obtained when we took the means of those hour-to-hour demand changes with the same sign as the median change within the cluster to give 23 hour-on-hour changes in demand, and set the remaining demand level so that the total demand across the profile equalled that within all the cluster's observations.

(3) Results

We find that the clustered data give a remarkably accurate simulation for many key variables, including the average cost of electricity, carbon emissions and generator revenues. Other variables, which are more dependent on extreme events, such as generator start-ups (often in response to a short-lived spike in demand) are less accurately modelled, and so the method is not applicable to every problem that we might wish to apply it to. Nonetheless, we find that using ten clusters to represent a year of data gives a remarkable saving in computer time, and recommend that this method be considered to allow the complexities of electricity dispatch to be represented in studies that require many repeated simulations of an electricity market.