

Ray Lindsay,  
Senior Data Miner  
Australian Taxation Office

## **Using ensembles of models to predict revenue for unlodged tax returns**

### **Approximating curves by many piecewise constants**

We will describe the statistical challenges that were overcome in one of the first deployments of Data Mining in the ATO. This model aimed to make predictions of revenue likely to be raised from a large number of outstanding returns. The challenges included a significant proportion of missing values, initially trying to model a step function, and issues related to the storage and processing of very large datasets required for scoring.

The presence of missing values meant that regression or neural network models could result in many records having the default prediction (mean for interval variables, most common level for nominal). Tree models are more robust to missing values but produce a relatively small number of distinct predicted values. When making predictions for millions of records, one needs some way of distinguishing amongst the top 1% (say) of predictions. When we combine the first (classification tree) and two second stage models (regression trees) we introduce some granularity as records in one node in one of the models are not necessarily in the same node in the other two. Much finer granularity in revenue prediction can be achieved by employing a relatively low number of tree models – where the features input into each stream are randomly selected and a random sample of training data is used, essentially of form of bagging. This approach is somewhat similar in philosophy to RandomForests, which could not be used in this problem due to technical issues, and non-compatibility with large numbers of missing values.