Visualisation of Regression Trees

Chris Brunsdon

April 8, 2007

The regression tree [1] has been used as a tool for exploring multivariate data sets for some time. As in multiple linear regression, the technique is applied to a data set consisting of a continuous *response* variable y and a set of *predictor* variables $\{x_1, x_2, ..., x_k\}$ which may be continuous or categorical. However, instead of modelling y as a linear function of the predictors, regression trees model y as a series of 'if-then-else' rules based on values of the predictors. For example consider a data set of details of house sales with the following variables

- *Price*: The response variable — the price at which a house was sold.
- *Type*: The type of the house — a categorical predictor variable taking one of the values *'Flat'*, *'Detached House'*, *'Semi Detached House'*, or *'Terraced House'*.
- *Period*: The period when the house was built - another categorical predictor with values *'Pre 1914'*, *'1914-1945'*,*'1946-1959'*,*'1960-1969'*,*'1970-1979'*,*'Post 1979'*.
- *area*: A continuous predictor variable, the floor area of the house in m^2.

For this data a potential regression tree model might be

```
if (area < 125.7) then
    if (type='Detached House') then
        price = 78,500
    else
        price = 53,600
    end if
else
   price = 113,000
end if
```

Here, the values assigned to price are the predicted values of the model. Note that rules are applied recursively, so that the 'if-then-else' blocks are nested. Although this example only nests rules to a maximum depth of two, in practice much deeper nestings often occur. Note also that in some cases the *price* variable can be assigned at a lower level of nesting than in others. Applying regression tree algorithms to identify optimal rule trees is a useful way of exploring and summarising structure in multivariate data. However, if the data are geographically referenced it may be useful to consider interaction between the tree and geographical space. This is best achieved using interactive visualisation methods. In this presentation two key issues will be addressed:

1. How to visualise regression trees
2. How to link these to dynamic maps

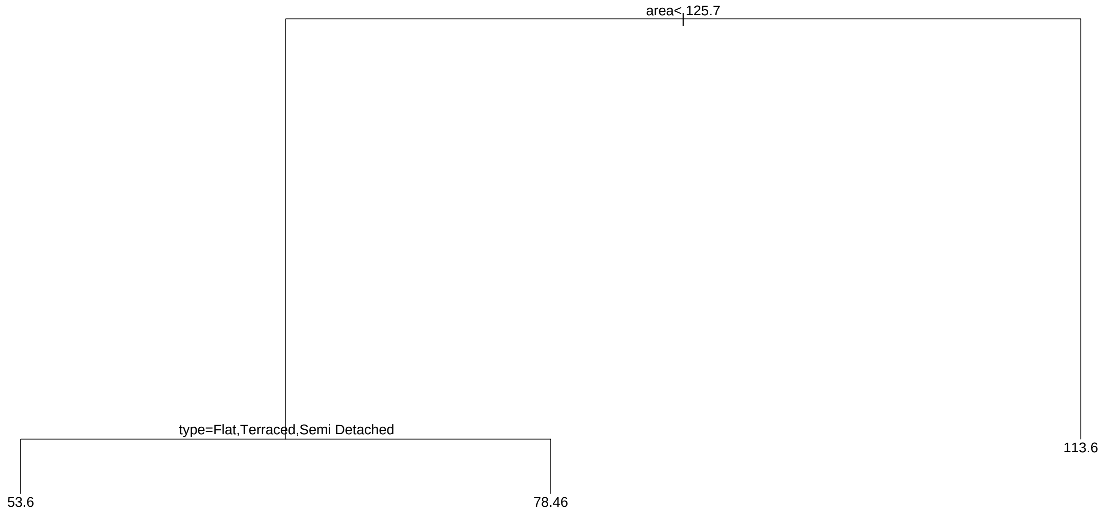

Figure 1: A typical ‘family tree’ visualisation of a regression tree.

The first issue is relevant to all regression tree applications, whilst the second is important for spatially referenced data. A typical approach to visualising trees is the ‘family tree’ approach - for the example above, this is shown in figure 1. Although conveying basic information, the approach has two shortcomings: firstly it requires very low $y : x$ aspect ratios as depth increases - since increasing depth by one level potentially doubles the number of branches on the final part of the tree. Secondly, although the predicted price values are printed on the graph, there is no other visual indication of their levels. For example, the x-location of a node on the tree does not signify the average y-value of data items associated with this node. An alternative visualisation is proposed here, using a part-circular scale (addressing the aspect ratio issue), having a set of rules for locating tree nodes according to y-values associated with them. An example (modelling the same data with a more complex tree) is given in figure 2. The writing in the top right hand corner appears interactively when clicking on a tree node: this shows the rule associated with the node, and the mean value of the response variable associated with the node. The location of each tree node, measured as an angle anticlockwise from the yellow part of the background, represents the mean value of all house prices that are ‘children’ of this node. The green lines connect final nodes to individual prices. A second, linked graphical window shows the locations of all of the data corresponding to this node - here shown in figure 3. From this it can be seen that houses associated with the selected node and its children tend to be located outside of London, but there are distinct clusters to the south-west and south east of London[1].

In the proposed presentation, the visualizatiuons will be fully explained, and choices made in their construction will be outlined. Issues in linking to the map will also be discussed. Also, an alternative tree visualisation in which branches are guaranteed not to cross will be demonstrated (figure 4). The cost here is that the angular locations of the nodes do not represent *absolute* information about mean prices. The relative merits of these two visualisations will also be considered in the talk.

[1] The ‘selected’ points are intentionally plotted after the others to avoid them being masked in the visualization.

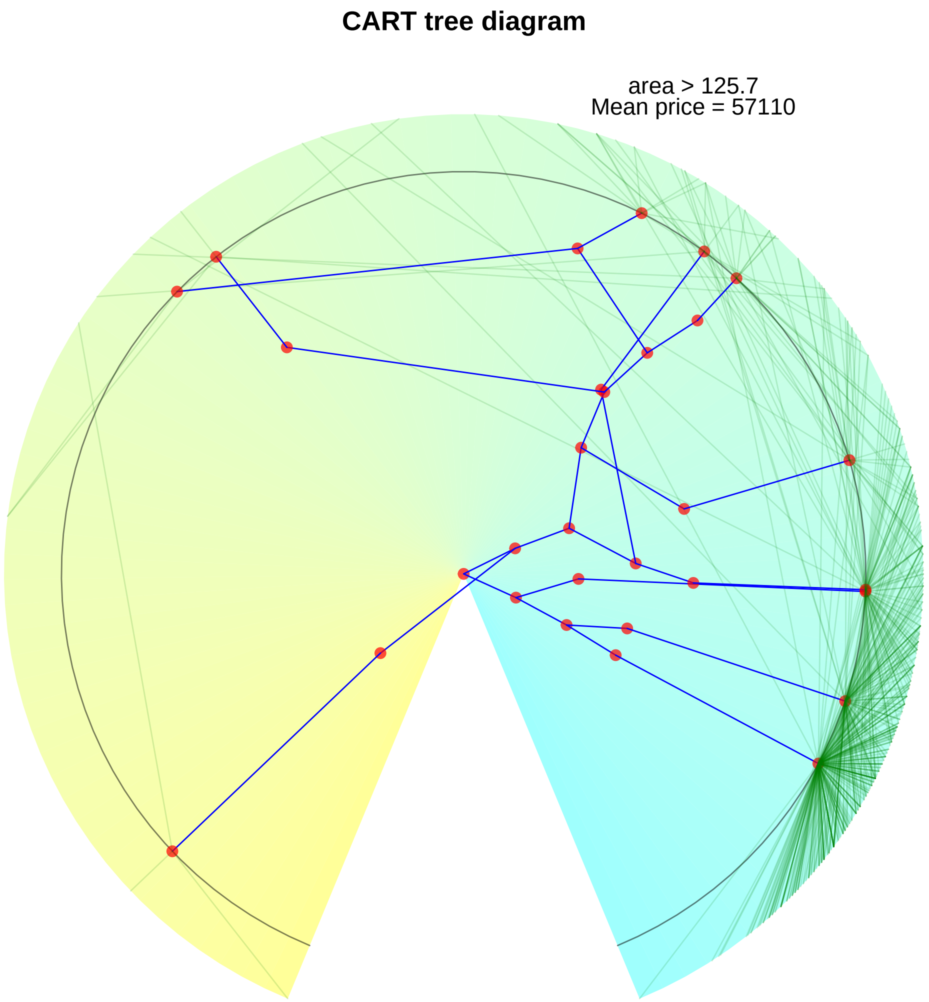

Figure 2: Proposed alternative visualization of regression tree

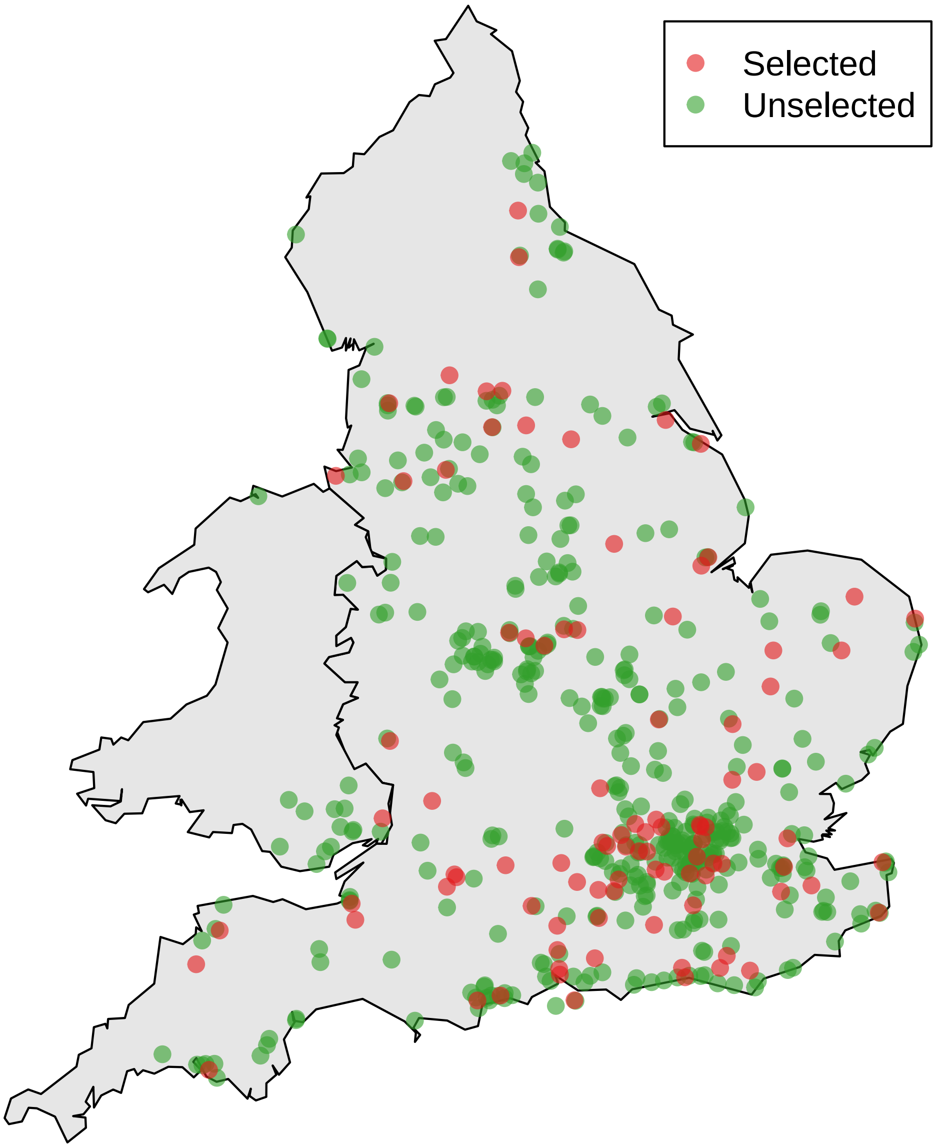

Figure 3: Map linked to regression tree visualisation in figure 2.

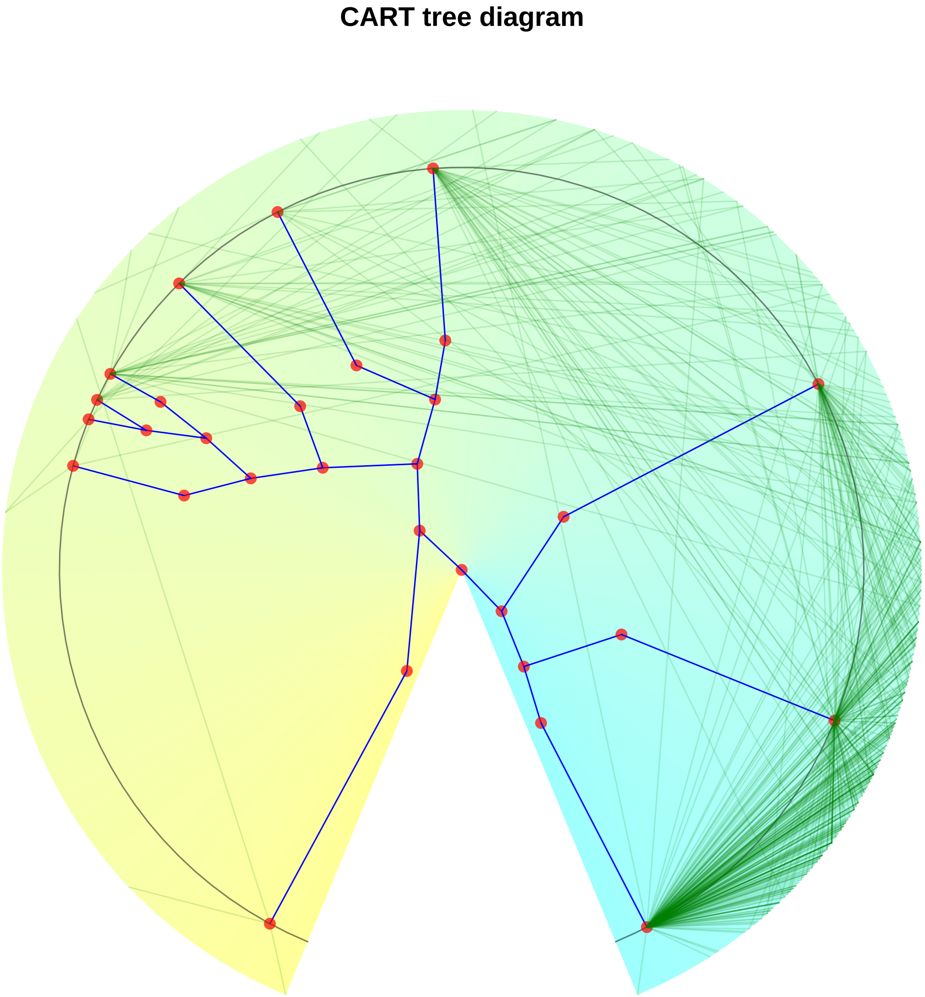

Figure 4: Alternative regression tree visualisation - here tree branches cannot cross, but y value information relative to node children, rather than absolute value, is depicted.

References

[1] Leo Breiman, Jerome Friedman, Charles Stone, and R. Olsen. *Classification and Regression Trees*. Wadsworth, London, 1984.