Notebook 06: Apply cartogram to data

Jon Minton

Martin Taulbut

Table of contents

# 1. Introduction

The aim of this notebook is to understand how straightforward and informative applying standard CART approaches to the UKHLS data, with a focus on understanding and presenting which factors are associated with different economic (in)activity statuses in particular waves.

# 2. Setup

Individual level responses

household level responses

Now to define the varnames of interest in the indresp files, along with which their variable type.

[1] "extracting file: /Users/JonMinton/repos/economic\_inactivity/big\_data/UKDA-6614-stata/stata/stata13\_se/ukhls/a\_indresp.dta"  
[1] "Attempting to find 17 variables"  
[1] "WARNING! Only 11 of the 17 requested have been found"  
[1] "read file in 0.324872016906738 seconds"  
[1] "slimming file..."  
[1] "extracted 11 variables in 0.42829418182373 seconds"  
[1] "extracting file: /Users/JonMinton/repos/economic\_inactivity/big\_data/UKDA-6614-stata/stata/stata13\_se/ukhls/b\_indresp.dta"  
[1] "Attempting to find 17 variables"  
[1] "All variables requested found"  
[1] "read file in 0.437664985656738 seconds"  
[1] "slimming file..."  
[1] "extracted 17 variables in 0.714245080947876 seconds"  
[1] "extracting file: /Users/JonMinton/repos/economic\_inactivity/big\_data/UKDA-6614-stata/stata/stata13\_se/ukhls/c\_indresp.dta"  
[1] "Attempting to find 17 variables"  
[1] "WARNING! Only 11 of the 17 requested have been found"  
[1] "read file in 0.641356945037842 seconds"  
[1] "slimming file..."  
[1] "extracted 11 variables in 0.374963045120239 seconds"  
[1] "extracting file: /Users/JonMinton/repos/economic\_inactivity/big\_data/UKDA-6614-stata/stata/stata13\_se/ukhls/d\_indresp.dta"  
[1] "Attempting to find 17 variables"  
[1] "All variables requested found"  
[1] "read file in 0.458340883255005 seconds"  
[1] "slimming file..."  
[1] "extracted 17 variables in 0.647369146347046 seconds"  
[1] "extracting file: /Users/JonMinton/repos/economic\_inactivity/big\_data/UKDA-6614-stata/stata/stata13\_se/ukhls/e\_indresp.dta"  
[1] "Attempting to find 17 variables"  
[1] "WARNING! Only 11 of the 17 requested have been found"  
[1] "read file in 0.535313129425049 seconds"  
[1] "slimming file..."  
[1] "extracted 11 variables in 0.346320867538452 seconds"  
[1] "extracting file: /Users/JonMinton/repos/economic\_inactivity/big\_data/UKDA-6614-stata/stata/stata13\_se/ukhls/f\_indresp.dta"  
[1] "Attempting to find 17 variables"  
[1] "All variables requested found"  
[1] "read file in 0.450836181640625 seconds"  
[1] "slimming file..."  
[1] "extracted 17 variables in 0.622936010360718 seconds"  
[1] "extracting file: /Users/JonMinton/repos/economic\_inactivity/big\_data/UKDA-6614-stata/stata/stata13\_se/ukhls/g\_indresp.dta"  
[1] "Attempting to find 17 variables"  
[1] "WARNING! Only 11 of the 17 requested have been found"  
[1] "read file in 0.494240045547485 seconds"  
[1] "slimming file..."  
[1] "extracted 11 variables in 0.397297143936157 seconds"  
[1] "extracting file: /Users/JonMinton/repos/economic\_inactivity/big\_data/UKDA-6614-stata/stata/stata13\_se/ukhls/h\_indresp.dta"  
[1] "Attempting to find 17 variables"  
[1] "All variables requested found"  
[1] "read file in 0.40286111831665 seconds"  
[1] "slimming file..."  
[1] "extracted 17 variables in 0.500993967056274 seconds"  
[1] "extracting file: /Users/JonMinton/repos/economic\_inactivity/big\_data/UKDA-6614-stata/stata/stata13\_se/ukhls/i\_indresp.dta"  
[1] "Attempting to find 17 variables"  
[1] "WARNING! Only 11 of the 17 requested have been found"  
[1] "read file in 0.550350189208984 seconds"  
[1] "slimming file..."  
[1] "extracted 11 variables in 0.284988164901733 seconds"  
[1] "extracting file: /Users/JonMinton/repos/economic\_inactivity/big\_data/UKDA-6614-stata/stata/stata13\_se/ukhls/j\_indresp.dta"  
[1] "Attempting to find 17 variables"  
[1] "All variables requested found"  
[1] "read file in 0.38463306427002 seconds"  
[1] "slimming file..."  
[1] "extracted 17 variables in 0.445006132125854 seconds"  
[1] "extracting file: /Users/JonMinton/repos/economic\_inactivity/big\_data/UKDA-6614-stata/stata/stata13\_se/ukhls/k\_indresp.dta"  
[1] "Attempting to find 17 variables"  
[1] "WARNING! Only 11 of the 17 requested have been found"  
[1] "read file in 0.459078073501587 seconds"  
[1] "slimming file..."  
[1] "extracted 11 variables in 0.263643026351929 seconds"  
[1] "extracting file: /Users/JonMinton/repos/economic\_inactivity/big\_data/UKDA-6614-stata/stata/stata13\_se/ukhls/l\_indresp.dta"  
[1] "Attempting to find 17 variables"  
[1] "All variables requested found"  
[1] "read file in 0.359766960144043 seconds"  
[1] "slimming file..."  
[1] "extracted 17 variables in 0.483259201049805 seconds"

[1] "Overall process took 0.191475816567739 minutes"

# A tibble: 7,058,281 × 4  
 pidp wave variable value   
 <dbl> <chr> <chr> <chr>   
 1 68001367 a sex male   
 2 68004087 a sex male   
 3 68006127 a sex female  
 4 68006135 a sex female  
 5 68006807 a sex female  
 6 68007487 a sex female  
 7 68007491 a sex male   
 8 68007495 a sex male   
 9 68007499 a sex male   
10 68008167 a sex female  
# ℹ 7,058,271 more rows

Didn’t I do something similar with household level variables (i.e. build an extractor?)

Yes. It’s in two parts…

# A tibble: 7,058,281 × 4  
 pidp wave variable value   
 <dbl> <chr> <chr> <chr>   
 1 68001367 a sex male   
 2 68004087 a sex male   
 3 68006127 a sex female  
 4 68006135 a sex female  
 5 68006807 a sex female  
 6 68007487 a sex female  
 7 68007491 a sex male   
 8 68007495 a sex male   
 9 68007499 a sex male   
10 68008167 a sex female  
# ℹ 7,058,271 more rows

To start with, let’s try to predict probability of being inactive in current wave given age and sex. I’m going to use [this webpage](https://www.r-bloggers.com/2021/04/decision-trees-in-r/) as a reference to borrow from.

|  |
| --- |
| Figure 1: Recursive partition decision tree showing algorithm for predicting whether someone is economically inactive given their age and sex |

Let’s try to interpret [Figure 1](#fig-rpart-inact-given-age-and-sex)

* If a person is under 20 years old, there is a 51% probability of being inactive. This group comprises 8% of the sample.
* If the person is 20 or older then:
  + If the person is 58 or older (at two waves before), there is a 51% change of being inactive. This group comprises 6% of the sample. (cumulative: 14%)
  + If the person is under 58 years of age (at two waves before):
    - If the person is male, there is a 9.5% probability of being inactive; this group comprises 37% of the sample. (cumulative 51%)
    - If the person is not male (i.e. female), there is a 24% probability of being inactive. This group comprises 49% of the sample (cumulative 100%)

## 2.1 Next steps

I promised to consider the following additional variables

* ethnicity
* number of dependent children
* equivalised household income
* previous wave’s economic status

We can do equivalised household income and number of dependent children now:

Now to do the CART…

|  |
| --- |
| Figure 2: Recursive partition decision tree showing algorithm for predicting whether someone is economically inactive given previous economic activity status and some other variables |

This model produces a warning, but a model still runs.

It illustrates how the model makes binary decisions when a variable has more than two states.

* The top level split is asking if the last status was either Employed or Unemployed.
  + If it was, then it goes to the second split
    - in the second split, it asks if the person was employed in the last wave:
    - if the person was employed, it predicts a probability of current wave inactivity of just 0.05, covering 71% of the population
    - if the person was not employed (so unemployed) it predicts a probability of 0.25 of being inactive in the current wave, covering 6% of the population.
  + it if was not, then it predicts inactivity in the current wave (a 0.78 probability, covering 22% of the data)
    - a no at the top level to employed or unemployed implies inactive in last wave
    - so inactivity in the last wave is the strongest predictor of inactivity in the current wave

Note 71%, 6%, and 22% do not add up to 100%. (Instead it adds up to 99%) This may be either due to missing data or to rounding error in showing %s to no decimal places. (I suspect it’s rounding error)

Note: even though this model was given more variables from which to recursively partition the data, it made use of just a single variable, the previous wave’s state. This highlights the ‘stickiness’ of economic inactivity states.

How about if I don’t give last wave’s status?

|  |
| --- |
| Figure 3: Recursive partition decision tree based on various variables excluding previous wave’s economic activity status |

It appears from [Figure 3](#fig-rpart-inact-no-last-status-etc) that if last wave’s status is not included in the model, other variables are more likely to be used as predictors. But to what extent are these variables proxies for economic activity status, especially equivalised monthly income?

Let’s now focus on probabilities of transition to economic inactivity given not economically inactive in last wave. This is a matter of just filtering the data then running the model again…. This is shown in **?@fig-rpart-inact-given-inact-etc**

|  |
| --- |
| Figure 4: Decision tree: prediction of economic activity given not economically inactive in previous wage |

Of course monthly income is in part (largely?) a consequence of economic activity status, so this might be a facetious predictor in this case. How about if we look at average equivalised monthly income in the previous wave? (The problem with this may be similar… equivalised monthly income in last wave may be determined quite a lot by economic activity status in last wave)

|  |
| --- |
| Figure 5: Decision tree: predicted probabilities of being economically inactive given not previously inactive, given various variables from last wave including monthly income |

Here only income in the previous wave matters as a predictor. An 11% probability of being inactive is predicted for those individuals with an equivalised income of less than £1020 pcm. Whereas a 3.8% probability is predicted for those with an income at or above this threshold.

Again, what happens if last period’s equivalised monthly income and last period were included in the model?

|  |
| --- |
| Figure 6: Decision tree: predicted probabilities of being economically inactive given not previously inactive, given various variables from last wave including monthly income and status last period |

Once the previous wave’s state is included, a combination of previous state and gender are preferred to equivalised income. If someone was employed in the previous wave, they are predicted to only have a 3.5% probability of being inactive in the current wave. Otherwise, if they are male they are predicted to have a 15% probability of being inactive, and if they are female they are predicted to have a 37% probability of being inactive.