OR60: *“You earn how much? You must be joking!”*

Author:
Colin Stewart, More Metrics Ltd colin.stewart@moremetrics.co.uk

September 2018

Acknowledgements:

Our models contain data provided by National Records of Scotland (Crown Copyright, OGL), Northern Ireland Statistics and Research Agency (Crown Copyright, OGL), Office of National Statistics (Crown Copyright, OGL).

Maps presented here contain data provided by OpenStreetMap (Copyright OSM contributors, CC-By) and Ordnance Survey (Crown Copyright & Database right, OGL)

What problem are we trying to solve?

- Business Context: *"You earn how much? You must be joking!"*
  - ***Situation:*** A loan provider (e.g. a Credit Union)
  - ***Requirement:*** Assess affordability (Income less outgoings)
  - ***Dilemma:*** How much checking of supplied income information should we do?

- Analytical Challenge:
  - Estimate earned income based on information collected routinely at application, such as:
    - What job do you do?
    - How many hours a week do you work on average?
    - Where do you live?

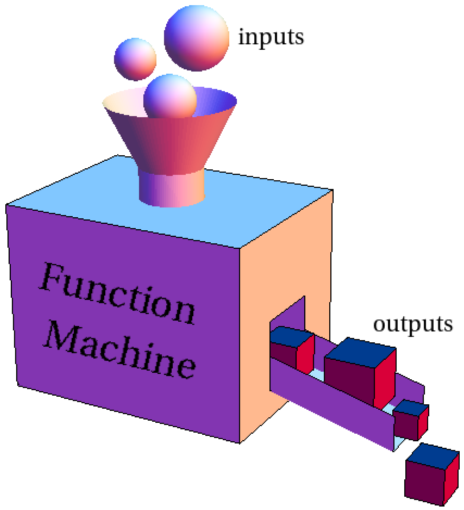
- Then make it a more interesting challenge:
  - Provide a model that will work for a new organisation with no data history
  - Provide income probability distributions rather than a single central estimate

Traditional modelling approach: Won’t work because there is no historic data to draw on

Select a sample population with historic data

Collate relevant personal data **and** income data

Build a model Score up new applicants



Disaggregation Model: Will work as it uses only open source data available to all (ASHE, census data, indices of deprivation etc.)

Source UK income data cut by different segments (ASHE)

Estimate local incomes by modelling on neighbourhood data

Estimate individual incomes by modelling on imputed micro data

Skilled / Routine
£20,000
40% of local pop

Professional
£40,000
30% of local pop

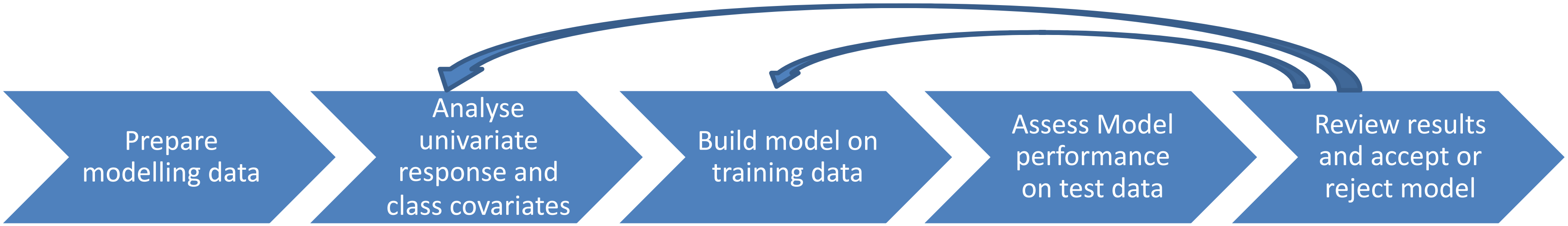
Manager
£50,000
30% of local pop

UK median
(Male/Female)
£28k / £19k

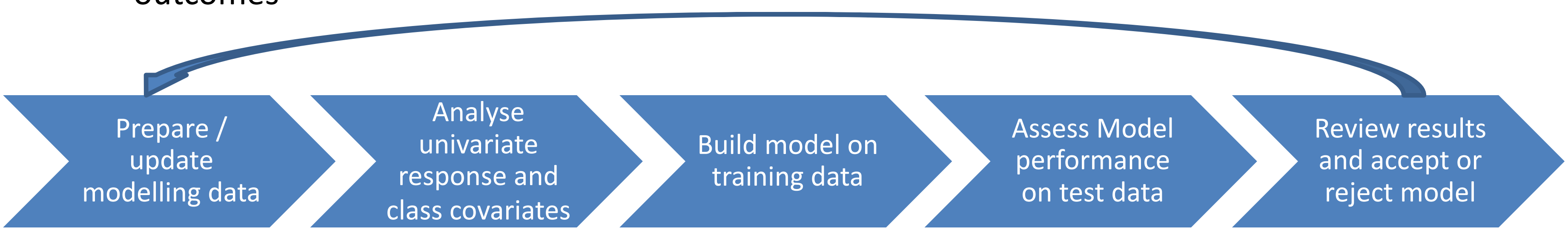
Local median
£35,000

Traditional and Disaggregation models compared.

- Traditional Model:
  - Individual outcomes are known.
  - Re-builds are done as needed to get a robust model with coefficients that "make sense"



- Disaggregation Model:
  - Individual outcomes are not known requiring an iterative approach (typically 20 plus iterations)
  - Aggregated outcomes are used as the initial target value
  - Model errors are apportioned after each iteration to update the target values
  - Process stops when the model is stable and predictions fit well to published aggregated outcomes



Case Study: Overview

- Model Hourly Pay using ASHE 2017 data as the target for people living in England

- Predictor dataset:
  - Index of Multiple Deprivation (IMD)
  - Output Area Classification (OAC) – ONS geo-demographic classification
  - Occupation mix (Soc2010 – 9 Major categories)

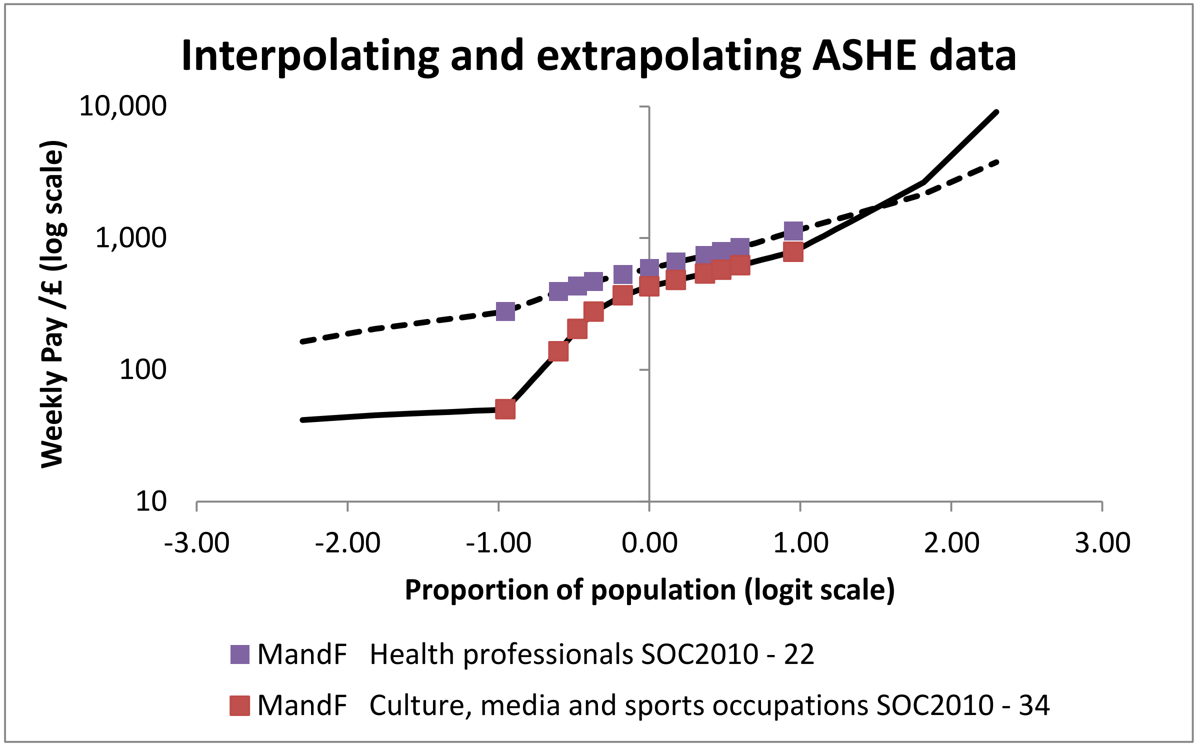
- Build disaggregation models to create scorecards that can be used to calculate mean, median and IQ range of hourly pay for an individual anywhere in England

Case Study: ASHE data overview

https://www.ons.gov.uk/surveys/informationforbusinesses/businesssurveys/annualsurveyofhoursandearningsashe

- The ONS Annual Survey of Hours and Earnings (ASHE) is well established with a long time series
  - Source is a c300,000 sample of employee jobs selected from HM Revenue & Customs (HMRC) PAYE records
  - Aggregated data is published in a range of tables split by percentiles plus a mean estimate
  - Percentiles provided are usually: 10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90

- It is superb data, but care is needed because pay data has extreme values as seen below



Case Study: Prepare the modelling dataset

Record Number	OA	Pcon (parliamentray constituency)	Region	IMD Decile	OAC	Occupation (Soc2010 Major)	Pcon Aggregated Value (ASHE)	Occupation Aggregated Value (ASHE)	Individual hourly pay (imputed)
1									?
2									?
3									?
4									?
.									?
53300									?

Data Type Key

Geographic Variables
Neighbourhood Variables
Individual Variables (imputed)
Hourly pay data (Target Values)

From the comfort of your home or office:

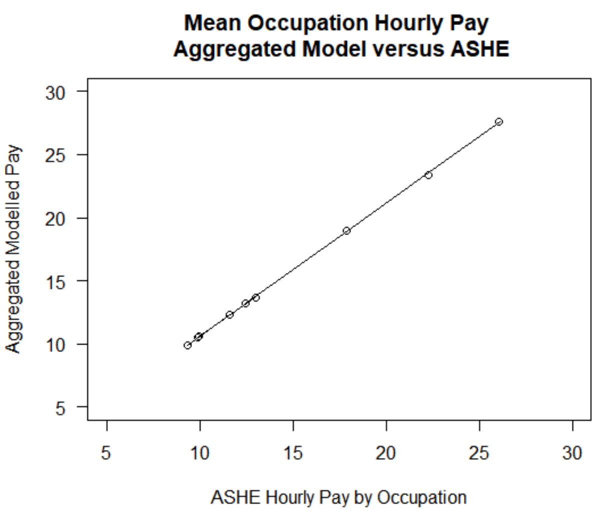
- "Knock on" a 100 front doors chosen at random in every parliamentary constituency (Pcon) across England. Record the Output Area (OA). This gives us lots of data to model on. A total of 53,300 cases.
- Source all the Geographic and Neighbourhood variables matched by OA
- Impute an occupation for one resident at each location using census data for the number of people with each type of occupation in the OA.
- Match in the aggregated pay data from ASHE by Parliamentary constituency and Occupation
- Impute an initial individual hourly pay estimate by using the Pcon aggregated value

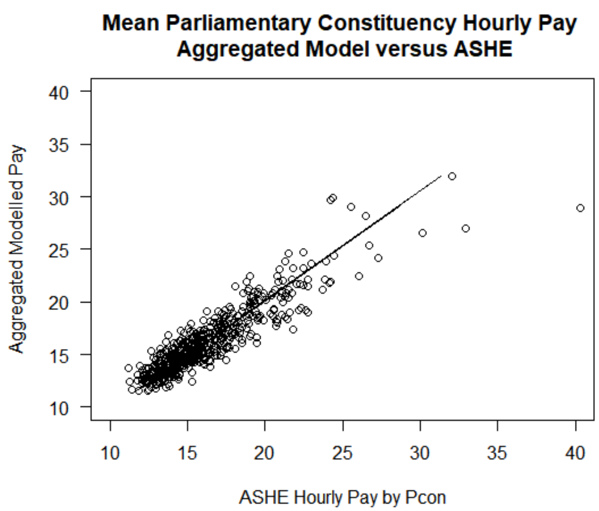
Case Study: Build the disaggregation model

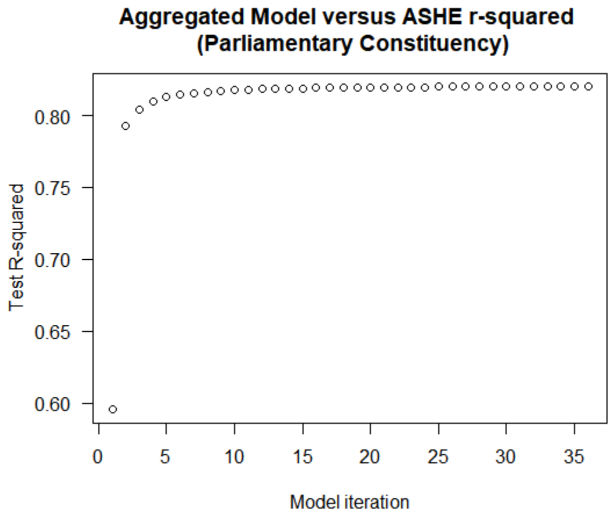
- Decide on the regression equation applied at each iteration. The R code might look like this:

```
Model <- lm( log(Hourly_Pay-5) ~ as.factor(IMD_Decile) + OAC + Occupation + Region,
             data = Train,
             weights = CaseWeight)
```

- Track Progress at each iteration by scoring up the test data: Example shown is for mean pay

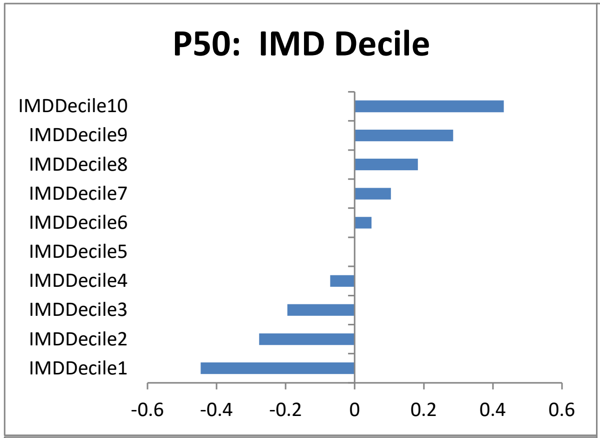


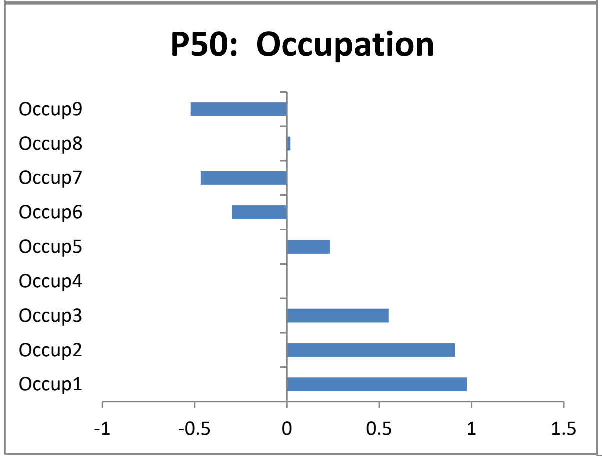


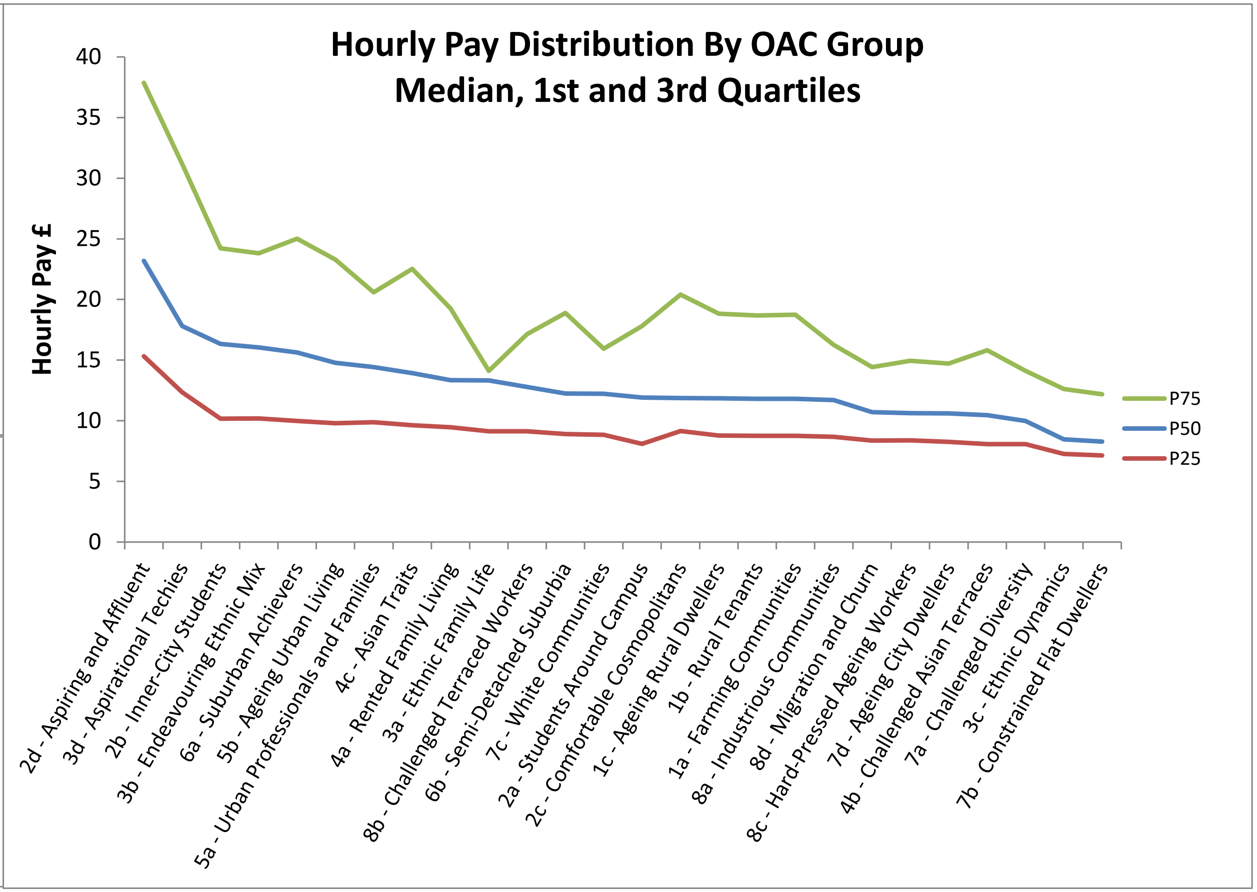


- Calculate the aggregated error terms by Pcon and Occupation at each iteration and update the target values. Stop when the Test R-squared value peaks or shows no material improvement

Case Study: Review model outputs including model coefficients and fitted values







Case Study: Apply our analysis to the real world

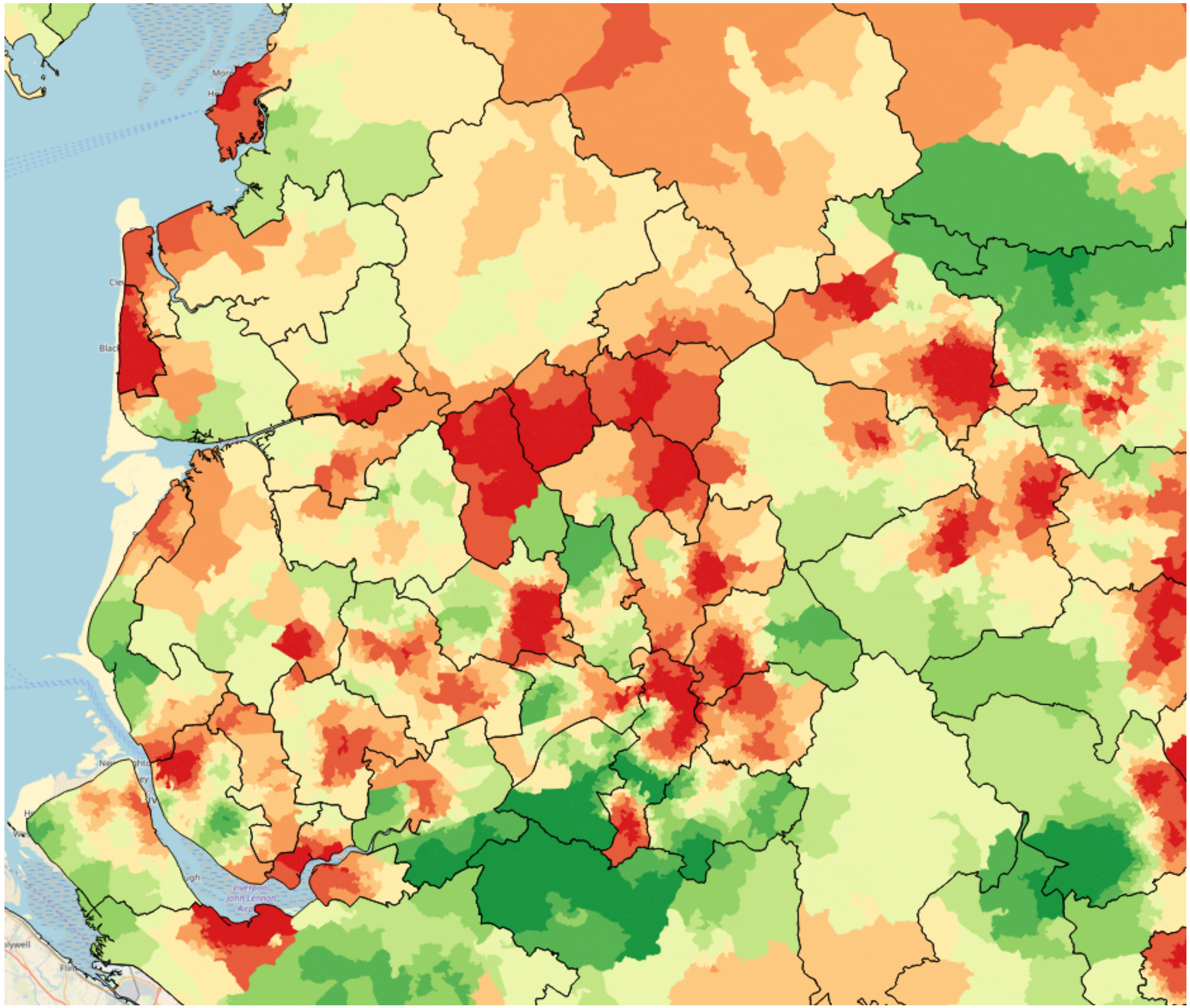
- Assessment of a Credit Union loan application: Using the Scorecard
  - ***Situation:*** Emma is a 19 year old trainee baker. Needs a car to get to work and is asking for £1000 to buy a second-hand car
  - ***Application:*** Use the "scorecard" to assess affordability having established outgoings. We can add a level of sophistication to this by using the IQ range to apply an appropriate age adjustment to the hourly pay rate estimate.

- Targeting Credit Union marketing activity: Leaflet Drop
  - ***Situation:*** Need to promote the Credit Union's activities to attract business on both sides of the balance sheet – providing loans and raising deposits
  - ***Application:*** Map the results for the median pay model, to help identify neighbourhoods that are in the right pay range for differently styled leaflets.

Case Study: Apply our analysis to the real world (cont)
Map of the North West. Which leaflet should we use and where?

Case Study: Apply our analysis to the real world (cont)
Our locally smoothed modelled income map provides some useful insight

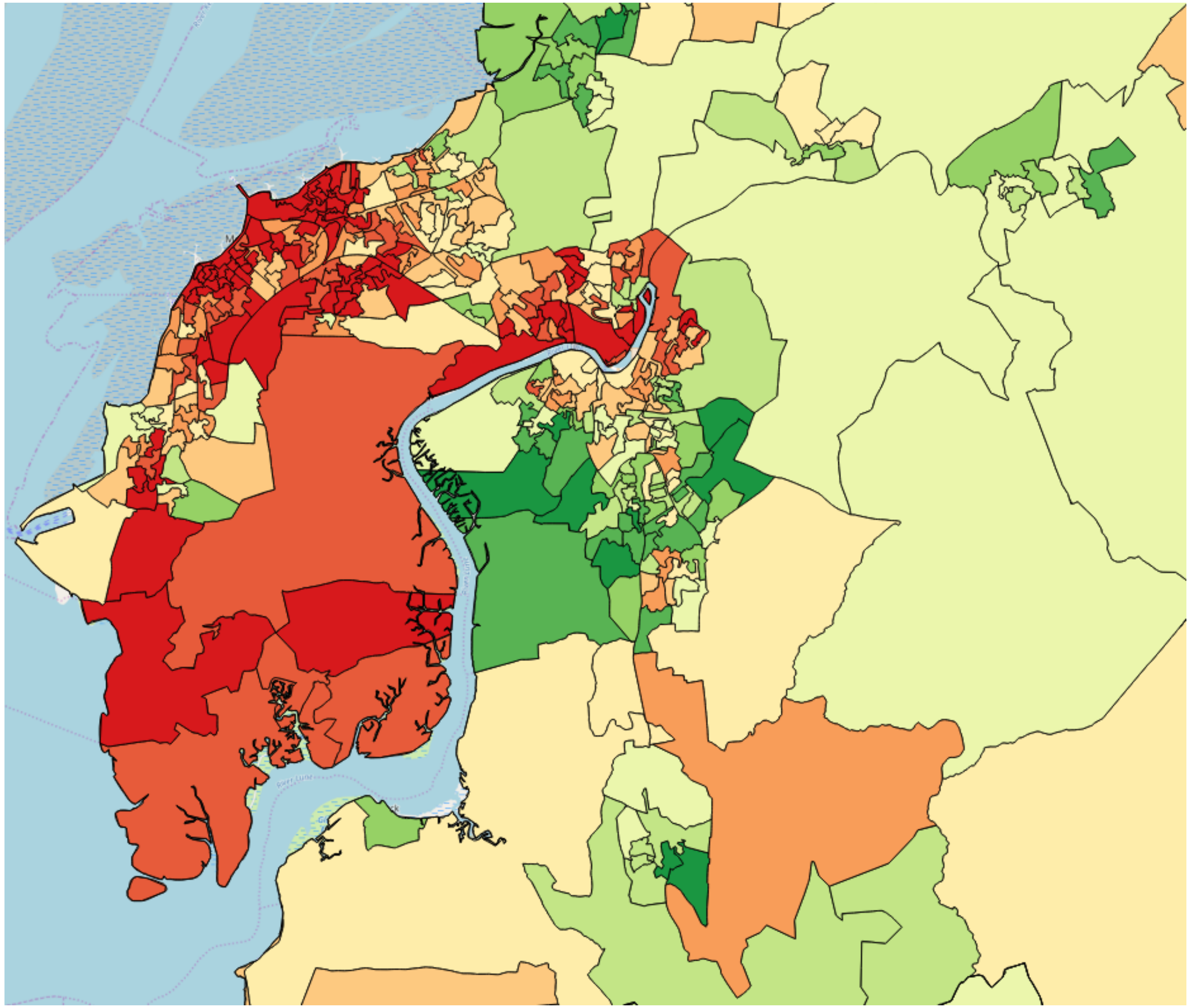
Median Hourly Pay
less than £10.54
£10.55 to £11.07
£11.08 to £11.56
£11.57 to £12.04
£12.05 to £12.59
£13.00 to £13.20
£13.21 to £14.00
£14.01 to £15.11
£15.12 to £16.60
£16.61 and over



Case Study: Apply our analysis to the real world (cont)
Map of Lancaster. Which leaflet should we use and where?

Case Study: Apply our analysis to the real world (cont)
Our raw modelled income map provides some useful insight

Median Hourly Pay	
	less than £9.55
	£9.56 to £10.34
	£10.35 to £11.02
	£11.03 to £11.69
	£11.70 to £12.41
	£12.42 to £13.21
	£13.22 to £14.21
	£14.22 to £15.54
	£15.55 to £17.61
	£17.62 and over



Case Study: Other Considerations

- Avoiding over-fitting:
  - Check for multicollinearity and use a suitable upper vif threshold
  - Use a process to add / remove model variables (e.g. "Step"). But be aware of performance issues if there are lots of predictors
  - Use coarse classing or consider converting categorical data to covariates to increase degrees of freedom
  - Use train / test datasets and a stopping rule applied to the test data

- Mitigating the reliance on 2011 census data
  - (Over) sample from Output Areas that are observed not to have changed (e.g. no change in postcodes; stable land registry data; stable mid-year population)
  - Supplement with more recent data (open source, and proprietary if available)
  - Build a series of models at different time points for the target variable starting with 2011 and track scorecard terms as you move away from 2011

Scaling up from the case study to provide more functionality

- More Metrics full income model covers:
  - Income distributions across 10 pay levels
  - Separate estimates for hourly pay and weekly pay
  - Split by 1300 imputed micro data combinations of Sex(2) x Occupation(25) x Age Band(13) x Work Pattern(2)
  - Further split by all UK Output Areas (230k)

  In total we calculate about 3 billion estimates each for hourly pay and annual pay which we re-configure to provide a variety of income datasets for end users

- An expanded set of about 170 predictor variables is also used

- Let's see what this looks like for a single Output Area in Bristol incorporating postcode BS7 8DN

An overview of income levels local to BS7 8DN using the More Metrics full income model

- Geographic locations that incorporate BS7 8DN
  - Parliamentary Constituency (E14000602) Bristol West
  - Lower Super Output Area (E01014670)
  - Output Area (E00074079)
    - Classified by ONS as "5a2. Multi-Ethnic Professionals with Families"

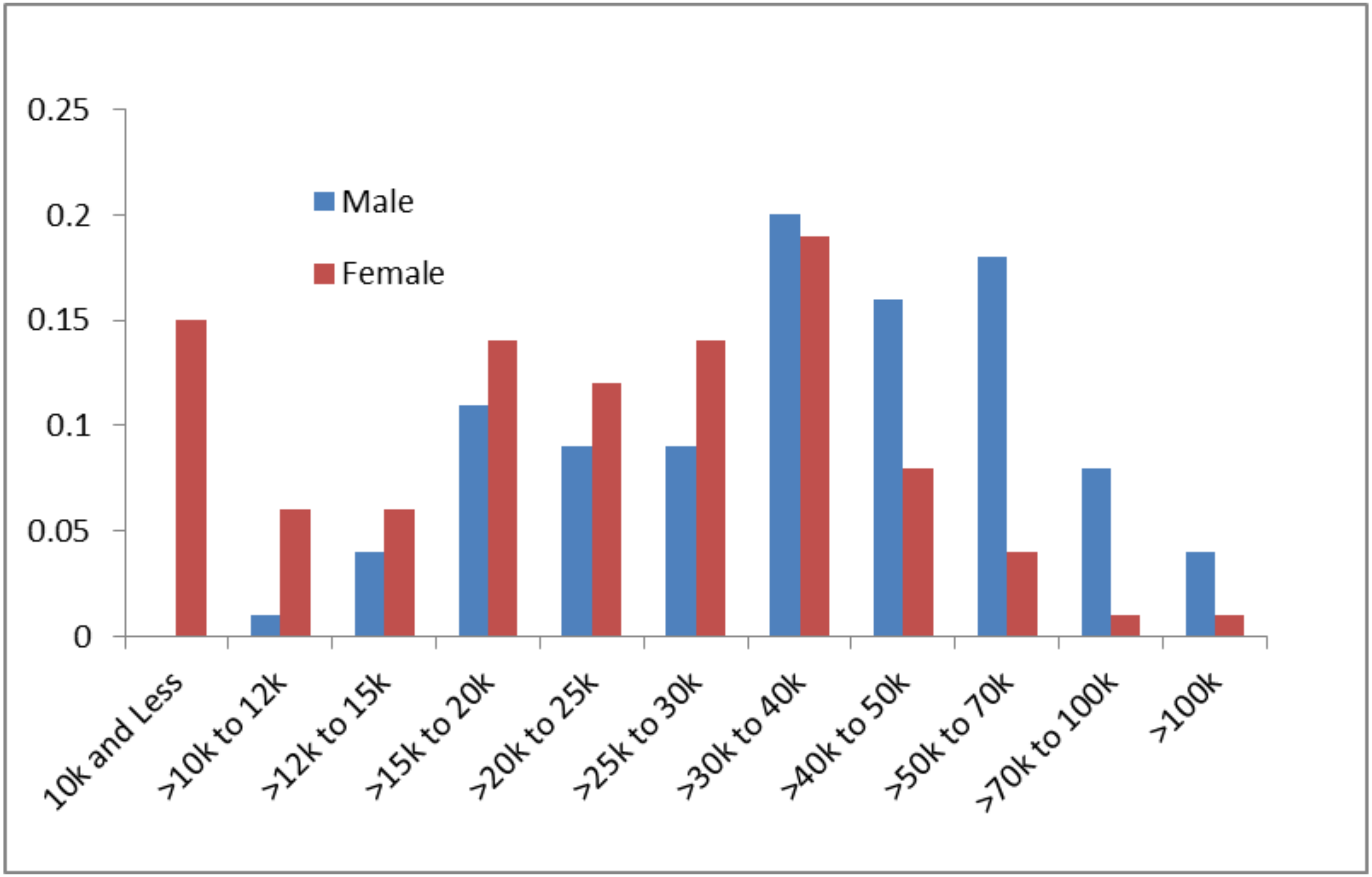
- Median Annual Pay for different geographies

Locations incorporating SE1 3UZ	Male median annual pay	Female median annual pay
UK*	£28,400	£18,700
Parliamentary Constituency*	£30,700	£22,500
OA**	£38,000	£23,800

Source: *ASHE survey (2017 provisional) and **More Metrics model value

Annual Pay Distribution for OA E00074079 (Males and Females)

Estimate of household earned income: 176 workers living in 111 households with median individual earnings of about £30,700 equates to median household earned income of c£49k (2011 census population + ASHE 2017 provisional).



Thank you for listening

- More Metrics is a start-up with 3 people involved part-time

- We provide UK-wide, small-area datasets and modelled output:
  - Mortality related (e.g. death rates, biological age)
  - Health and lifestyle related (e.g. obesity, smoker)
  - Income and wealth (e.g. earned income, pensioner income, inheritance tax)
  - Other (e.g. university entry rates, fuel poverty)

- Distribution is through selected partners and direct to end users

- We are interested in working with Credit Unions. We are also open to working with academics and the wider OR community

- Contact: colin.stewart@moremetrics.co.uk