CITY UNIVERSITY LONDON

Predicting Non-Response with Small-Area Auxiliary Data

Rainer Schnell Kathrin Thomas

Centre for Comparative Social Surveys,
City University London

May 26, 2016

Acknowledgements

We would like to thank:

- Rory Fitzgerald, PI of the ADDResponse Project, for securing the funding and making the ADDResponse data available to us,
- Sarah Butt, ADDResponse Project Manager, and
- Kaisa Lahtinen, ADDResponse Researcher, for the data collection and processing as well as for their helpful support and their patience in answering numerous and tedious questions.

CITY UNIVERSITY LONDON

Background

- Surveys increasingly suffer from unit non-response
- Auxiliary data may be useful to develop predictive models and/or corrective weights given that two core criteria are met (Schnell, 1993; Bethlehem, 2009):
  1. Records are available for (non-)respondents
  2. The auxiliary variables highly correlate with substantive survey questions

Types of Auxiliary Data

Individual or aggregate level...

1. Administrative data, but privacy concerns and risk of deductive disclosure
2. Commercial data, but in addition to potential privacy concerns also worries about completeness, accuracy, and processing of these data (Pasek et al., 2014) and financial constraints

The ADDResponse Data

- UK sample of the ESS Round 6 (2012/2013)
  - Random population sample based on the Royal Mail Postal Address File
  - 226 Primary Sampling Units (PSUs)
  - 4,520 individual records: 2,289 respondents (50.6%), 1,676 non-respondents (37.1%), 555 ineligible records (12.3%)
- ADDResponse merged aggregate administrative and commercial data

CITY UNIVERSITY LONDON

Hard to Count Measures (HTC)

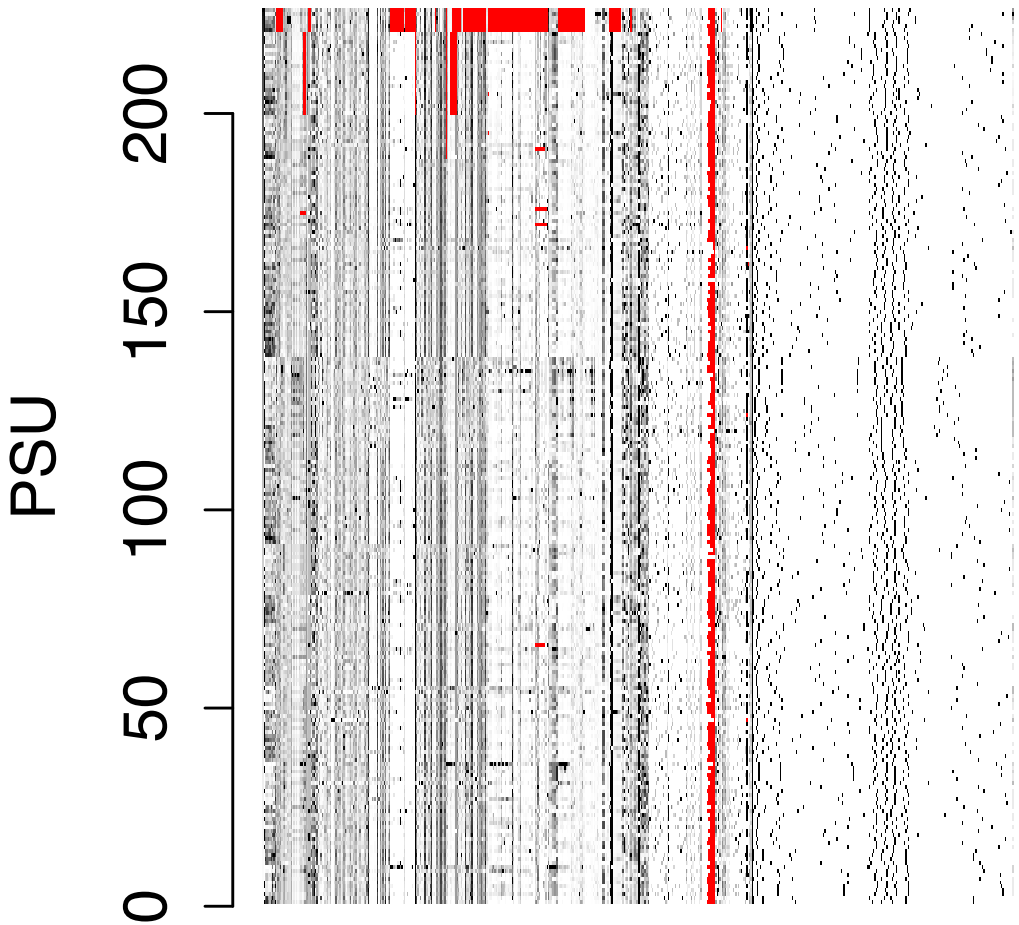
- HTC="national categorisation of areas designed to predict the level of non-response" (Abbott and Compton, 2014)
- Allowed effective allocation of resources in 2011 Census
- Separate HTC scores for England and Wales (ONS), Scotland (NRS) and Northern Ireland (NISRA)
  - England, Wales, Scotland: LSOA level, identical procedures
  - Northern Ireland: OA level, amended procedure
- We have rescaled the HTC measures using the respective quartiles in the devolved region

Empirical Strategy

1. Systematic exploration of the data quality
2. Recoding of core variables
3. Dimension reduction
4. Area-classification
5. Clustering
6. Predictive modelling using classification trees
7. Development of propensity weights
8. Validation of the propensity weights

CITY UNIVERSITY LONDON

Missing Data Patterns after Imputation (1)

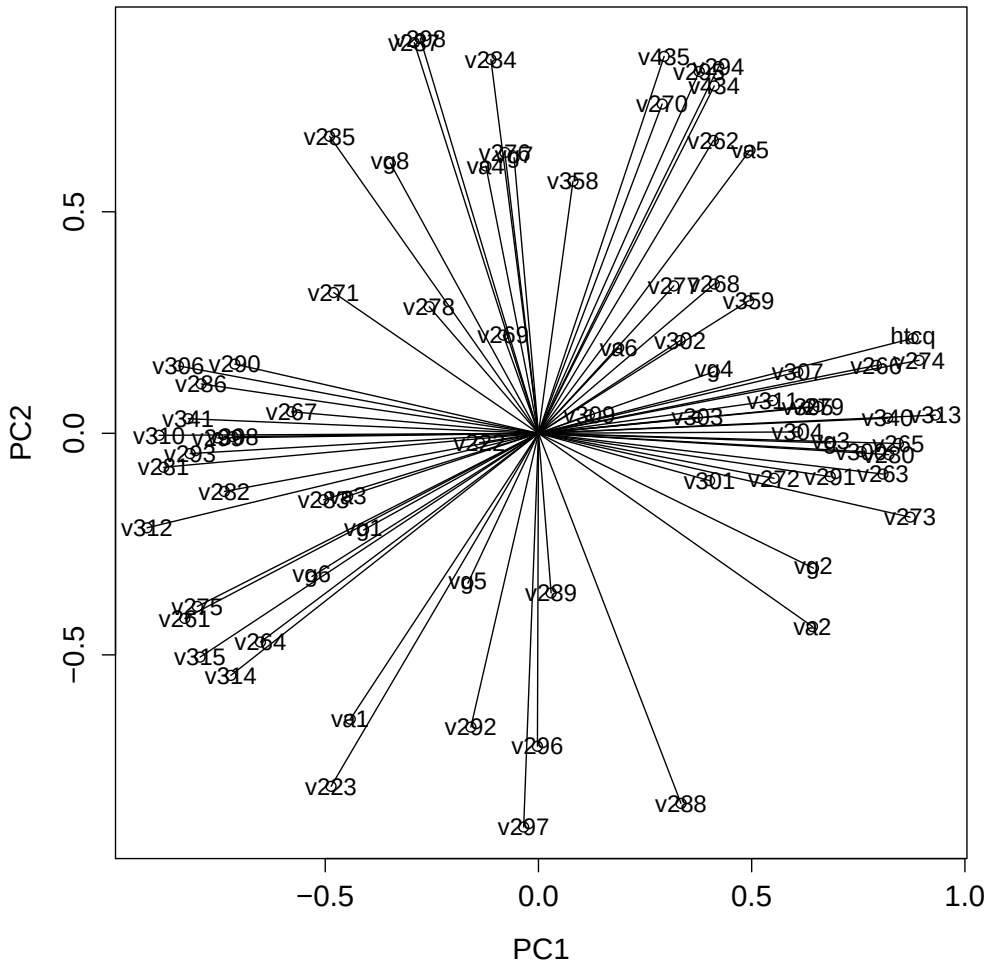


Missing Data Patterns after Imputation (2)

- Data aggregated on different levels
- But, also systematic missingness due to devolved government
  - Imputation from higher aggregation level whenever possible
  - Imputation of regional averages, e.g., proportion of Muslims in Northern Ireland (only 6 PSUs)
  - Exclusion of commercial and Points of Interest data

CITY UNIVERSITY LONDON

Dimension Reduction: Loadings Plot (1)

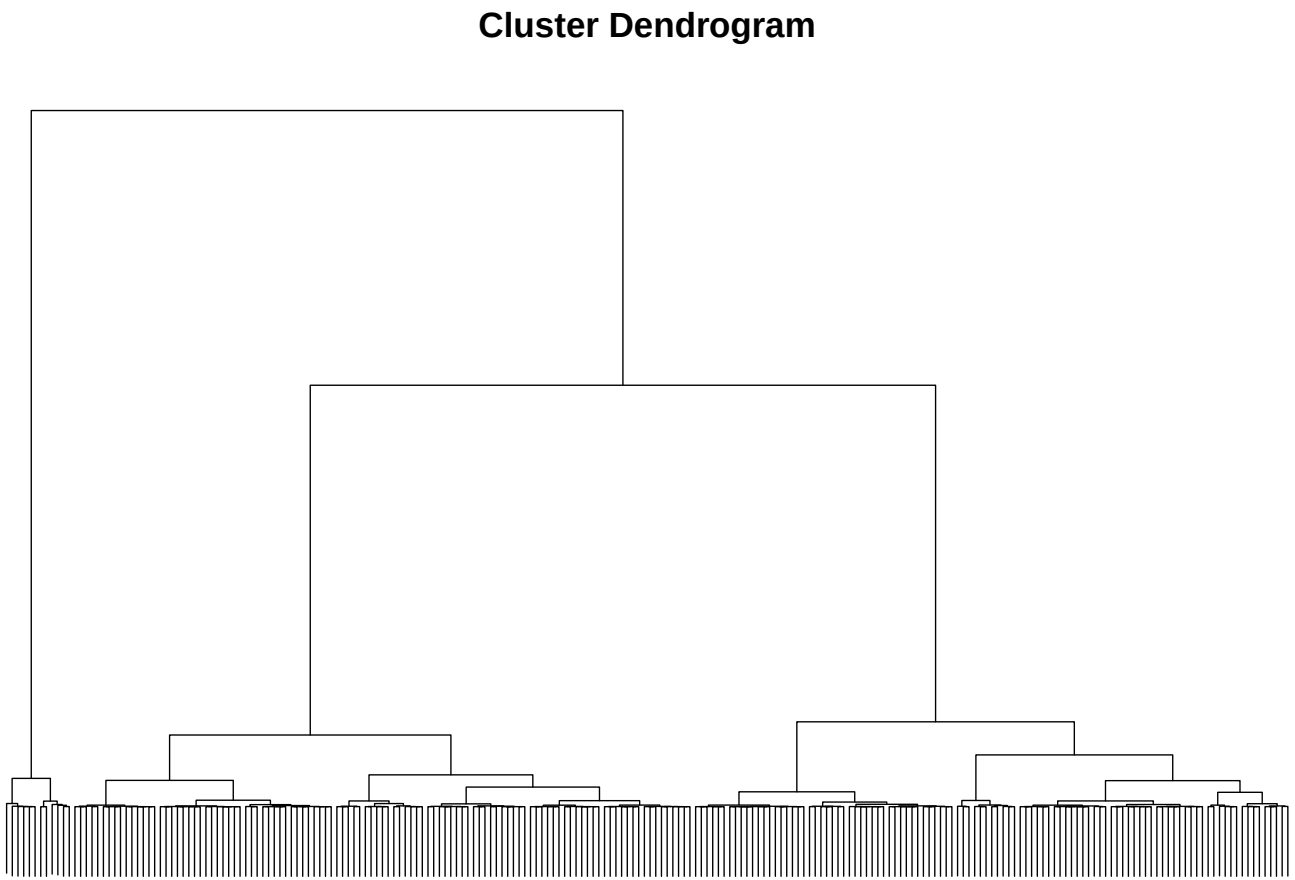


Dimension Reduction: Loadings Plot (2)

- PCA on the pairwise correlation matrix revealed different dimensions, which we have plotted in a loadings plot
- The smaller the angle between to lines is, the higher is the correlation between the variables
  - E.g., high correlation of HTC, v307, v266, v274, and v311, that is the HTC, proportion of Muslims, flats, single HH, and commuters (bike/foot)
  - E.g., high correlation of v434, v435, v270, v294, and v295, that is the proportions of (out-of-)job benefits receivers, (long-term) unemployed, single HH with dependent children

CITY UNIVERSITY LONDON

Area Classification: Cluster Analysis (1)



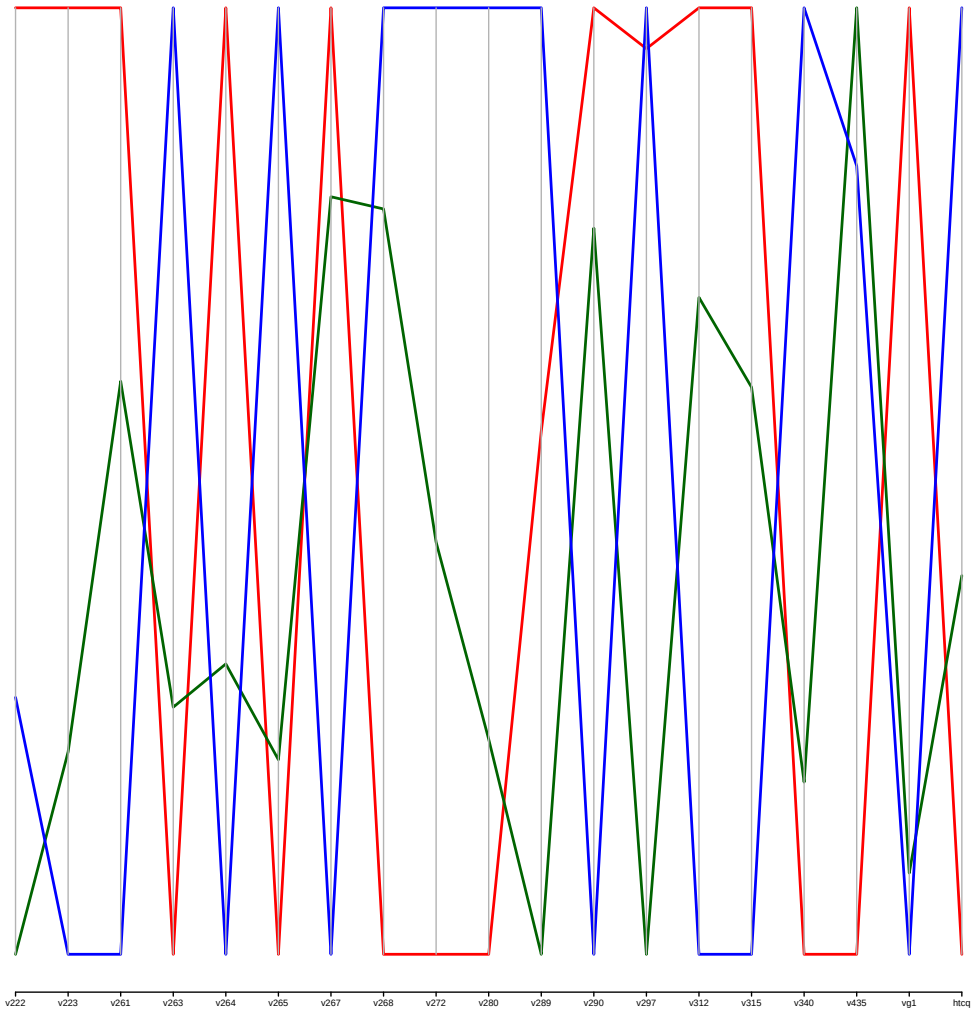
CITY UNIVERSITY LONDON

Area Classification: Cluster Analysis (2)

- Visual inspection of the dendrogram suggests 3, maybe even 5 clusters
- Cluster analysis using R's NbClust allows calculating 24 different indices and provides us with a poll identifying the number of clusters in our data
- Five indices each suggested a 2- or a 3-cluster solution, another 4 indices proposed 4, 12, 14, or 15 clusters, 10 did not converge
- We opted for a 3-cluster solution in our analysis:
  - Cluster 1: 109 PSUs (48.2%)
  - Cluster 2: 105 PSUs (46.5%)
  - Cluster 3: 12 PSUs (5.3%)

CITY UNIVERSITY LONDON

Area Classification: Cluster Identification (1)



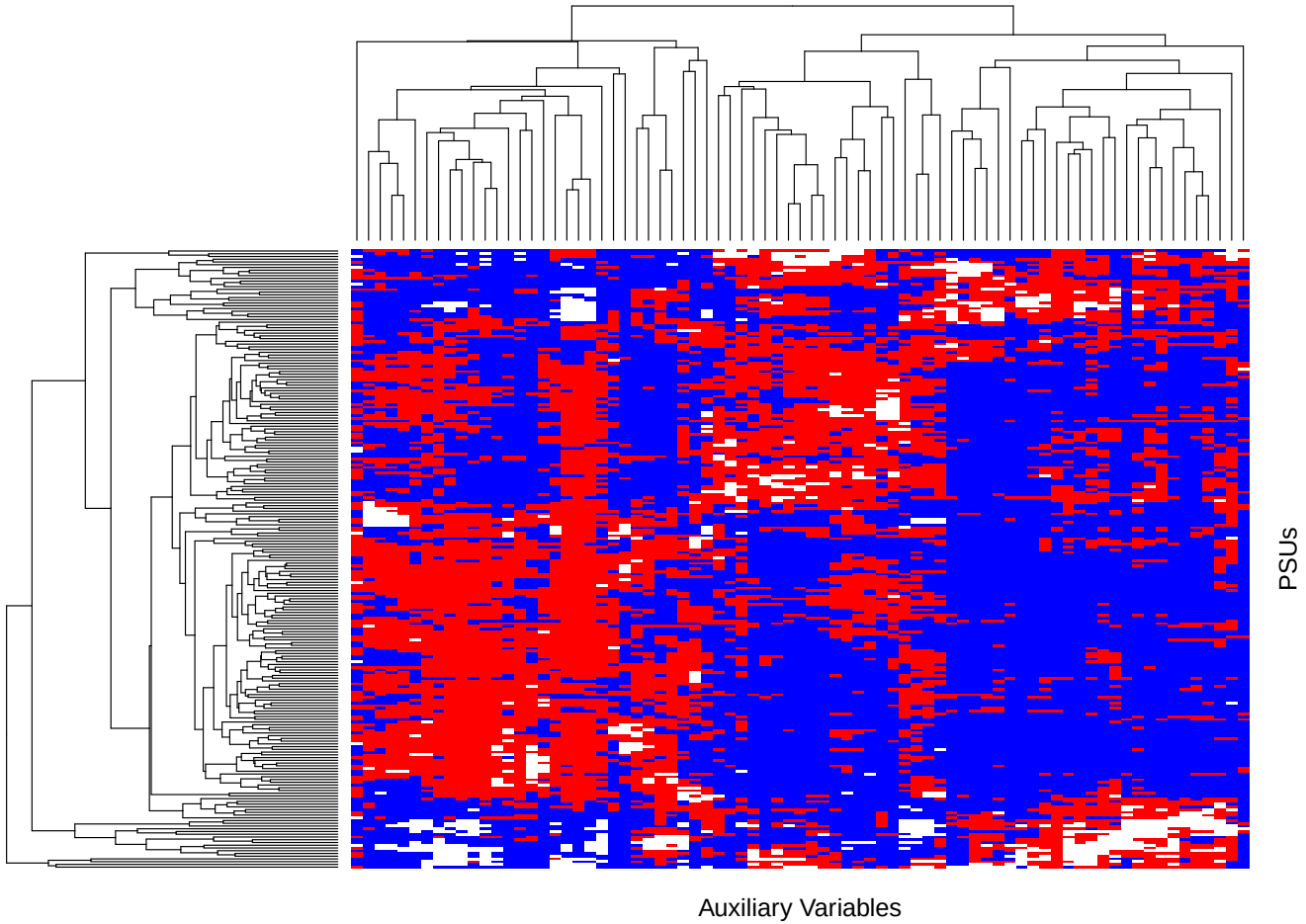
CITY UNIVERSITY LONDON

Area Classification: Cluster Identification (2)

- Cluster 1 (n=109, red line):
  Easy to reach, better off, predominantly retired residents (65+)
- Cluster 2 (n=105, blue line):
  Harder to reach working population in densely populated areas
- Cluster 3 (n=12, green line):
  Comparatively easy to reach mix of working and job seeking population in less suburban areas

CITY UNIVERSITY LONDON

Clustering: Heatmap (1)



CITY UNIVERSITY LONDON

Clustering: Heatmap (2)

- The graph pots the values represented in the data matrix and dendrograms for the PSUs and auxiliary variables
- Suggests similar number of PSU clusters
  - Two more distinct, large PSU clusters and a more heterogeneous third PSU cluster
- Likewise, approximately three larger clusters of auxiliary variables
  - Once again two more distinct larger clusters and a third more heterogeneous cluster

CITY UNIVERSITY LONDON

Preliminary Conclusions (1)

- Data quality, esp. missingness, was an issue, which we could effectively deal with
- Analysis on the PSU level suggests common dimensions in the data
- Identification of 3 PSU- and 3 auxiliary variable-clusters

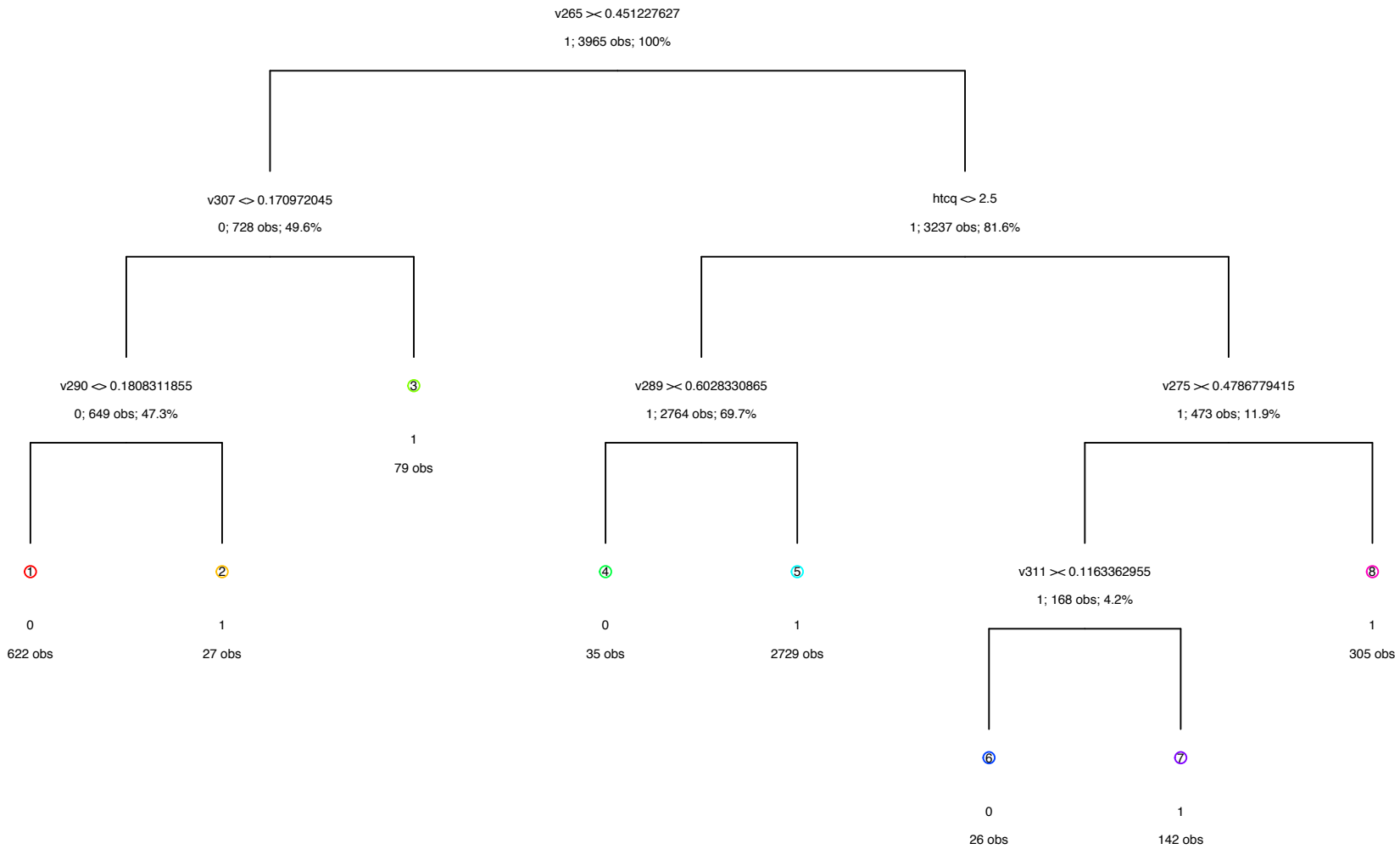
CITY UNIVERSITY LONDON

Predictive Modelling: Classification Trees

- DV: individual level binary response code excluding all ineligible records (response=1, n=2,289; non-response=0, n=1,676; RR=57.7%)
- Method: Classification Tree (CT) using R's rpart and maptree packages
  - CTs allow classifying large data into set outcome categories (*here: response vs. non-response*)
  - But, drawback is over-fitting
  - Yet, preferable to logistic regression (LR) as analysis with LR would be more difficult due to multi-collinearity issues and potential interactions
- CT for the full data set as well as by PSU cluster

CITY UNIVERSITY LONDON

Predictive Modelling: Classification Trees



Total classified correct = 55.8 %

CITY UNIVERSITY LONDON

Predictive Modelling: Classification Trees

- Root node: proportion of flats
- Left subtree: child nodes for proportion of Muslim population, PT employment
- Right subtree: child nodes for HTC, proportions of FT employment, married and commuters (foot/bike),
- Correctly classified observations: 55.8% (naive estimate)

Predictive Modelling: Classification Tree by Cluster

- Cluster 1 (n=1,978; RR: 59.4%)
  - Root node: proportion of receivers of job seeker allowance
  - Only a left subtree with two child nodes: proportions of PT and FT employment
    - Another left subtree for FT employment with two child nodes: proportion of people in managerial positions (aged 16-74)
  - But, only 39.0% of cases correctly classified

Predictive Modelling: Classification Tree by Cluster

- Cluster 2 (n=1,813; RR=56.0%)
  - Root node: proportion of flats
  - Right subtree with 1 child node: proportions of HH with dependent children
  - Left subtree with 1 child node private rentals
    - Another left subtree for HH with dependent children with 3 child nodes: proportions of married people, PT employment, and single HH (aged 35)
  - 69.8% (!) of the observations correctly classified

CITY UNIVERSITY LONDON

Predictive Modelling: Classification Tree by Cluster

- Cluster 3 (n=174; RR=56.9%)
  - Root node: index of multiple deprivation
  - child nodes: proportion of married, people aged 44-65, young children (5-15 year olds)
  - But, only 45.7% correctly classified

Predictive Modelling: Preliminary Conclusions (2)

- Classification tree for the full data set suggested partitioning by 8 core auxiliary variables, among them the HTC score
  - But, relatively low proportion of correctly classified observations (55.8%) considering the marginal probability of the outcome code
- Looking at CT by PSU cluster:
  - Different auxiliary variables predict (non-)response
  - But, rather poor fit of the trees given the marginal probability of the outcome code
  - Exception: Cluster 2 Tree (almost 70%)

Next steps: Propensity Weights and Validation

- Validation of the results on the basis of the CTs
- Construction of propensity weights for non-response
- Validation of weights by looking at variables not used for weighting or clustering, such as labour force participation or youth unemployment

CITY UNIVERSITY LONDON

Conclusion and Discussion

- Value of auxiliary variables limited
- Trade-off between cost and benefits
  - Collecting all available information vs. systematically collecting few, but complete and high quality indicators
- Useful exercise to further explore non-response and the use of auxiliary data in the European context
- Perhaps survey participation is after all a low cost decision, which is just hard to predict using area classifications

References

Abbott, O. and Compton, G. (2014). Hard-to-survey populations. In Tourangeau, R., Edwards, B., Johnson, T. P., Bates, N., and Wolter, K. M., editors, *Counting and estimating hard-to-survey populations in the 2011 Census*, pages 58–81. Cambridge University Press.

Bethlehem, J. (2009). *Applied survey methods: A statistical perspective*, volume 558. John Wiley & Sons.

Pasek, J., Jang, S. M., Cobb, C. L., Dennis, J. M., and Disogra, C. (2014). Can marketing data aid survey research? examining accuracy and completeness in consumer-file data. *Public Opinion Quarterly*, 78(4):889–916.

Schnell, R. (1993). Die Homogenität sozialer Kategorien als Voraussetzung für Repräsentativität und Gewichtungsverfahren. *Zeitschrift für Soziologie*, 22(1):16–32.

Contact Information

This presentation is the basis for a prospective publication. For further suggestions, information, or enquiries about our work please contact:

kathrin.thomas@city.ac.uk

Thank you very much for your attention!