SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

Utility of synthetic microdata generated using tree-based methods

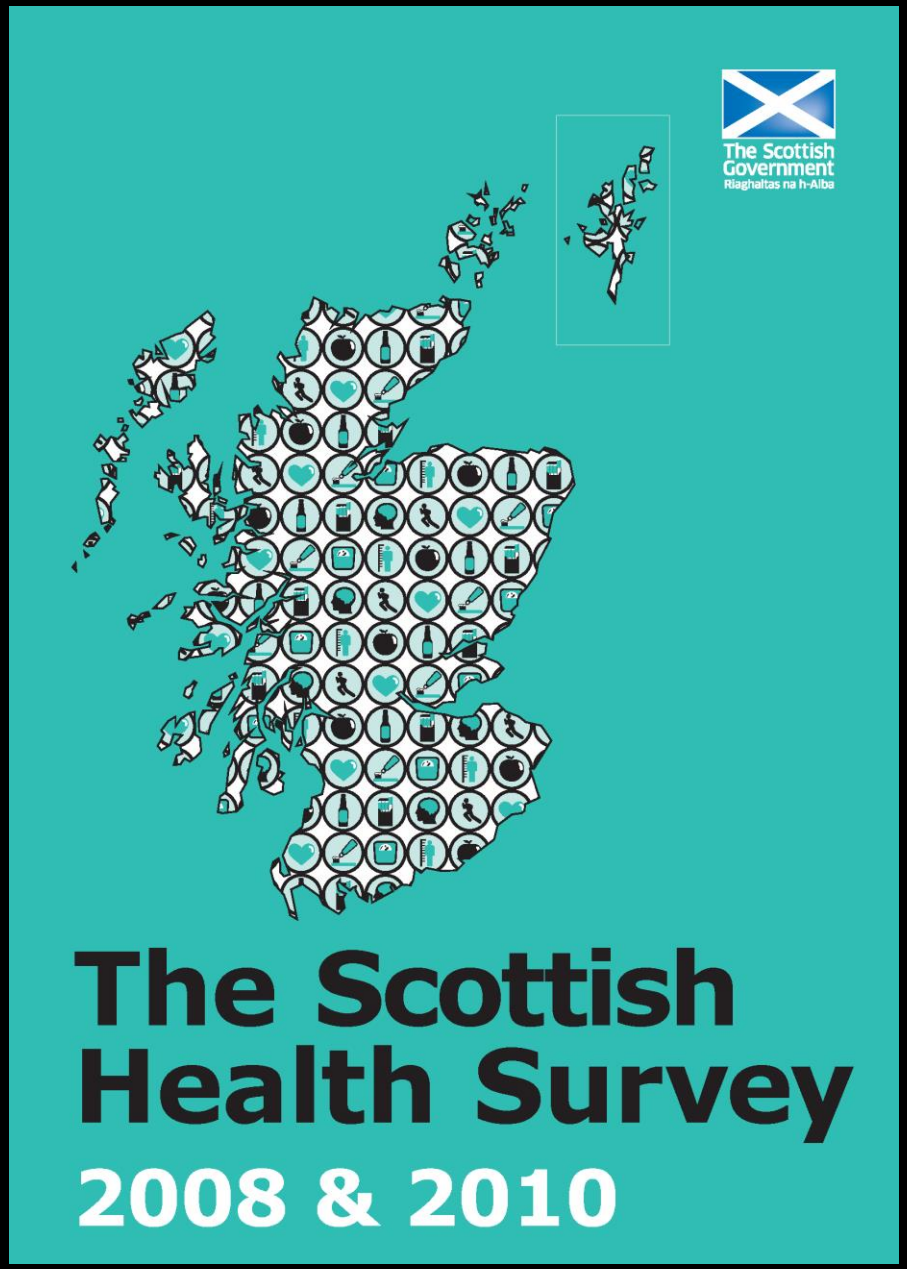
Beata Nowok

Administrative Data Research Centre – Scotland

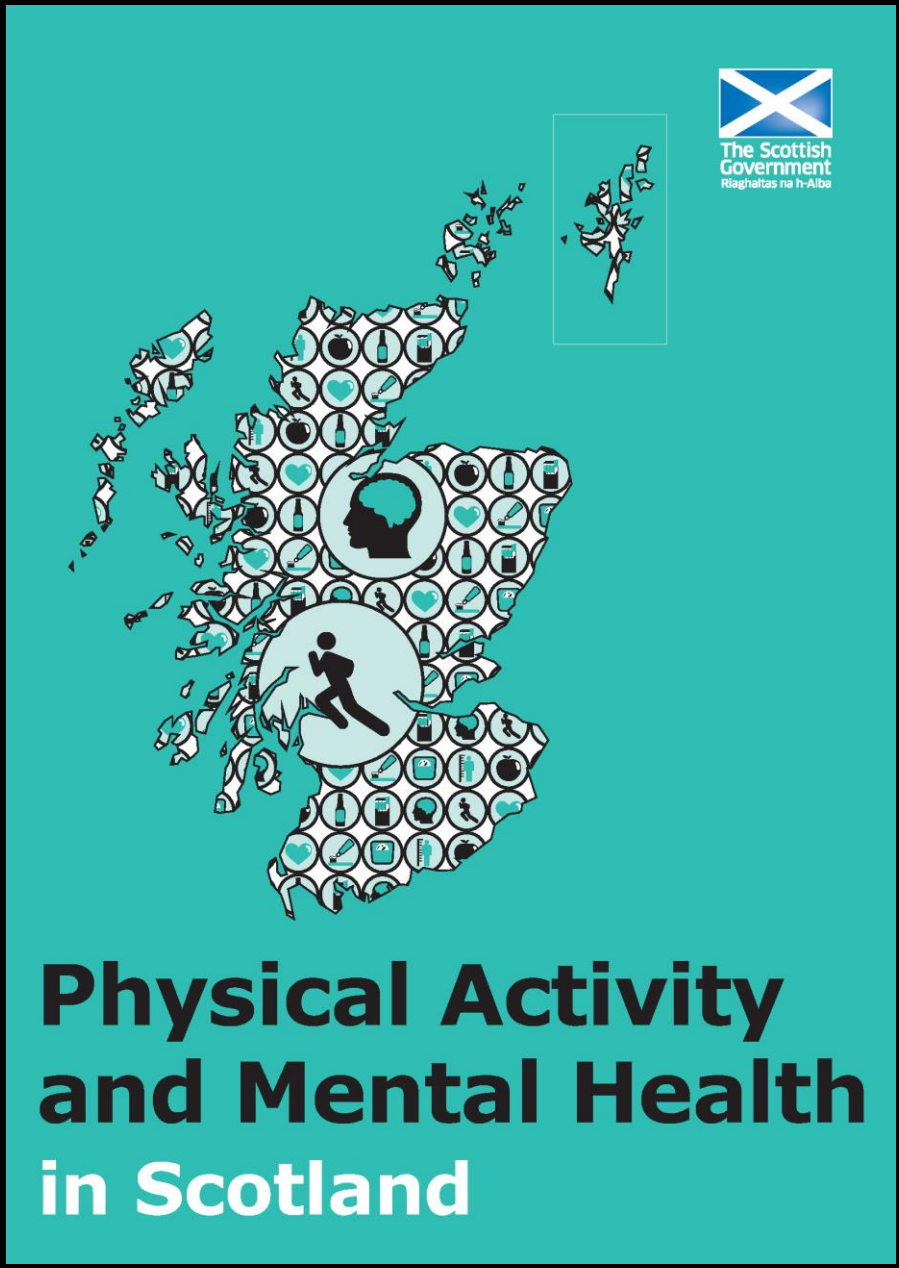
Administrative Data Research Network

An ESRC Data Investment

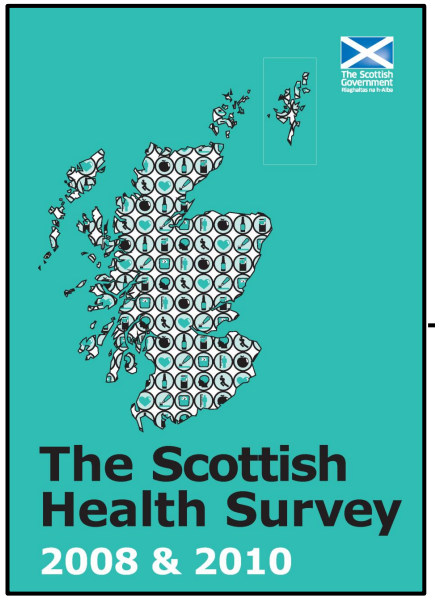
DATA


The Scottish Government
Riaghaltas na h-Alba
The Scottish Health Survey
2008 & 2010

RESEARCH TOPIC



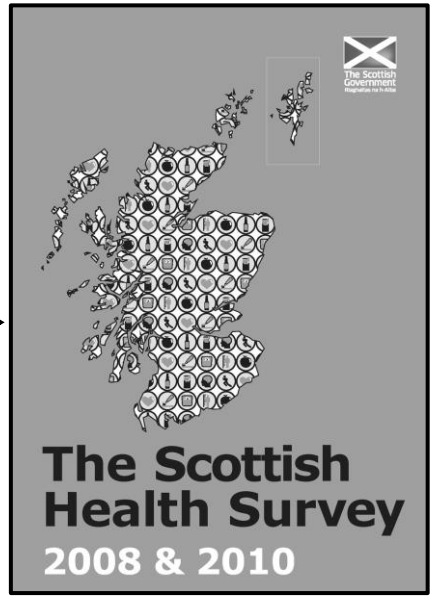
observed



Generating synthetic data

$$Y_j \sim (Y_0, Y_1, \ldots, Y_{j-1})$$

Tree-based methods



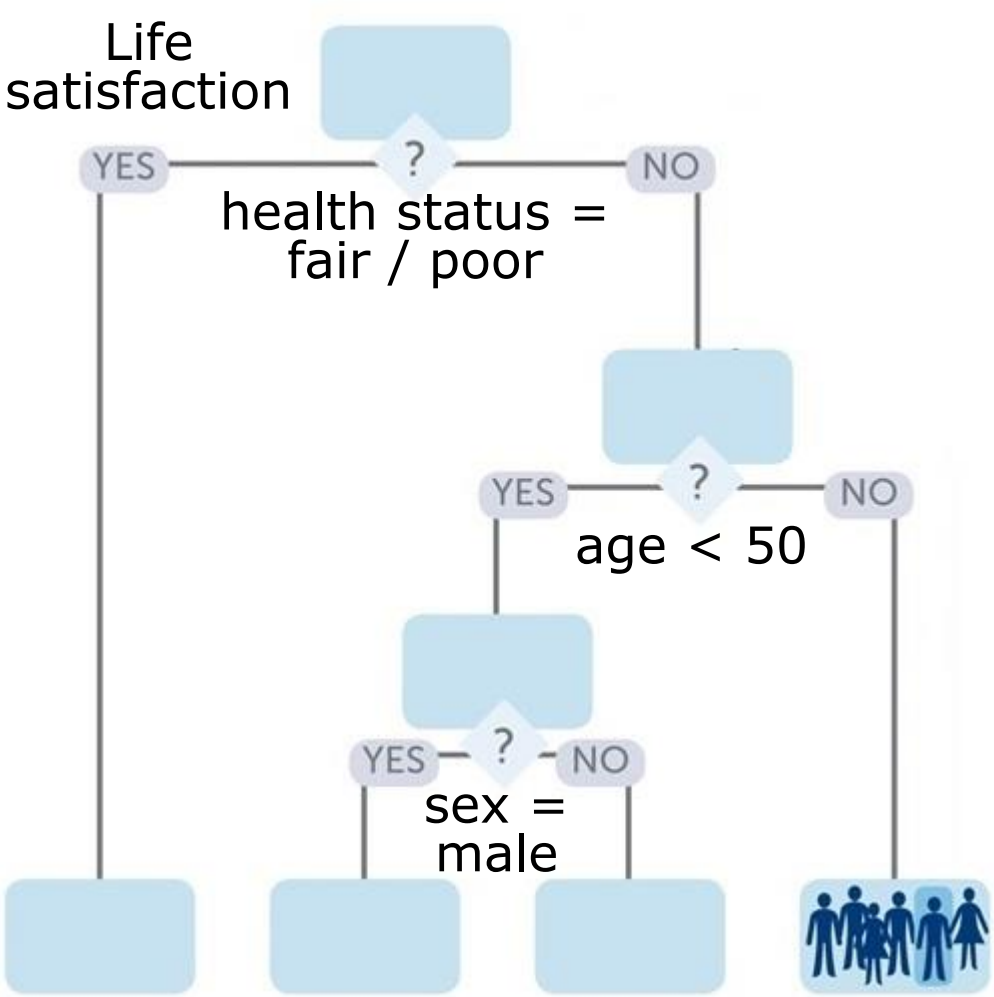
synthetic

Utility evaluation

Tree-based methods

- Classification and regression trees (CART)
- Bagging
- Random forest

CART

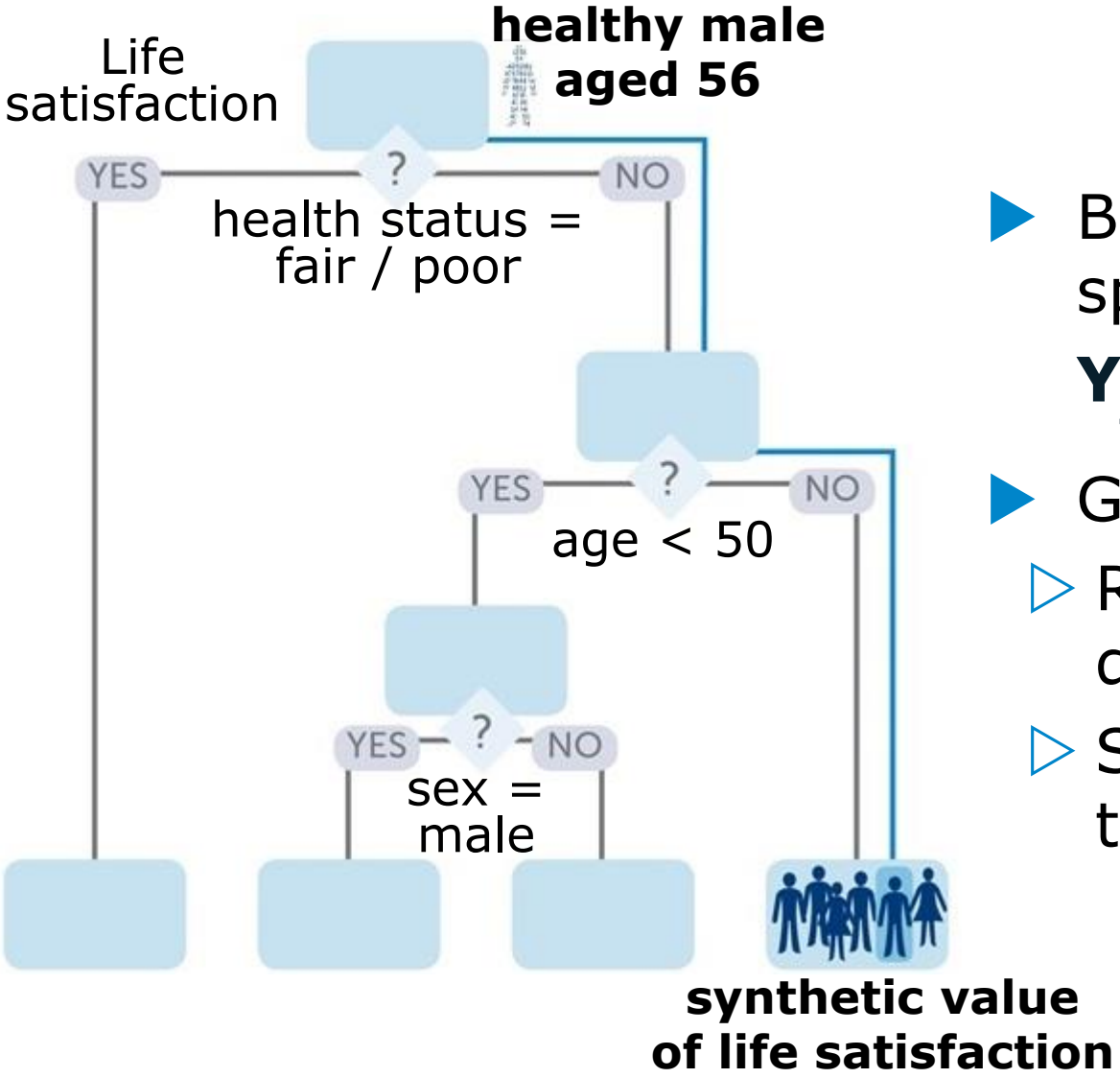


- Build a tree (binary splits)

  $\mathbf{Y_j \sim (Y_0, \ldots, Y_{j-1})}$

- Generate $\mathbf{Y_j}$ by:
  - Running $\mathbf{Y_0, \ldots, Y_{j-1}}$ down the tree
  - Sampling $\mathbf{Y_j}$ from the leaves

CART



- Build a tree (binary splits)

  $\mathbf{Y_j \sim (Y_0, \ldots, Y_{j-1})}$

- Generate $\mathbf{Y_j}$ by:
  - Running $\mathbf{Y_0, \ldots, Y_{j-1}}$ down the tree
  - Sampling $\mathbf{Y_j}$ from the leaves

CART splitting rules

- Association with the response variable
- Gini index and deviance

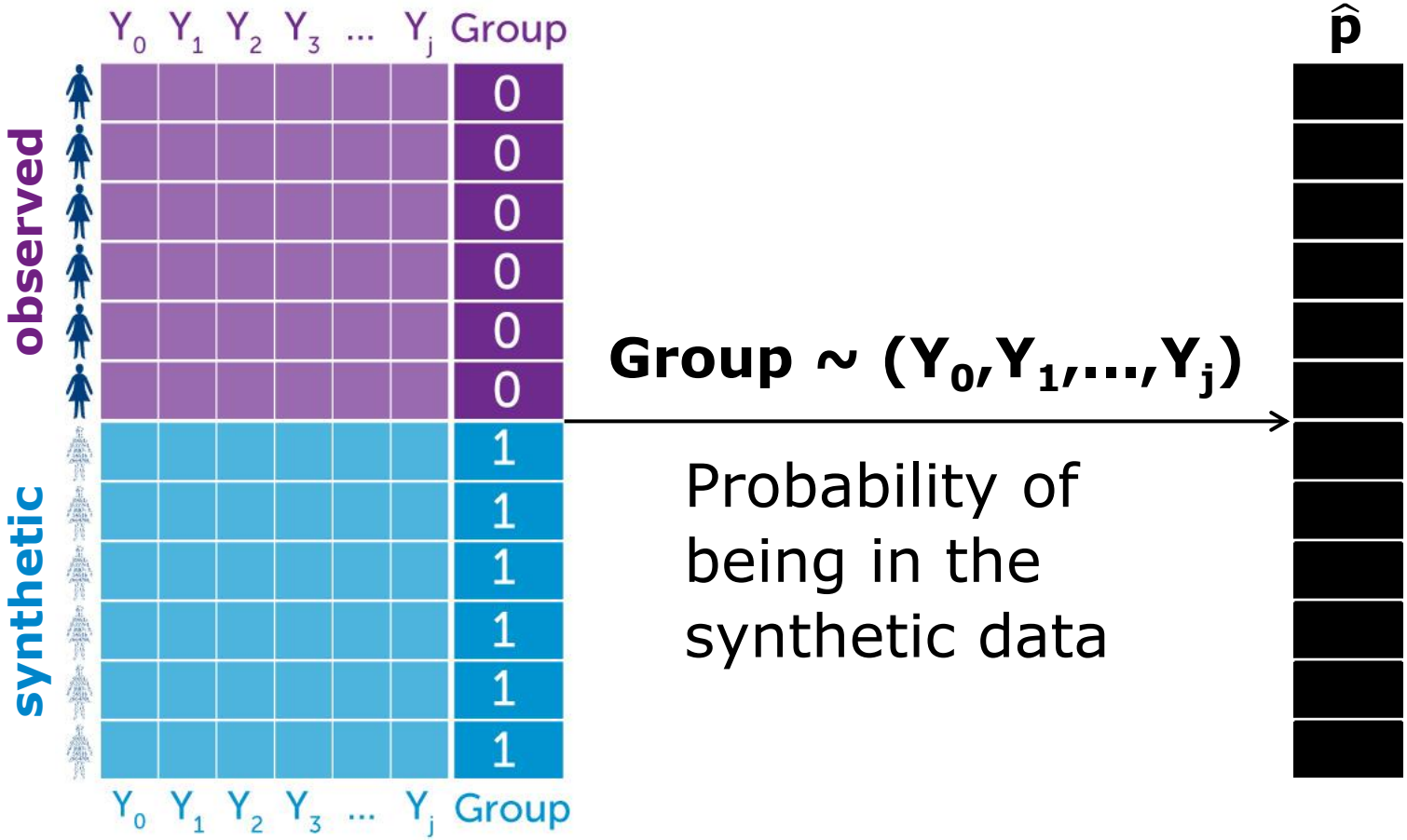
Random forest and bagging

- Collection of trees:
  - Bootstrap samples from the original data set
  - Sample of predictors as split candidates

Synthetic data utility

- **General** measures:
  - consider overall microdata similarity
  - predict effectiveness without fitting all models
- **Specific** measures:
  - consider analysis-specific similarity
  - high specific utility is desirable for researchers in preliminary analysis

General utility measures



Analysis-specific utility measures

- Comparison of models fit with both the observed data and the synthetic data
  - Overlap of coefficient confidence intervals
  - Standardized difference of coefficient estimates


synthpop®

Synthetic data

- SHeS 2008 & 2010, 46 variables
- Replacing all values for all variables
- 10 synthetic data sets for each method:
  - $CART_1$ (correlation)
  - $CART_2$ (Gini index and deviance)
  - [RF] Random forest (100 trees)
  - [BAG] Bagging (100 trees)
  - [PARA] Parametric
  - [SAMP] Sampling

Estimated regression models

- Linear: mental well-being

  Ywemwbs ~ Sex + age + age2 + sptinPAg + hrstot10 + AllNature + Gym + Other + Longill08 + simd15 09 + maritalg

- Logistic: life satisfaction

  Ylsg ~ Sex + ag16g10 + ParentInf + hrsptg10 + eqvinc + maritalg + XOwnRnt08 + econac08 + drkcat3 + cigst3 + porftvg3 + AllNature

- Poisson: time of sport activities

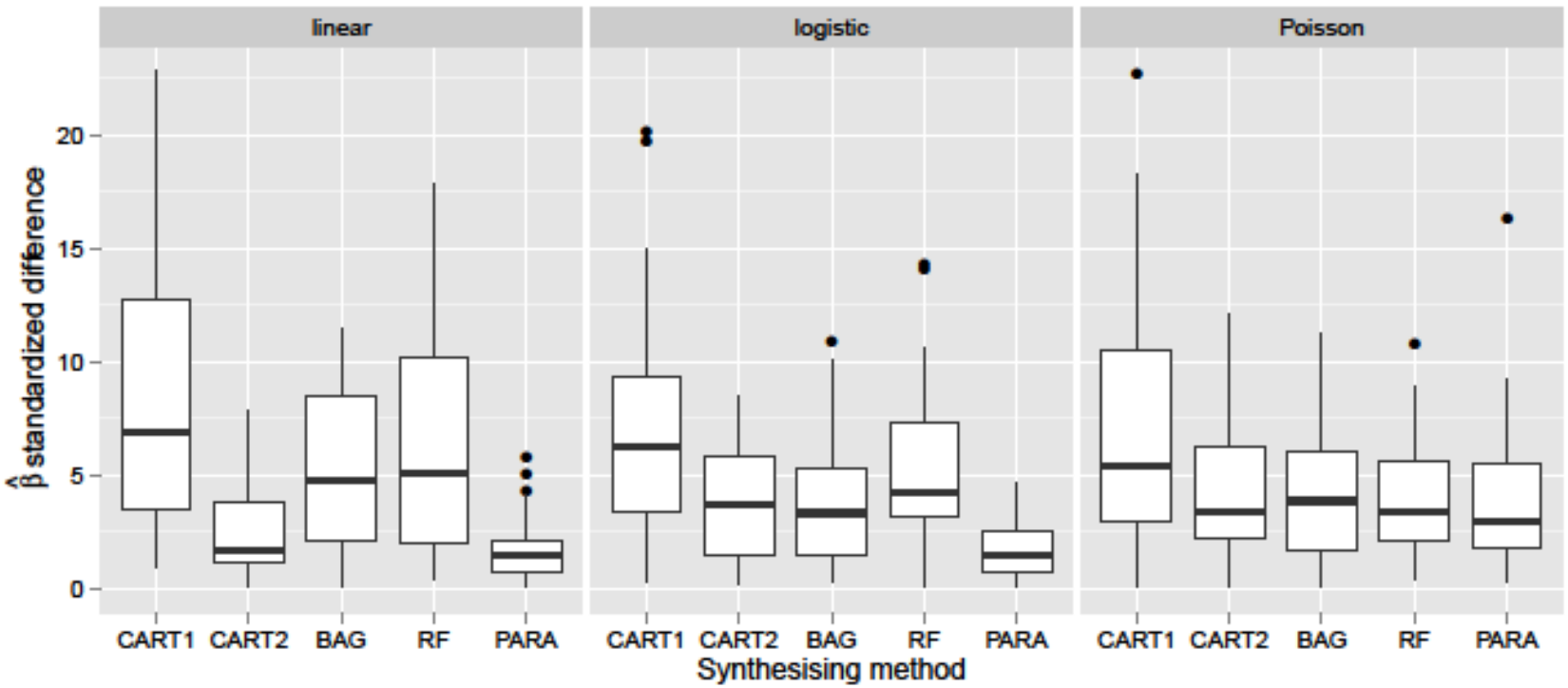
  hrsspt10 ~ Sex + ag16g10 + hrstot10 + URINDSC + XCare + Longill08 + eqvinc + maritalg + Active + ParentInf + econac08

General and analysis-specific utility

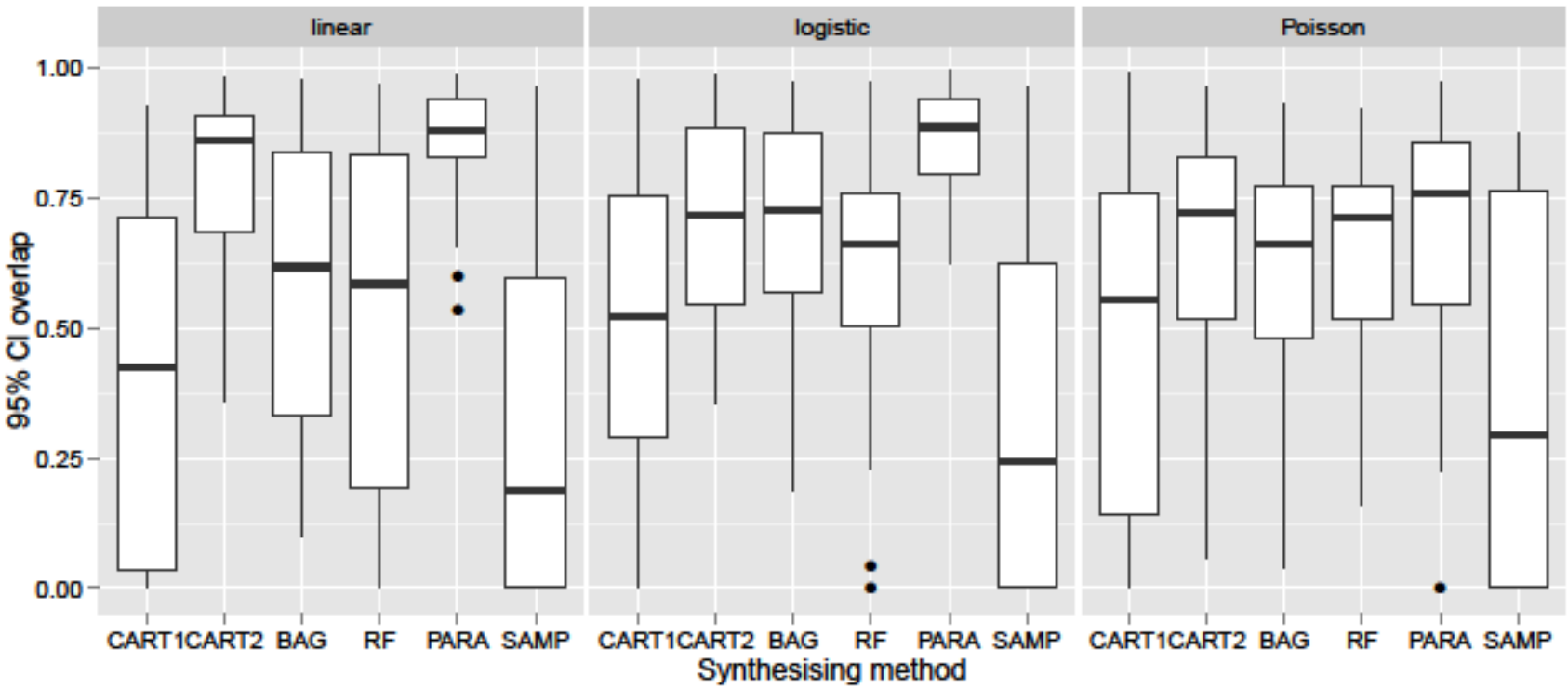
	$CART_1$	$CART_2$	BAG	RF	$PARA$	$SAMP$
General utility						
Propensity score measure	0.14	**0.11**	0.12	0.13	0.15	0.21
Analysis-specific utility						
Model 1: linear						
Mean $\hat{\beta}$ stand. difference	8.40	2.76	4.93	6.51	**1.77**	11.63
Mean 95% CI overlap	0.42	0.77	0.60	0.52	**0.86**	0.31
Model 2: logistic						
Mean $\hat{\beta}$ stand. difference	6.95	3.72	3.73	5.30	**1.72**	10.98
Mean 95% CI overlap	0.50	0.71	0.70	0.60	**0.86**	0.33
Model 3: Poisson						
Mean $\hat{\beta}$ stand. difference	7.32	4.29	4.04	4.01	**3.89**	9.48
Mean 95% CI overlap	0.47	0.66	0.62	0.64	**0.69**	0.39

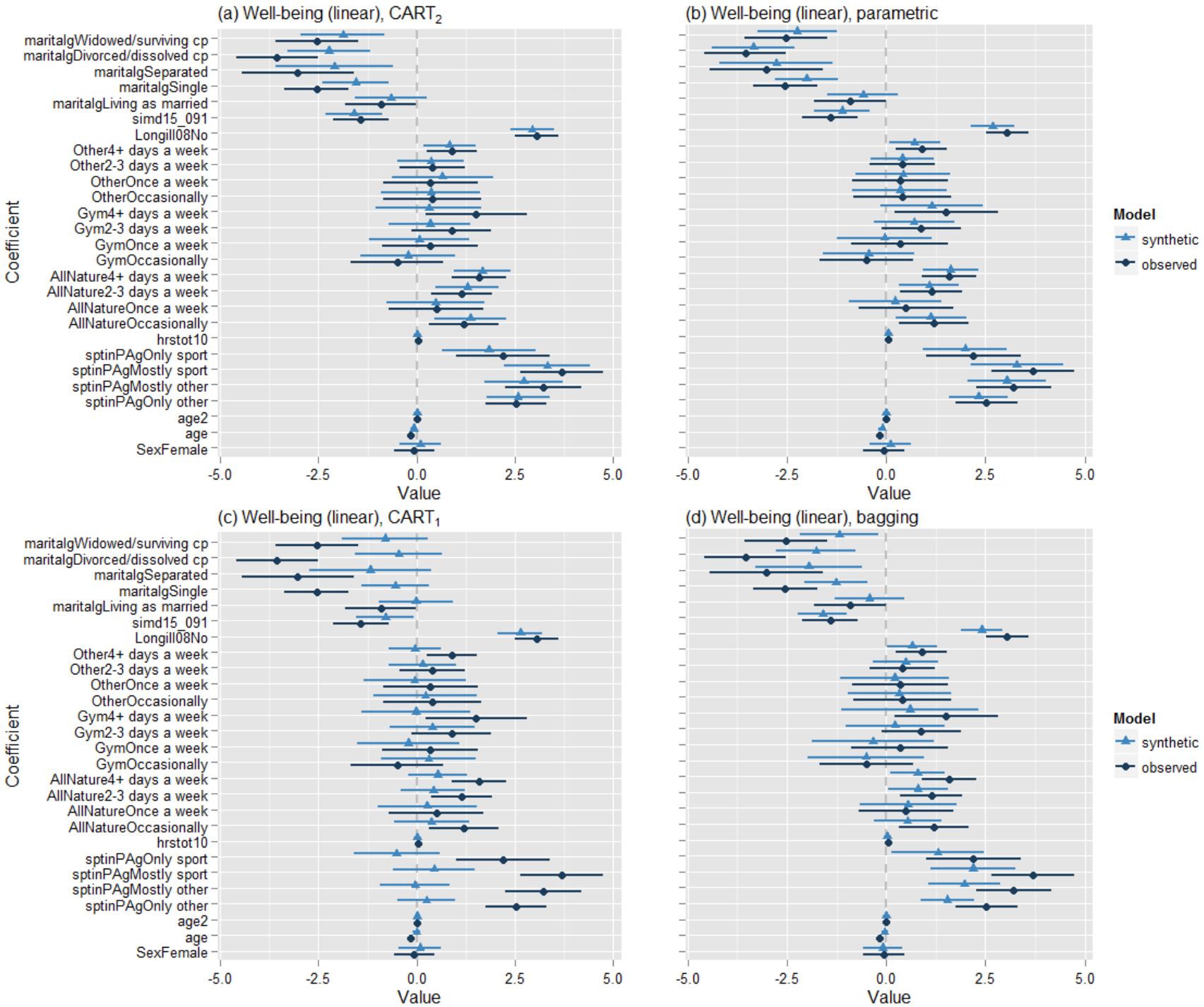
NOTE: *BAG* - bagging, *RF* - random forest, *PARA* - parametric, *SAMP* - random sampling with replacement; bold indicates best model

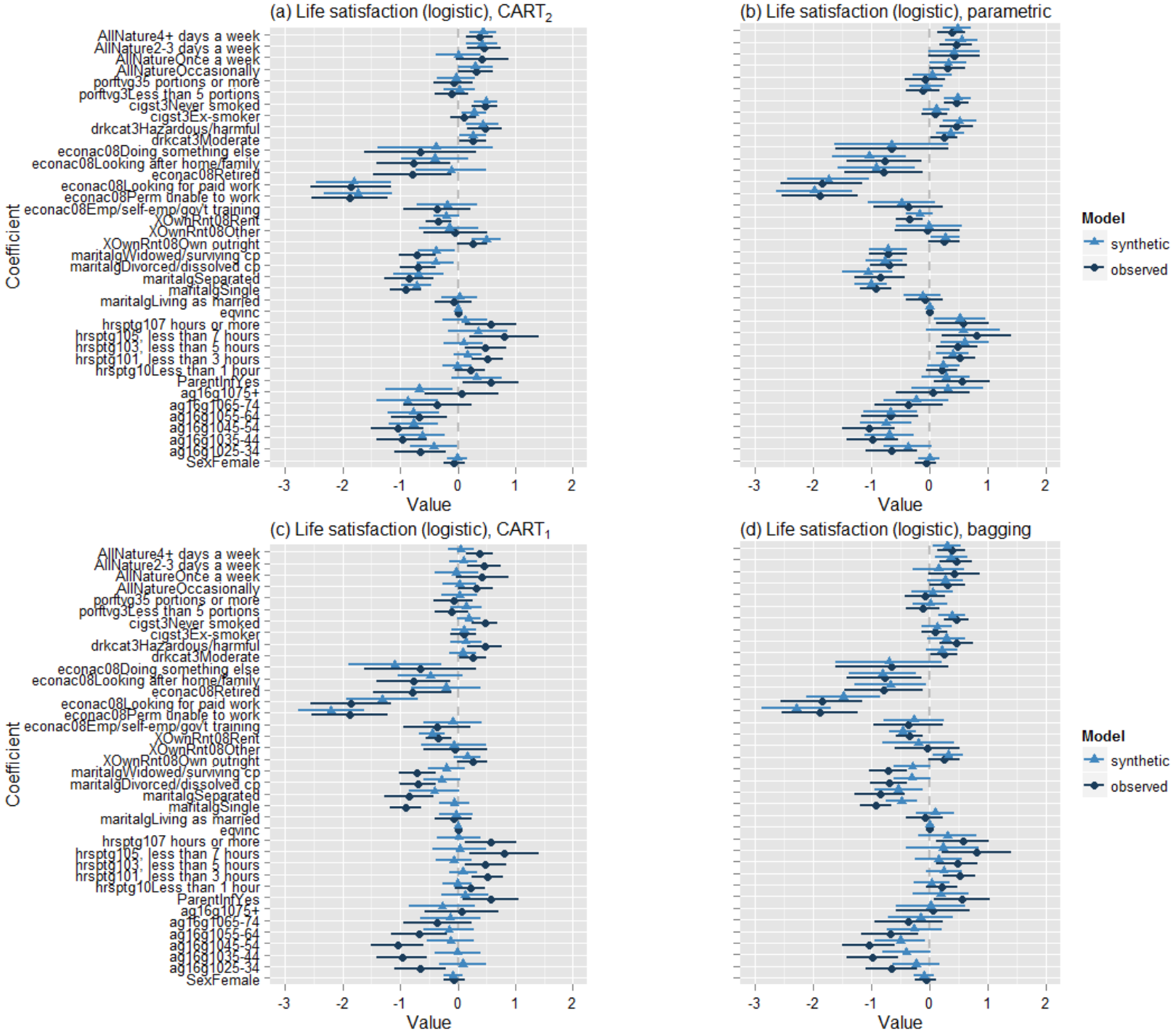
Analysis-specific utility

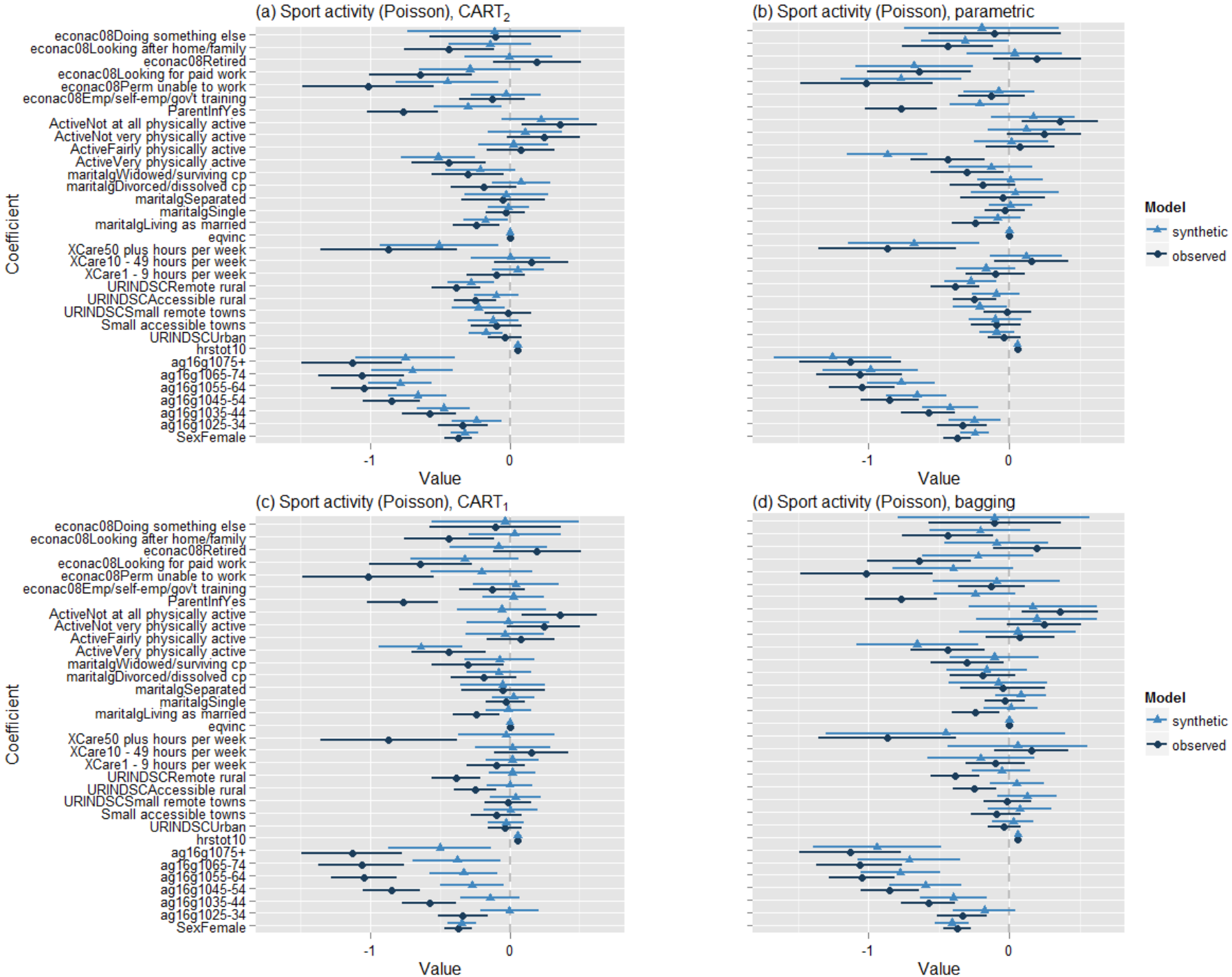


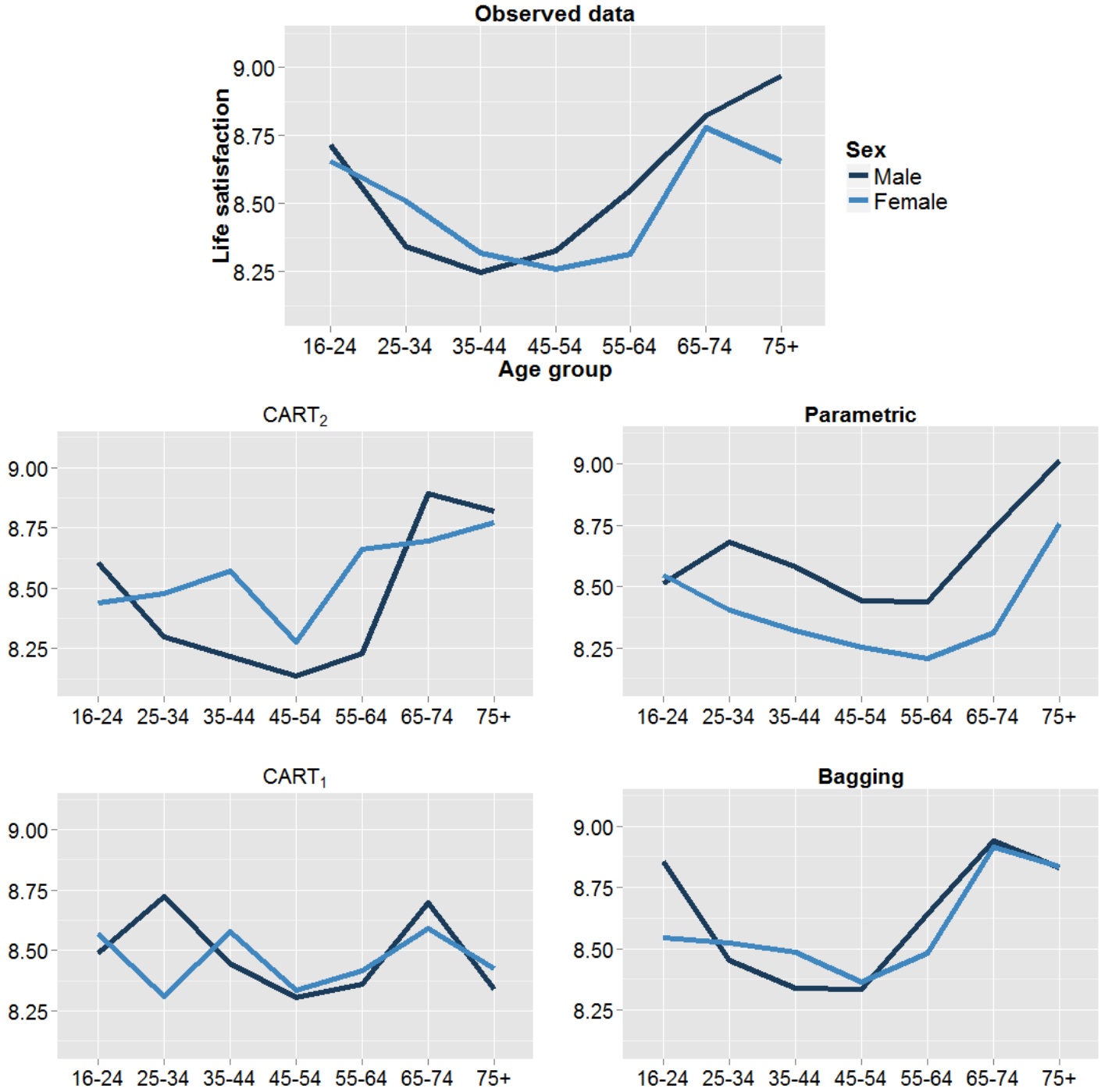
Analysis-specific utility




(a) Well-being (linear), CART2
(b) Well-being (linear), parametric
(c) Well-being (linear), CART1
(d) Well-being (linear), bagging
Coefficient
maritalgWidowed/surviving cp
maritalgDivorced/dissolved cp
maritalgSeparated
maritalgSingle
maritalgLiving as married
simd15_091
Longill08No
Other4+ days a week
Other2-3 days a week
OtherOnce a week
OtherOccasionally
Gym4+ days a week
Gym2-3 days a week
GymOnce a week
GymOccasionally
AllNature4+ days a week
AllNature2-3 days a week
AllNatureOnce a week
AllNatureOccasionally
hrstot10
sptinPAgOnly sport
sptinPAgMostly sport
sptinPAgMostly other
sptinPAgOnly other
age2
age
SexFemale
-5.0
-2.5
0.0
2.5
5.0
Value
Model
synthetic
observed


(a) Life satisfaction (logistic), CART2
(b) Life satisfaction (logistic), parametric
(c) Life satisfaction (logistic), CART1
(d) Life satisfaction (logistic), bagging
Coefficient
AllNature4+ days a week
AllNature2-3 days a week
AllNatureOnce a week
AllNatureOccasionally
porftvg35 portions or more
porftvg3Less than 5 portions
cigst3Never smoked
cigst3Ex-smoker
drkcat3Hazardous/harmful
drkcat3Moderate
econac08Doing something else
econac08Looking after home/family
econac08Retired
econac08Looking for paid work
econac08Perm unable to work
econac08Emp/self-emp/gov't training
XOwnRnt08Rent
XOwnRnt08Other
XOwnRnt08Own outright
maritalgWidowed/surviving cp
maritalgDivorced/dissolved cp
maritalgSeparated
maritalgSingle
maritalgLiving as married
eqvinc
hrsptg107 hours or more
hrsptg105, less than 7 hours
hrsptg103, less than 5 hours
hrsptg101, less than 3 hours
hrsptg10Less than 1 hour
ParentInfYes
ag16g1075+
ag16g1065-74
ag16g1055-64
ag16g1045-54
ag16g1035-44
ag16g1025-34
SexFemale
-3
-2
-1
0
1
2
Value
Model
synthetic
observed


(a) Sport activity (Poisson), CART2
(b) Sport activity (Poisson), parametric
(c) Sport activity (Poisson), CART1
(d) Sport activity (Poisson), bagging
Coefficient
econac08Doing something else
econac08Looking after home/family
econac08Retired
econac08Looking for paid work
econac08Perm unable to work
econac08Emp/self-emp/gov't training
ParentInfYes
ActiveNot at all physically active
ActiveNot very physically active
ActiveFairly physically active
ActiveVery physically active
maritalgWidowed/surviving cp
maritalgDivorced/dissolved cp
maritalgSeparated
maritalgSingle
maritalgLiving as married
eqvinc
XCare50 plus hours per week
XCare10 - 49 hours per week
XCare1 - 9 hours per week
URINDSCRemote rural
URINDSCAccessible rural
URINDSCSmall remote towns
Small accessible towns
URINDSCUrban
hrstot10
ag16g1075+
ag16g1065-74
ag16g1055-64
ag16g1045-54
ag16g1035-44
ag16g1025-34
SexFemale
-1
0
Value
Model
synthetic
observed


Observed data
Life satisfaction
9.00
8.75
8.50
8.25
16-24
25-34
35-44
45-54
55-64
65-74
75+
Age group
Sex
Male
Female
CART2
Parametric
CART1
Bagging

Conclusions

- Useful completely synthetic data can be generated using automated methods
- More research needed to validate synthesising methods in complex settings
- Tree-building procedure matters