Mobility Model Comparison (Next Frequency OD)

Andrew J K Conlan

16/03/2020

Data Sets

In this report we analyse origin-destination flux matrices based on the next (most frequent) definition. We fit each model to the total BBC mobility data set (Ω^{T}) and three stratifications by employment status (Ω^{U},Ω^{Ed}, Ω^{Em}, Ω^{N}), age of user (Ω^{U},Ω^{18-30}, Ω^{30-60}, Ω^{60-100}) and member nation of the UK (Ω^{E},Ω^{W}, Ω^{S}, Ω^{NI}).

We compare the estimated mobility models to estimates from the 2011 census commuting flow data for England (Ω^{CE}), Wales (Ω^{CW}), Scotland (Ω^{CS}) and Northern Ireland (Ω^{CNI}).

We estimate posterior distributions for each model using hamiltonian MCMC (as implemented by the Stan package http://mc-stan.org/). To assess model fit and provide a basis for model selection we use approximate leave-one-out cross validation as implemented in the loo package (doi:10.1007/s11222-016-9696-4).

The per capita probability of moving to a different LAD each day varies by category:

```
## # A tibble: 5 x 5
##   employment_cat     N movers p_move cat_prop
##   <chr>          <int>  <int>  <dbl>    <dbl>
## 1 Under 18        2955    724  0.245   0.0683
## 2 Education       3511   1101  0.314   0.0811
## 3 Employed       30500  14710  0.482   0.705
## 4 NEET            6325   1642  0.260   0.146
## 5 Total          43291  18177  0.420  NA
```

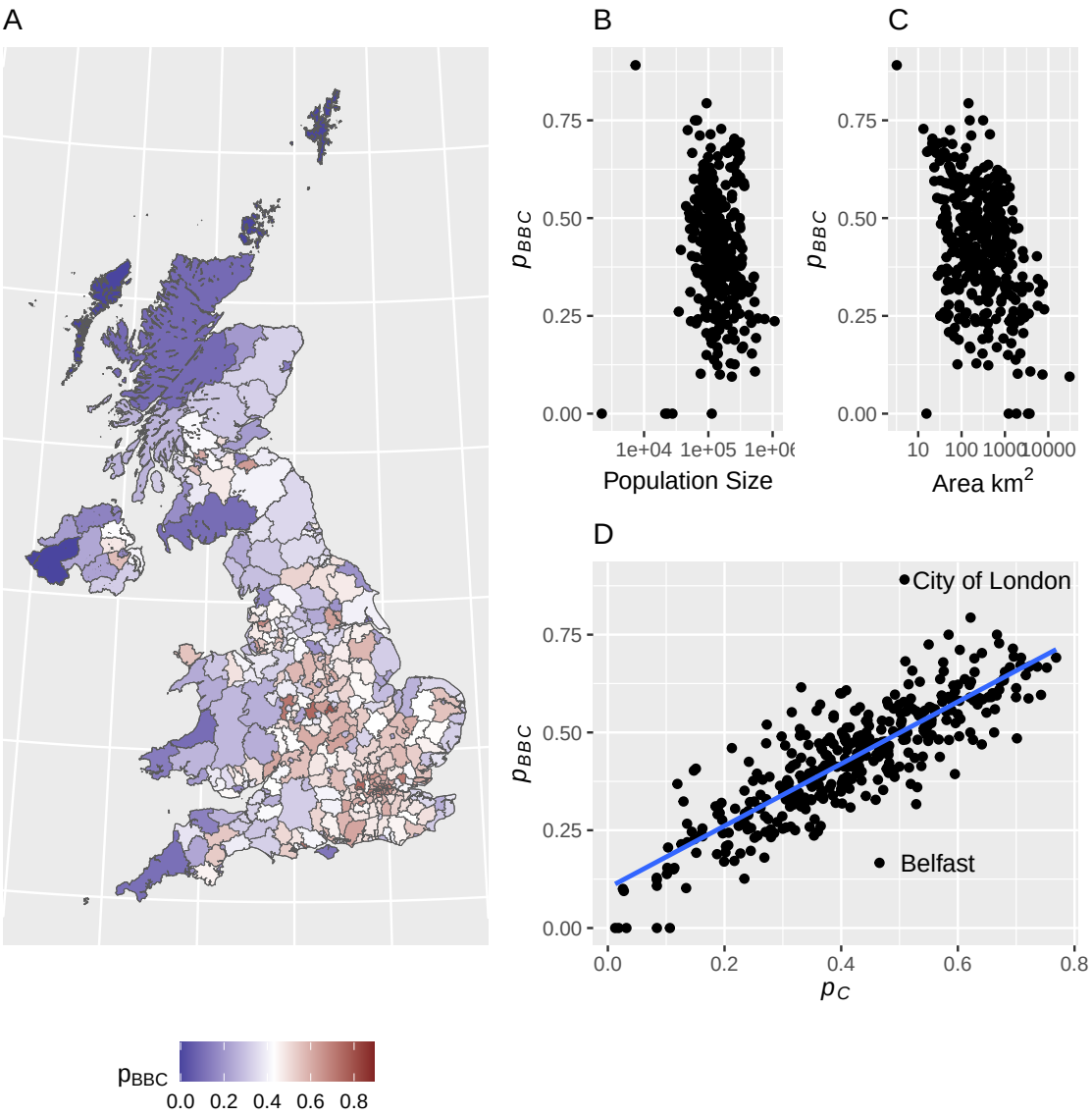
```
## # A tibble: 4 x 5
##   age_cat      N movers p_move cat_prop
##   <chr>    <int>  <int>  <dbl>    <dbl>
## 1 Under 18  2955    724  0.245   0.0683
## 2 18-30     9611   4015  0.418   0.222
## 3 30-60    26009  11948  0.459   0.601
## 4 60-100    4716   1490  0.316   0.109
```

but also by LAD.

```
##
## Call:
## lm(formula = df$p_move ~ df$census_p_move)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30521 -0.04598 -0.00478  0.04149  0.38498
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       0.10242    0.01129    9.07   <2e-16 ***
## df$census_p_move  0.79342    0.02551   31.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07962 on 389 degrees of freedom
## Multiple R-squared:  0.7133, Adjusted R-squared:  0.7125
## F-statistic: 967.7 on 1 and 389 DF,  p-value: < 2.2e-16
```

```
##                       2.5 %    97.5 %
## (Intercept)      0.08022039 0.1246244
## df$census_p_move 0.74327115 0.8435639
```


A
B
C
D
0.75
0.50
0.25
0.00
p_{BBC}
1e+04 1e+05 1e+06
Population Size
10 100 1000 10000
Area km^2
City of London
Belfast
0.0 0.2 0.4 0.6 0.8
p_C
pBBC
0.0 0.2 0.4 0.6 0.8

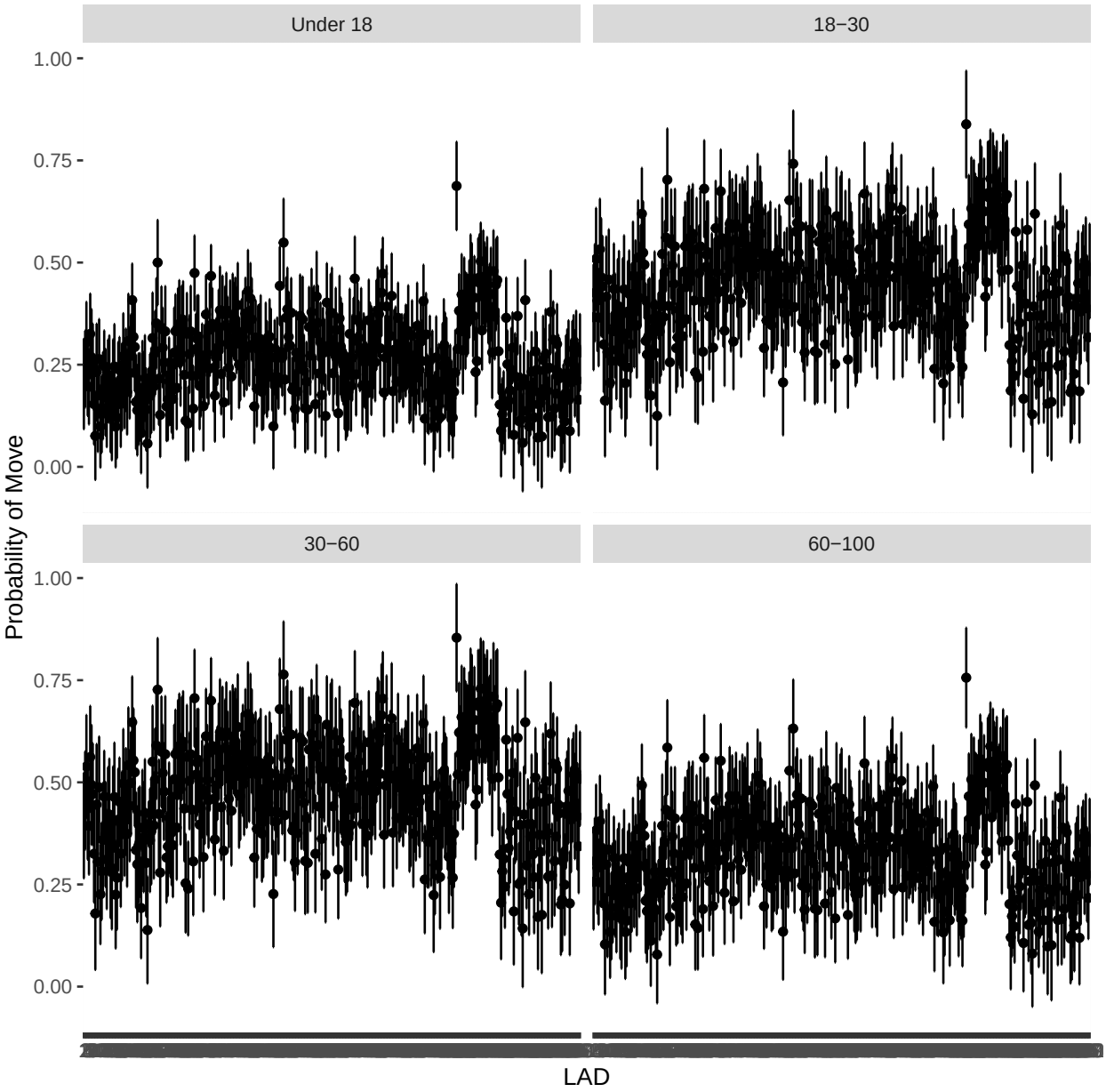
A linear regression of p_{BBC} against p_C demonstrates a strong linear relationship between the probability of moving as estimated from census data and the BBC total data set (Adjusted R-squared 0.71). The probability of moving to a different LAD per day is $\sim 10\%$ (7-12% 95% CI) greater in the BBC data set.

The coverage of the BBC mobility data set - with a median of 81 users per LAD (range 2-948) - means for the majority of LADS the raw data is too sparse to estimate movement rates for each strata of the BBC model. To address this, we estimate a generalised linear model (with logit link and random effects at the LAD level) to model the per LAD probability of moving and how this is adjusted for each strata (age or employment status). We estimate the random effects models using the lme4 package.

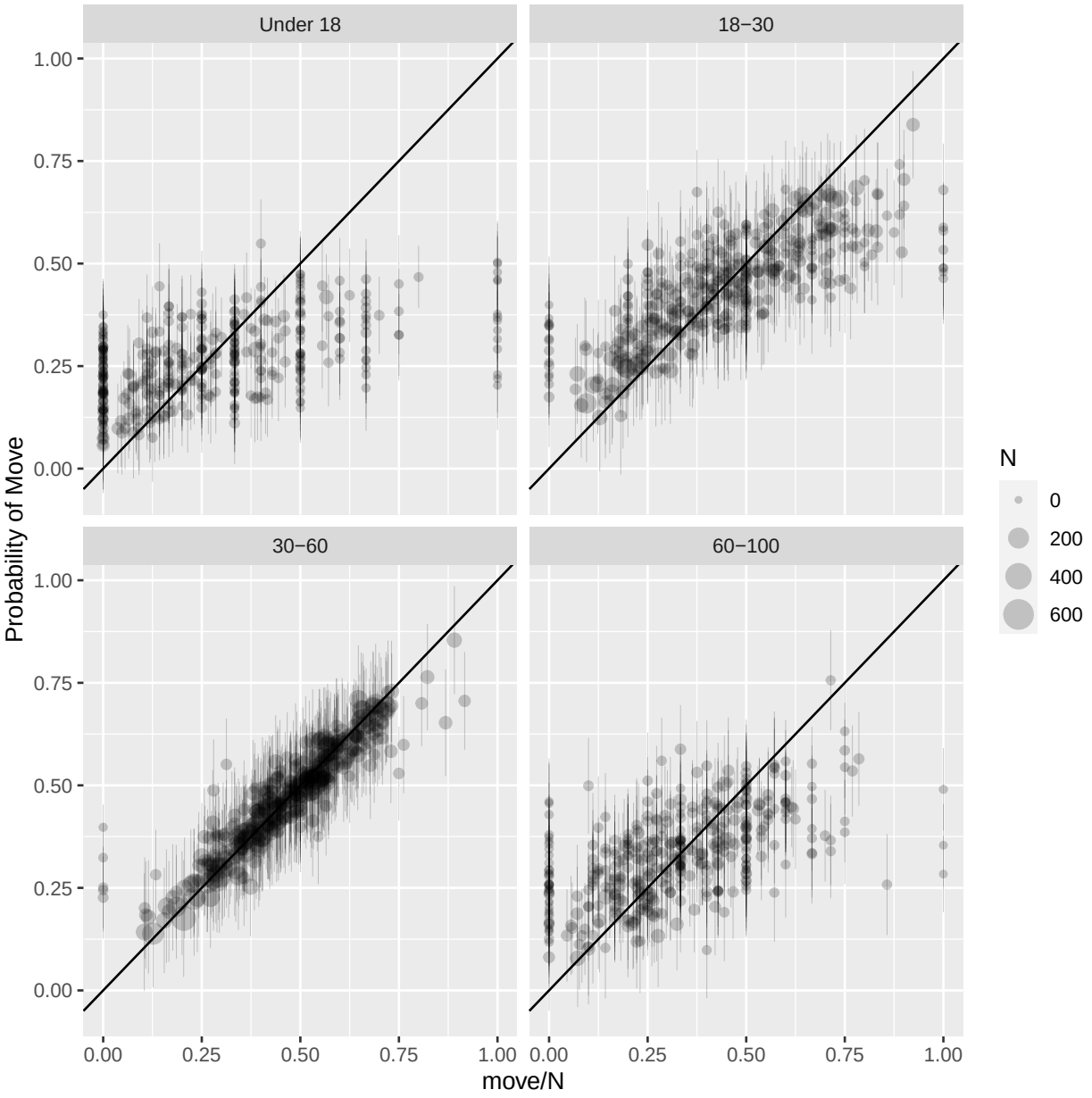
$$p_{BBC} \sim group + (1|LAD)$$

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: cbind(move, N - move) ~ group + (1 | LAD)
##    Data: age_df
##
##      AIC      BIC   logLik deviance df.resid
##   7066.5   7093.3  -3528.2   7056.5     1559
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1694 -0.6848 -0.0190  0.6354  3.6237
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  LAD    (Intercept) 0.3552   0.596
## Number of obs: 1564, groups:  LAD, 391
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.11036    0.05401 -20.560  < 2e-16 ***
## group18-30   0.85922    0.04981  17.249  < 2e-16 ***
## group30-60   0.97816    0.04626  21.143  < 2e-16 ***
## group60-100  0.34333    0.05503   6.238 4.42e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) g18-30 g30-60
## group18-30  -0.730
## group30-60  -0.785  0.858
## group60-100 -0.660  0.715  0.770
```

```
## # A tibble: 3 x 2
##   group  or
##   <chr>  <chr>
## 1 18-30  2.36 (2.14,2.6)
## 2 30-60  2.66 (2.43,2.91)
## 3 60-100 1.41 (1.27,1.57)
```



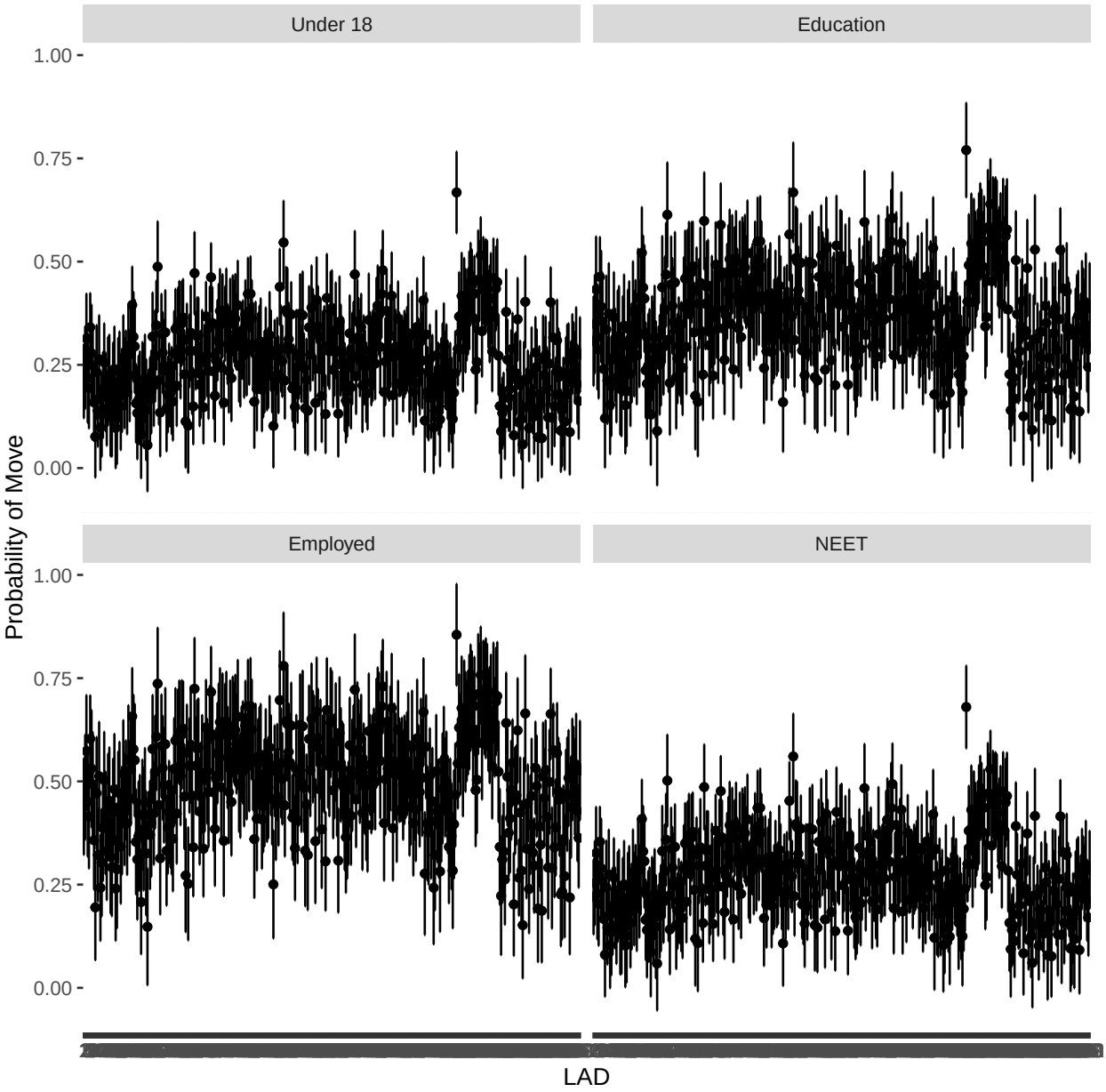

```
## Warning: Removed 18 rows containing missing values (geom_point).
```



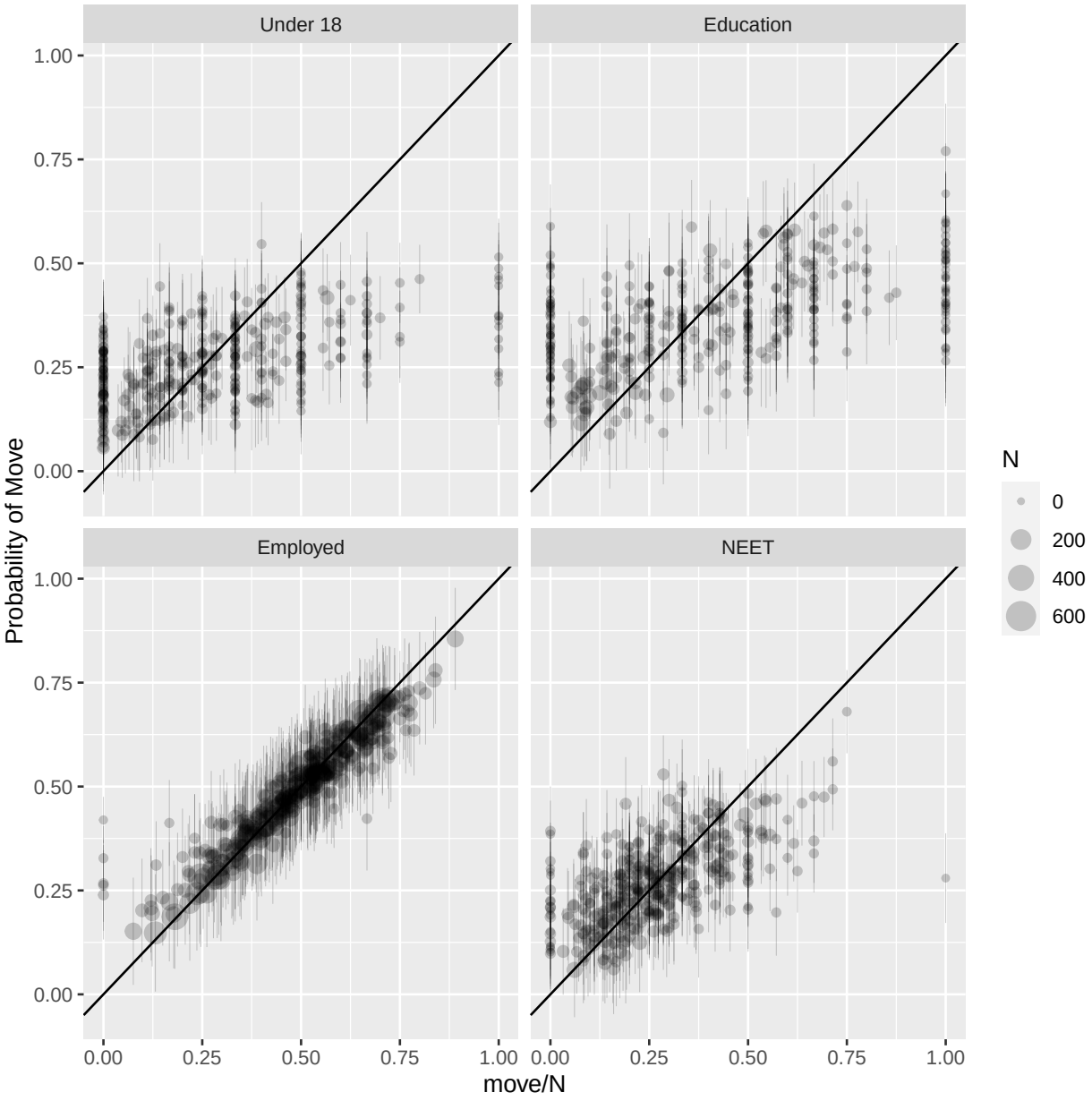

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: cbind(move, N - move) ~ group + (1 | LAD)
##    Data: emp_df
##
##      AIC      BIC   logLik deviance df.resid
##   6713.9   6740.7  -3352.0   6703.9     1559
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3905 -0.6224  0.0000  0.6582  3.4142
##
## Random effects:
```

```
##  Groups Name        Variance Std.Dev.
##  LAD    (Intercept) 0.3502   0.5918
## Number of obs: 1564, groups:  LAD, 391
##
## Fixed effects:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.10849    0.05388 -20.572   <2e-16 ***
## groupEducation  0.51151    0.05903   8.665   <2e-16 ***
## groupEmployed   1.07800    0.04604  23.416   <2e-16 ***
## groupNEET       0.05724    0.05333   1.073    0.283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) grpEdc grpEmp
## groupEductn -0.616
## groupEmplyd -0.790  0.728
## groupNEET   -0.683  0.622  0.799
```

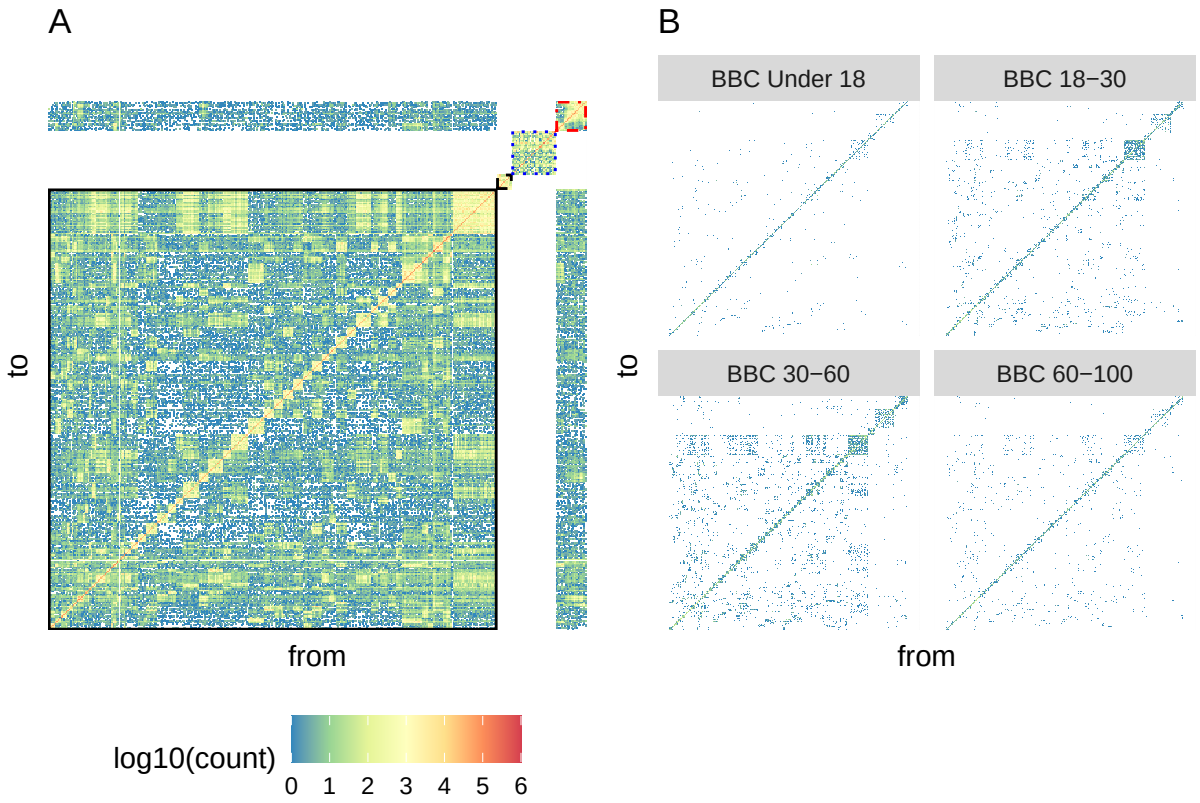
```
## # A tibble: 3 x 2
##   group     or
##   <chr>     <chr>
## 1 Education 1.67 (1.49,1.87)
## 2 Employed  2.94 (2.69,3.22)
## 3 NEET      1.06 (0.95,1.18)
```

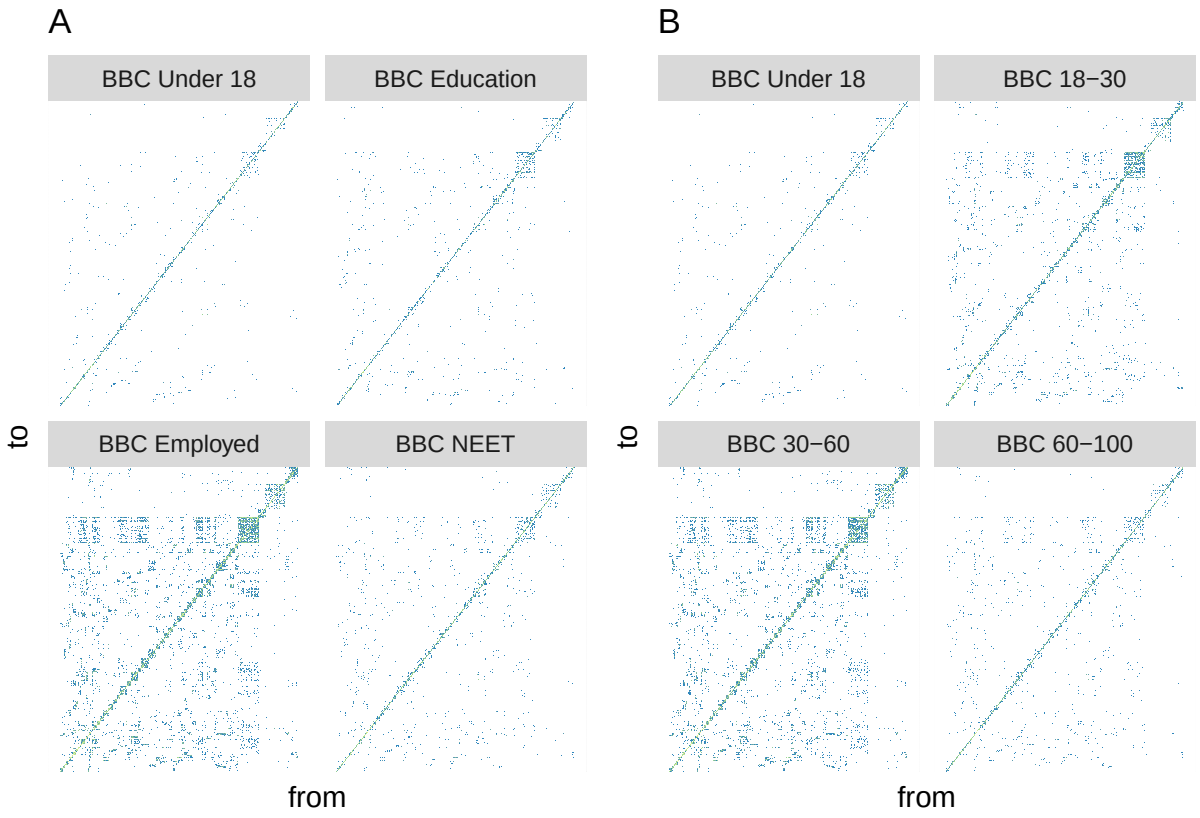



```
## Warning: Removed 32 rows containing missing values (geom_point).
```



Raw flux matrices comparing the census work-flow data set (A) to stratified BBC flux matrices.


A
B
BBC Under 18
BBC 18−30
BBC 30−60
BBC 60−100
to
from
to
from
log10(count)
0 1 2 3 4 5 6

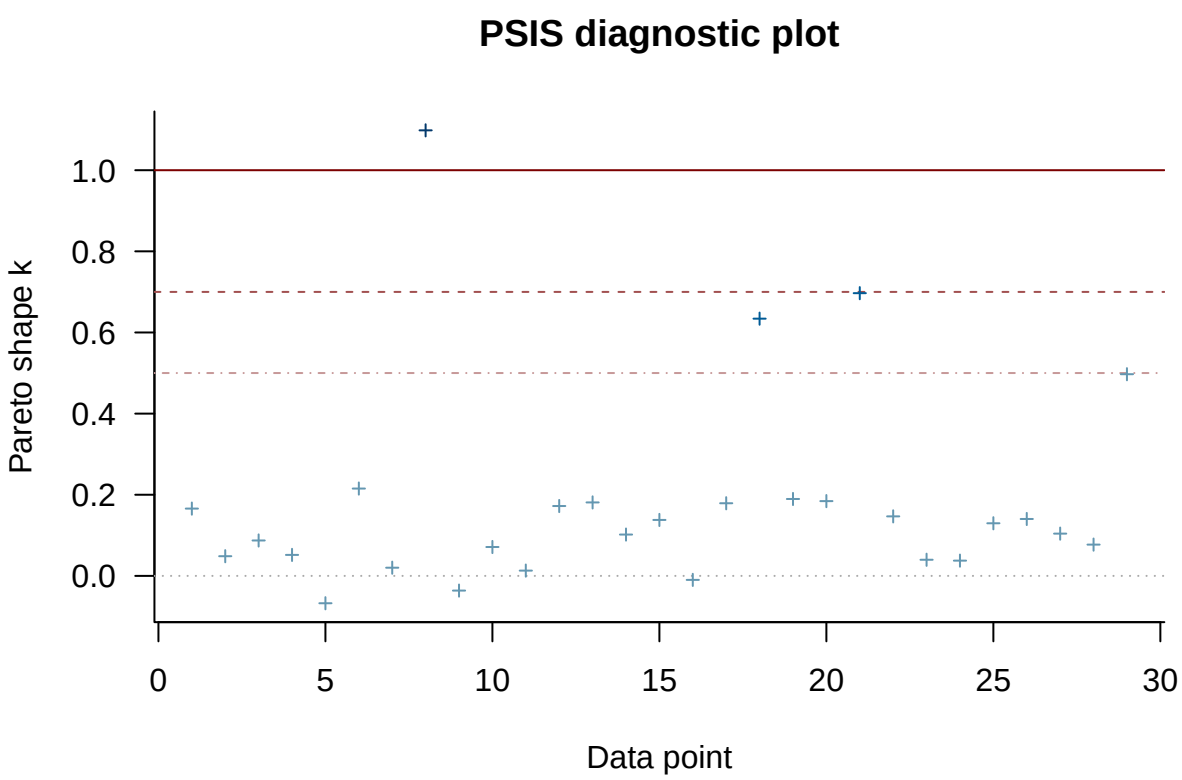


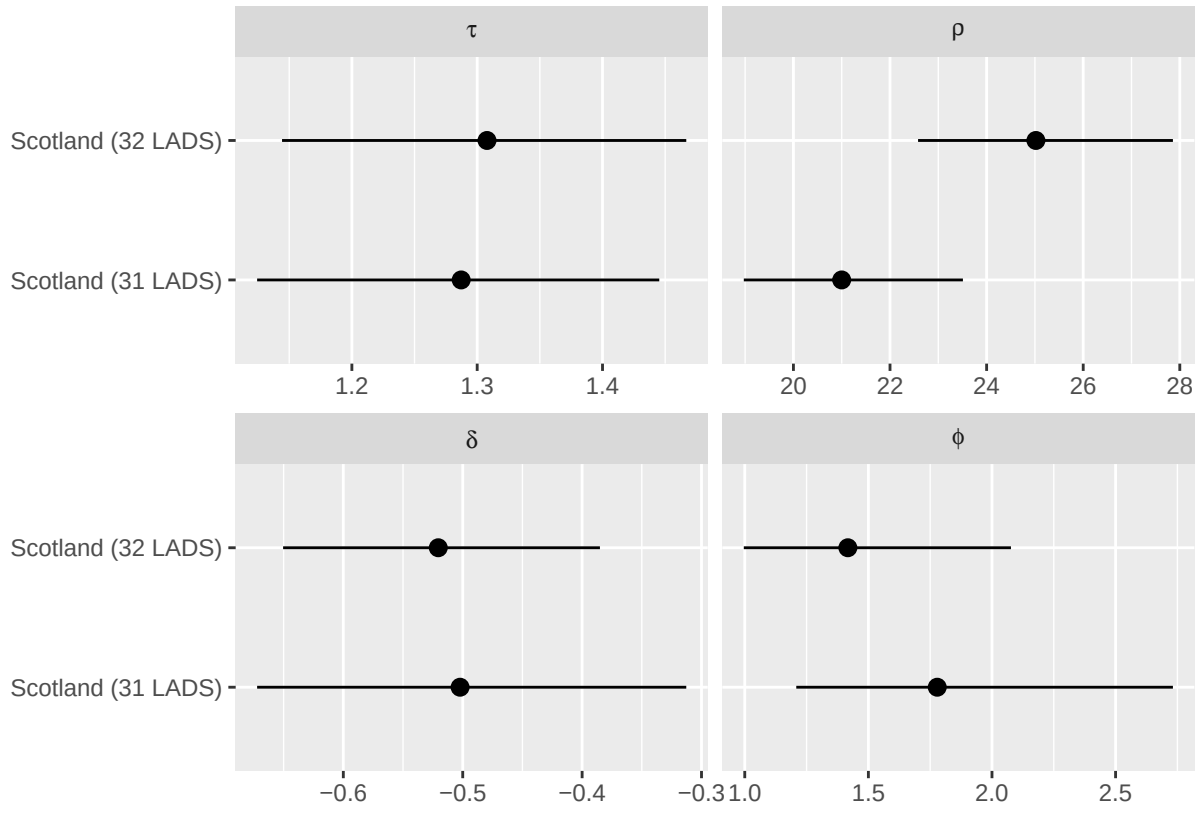
PSIS diagnostic plots

The approximate leave-one-out cross validation (LOO) method uses pareto smoothed importance sampling (PSIS-LOO) to efficiently estimate the predictive accuracy of a model (expected log pointwise predictive density, $\widehat{elpd}$) and as a basis for model comparison and selection. The estimated shape parameter $\hat{k}$ can be used to judge the reliability of the estimate of $\widehat{elpd}$ for each data point (or in our case for each LAD corresponding to a row of Ω_{ji}). The estimate of $\widehat{elpd}$ is considered reliable (quick convergence) for $\hat{k} < 0.5$, performance may still be reliable for values of $\hat{k}$ up to 0.7. Values of $\hat{k} > 0.7$ suggest that the data points are highly influential to the estimated posterior and potentially introducing bias.

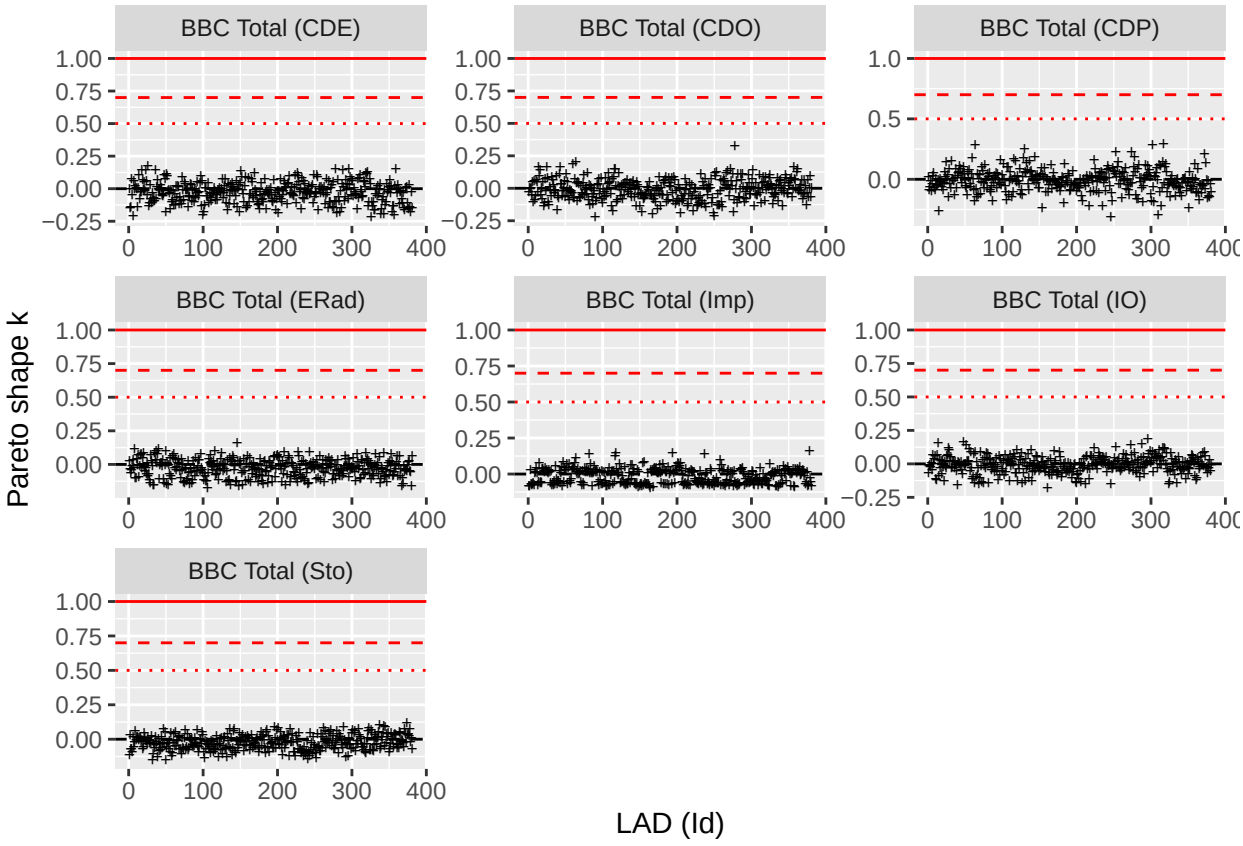
Impact of Highland LAD on CDE model

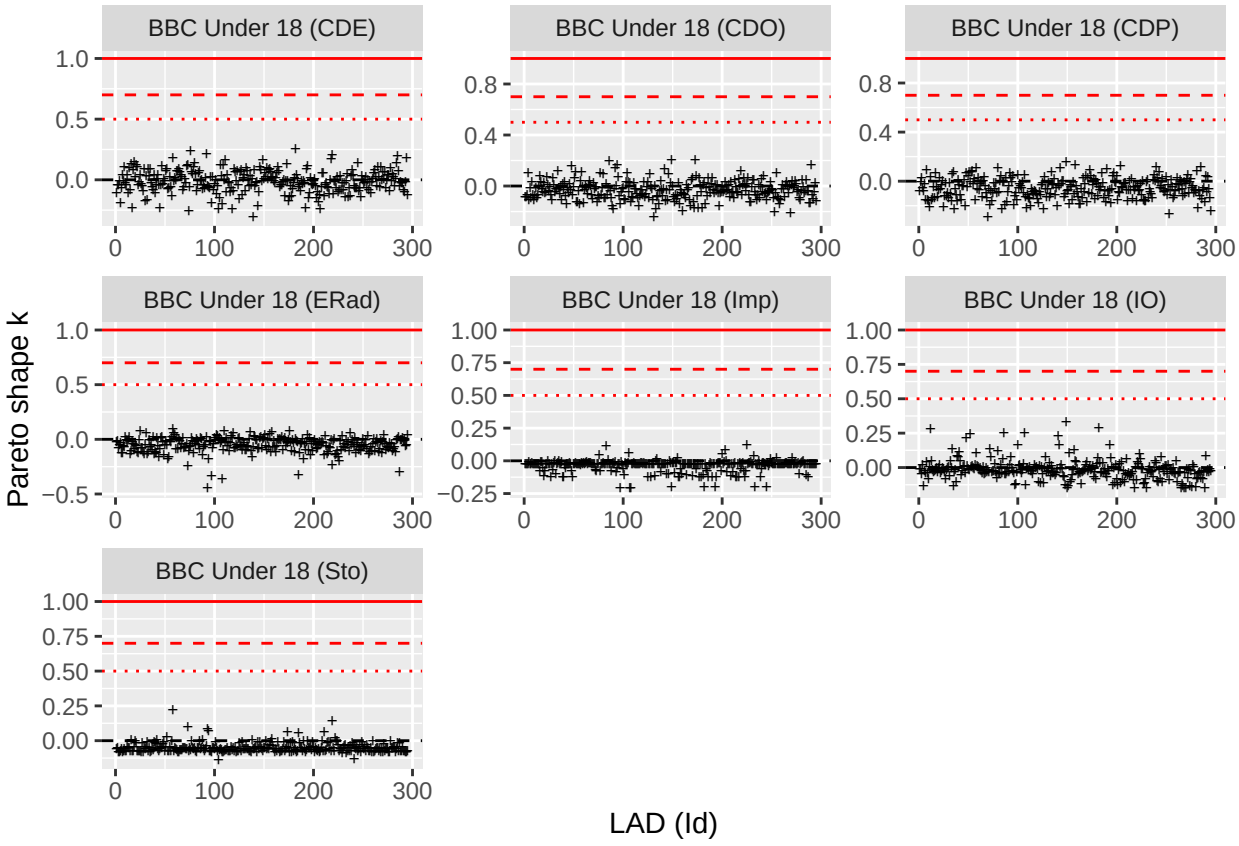
The highland LAD fails PSIS diagnostic checks with a value of $\hat{k} > 7$.

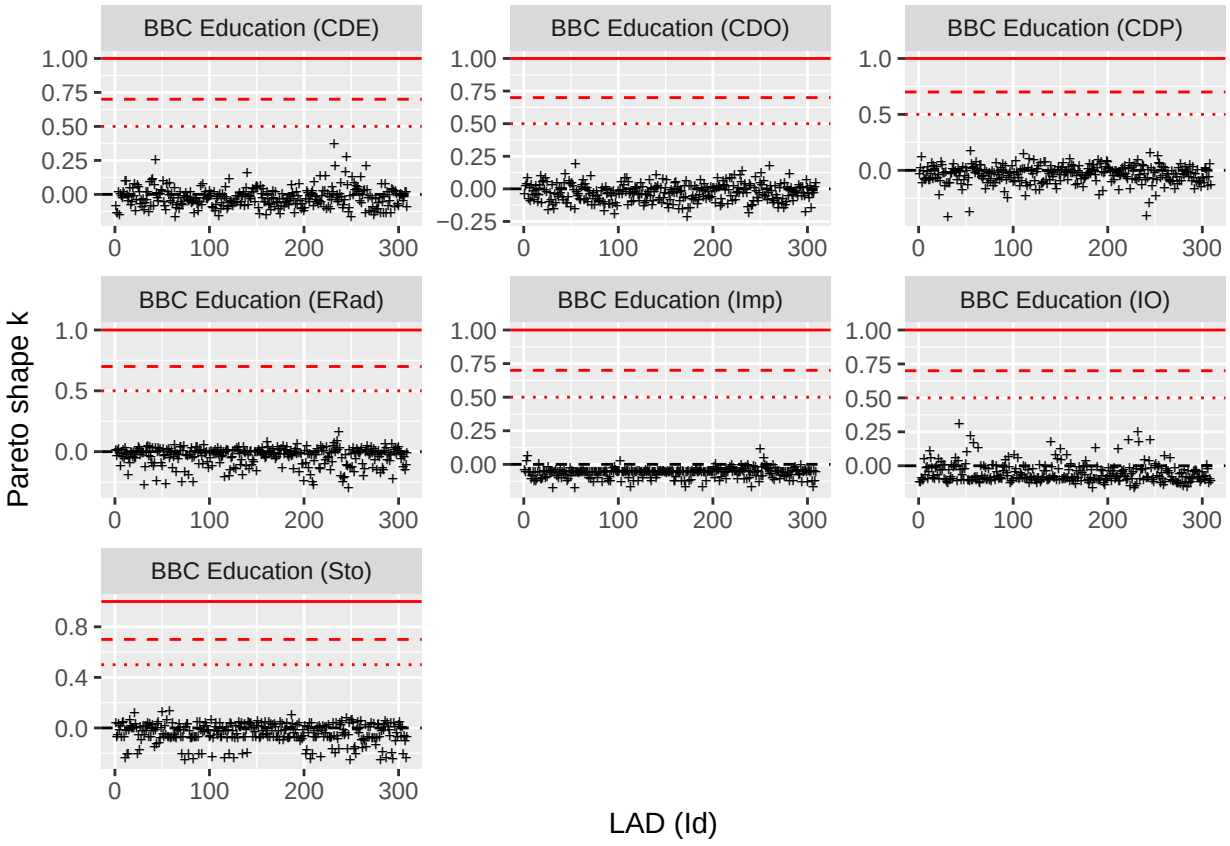

PSIS diagnostic plot
Pareto shape k
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
Data point

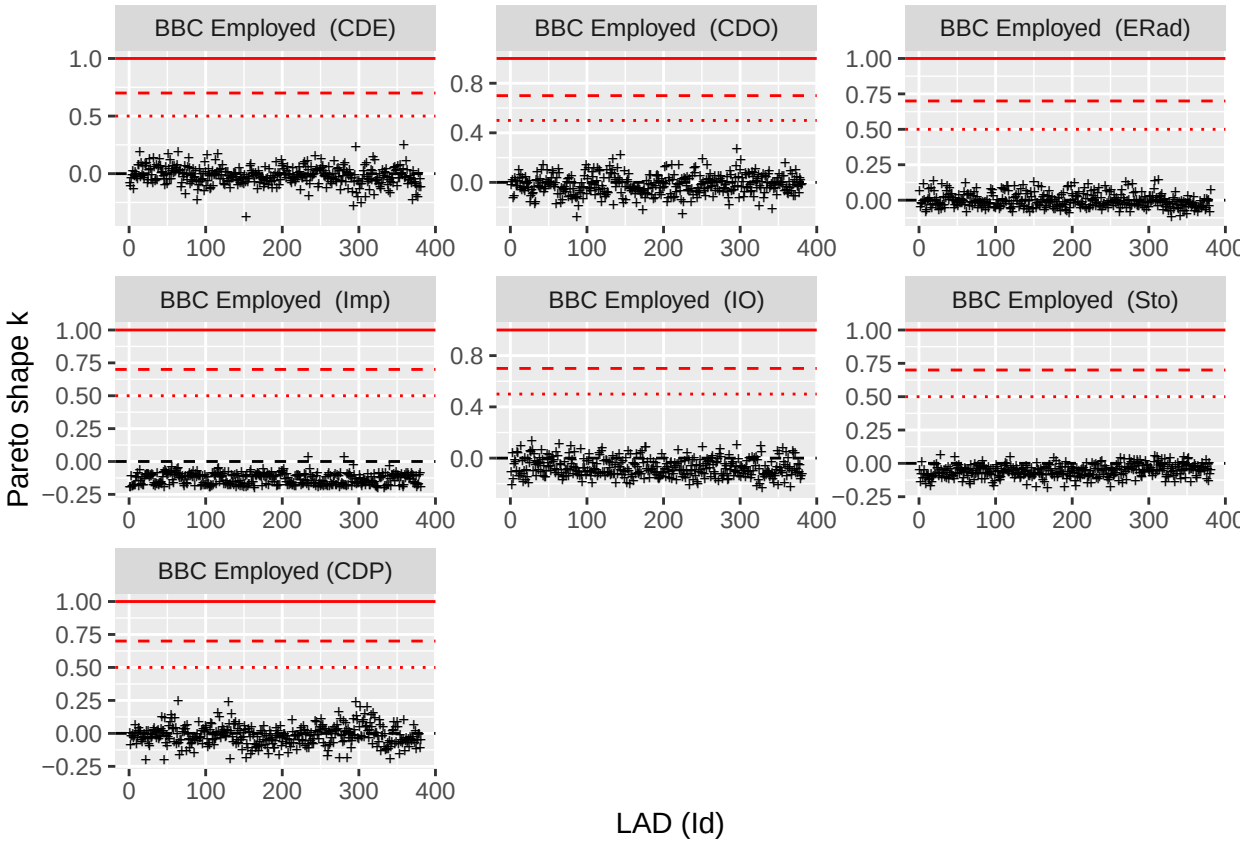


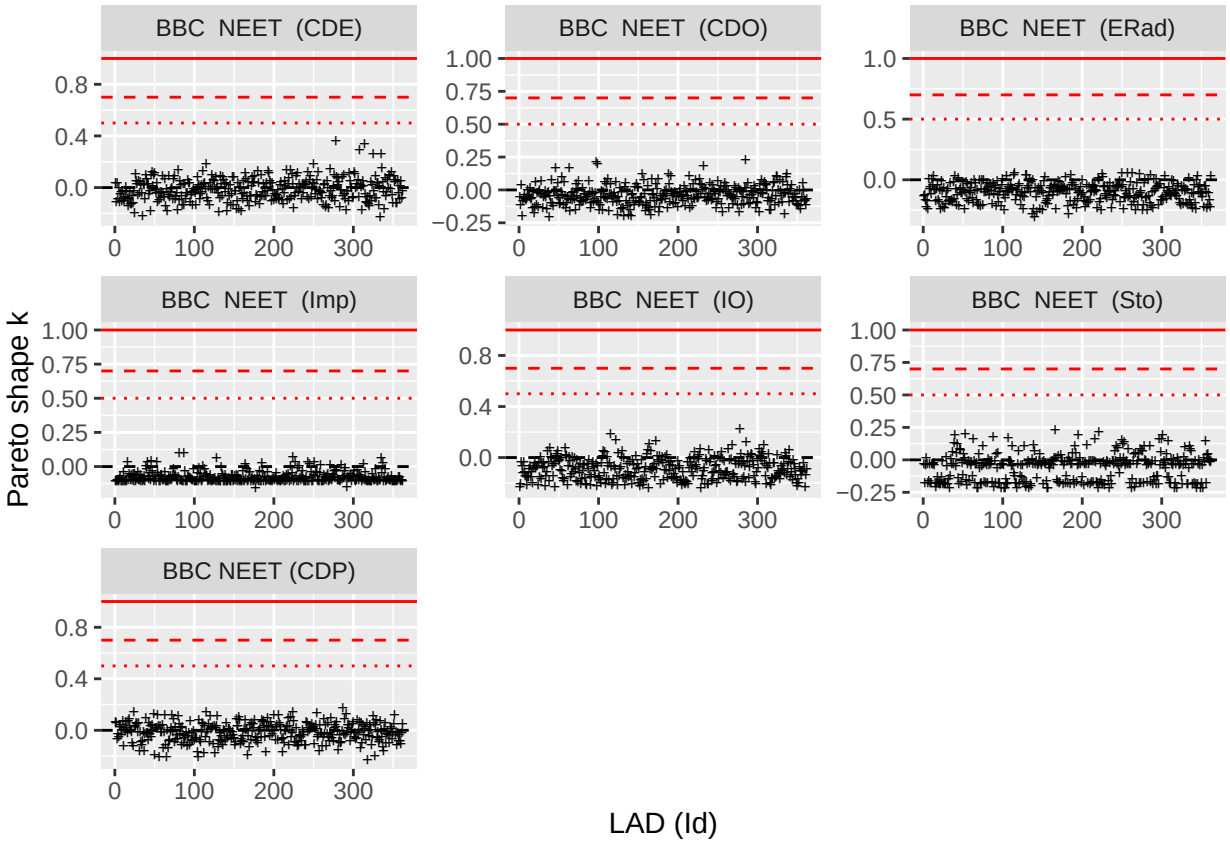
Comparison of a model fitted to the full 32 Scottish LADS to a reduced data set with Highlands removed (31 LADS) illustrates the systematic bias introduced on the distance scaling (ρ parameter). Although posterior distributions are overlapping we remove the highland LAD from the data set for inference to avoid affecting model comparison.

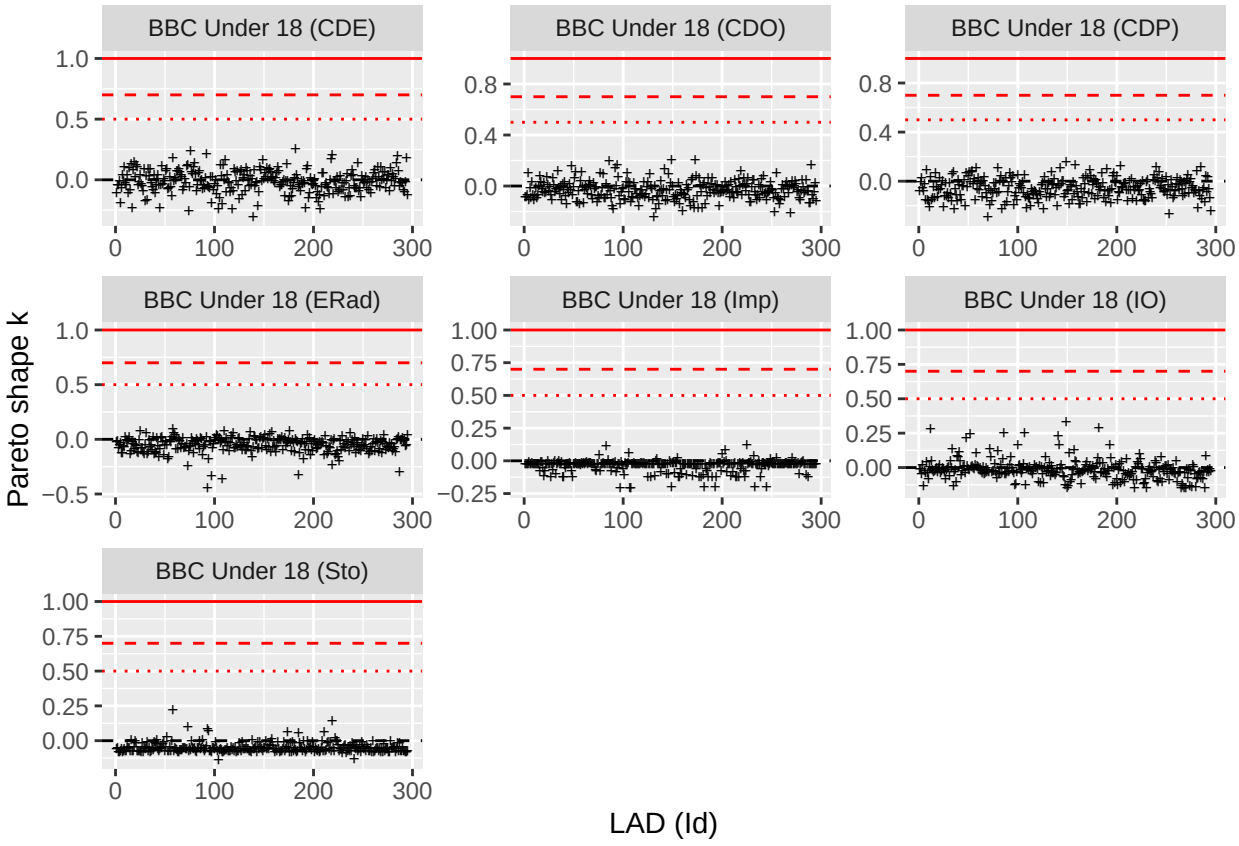

BBC Total (CDE)
BBC Total (CDO)
BBC Total (CDP)
BBC Total (ERad)
BBC Total (Imp)
BBC Total (IO)
BBC Total (Sto)
Pareto shape k
LAD (Id)

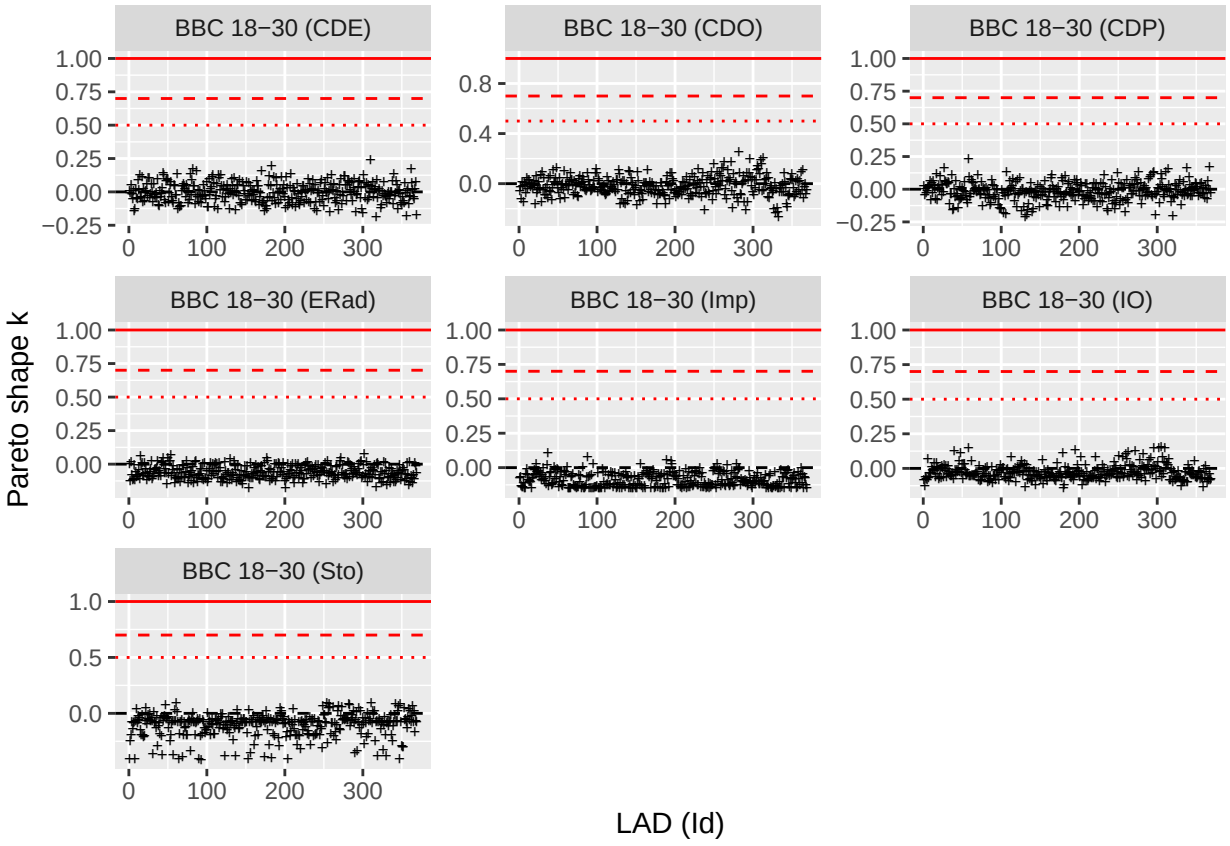

BBC Under 18 (CDE)
BBC Under 18 (CDO)
BBC Under 18 (CDP)
BBC Under 18 (ERad)
BBC Under 18 (Imp)
BBC Under 18 (IO)
BBC Under 18 (Sto)
Pareto shape k
LAD (Id)

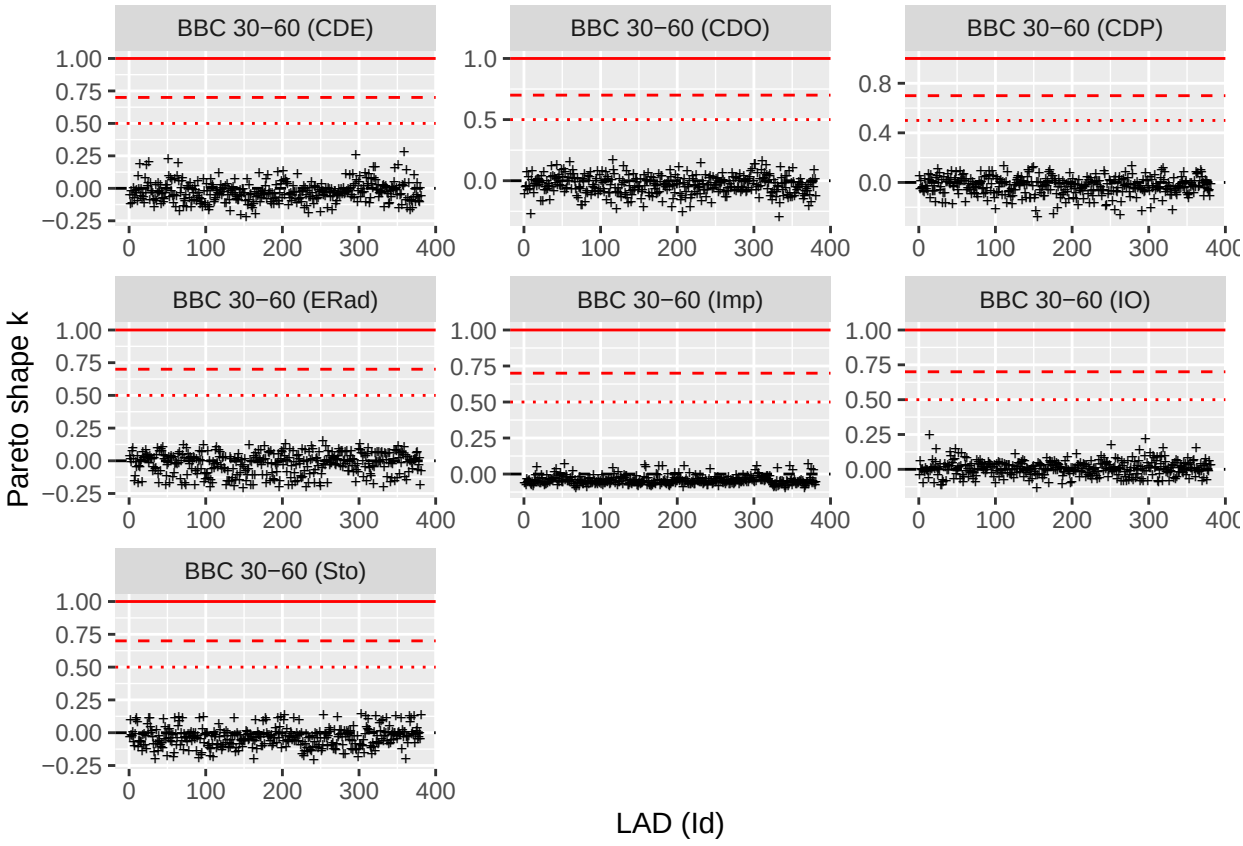

BBC Education (CDE)
BBC Education (CDO)
BBC Education (CDP)
BBC Education (ERad)
BBC Education (Imp)
BBC Education (IO)
BBC Education (Sto)
Pareto shape k
LAD (Id)

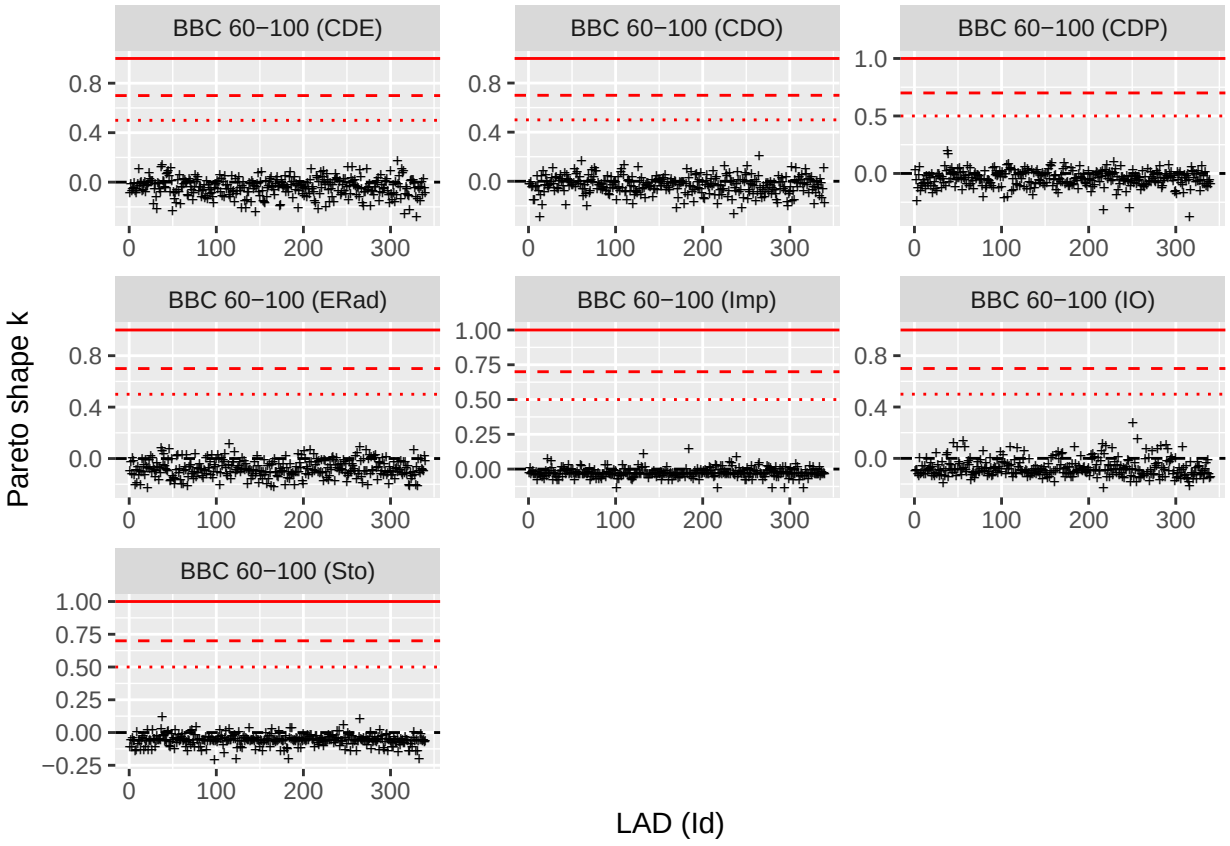

BBC Employed (CDE)
BBC Employed (CDO)
BBC Employed (ERad)
BBC Employed (Imp)
BBC Employed (IO)
BBC Employed (Sto)
BBC Employed (CDP)
Pareto shape k
LAD (Id)


BBC NEET (CDE)
BBC NEET (CDO)
BBC NEET (ERad)
BBC NEET (Imp)
BBC NEET (IO)
BBC NEET (Sto)
BBC NEET (CDP)
Pareto shape k
LAD (Id)


BBC Under 18 (CDE)
BBC Under 18 (CDO)
BBC Under 18 (CDP)
BBC Under 18 (ERad)
BBC Under 18 (Imp)
BBC Under 18 (IO)
BBC Under 18 (Sto)
Pareto shape k
LAD (Id)


BBC 18−30 (CDE)
BBC 18−30 (CDO)
BBC 18−30 (CDP)
BBC 18−30 (ERad)
BBC 18−30 (Imp)
BBC 18−30 (IO)
BBC 18−30 (Sto)
Pareto shape k
LAD (Id)


BBC 30−60 (CDE)
BBC 30−60 (CDO)
BBC 30−60 (CDP)
BBC 30−60 (ERad)
BBC 30−60 (Imp)
BBC 30−60 (IO)
BBC 30−60 (Sto)
Pareto shape k
LAD (Id)



Model comparison

The difference between $\widehat{elpd}$ for alternative models fitted to the same data provides a measure of their relative predictive accuracy.

```
##     model          elpd   model        elpd   model       elpd   model
## 1    ERad         0 (0)     CDO       0 (0)     CDO      0 (0)     CDO
## 2     CDO -11500 (1100)     CDP -23.2 (7.8)     CDP -16.4 (10)     CDE
## 3     CDP -12300 (1100)     CDE  -82.5 (17)      IO  -239 (51)     CDP
## 4      IO  -35400 (640)      IO   -278 (32)    ERad  -256 (45)      IO
## 5     CDE  -41400 (850)    ERad   -302 (26)     CDE  -351 (37)    ERad
## 6     Imp  -60900 (880)     Imp   -392 (31)     Imp  -465 (44)     Imp
## 7 Stoufer  -79300 (960) Stoufer   -486 (33) Stoufer  -644 (44) Stoufer
##          elpd
## 1       0 (0)
## 2 -4.18 (3.2)
## 3 -32.6 (9.4)
## 4 -50.9 (8.2)
## 5   -65 (7.2)
## 6 -78.9 (5.1)
## 7 -98.8 (6.3)
```

```
##     model          elpd   model        elpd   model       elpd   model
## 1     CDO        0 (0)     CDO       0 (0)     CDO      0 (0)      IO
## 2     CDE    -287 (41)     CDE -3.64 (2.7)     CDE -12.5 (8.6)     CDP
## 3    ERad    -319 (76)     CDP -9.56 (4.4)     CDP  -14.3 (10)     CDO
```

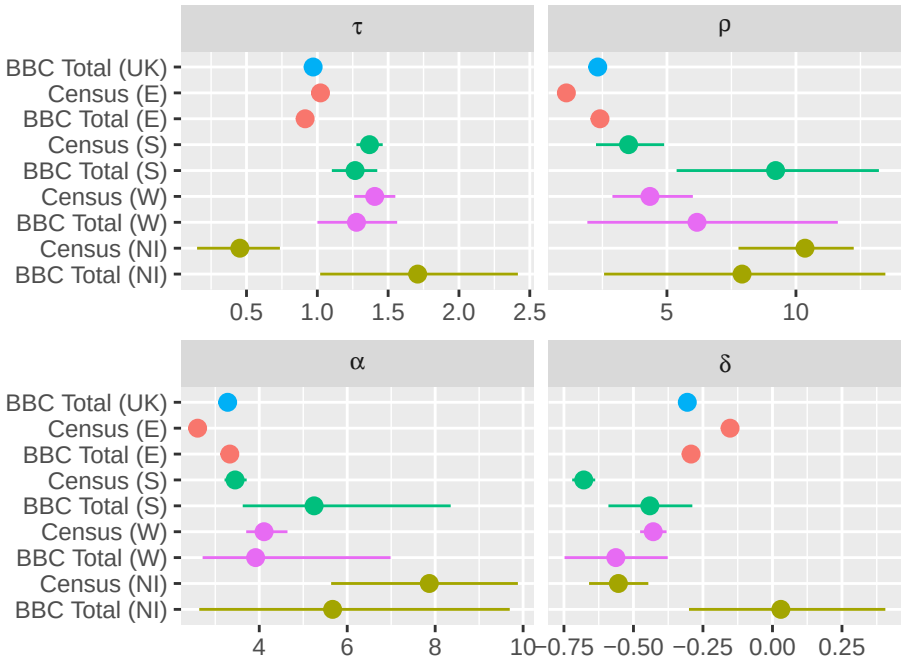
```
## 4     CDP  -2270 (110)      IO -46.6 (9.2)      IO  -34.1 (16)    ERad
## 5      IO  -3230 (110)    ERad -51.2 (8.2)    ERad  -45.6 (19)     CDE
## 6     Imp  -6330 (160)     Imp  -80.1 (11)     Imp   -123 (23)     Imp
## 7 Stoufer -10400 (220) Stoufer   -151 (14) Stoufer   -220 (27) Stoufer
##           elpd
## 1        0 (0)
## 2 -0.562 (2.7)
## 3   -1.9 (2.9)
## 4   -4.9 (2.1)
## 5  -4.96 (3.8)
## 6      -27 (5)
## 7  -42.2 (6.3)
```

```
##     model         elpd   model        elpd   model        elpd   model
## 1     CDO        0 (0)     CDO       0 (0)     CDO       0 (0)     CDO
## 2     CDE    -299 (43)     CDE -21.1 (7.1)     CDE -19.9 (6.4)     CDE
## 3    ERad    -438 (81)    ERad    -67 (18)    ERad  -33.7 (20)    ERad
## 4     CDP  -2950 (150)     CDP   -154 (28)     CDP   -298 (38)     CDP
## 5      IO  -4130 (120)      IO   -395 (48)      IO   -530 (44)      IO
## 6     Imp  -7580 (170)     Imp   -795 (42)     Imp   -803 (44)     Imp
## 7 Stoufer -12500 (250) Stoufer  -1340 (66) Stoufer  -1660 (80) Stoufer
##           elpd   model        elpd
## 1        0 (0)     CDO       0 (0)
## 2    -297 (40)     CDE -27.9 (9.6)
## 3    -449 (73)    ERad  -98.7 (26)
## 4  -2410 (140)     CDP   -525 (52)
## 5  -3520 (120)      IO   -872 (52)
## 6  -6600 (160)     Imp  -1230 (57)
## 7 -11300 (240) Stoufer  -2560 (91)
```

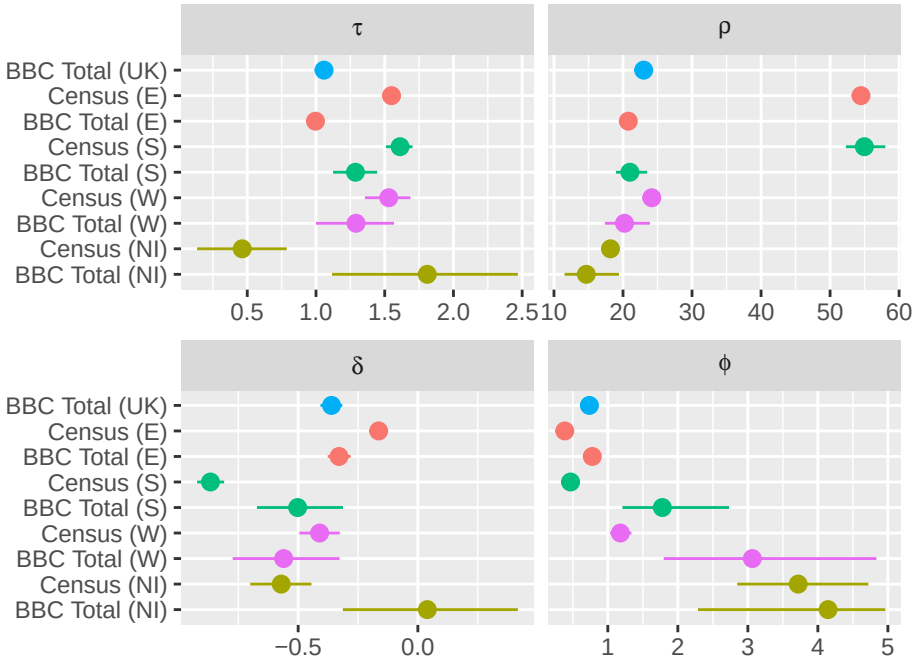
```
##     model        elpd   model        elpd   model         elpd   model
## 1     CDO       0 (0)     CDO       0 (0)     CDO        0 (0)     CDO
## 2     CDE -21.1 (7.1)     CDE  -95.8 (18)     CDE    -247 (35)     CDE
## 3    ERad    -67 (18)    ERad   -182 (38)    ERad    -336 (67)    ERad
## 4     CDP   -154 (28)     CDP   -843 (89)     CDP  -2100 (130)     CDP
## 5      IO   -395 (48)      IO  -1350 (82)      IO  -3130 (110)      IO
## 6     Imp   -795 (42)     Imp  -2450 (94)     Imp  -5710 (150)     Imp
## 7 Stoufer  -1340 (66) Stoufer -4850 (170) Stoufer -10100 (220) Stoufer
##          elpd
## 1       0 (0)
## 2 -30.3 (9.4)
## 3   -115 (21)
## 4   -397 (41)
## 5   -749 (48)
## 6  -1080 (52)
## 7  -2330 (88)
```

Posterior Estimates (Favoured model - CDO)

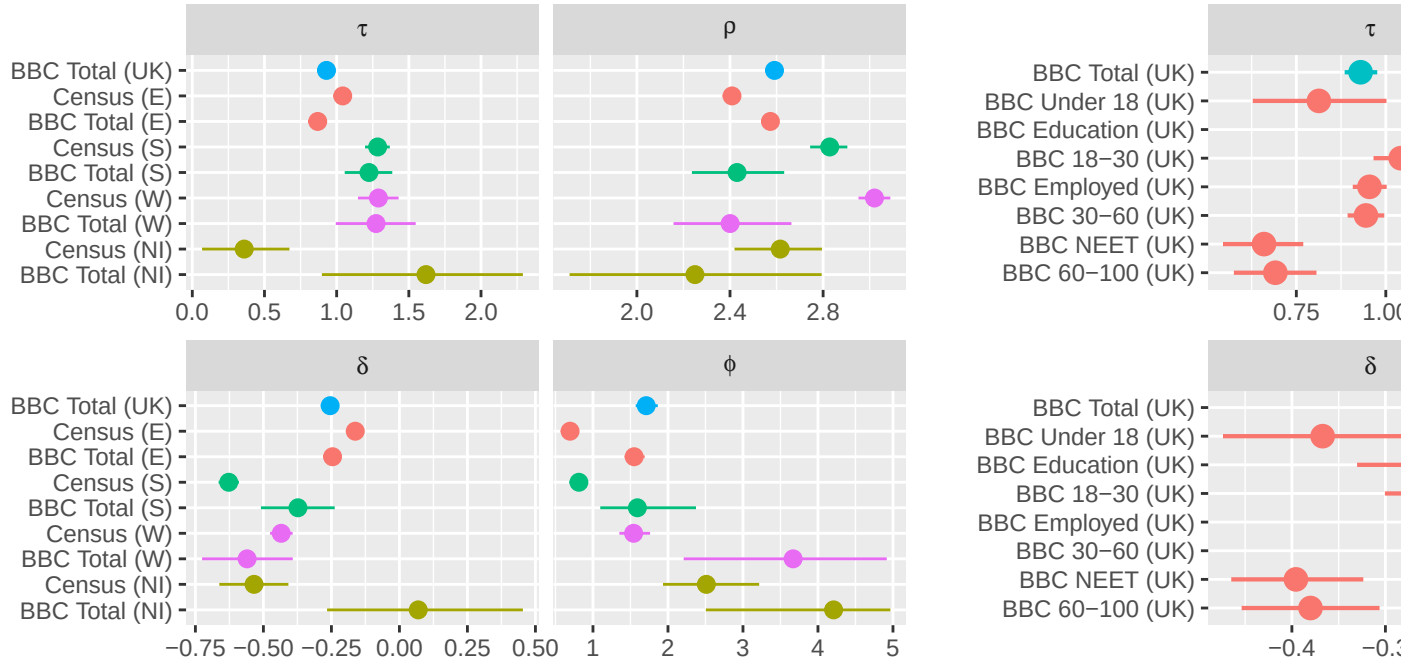
The CDO model (Competing destinations with offset) is favoured (or has indistinguishable predictive performance) for each data set.



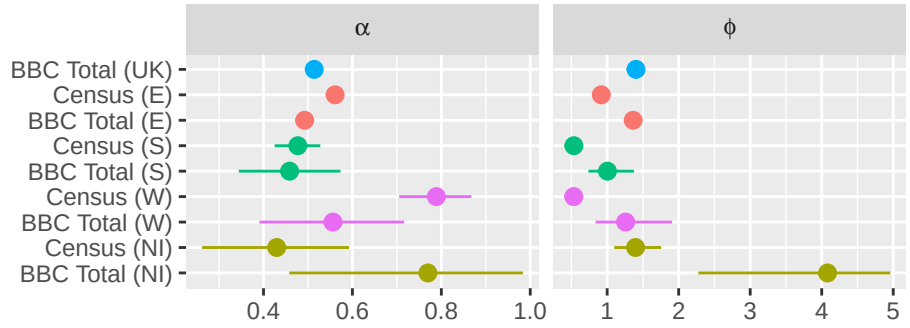
Posterior Estimates (CDE)

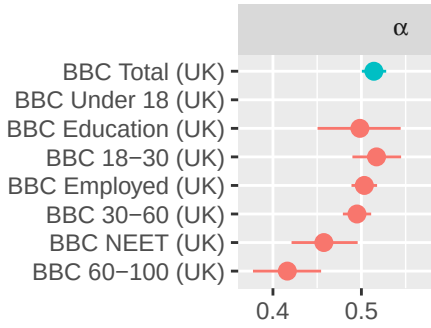


Posterior Estimates (CDP)

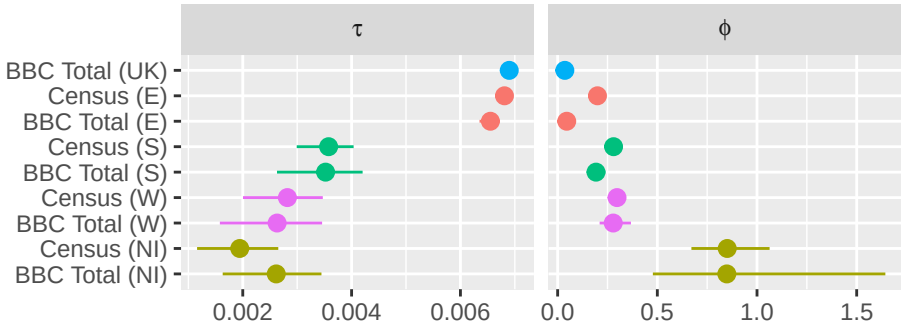


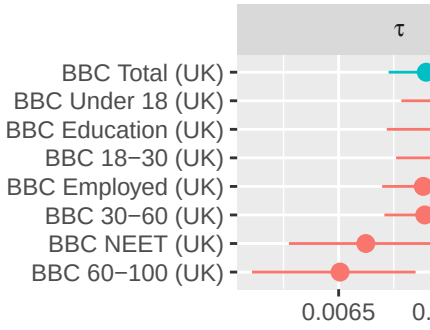
Posterior Estimates (ERad)



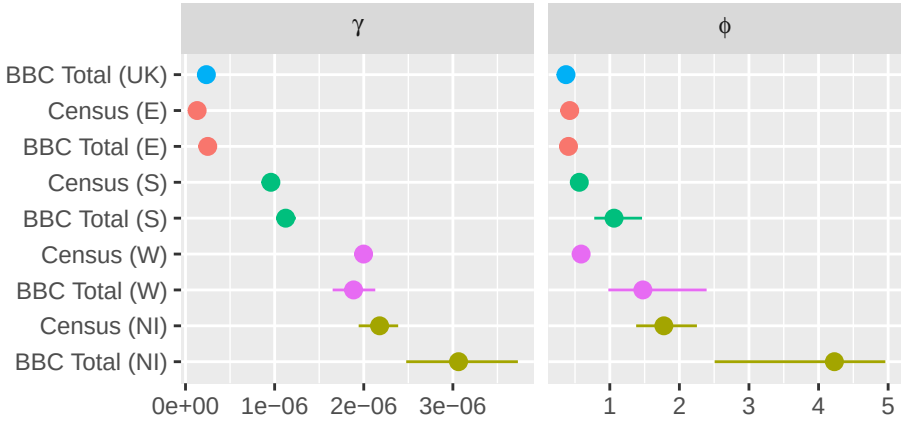


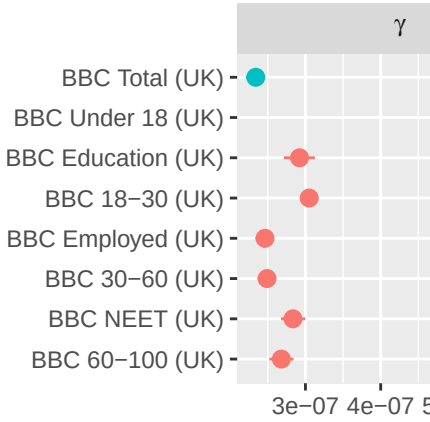
Posterior Estimates (Stoufer)



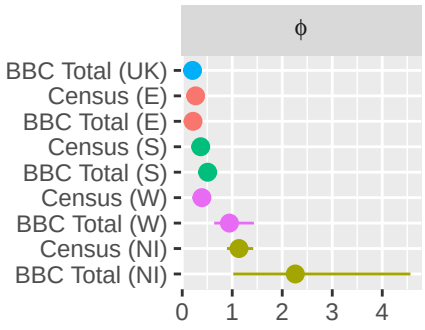


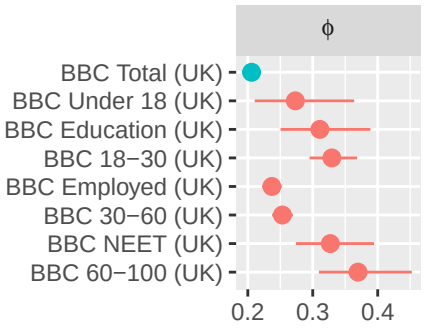
Posterior Estimates (IO)



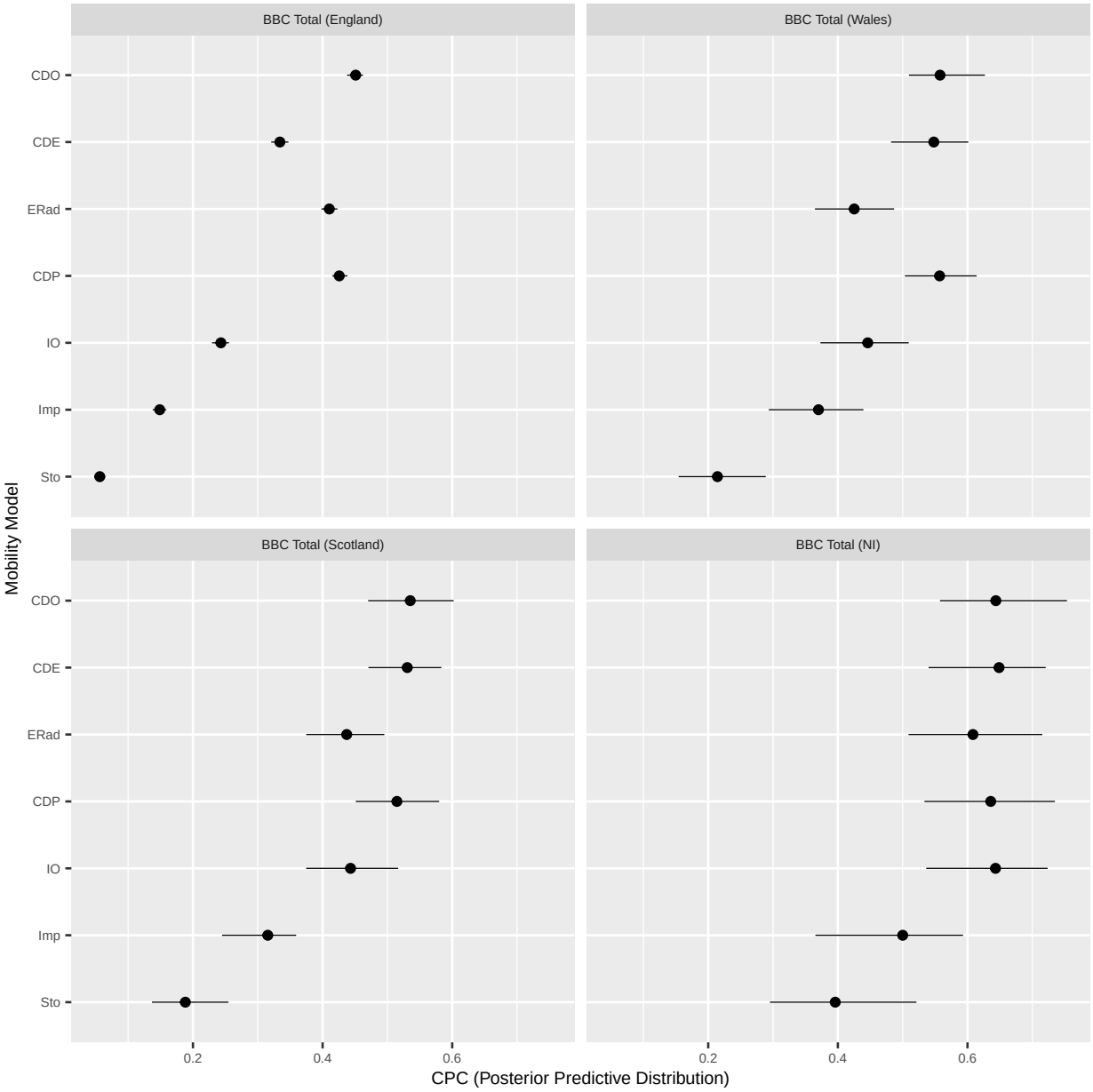


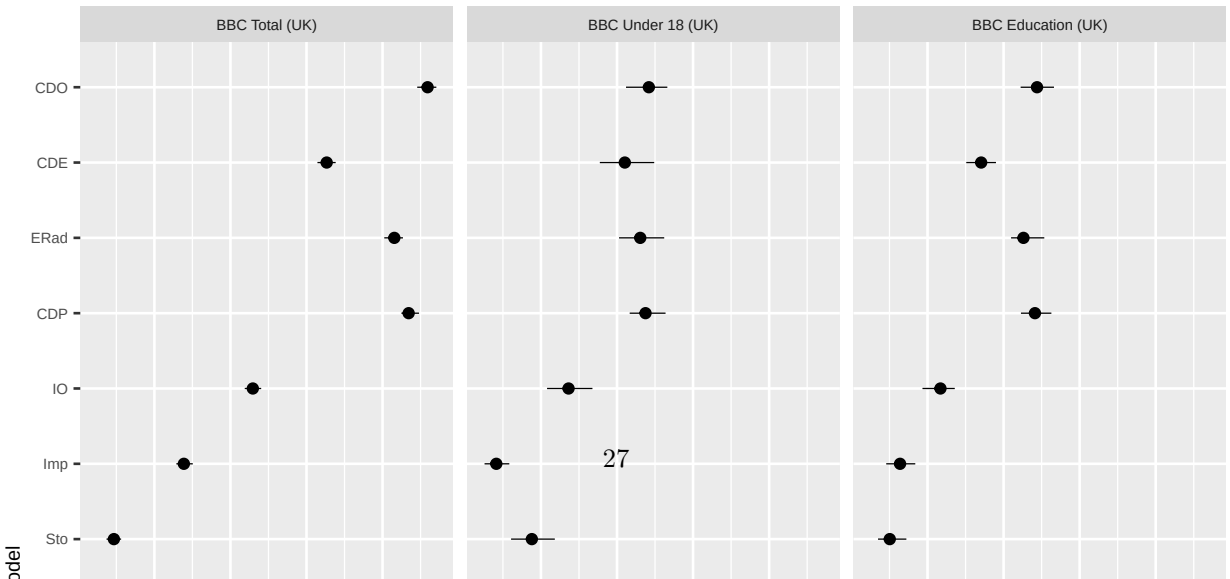
Posterior Estimates (Imp)

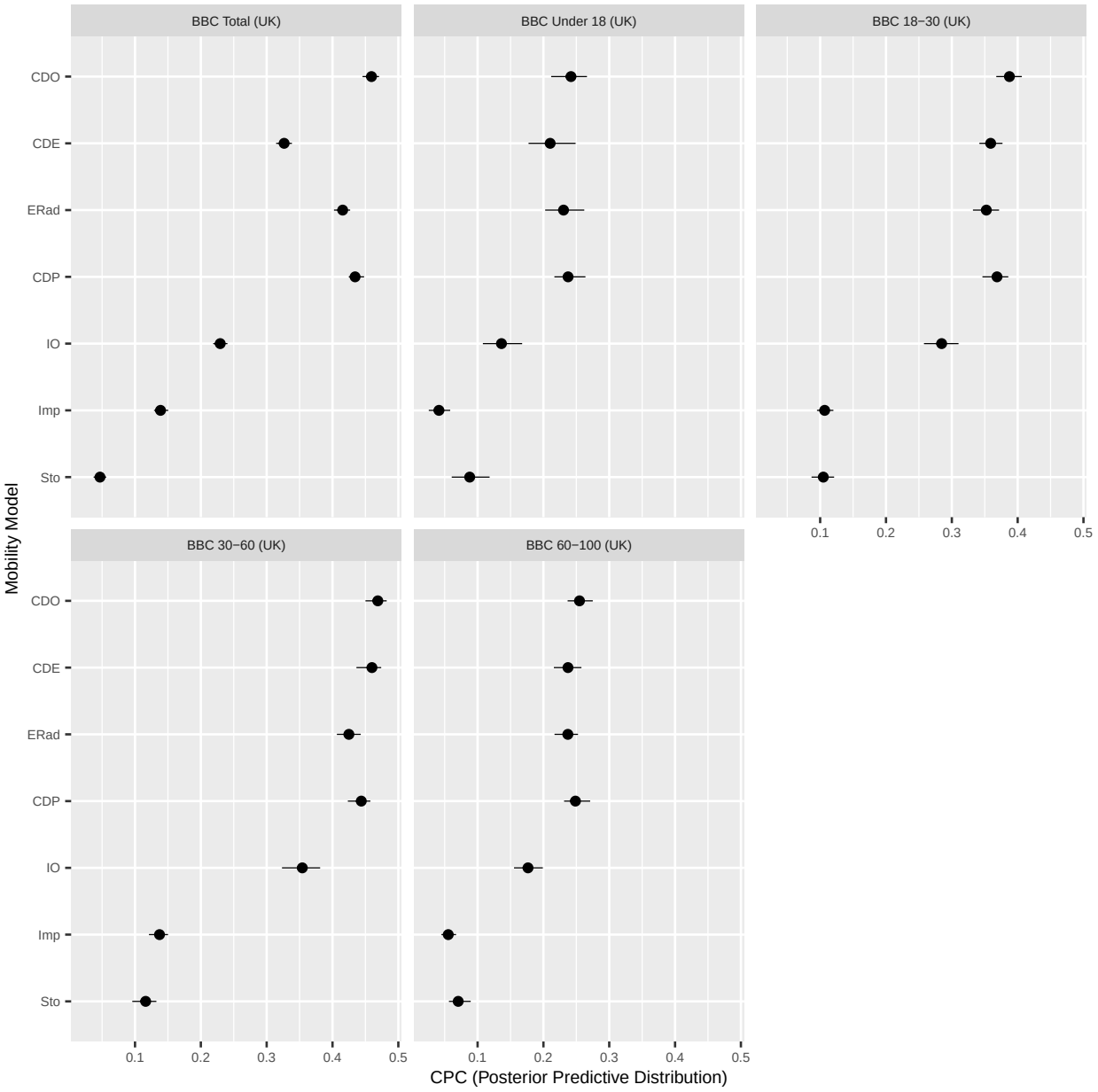




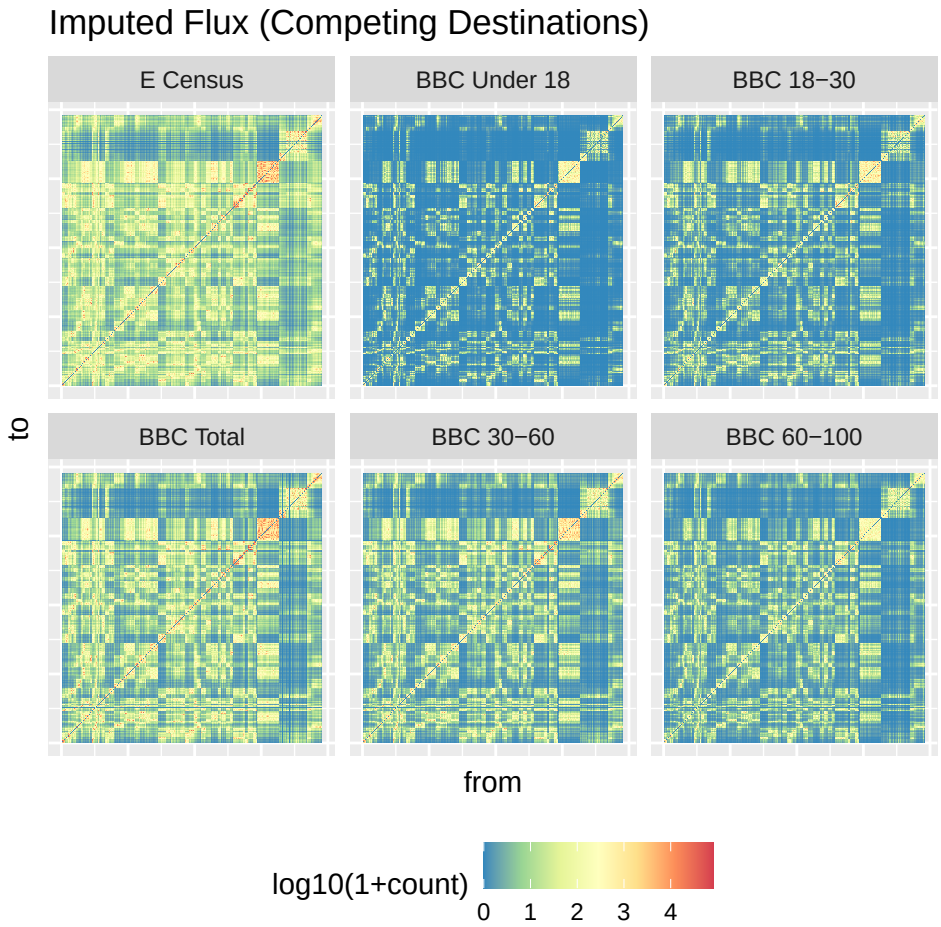
Posterior Predictive Checks (CPC)



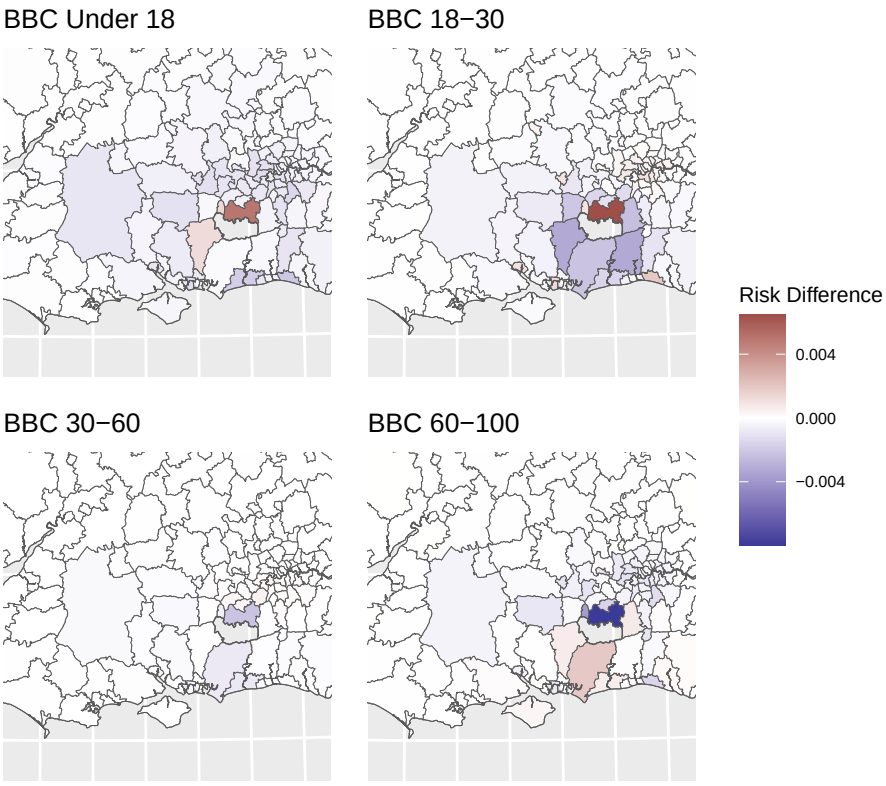


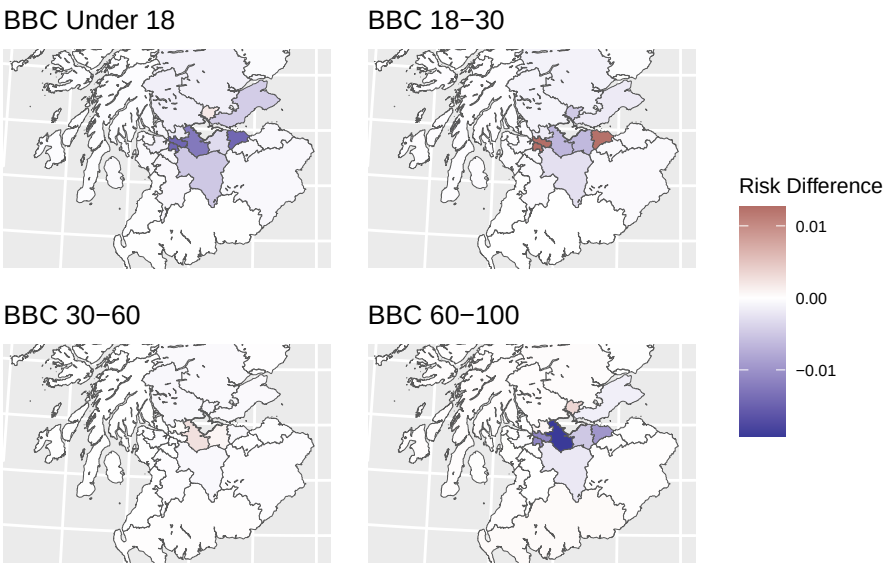

BBC Total (UK)
BBC Under 18 (UK)
BBC 18−30 (UK)
BBC 30−60 (UK)
BBC 60−100 (UK)
CDO
CDE
ERad
CDP
IO
Imp
Sto
Mobility Model
0.1
0.2
0.3
0.4
0.5
CPC (Posterior Predictive Distribution)

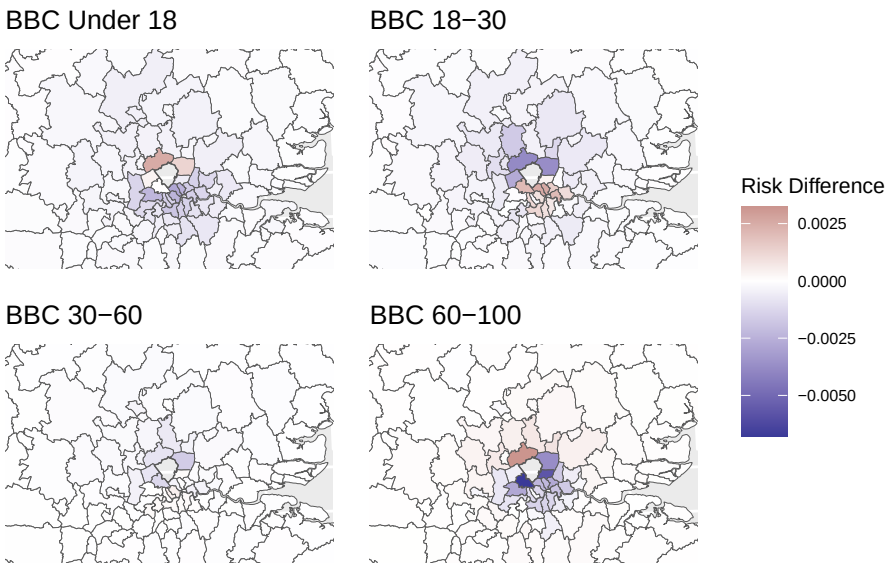
Imputed BBC flux matrices from CDO Model



Risk difference between imputed Census and BBC commuter flows




BBC Under 18
BBC 18−30
BBC 30−60
BBC 60−100
Risk Difference
0.01
0.00
−0.01


BBC Under 18
BBC 18−30
BBC 30−60
BBC 60−100
Risk Difference
0.0025
0.0000
−0.0025
−0.0050