Testing different imputation methods on PUMS (MCAR) - RandomForest

---

```
# load dataset: df
load('../Datasets/ordinalPUMS.Rdata')

# take 10,000 samples: df
n = 10000
sample <- sample(nrow(df), size = 10000)
df <- df[sample,]

# create MCAR scneario with 30% chance of missing: df_observed
set.seed(0)
missing_prob = 0.3
df_observed <- df
missing_col = colnames(df)[c(1,3,5,7,9,11)]
for (col in missing_col) {
  missing_ind <- rbernoulli(n,p = missing_prob)
  df_observed[missing_ind, col] <- NA
}
```

MICE-CART

```
library(mice)
```

```
##
## Attaching package: 'mice'
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
df.imp <- missForest(df_observed, verbose = TRUE)
```

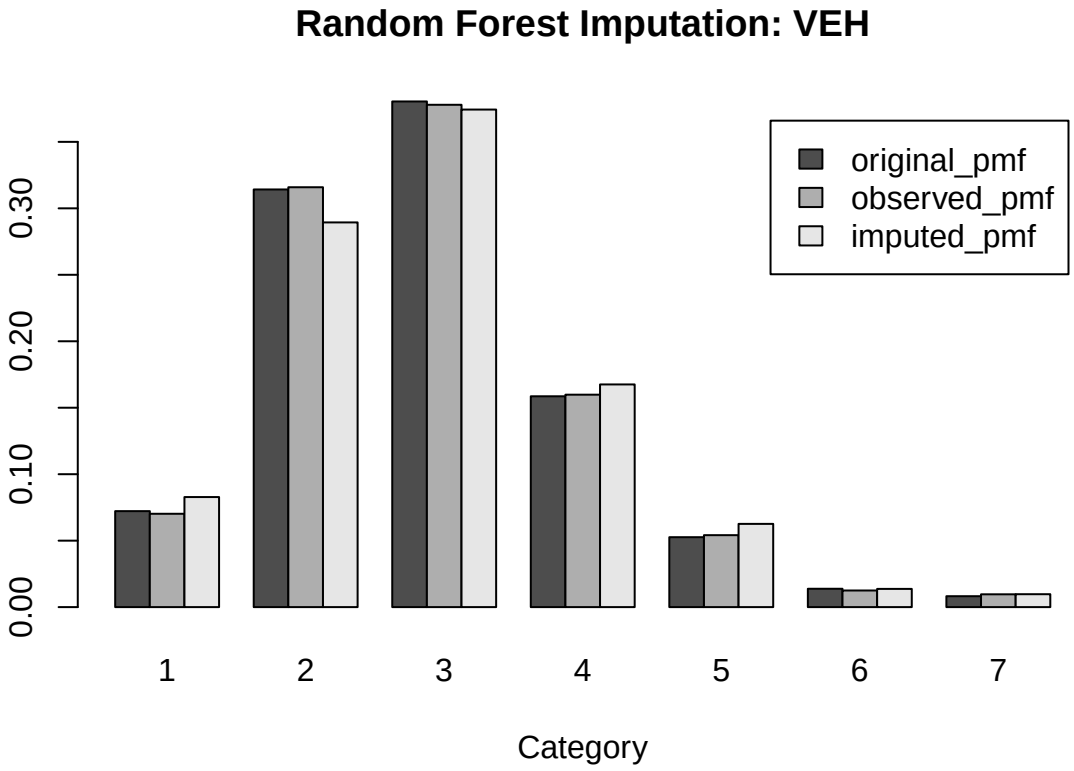
```
##   missForest iteration 1 in progress...done!
##     estimated error(s): 0.3071119
##     difference(s): 0.1048364
##     time: 5.972 seconds
##
##   missForest iteration 2 in progress...done!
##     estimated error(s): 0.2633827
##     difference(s): 0.0407
##     time: 5.461 seconds
##
##   missForest iteration 3 in progress...done!
##     estimated error(s): 0.2566451
##     difference(s): 0.02505455
##     time: 5.473 seconds
##
##   missForest iteration 4 in progress...done!
##     estimated error(s): 0.2544474
```

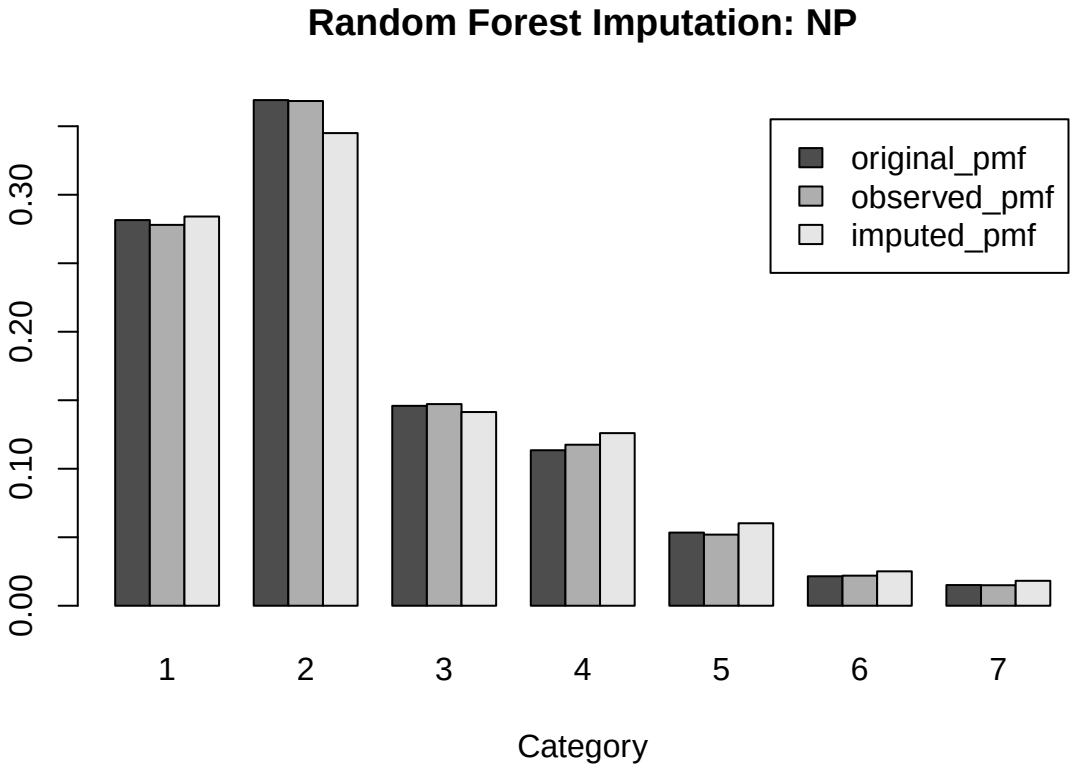
```
##     difference(s): 0.02060909
##     time: 5.442 seconds
##
##   missForest iteration 5 in progress...done!
##     estimated error(s): 0.2524806
##     difference(s): 0.01999091
##     time: 5.229 seconds
##
##   missForest iteration 6 in progress...done!
##     estimated error(s): 0.2513866
##     difference(s): 0.01840909
##     time: 5.449 seconds
##
##   missForest iteration 7 in progress...done!
##     estimated error(s): 0.2511867
##     difference(s): 0.0174
##     time: 5.029 seconds
##
##   missForest iteration 8 in progress...done!
##     estimated error(s): 0.2499469
##     difference(s): 0.01743636
##     time: 5.255 seconds
```

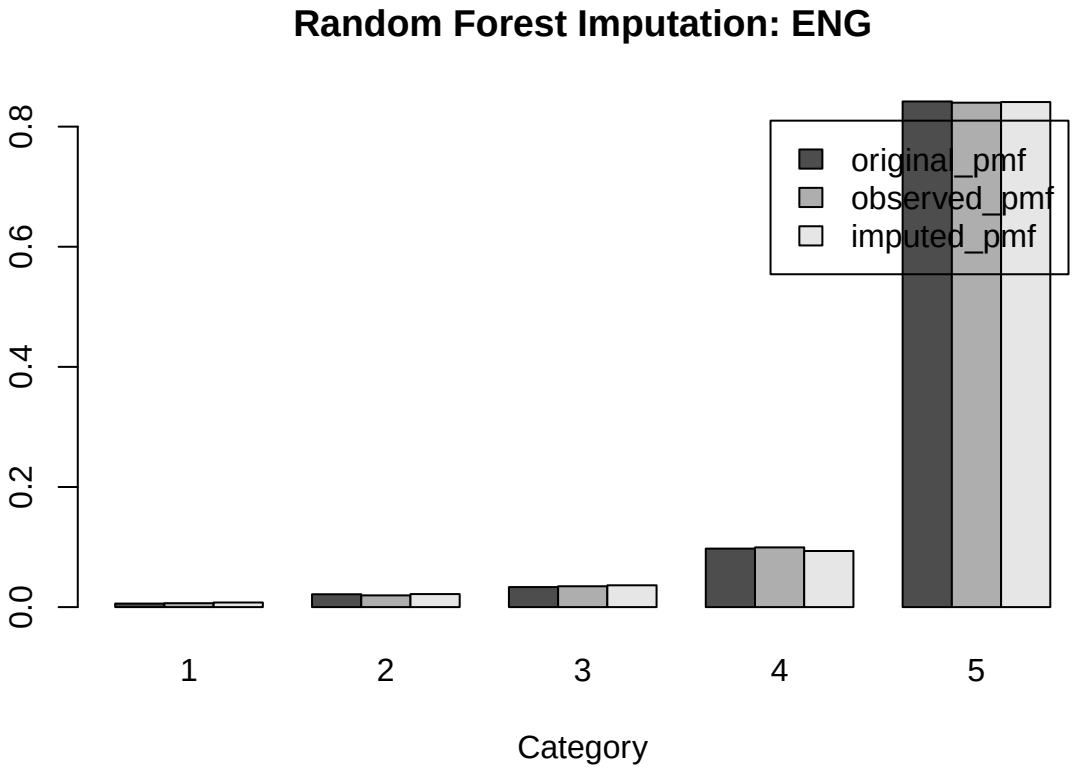
```
imputed_df <-  df.imp$ximp
```

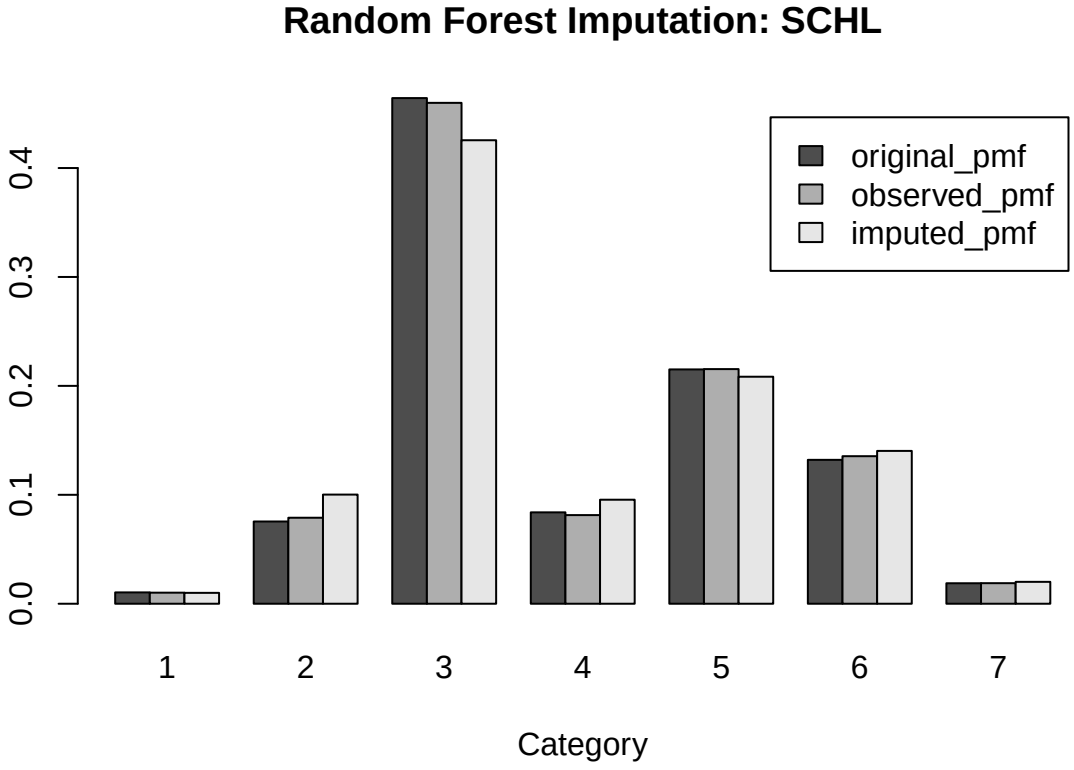
Diagnostics

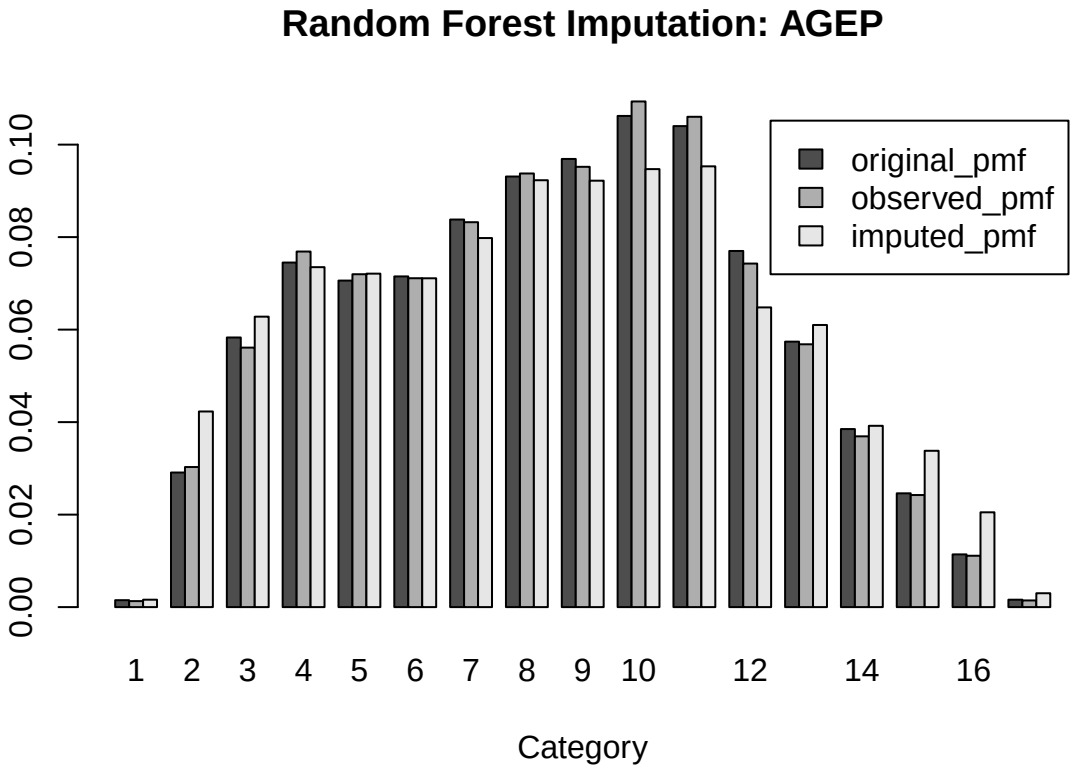
```
for (var_index in c(1,3,5,7,9,11)) {
  y_original = df[,var_index]
  original_pmf = table(y_original)/length(y_original)

  # Observed distribution
  missing_indicator = is.na(df_observed)[,var_index]
  y_observed = y_original[!missing_indicator]
  observed_pmf = table(y_observed)/length(y_observed)

  # Marginal distribution after imputation
  imputed_pmf = table(imputed_df[, var_index])/sum(table(imputed_df[, var_index]))

  results = rbind(original_pmf,observed_pmf,imputed_pmf)
  colnames(results)<- 1:dim(imputed_pmf)
  barplot(results, xlab = 'Category', beside = TRUE,
          legend = TRUE,
          main = paste('Random Forest Imputation:', colnames(df)[var_index]))
}
```

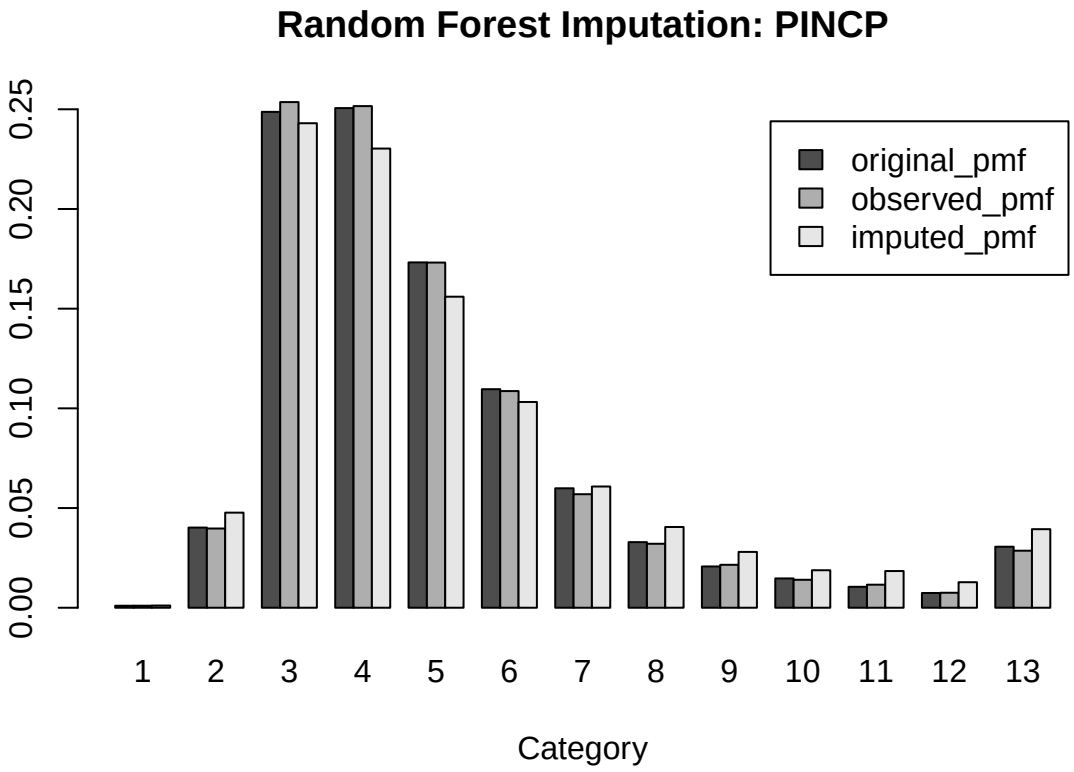

Random Forest Imputation: VEH
original_pmf
observed_pmf
imputed_pmf
0.00
0.10
0.20
0.30
1
2
3
4
5
6
7
Category


Random Forest Imputation: NP
original_pmf
observed_pmf
imputed_pmf
0.00
0.10
0.20
0.30
1
2
3
4
5
6
7
Category


Random Forest Imputation: ENG
0.0
0.2
0.4
0.6
0.8
original_pmf
observed_pmf
imputed_pmf
1
2
3
4
5
Category


Random Forest Imputation: SCHL
0.0
0.1
0.2
0.3
0.4
original_pmf
observed_pmf
imputed_pmf
1
2
3
4
5
6
7
Category


Random Forest Imputation: AGEP
original_pmf
observed_pmf
imputed_pmf
0.00
0.02
0.04
0.06
0.08
0.10
1
2
3
4
5
6
7
8
9
10
12
14
16
Category


Random Forest Imputation: PINCP
original_pmf
observed_pmf
imputed_pmf
0.00
0.05
0.10
0.15
0.20
0.25
1
2
3
4
5
6
7
8
9
10
11
12
13
Category