Individual-level analysis: data imputation

Part of the final project for AQMSS II

Stepan Polikanov Vera Okisheva

Imputation

In this section we impute values for the dataset values on the individual level. The number of predictors for high-quality imputations is limited.

```
source(here::here("utilities", "check_packages.R"))
```

```
ep_raw_dep <- read_rds(here("data", "data_built", "ep_raw_dep.rds"))
```

We first select relevant variables. Note that we do not impute the "vote" variable. This is because it does not include missing values other than "Declined to answer". It might be an interesting exercise to impute vote choices and the literature indicates that imputing the outcome is not problematic in most scenarios (Woods et al., 2024). However the censoring of the outcome is interesting in and of itself, which is why we choose to model it explicitly. [1]

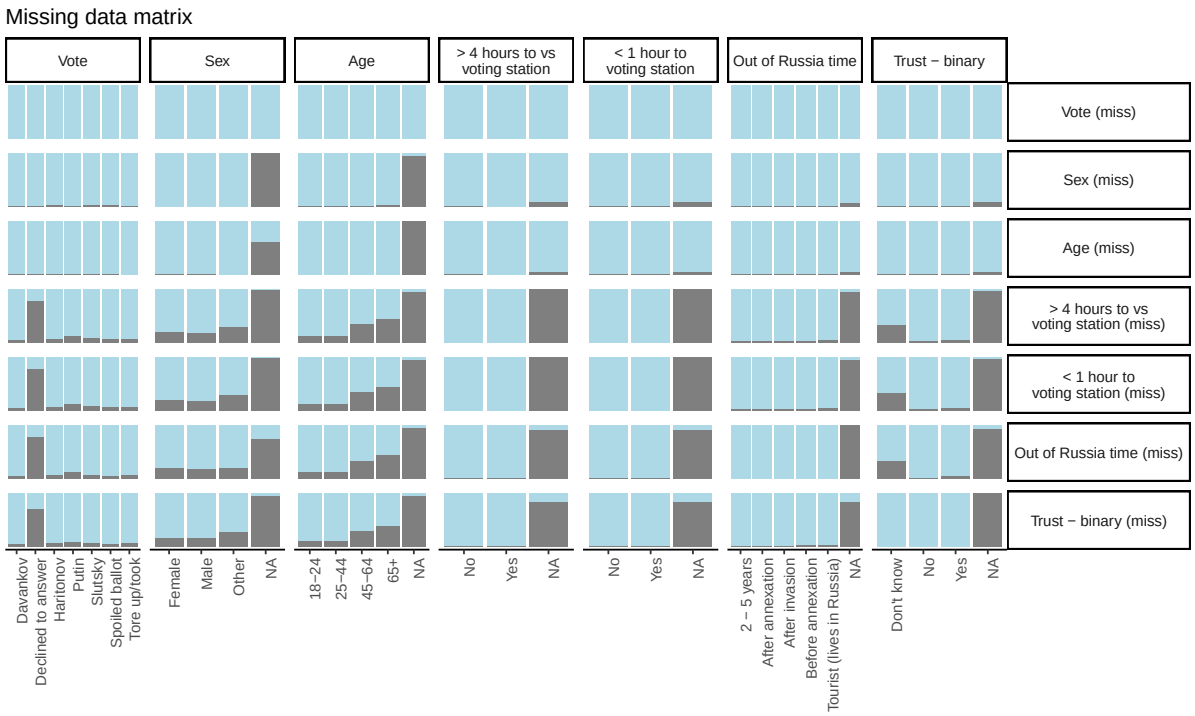
```
ep_mice_prep <- ep_raw_dep |>
  select(
    # Identificators and fixed effects
    countryname_en, countrycode_c, countrycode_n, city_en, voting_station,
    # Variables to be imputed
    vote, sex, age_bin, out_of_Russia_time, time_to_vs.less_than_hour,
    time_to_vs.more_than_4hours, result_trust_bin,
    # Auxiallary variables
    time_to_vs, result_trust) |>
  mutate(across(c(-vote), ~ as.factor(if_else(. %in% c("No Data",
                                                       "Declined to answer"),
                                              NA, .))),
         city_en = as.integer(city_en))
```

Upon examining the distributions of missingness, with x-axis indicating the variables and y-axis the missing dummies in them, it becomes apparent that the missingness is primarily

[1] In particular, we are interested in a scenario where all "Declined to answer" come from the supporters of the incumbent.

dependent on the "Declined to answer" category of the vote question. This means that people that refused to disclose their choice of vote also didn't share their socio-demographic data. This means that we are unlikely to get unbiased estimates for that category, or, indeed for its combination with the honest incumbent vote answers.

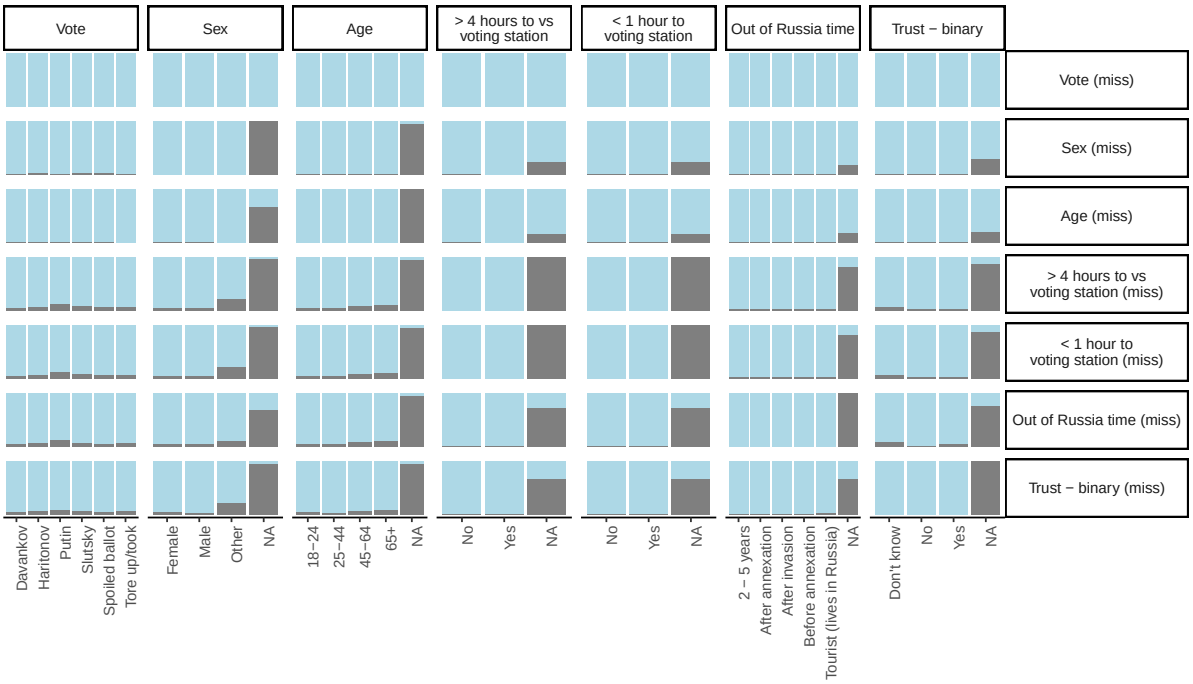
```
labelled::var_label(ep_mice_prep) <- list(
  vote = "Vote", sex = "Sex", age_bin = "Age",
  time_to_vs.more_than_4hours = "> 4 hours to vs\nvoting station",
  time_to_vs.less_than_hour = "< 1 hour to\nvoting station",
  out_of_Russia_time = "Out of Russia time",
  result_trust_bin = "Trust - binary",
  countryname_en = "Country")

ep_mice_prep |>
  missing_pairs(dependent = "vote",
                explanatory = c("sex", "age_bin",
                                "time_to_vs.more_than_4hours",
                                "time_to_vs.less_than_hour",
                                "out_of_Russia_time",
                                "result_trust_bin"),
                position = "fill") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1),
          strip.text.y = element_text(angle = 0))
```




```
ep_mice_prep |>
  filter(vote != "Declined to answer") |>
  missing_pairs(dependent = "vote",
                explanatory = c("sex", "age_bin",
```

```
                                "time_to_vs.more_than_4hours",
                                "time_to_vs.less_than_hour",
                                "out_of_Russia_time",
                                "result_trust_bin"),
                    position = "fill") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1),
          strip.text.y = element_text(angle = 0))
```

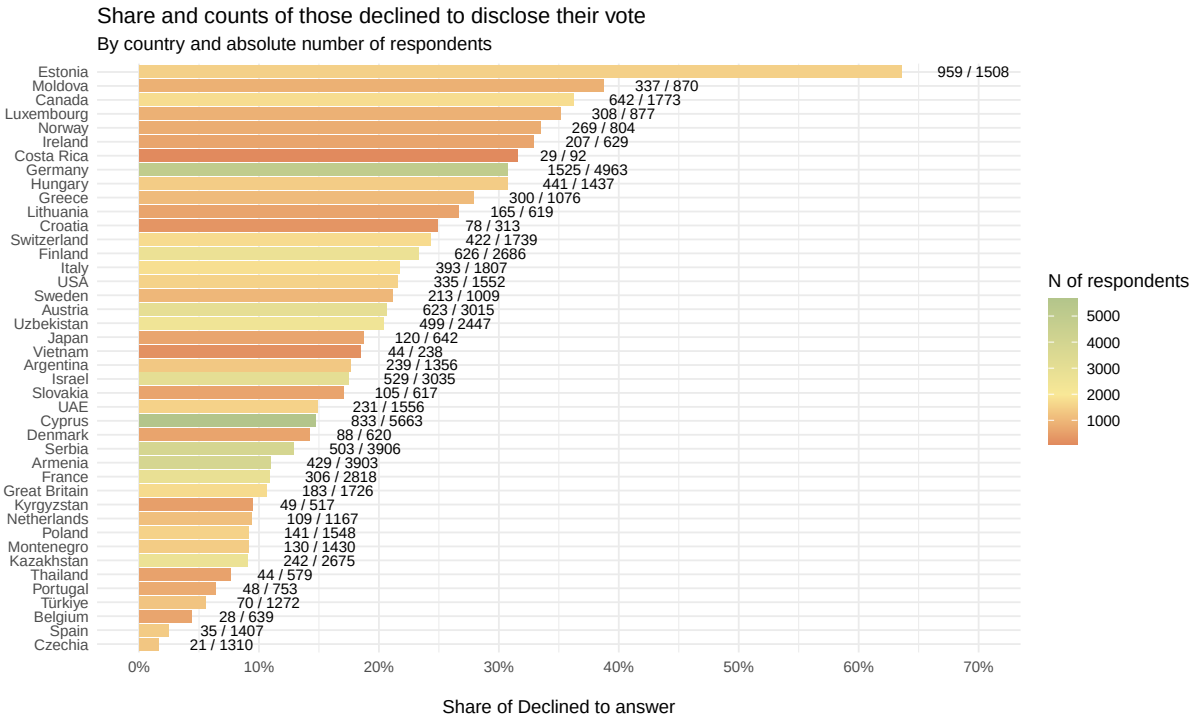
Missing data matrix



There remain (very) significant differences between countries in terms of vote non-disclosure rates - from less than 1 to more than 60. This points to the relevancy of country-level characteristics, which are the main focus of this paper.

```
ep_mice_prep |>
  filter(!countryname_en %in% c("New Zealand", "Australia")) |>
  group_by(countryname_en) |>
  summarize(vote_na = sum(vote == "Declined to answer"),
            vote = n()) |>
  mutate(share = vote_na/vote) |>
  ggplot(aes(x = share, y = reorder(countryname_en, share),
             label = paste0(vote_na, " / ", vote), fill = vote)) +
    geom_bar(stat = "identity") +
    geom_text(size = 3, hjust = -0.5, position = position_dodge(width = 1),
              inherit.aes = TRUE) +
    scale_x_continuous(limits = c(0, 0.7),
                       breaks = seq(0, 0.7, 0.1),
                       labels = scales::label_number(scale = 100,
                                                     suffix = "%")) +
    scale_fill_gradient2(low = "#bf212f", mid = "#F9E897", high = "#b3c58b",
                         midpoint = 2000) +
```

```
  labs(x = "\nShare of Declined to answer",
       y = NULL, fill = "N of respondents",
       title = "Share and counts of those declined to disclose their vote",
       subtitle = "By country and absolute number of respondents") +
  theme_minimal()
```



We then set up the imputation itself. There are two id categorical variables for cities (which in the exit poll sample equal voting stations) and countries - I retain those as predictors as they serve as fixed effects capturing the inherent country- and city- level group (cluster) differences. There are no methods in the mice family of packages to impute polynomial data with mixed effects modelling techniques, so we will have to do without the helpful properties of partial pooling.

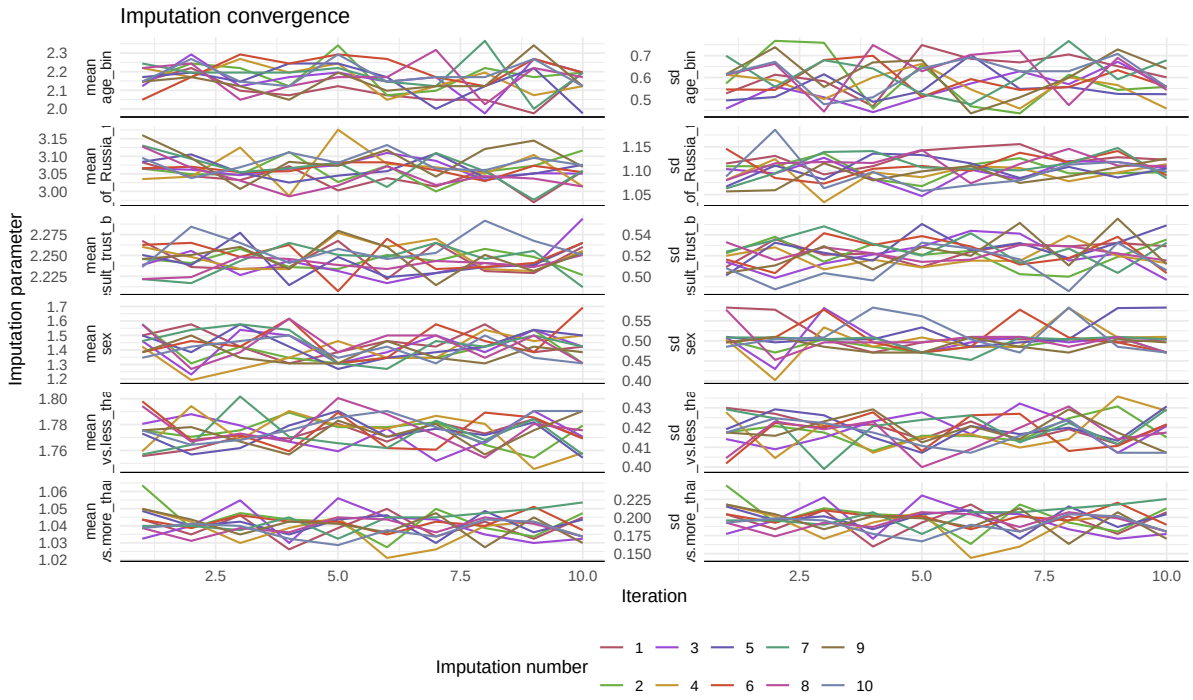
Apart from the variables we are imputing there are also two auxiliary variables - those are untransformed `result_trust` and `time_to_vs` variables. We have modified them when creating variables (in the `variables_descriptive` script), however for imputation we prefer to use untransformed (apart from NA definitions) versions that contain more information. We thus set these variables and not the modified ones as predictors in the imputation - this avoids collinearity in the models. On the other hand we disallow them to predict their modified variants - this avoids circularity in the predictions.

```
# Dry run
imp1dry <- mice(ep_mice_prep, seed = 1, maxit = 0)
```

```r
# Extract predictor matrix
pred <- imp1dry$predictorMatrix

# Modify predictor matrix

  ## Exclude duplicate variables
  pred[, "countrycode_c"] <- 0
  pred[, "countrycode_n"] <- 0
  pred[, "voting_station"] <- 0

  ## Use full-information variables for imputation
  pred[, "time_to_vs.less_than_hour"] <- 0
  pred[, "time_to_vs.more_than_4hours"] <- 0
  pred[, "result_trust_bin"] <- 0

  ## Do not impute time_to_vs and result_trust with their (collinear)
  ## full-information alternatives
  pred["time_to_vs.less_than_hour", "time_to_vs"] <- 0
  pred["time_to_vs.more_than_4hours", "time_to_vs"] <- 0
  pred["result_trust_bin", "result_trust"] <- 0


imp1 <- futuremice(ep_mice_prep, method = c(
  # Identificators
  "", "", "", "", "",
  # Variables to be imputed
  "", "polyreg", "polyreg", "polyreg", "logreg", "logreg", "polyreg",
  # Aux
  "", ""),
  predictorMatrix = pred, m = 10, maxit = 10,
  parallelseed = 1, ncore = parallel::detectCores() - 1)


load(here("data", "data_raw", "imp1.RData"))
```

We first look at convergence. The trace plot shows that imputations mix for all variables nicely, which leads us to conclude that convergence is good!

```r
plot_trace(imp1) +
  scale_color_manual(values = c("#b15268", "#69b040", "#9c49d6", "#c79830",
                                "#6558b1", "#cc4d33", "#549e78", "#ba4c9d",
                                "#8a7443", "#7484ac")) +
  labs(title = "Imputation convergence") +
  theme_minimal() +
  theme(legend.position = "bottom")
```



We then turn our attention to a probably most important issue with these imputations - those being the number of actual imputed values.

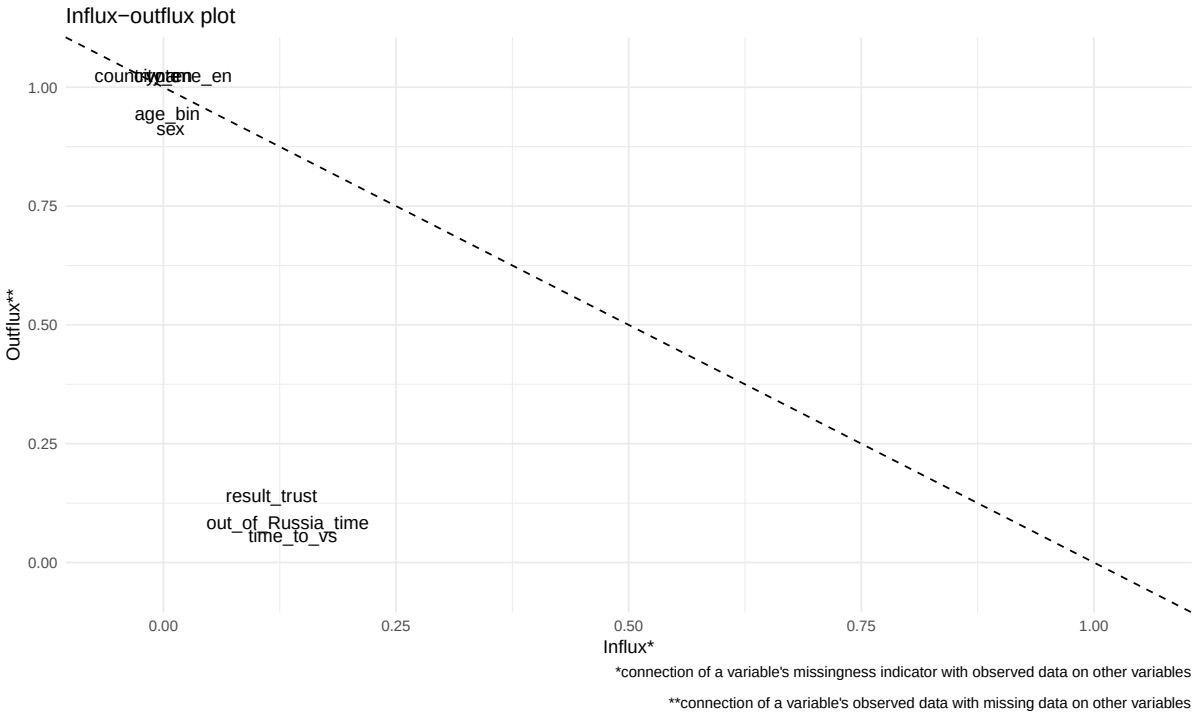
```
imp1_cmp <- complete(imp1, "long", include = T) |>
  mutate(imputed = factor(if_else(`.imp` > 0, 1, 0),
                          labels = c("Observed", "Average Imputed")),
         imp = factor(if_else(`.imp` == 0, "Observed", as.character(`.imp`)))) |>
  filter(!countryname_en %in% c("Australia", "New Zealand"))

imp1_cmp |>
  group_by(imp) |>
  summarize(across(c(sex, age_bin, out_of_Russia_time,
                     time_to_vs.less_than_hour, time_to_vs.more_than_4hours,
                     result_trust_bin), ~ sum(is.na(.)))) |>
  select(-imp) |>
  distinct() |>
  rownames_to_column() |>
  pivot_longer(cols = c(-rowname)) |>
  mutate(name = factor(name,
                       levels = c("sex", "age_bin", "out_of_Russia_time",
                                  "time_to_vs.less_than_hour",
                                  "time_to_vs.more_than_4hours",
                                  "result_trust_bin"),
                       labels = c("Sex", "Age", "Time out of Russia",
                                  "Time to voting station < 1 hour",
                                  "Time to voting station > 4 hours",
                                  "Trust in the result")),
         rowname = factor(rowname, levels = c(1, 2), labels = c("Imputed",
                                                                "Observed"))) |>
  ggplot(aes(x = value, y = rowname, fill = rowname)) +
    geom_bar(stat = "identity", width = 0.5) +
```

```
  facet_wrap(~ name, scales = "free_x", ncol = 2) +
  scale_fill_manual(values = c("#b3c58b", "#cc4d33")) +
  labs(y = NULL, x = "\nNumber of missing values",
       title = "Imputation: number of missing values before and after") +
  theme_minimal() +
  theme(legend.position = "none")
```

The number of values that the model imputes is very low - this is because if one of the predictors contains missing values, the outcome of the imputation is also missing. There is essentially no way to correct for this issue with the data at hand - we are wary of deleting informative predictors from the imputation as the variables are already quite limited and it may lead to imprecise imputations.

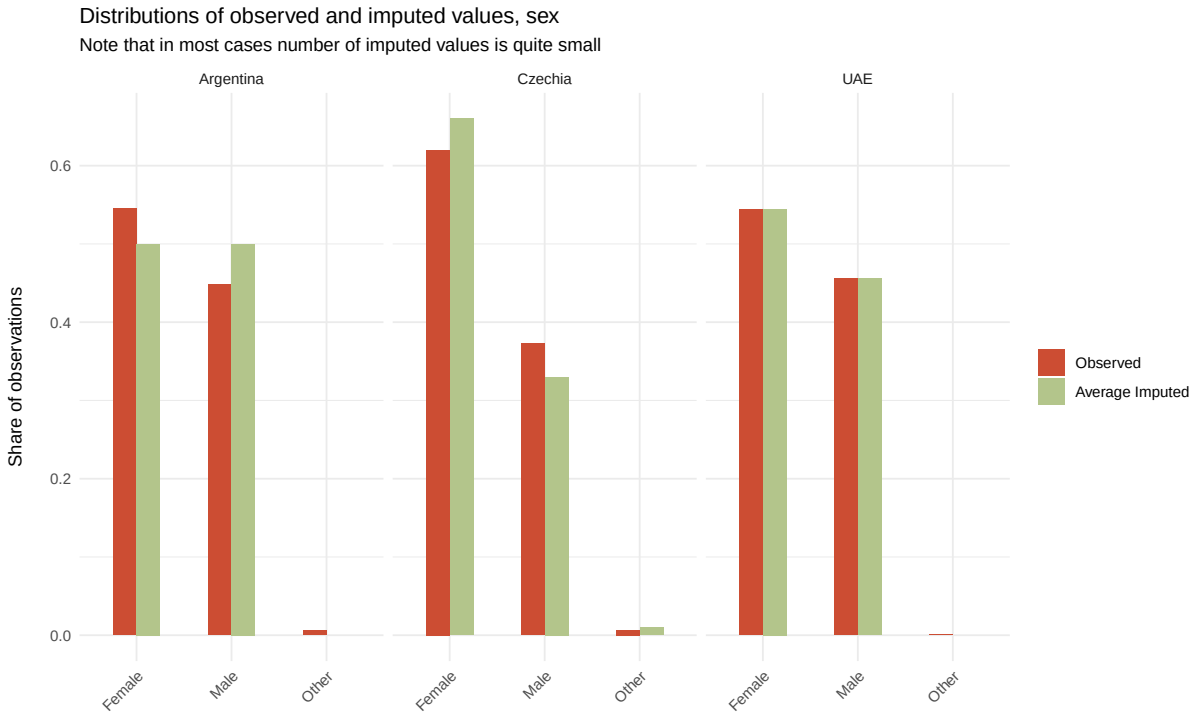
Consider the following influx-outflux plot: the less informative variables are those with more missing values - but they also vary systematically by age and sex, so there is no way to model the latter correctly.

```
ep_mice_prep |>
  select(-countrycode_c, -countrycode_n, -voting_station, -time_to_vs.more_than_4hours,
         -time_to_vs.less_than_hour, -result_trust_bin) |>
  plot_flux() +
    labs(title = "Influx-outflux plot") +
    theme_minimal()
```

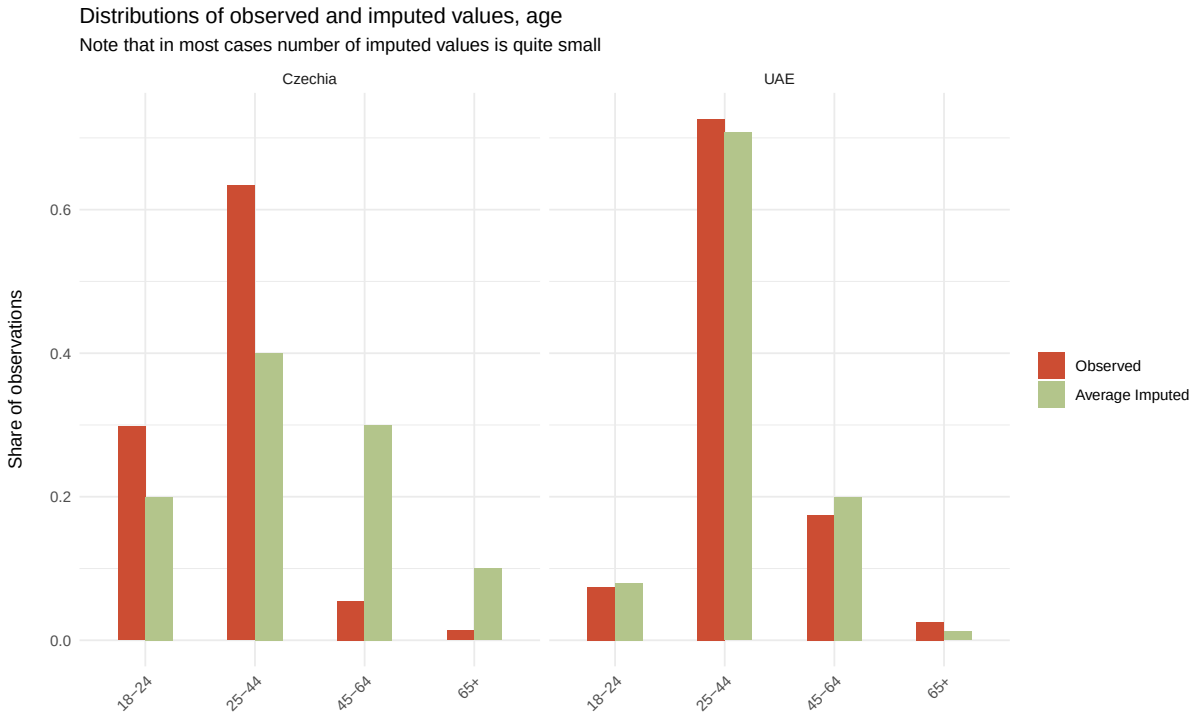


Lastly, it makes sense to examine the quality of the imputation. We plot the average share of each category across the 10 imputations performed against the distribution of observed values for each variable. The data is presented by country as we want to ensure that aggregations will be correct not only on the whole but also by country.

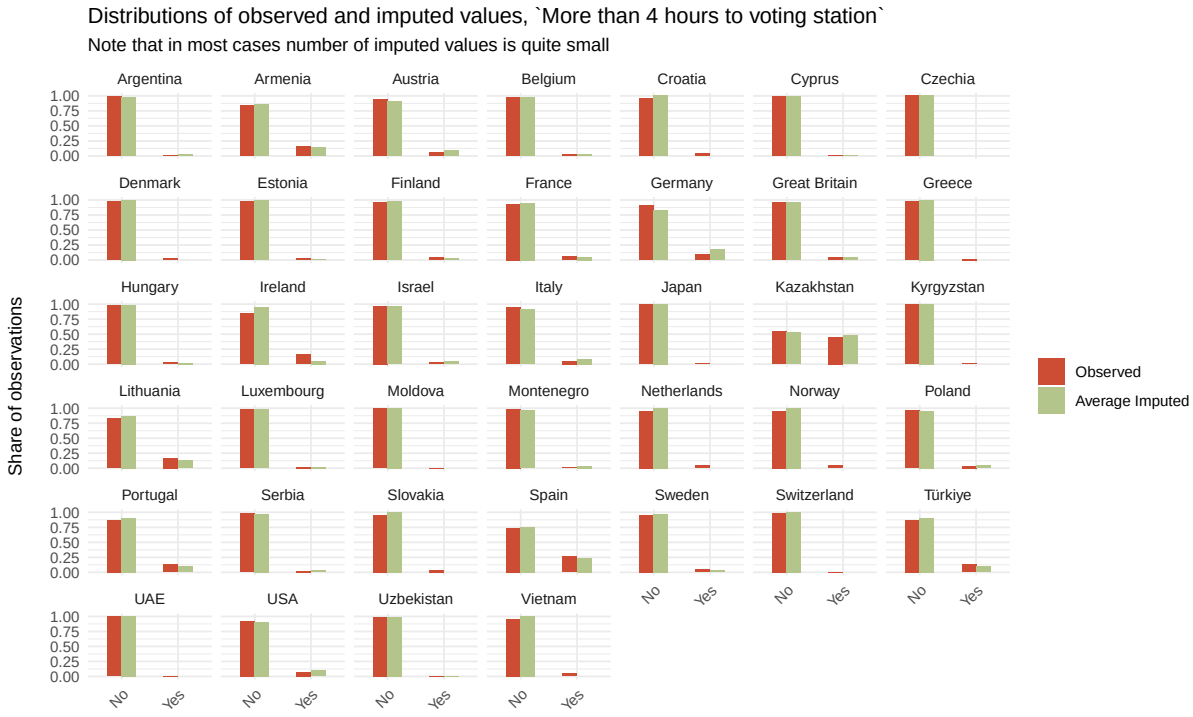
```
# Sex
imp1_cmp |>
  mutate(na = if_else(is.na(sex)[imp == "Observed"] & !is.na(imp == "1"), 1, 0)) |>
  filter(na == TRUE | imp == "Observed") |>
  group_by(countryname_en, imputed, sex) |>
  summarize(n = n(),
            country_na = any(na == 1)) |>
  drop_na() |>
  group_by(countryname_en) |>
  filter(!all(country_na == FALSE)) |>
  group_by(countryname_en, imputed) |>
  mutate(share = n/sum(n)) |>
  ggplot(aes(x = sex, y = share, fill = imputed)) +
    geom_bar(stat = "identity", position = position_dodge(preserve = "single"),
             width = 0.5) +
    facet_wrap(~ countryname_en) +
    scale_fill_manual(values = c("#cc4d33", "#b3c58b")) +
    labs(x = NULL, y = "Share of observations\n",
         title =
           "Distributions of observed and imputed values, sex",
         subtitle =
           "Note that in most cases number of imputed values is quite small",
         fill = NULL) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



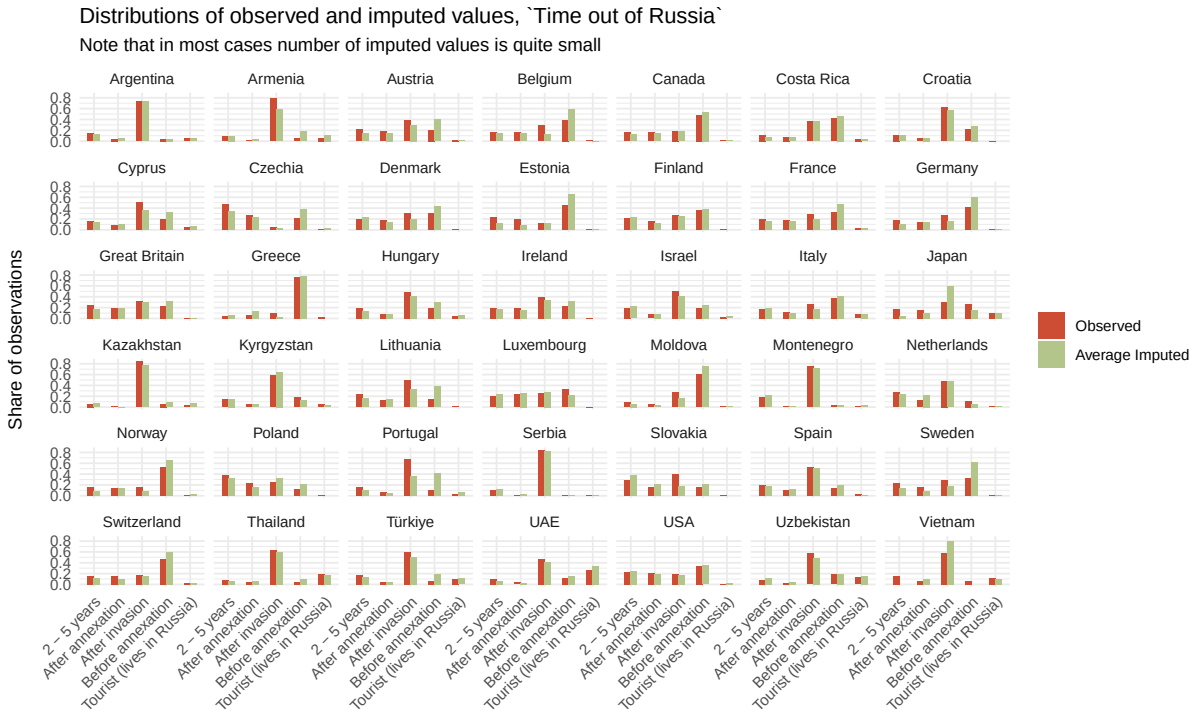

```
# Age
imp1_cmp |>
  mutate(na = if_else(is.na(age_bin)[imp == "Observed"] & !is.na(imp == "1"), 1, 0)) |>
  filter(na == TRUE | imp == "Observed") |>
  group_by(countryname_en, imputed, age_bin) |>
  summarize(n = n(),
            country_na = any(na == 1)) |>
  drop_na() |>
  group_by(countryname_en) |>
  filter(!all(country_na == FALSE)) |>
  group_by(countryname_en, imputed) |>
  mutate(share = n/sum(n)) |>
  ggplot(aes(x = age_bin, y = share, fill = imputed)) +
    geom_bar(stat = "identity", position = position_dodge(preserve = "single"),
             width = 0.5) +
    facet_wrap(~ countryname_en) +
    scale_fill_manual(values = c("#cc4d33", "#b3c58b")) +
    labs(x = NULL, y = "Share of observations\n",
         title =
           "Distributions of observed and imputed values, age",
         subtitle =
           "Note that in most cases number of imputed values is quite small",
         fill = NULL) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



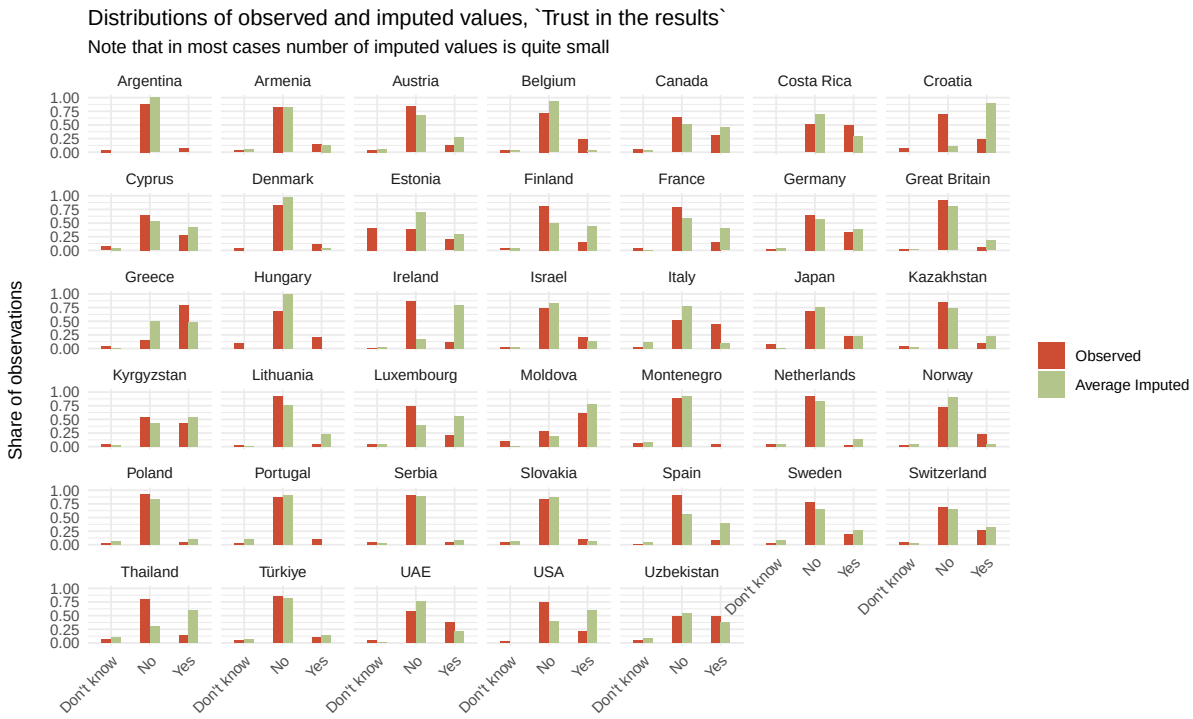

```
# More than 4
imp1_cmp |>
  mutate(na = if_else(is.na(time_to_vs.more_than_4hours)[imp == "Observed"]
                      & !is.na(imp == "1"), 1, 0)) |>
  filter(na == TRUE | imp == "Observed") |>
  group_by(countryname_en, imputed, time_to_vs.more_than_4hours) |>
  summarize(n = n(),
            country_na = any(na == 1)) |>
  drop_na() |>
  group_by(countryname_en) |>
  filter(!all(country_na == FALSE)) |>
  group_by(countryname_en, imputed) |>
  mutate(share = n/sum(n)) |>
  ggplot(aes(x = time_to_vs.more_than_4hours, y = share, fill = imputed)) +
    geom_bar(stat = "identity", position = position_dodge(preserve = "single"),
             width = 0.5) +
    facet_wrap(~ countryname_en) +
    scale_fill_manual(values = c("#cc4d33", "#b3c58b")) +
    labs(x = NULL, y = "Share of observations\n",
         title =
           "Distributions of observed and imputed values, `More than 4 hours to voting station`",
         subtitle =
           "Note that in most cases number of imputed values is quite small",
         fill = NULL) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Out of Russia
imp1_cmp |>
  mutate(na = if_else(is.na(out_of_Russia_time)[imp == "Observed"]
                      & !is.na(imp == "1"), 1, 0)) |>
  filter(na == TRUE | imp == "Observed") |>
  group_by(countryname_en, imputed, out_of_Russia_time) |>
  summarize(n = n(),
            country_na = any(na == 1)) |>
  drop_na() |>
  group_by(countryname_en) |>
  filter(!all(country_na == FALSE)) |>
  group_by(countryname_en, imputed) |>
  mutate(share = n/sum(n)) |>
  ggplot(aes(x = out_of_Russia_time, y = share, fill = imputed)) +
    geom_bar(stat = "identity", position = position_dodge(preserve = "single"),
             width = 0.5) +
    facet_wrap(~ countryname_en) +
    scale_fill_manual(values = c("#cc4d33", "#b3c58b")) +
    labs(x = NULL, y = "Share of observations\n",
         title =
           "Distributions of observed and imputed values, `Time out of Russia`",
         subtitle =
           "Note that in most cases number of imputed values is quite small",
         fill = NULL) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Trust in result
imp1_cmp |>
  mutate(na = if_else(is.na(result_trust_bin)[imp == "Observed"]
                      & !is.na(imp == "1"), 1, 0)) |>
  filter(na == TRUE | imp == "Observed") |>
  group_by(countryname_en, imputed, result_trust_bin) |>
  summarize(n = n(),
            country_na = any(na == 1)) |>
  drop_na() |>
  group_by(countryname_en) |>
  filter(!all(country_na == FALSE)) |>
  group_by(countryname_en, imputed) |>
  mutate(share = n/sum(n)) |>
  ggplot(aes(x = result_trust_bin, y = share, fill = imputed)) +
    geom_bar(stat = "identity", position = position_dodge(preserve = "single"),
             width = 0.5) +
    facet_wrap(~ countryname_en) +
    scale_fill_manual(values = c("#cc4d33", "#b3c58b")) +
    labs(x = NULL, y = "Share of observations\n",
         title =
           "Distributions of observed and imputed values, `Trust in the results`",
         subtitle =
           "Note that in most cases number of imputed values is quite small",
         fill = NULL) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



As noted in the graphs at times the number of imputed values is so low that the comparison is rather unfair. Overall we are satisfied with the imputation quality.

Conclusion

Based on these results, it appears that we cannot effectively solve the missingness pattern of the data. It clearly indicates that it depends on candidate choice. However since data is decidedly Missing At Random (MAR) this needn't be an issue - except for the Declined to answer category we can still recover unbiased estimates. While we perform multiple imputations of the missing values the systemic character of the pattern precludes the process from making any sort of difference. We therefore proceed with listwise deletion of missing observations for simplicity.