sad

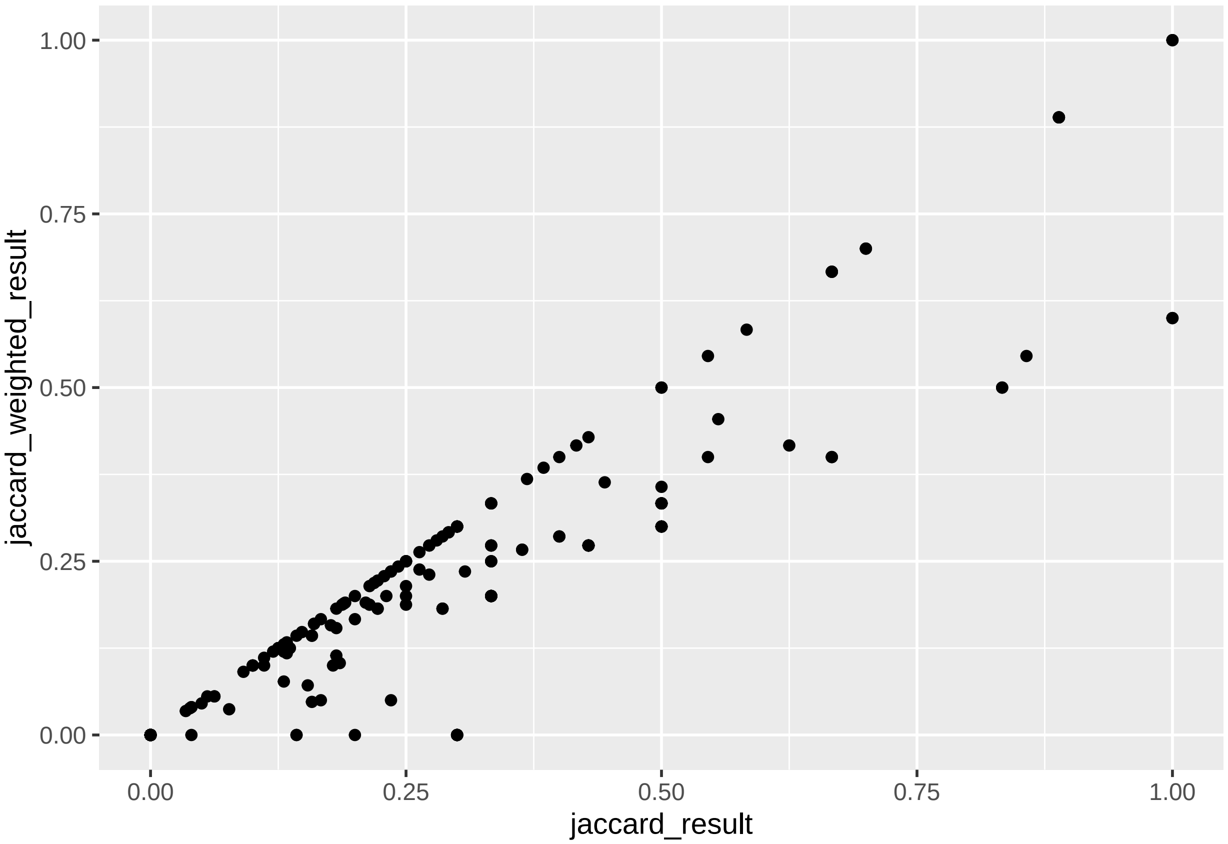
Marcin Stopka, Franciszek Sobota, Grzegorz Dolanowski

2026-01-10

Graph Metrics

We can see that the two graph metrics are very similar on the following plot.

```
ggplot(data, aes(x=jaccard_result, y=jaccard_weighted_result)) + geom_point()
```



they are even exactly the same most of the time

```
mean(data$jaccard_result == data$jaccard_weighted_result)
```

```
## [1] 0.6125
```

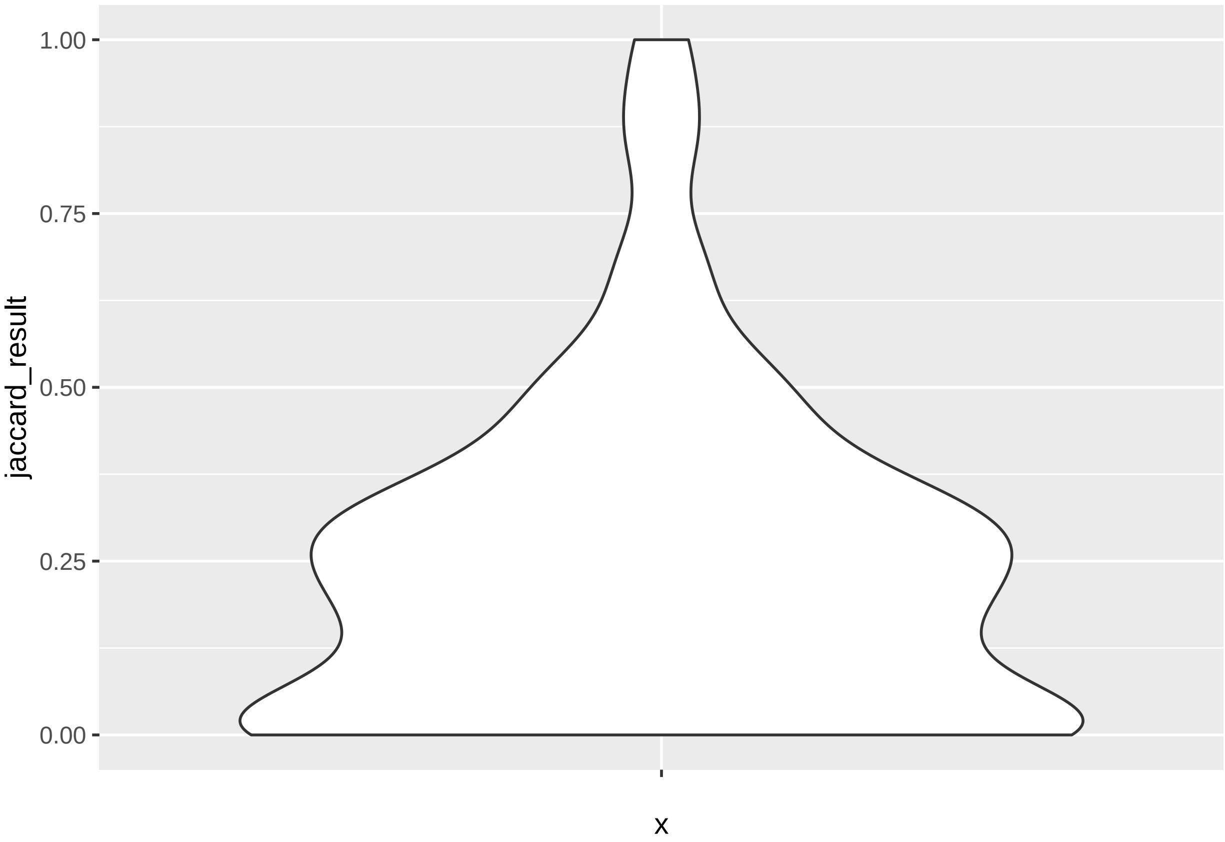
Achieved results

The achieved results aren't very good, the median is low.

```
summary(data$jaccard_result)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.0000  0.2000  0.2370  0.3333  1.0000
```

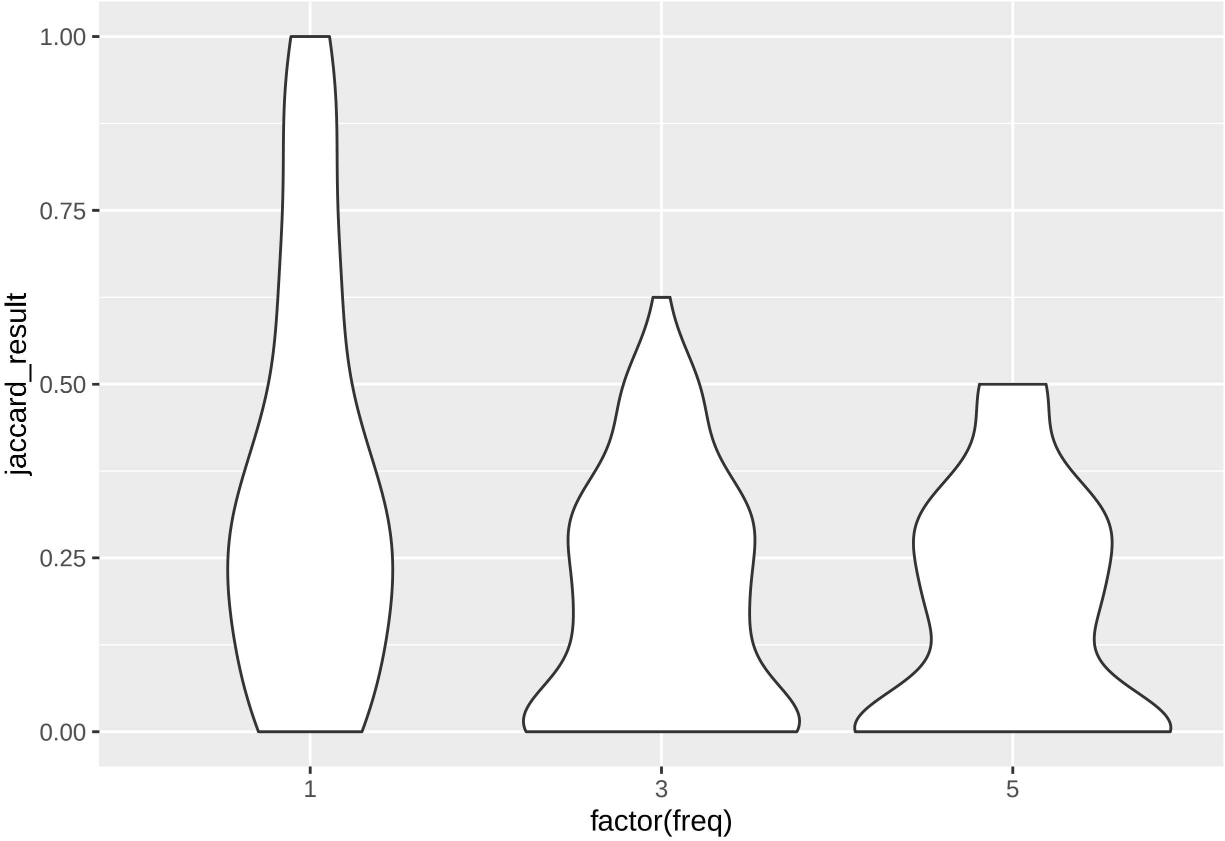
```
ggplot(data, aes(x="", y=jaccard_result)) +
  geom_violin()
```



Effect of sampling frequency

We can see on the plots that if time between sampling increases, the results get worse.

```
ggplot(data, aes(x=factor(freq), y=jaccard_result)) +
  geom_violin()
```



We can test if the correlation is negative and we get

```
tst <- cor.test(data$freq, data$jaccard_result, method = "spearman")
```

```
## Warning in cor.test.default(data$freq, data$jaccard_result, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
tst
```

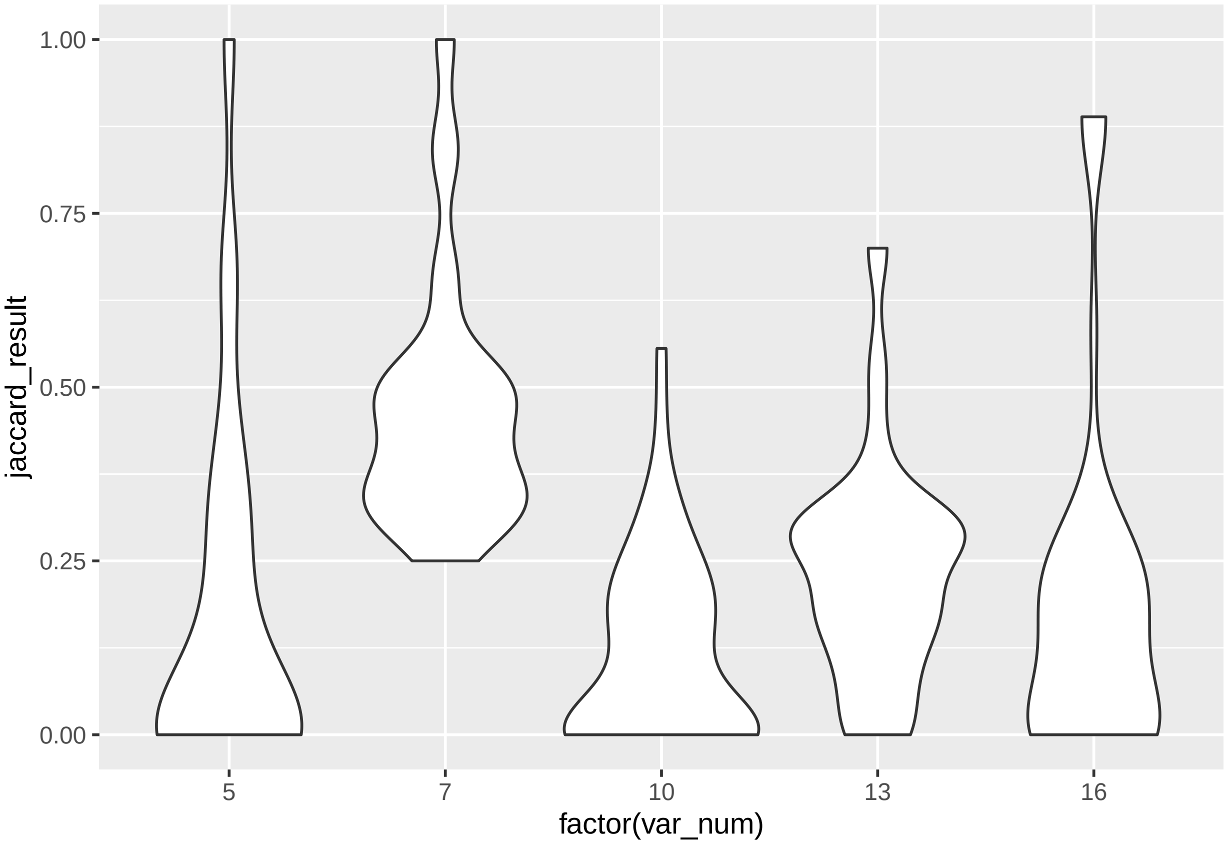
```
##
##  Spearman's rank correlation rho
##
## data:  data$freq and data$jaccard_result
## S = 3005774, p-value = 1.516e-06
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.304612
```

with low p-value of $1.5155853 \times 10^{-6‘}$

Effect of number of variables

For some reason the results are best when using 7 variables.

```
ggplot(data, aes(x=factor(var_num), y=jaccard_result)) +
  geom_violin()
```



but the correlation is test does not give anything conclusive

```
cor.test(data$var_num, data$jaccard_result, method = "spearman")
```

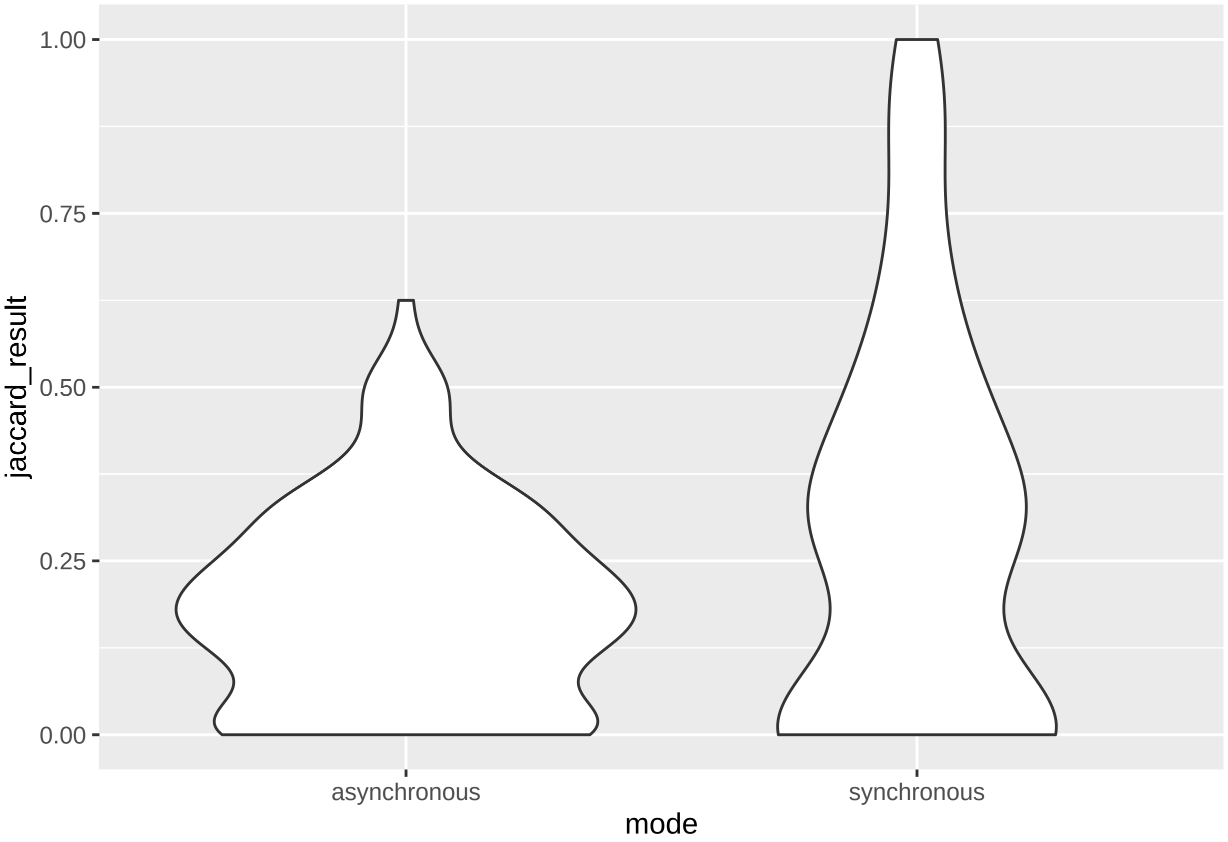
```
## Warning in cor.test.default(data$var_num, data$jaccard_result, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$var_num and data$jaccard_result
## S = 2551798, p-value = 0.09639
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.1075705
```

Effect of mode

Intuitively the asynchronous should give better results because it can give more data and synchronous can get stuck in a loop. But the plot shows that the opposite is true and synchronous has better scores.

```
ggplot(data, aes(x=mode, y=jaccard_result)) +
  geom_violin()
```



We can see also that lots of scores are zeros. We can test if the mode is independent of the score being zero

```
data_with_zero_jaccard <- data %>% mutate(jaccard_zero = (jaccard_result==0))
contingency_table <- table(data_with_zero_jaccard$mode, data_with_zero_jaccard$jaccard_zero)
chisq.test(contingency_table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table
## X-squared = 6.5904, df = 1, p-value = 0.01025
```

```
contingency_table
```

```
##
##                FALSE TRUE
##   asynchronous    95   25
##   synchronous     76   44
```

But we can actually see that synchronous gathering actually increases the risk of having zero score.