2_Respondents_Profiling

Generally, this is the variation of the software classification, though, in the center of it respondents and instead of MDS and K-mean Latent Profile Analysis is used. Firstly, it is convenient wrapper, secondly, it is specifically suited for the survey kind of data. What I haven't used yet, though, might be useful is fixed/freed/random means across various profiles - they might represent our expectations about the profiles characteristic, such as mean age or typical occupation.

What else could it be used for: in the next report I will try to make LPA on Occupation+Education+Age and use profiles probability to map software

Showing job shifts in connection with associated software is not possible due to NAs in Q21_1_open (previous job): there is `sum(df_clean$Q22 == "")` NAs, while the total complete sample is `nrow(df_clean)`.

TODO: what I need to do is to consider public/private sector
Q19 - do you work in a public/private sector

Description of items for LPA

```
# Q1 - language

# Q3_1 - Desktop
# Q3_2 - Laptop
# Q3_3 - Tablet
# Q3_4 - Mobile phone/Smartphone
#

# Q8_v2_1 - Built-in default settings
# Q8_v2_2 - Plugins/add-ons/extensions
# Q8_v2_3 - Script to extend Software
# Q8_v2_4 - I reprogram

# Traditional Surveys
# Q9 - Q10 - Q11 - Q12 - Q13 - Q14 - Q15 - Q16 - Q16

# Q17_R - employment status
```

Latent Profile Analysis is used in social science and educational research. It suits the needs to aggregate reprogrammability items, though, need to check how it will deal with Occupations and SES.

Number of profiles to choose - analytically - less BIC is better.

n_profiles	Constrained variance, fixed covariance	Freed variance, fixed covariance	Constrained variance, constrained covariance	Freed variance, freed covariance
1	-61164.26	-59565.16	-55664.78	-55664.78
2	-59305.92	NA	-55313.01	NA
3	-58064.10	NA	-54427.77	NA
4	-56775.90	NA	-54683.45	NA
5	-56403.10	NA	-54388.93	NA
6	-56230.51	NA	-54439.99	NA

Number of profiles to choose - visually.

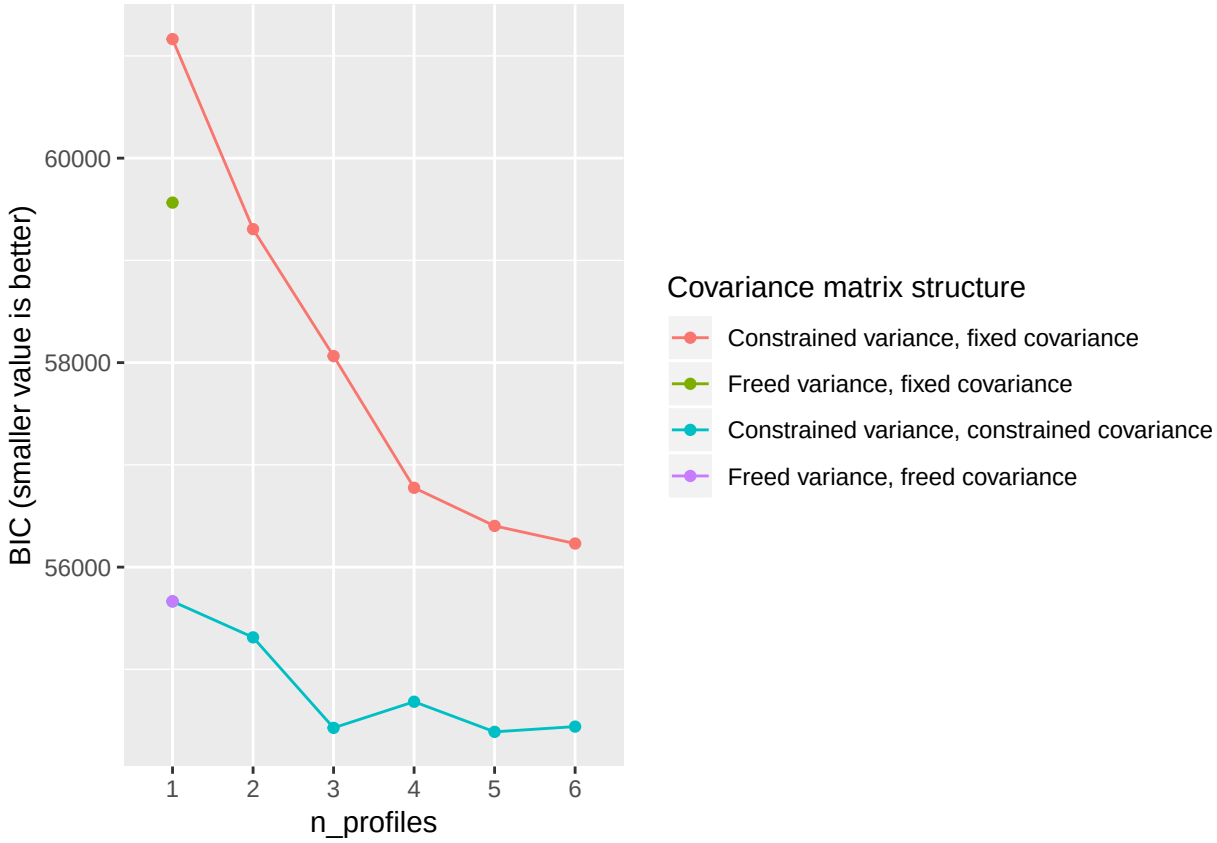
Freed variance - variance might vary across the groups. Constrained variance means that profiles are chosen according to the assumption that variance should be closely the same in all of the profiles. Fixed variance indicates the strength of this assumption, so, variance is less then in case of Constrained Variance.

In it's turn, covariance indicate the degree of variation for interconnection of variables across profiles. Basic example - income should be equally connected with education across all profiles in case of fixed covariance.

Two models are plotted:

- Constrained variance, fixed covariance
- Constrained variance, constrained covariance (+)

Based on the BIC, we should choose EEE = Constrained variance, constrained covariance model = model 2.



Choosing the model with 5 profiles.

Plotting

Here just a brief description I made in Overleaf some time ago.

Using Latent Profile Analysis 5 consistent groups of respondents were extracted. 1st profile is the largest one (588), respondents in other profiles are distributed more or less uniformaly.

Figure 1 demonstrates profile differences of responses about extension or reprogrammability of software. Firstly, it should be noted that respondents across all of the profiles tend to change built-in settings of the software they are using. Notably, the the respondents from *1 profile* have one of the lowest dispersion on this item. This might mean that while respondents from 1 profile are not using plugins, scripting or reprogram their software, their generally tend to fit the programms to their own need exploiting the highest possible level of it (settings).

The respondents stressed out as being out of *2nd profile* are tend to use scripting for software extension in much larger extent than respondents of other profiles. Generally, the pattern is that based on all 4 response items regarding the extensibility of software they tend to have higher medians. ICT employees are more highly prominent in *profile 2*.

Typical respondent of *Profile 3* generally tweak software much less, than respondents from other profiles, though, it is the only case where users are equally not changing settings nor using plugins. Based on the pearson residuals, health

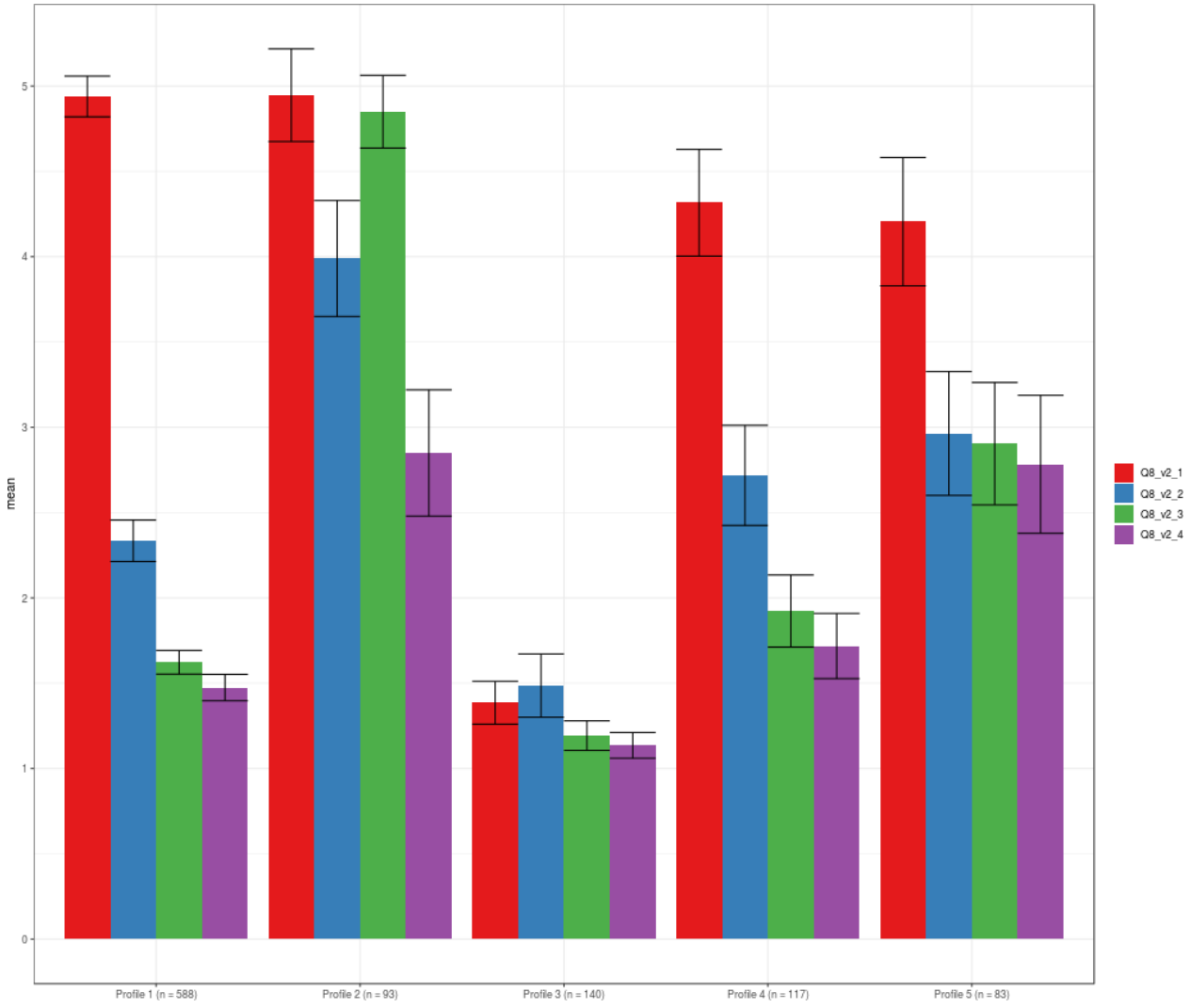


Figure 1: Reprogrammability Items Across Profiles: Will make a bit more sense later in conjunction with Occupations

workers are prominently highlighted to be in the 3rd profile (the percentage of health workers tend to be highest, while absolute number is still twice as low as in 1rd profile). While for users of some occupations changing settings might not lead damaging consequences, in health service one should have a strong understanding of how the software works in order to obtain predictable outcomes. (Soft. systems are black boxes, probably, for most users. Tweaking them in health service might be dangerous. Or it might be that they are principally hardly accessible for changing defaults. Or health guys just do not posses enough knowledge. Anyway, it is just a one point for discussion about different standards for software extensibility across industries).

Figure 2 indicates that chief executives even though being the one of the less smallest group of respondents are not typical for the 1 profile, though, prominent for the *4th* and less for *5th*, both in relative and absolute numbers. Those two profiles tend to change defaults less than users from *profile 1 and profile 2*. One interesting point is that a respondent from profile 5 is, to some extent, balanced in using different levels of changing the software, which might be seen from the equal medians on items about plugin, scripts and reprogramming use.

Figuring out which Occupations prominent for each of the profile

```
## null device
##           1
```

Statistical Testing Block

Here I am constructing simple variables like number of:

- unique software used = n_unique_soft
- unique software use per device

- software used out of the top-10 popular items = n_unique_soft_non_pop
- software use out of the top-10 popular per device = n_unique_soft_per_q_non_pop[1].

I do not completely understand which theoretical constructs they do represent yet, but found them useful in testing differences between profiles later.

Testing difference between reprogrammibility capabilities across Profiles

In the current case when the LPA was accomplished considering difference in the Q8_v2_ items it doesn't make much sense to compare them statistically across profiles, but still might give some information about profiles/

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Q8_v2_1 by profile
## Kruskal-Wallis chi-squared = 340.15, df = 4, p-value < 2.2e-16
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Q8_v2_2 by profile
## Kruskal-Wallis chi-squared = 156.18, df = 4, p-value < 2.2e-16
```

[1] I used camelCase versions of the variables later - latex does not get well with "__" symbols

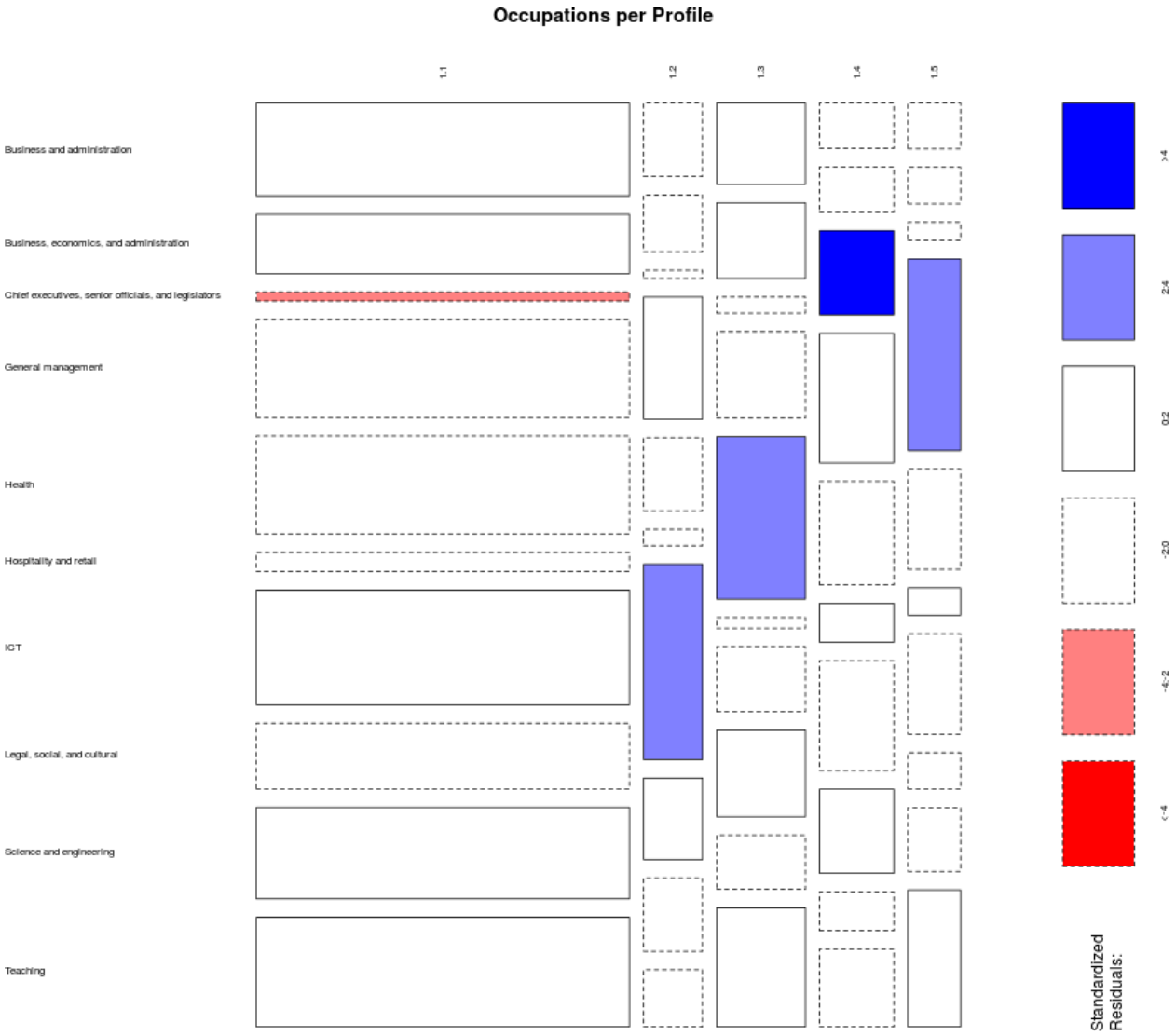


Figure 2: MosaicPlot - blue color indicates that occupation is prominent more than expected, red - less than expected

Table 1: m3$Q8_v2_1 and m3$profile

	1	2	3	4
2	1.000	NA	NA	NA
3	0.000	0.000	NA	NA
4	0.005	0.115	0	NA
5	0.005	0.059	0	1

Table 2: m3$Q8_v2_2 and m3$profile

	1	2	3	4
2	0.000	NA	NA	NA
3	0.000	0.000	NA	NA
4	0.104	0.000	0	NA
5	0.007	0.002	0	1

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Q8_v2_3 by profile
## Kruskal-Wallis chi-squared = 362.54, df = 4, p-value < 2.2e-16
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Q8_v2_4 by profile
## Kruskal-Wallis chi-squared = 148.82, df = 4, p-value < 2.2e-16
```

Post hoc testing

While we know, that there is a difference across profiles, we don't know which profiles have differences (all of them?).

Plot the differences across profiles

TODO: put significance levels

Turning Back

Let's return to the variables we've constructed earlier - since they were not included into the LPA it makes much more sense to include them in profile difference hypotheses testing.

We use bonferroni p-value adjustment since we simultaneously testing several hypotheses. P-values smaller than .05 indicate difference in groups given variable. I have not provided interpretation yet, since LPA and grouping might and probably will be changed.

Table 3: m3$Q8_v2_3 and m3$profile

	1	2	3	4
2	0.00	NA	NA	NA
3	0.00	0	NA	NA
4	0.22	0	0	NA
5	0.00	0	0	0

Table 4: m3$Q8_v2_4 and m3$profile

	1	2	3	4
2	0.000	NA	NA	NA
3	0.000	0	NA	NA
4	0.035	0	0	NA
5	0.000	1	0	0.001

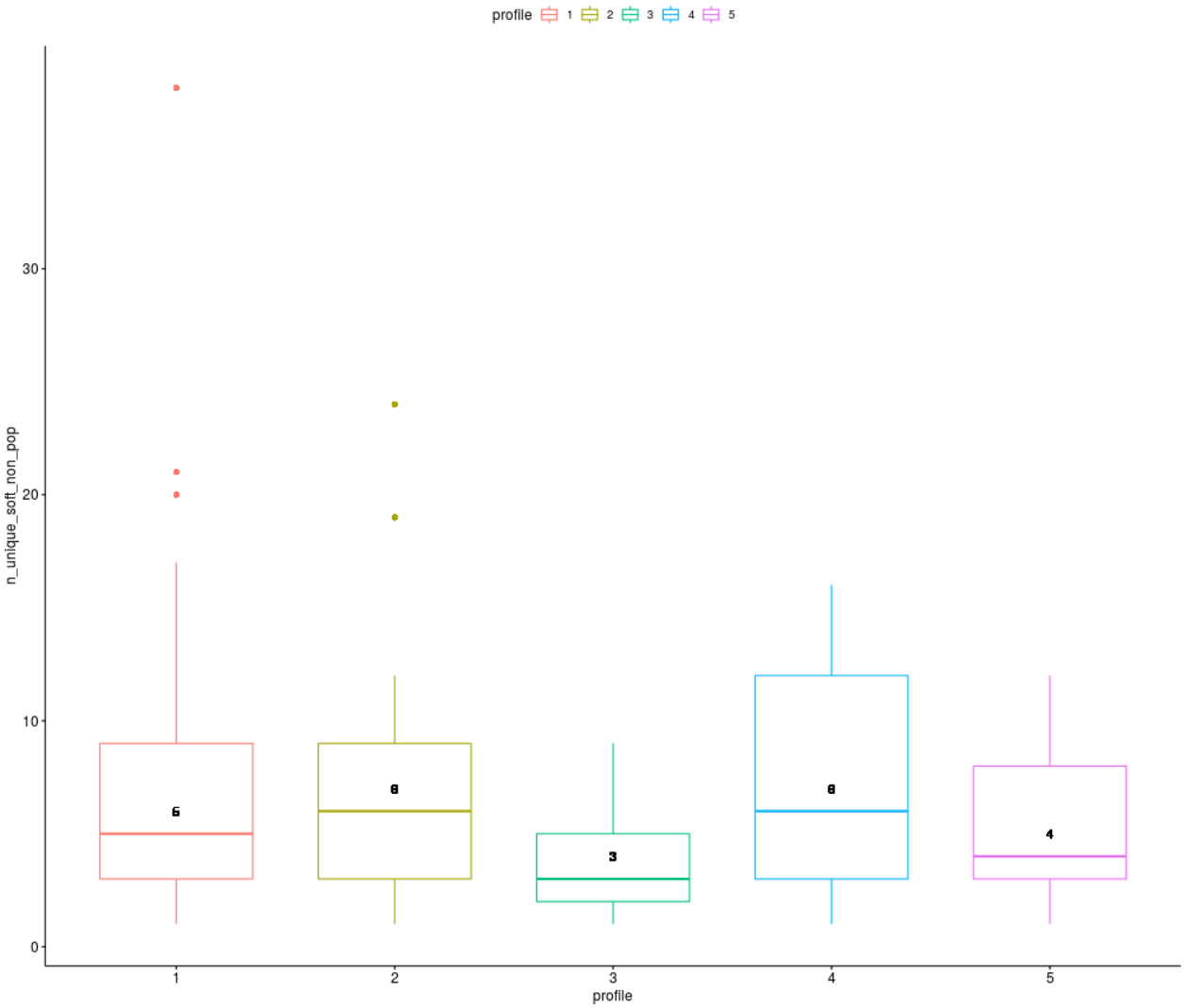


Figure 3: Difference in Soft Used out Of Top-10 Across Profiles

Table 5: tmp$nUniqueSoftPerqNonPop and tmp$profile

	1	2	3	4
2	1.000	NA	NA	NA
3	0.000	0.000	NA	NA
4	0.327	0.632	0.000	NA
5	0.236	0.308	0.305	1

Table 6: tmp$nUniqueSoftNonPopandtmp$profile

	1	2	3	4
2	1.000	NA	NA	NA
3	0.000	0.000	NA	NA
4	1.000	1.000	0	NA
5	0.079	0.062	0	0.157

Table 7: tmp1$nUniqueSoftandtmp$profile

	1	2	3	4
2	1.000	NA	NA	NA
3	1.000	1.000	NA	NA
4	0.427	0.053	0.153	NA
5	0.007	0.002	0.005	0.672