умомроссиюнепонять аршиномобщимнеизмерить

Gender lag in a Russian village

Ustja River Basin Corpus Project Team

Generalities on gender in sociodialectology

- Western social dialectology:

- Women more innovative than men
- Women more standard than men
- Men more vernacular than women

- Tons of interpretations…

- Upward mobility aspirations among women
- Women's more using language as social resource
- Men more integrated into local social networks

Now against Russian background

- **Common belief**, contra Western SL: women more dialectal than men. Bubrikh 1914
- “Наблюдения над *ц* позднее позволили мне убедиться, что различие в характере ц в зависимости от его положения свойственно лишь мужскому произношению, в произношении же женщин и детей, которое надо считать неиспорченным пустошенским произношением, ц одинаково твердо во всех положениях.”
- **Interpretation**: men more mobile ⇒ more exposed to standard (especially military service and war)

- **In this talk**: checking Bubrikh stipulation against Russian dialect corpus data - Ustja River Basin Corpus

Ustja River Basin Corpus

○ A corpus of a dialect (Vologda group) in northern Russia, over 750,000 tokens, 80 hours

○ Geographically focused (several neighboring villages in the south of Arkhangelskaja oblast)

○ A team project involving BA students and their supervisors

Nina Dobrushina, Ruprecht von Waldenfels, Michael Daniel, Alexandra Ter-Avanesova, Maria Ovsjannikova, Sergej Saj; and Tatiana Gavrilova, Ekaterina Gerasimenko, Darya Ignatenko, Polina Kazakova, Aleksandra Khadzhiyskaya, Elena Kilchitskaya, Darya Kosheleva, Ivan Levin, Ekaterina Makhlina, Anna Rusanova, Anastasia Safonova, Marija Sapozhnikova, Pavel Taratynov, Maria Tsfasman, Samira Verhees, Lesha Vinyar, Vasilisa Zhigulskaya, Elena Zinchenko

Ustja River Basin Corpus

○ some 50 speakers, whose representation is very different, from one thousand to dozens of thousands of tokens

○ aged between 1922 and 1996, with a bias towards older speakers (median 1952, about one third of all tokens coming from people born before 1930)

Ustja River Basin Corpus

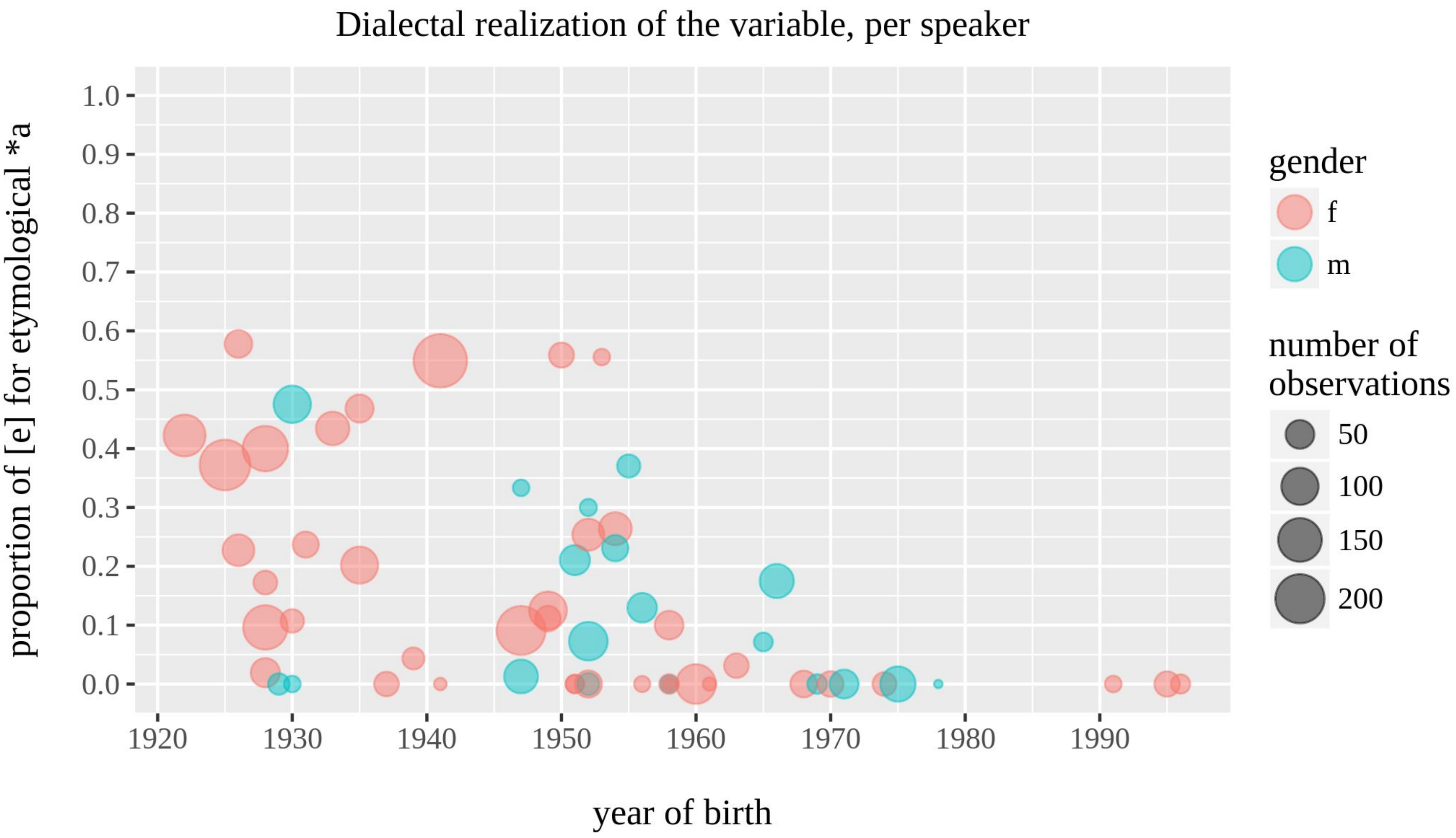
- change in progress: dialect loss / convergence with the standard; clear correlation with age

- 11 phonological and morphophonological variables, binary coded (dialect vs. standard), 722 to 2963 observations per variable

- observations binary coded by the members of the corpus team and partly cross-validated by other members

- example: the reflex of *a between palatalized consonants - dialect [e] or standard [a]

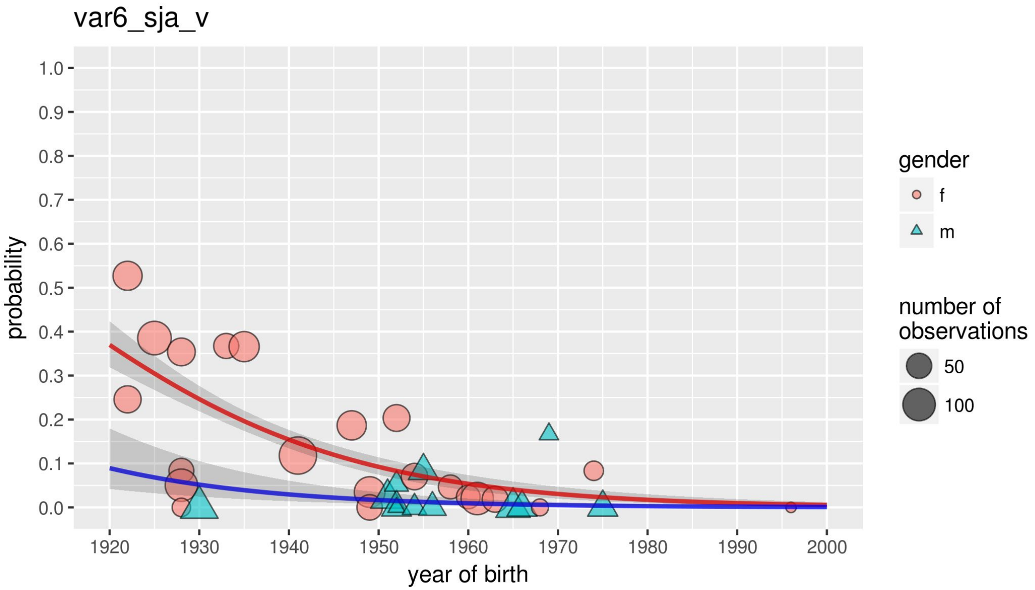
Ustja River Basin Corpus

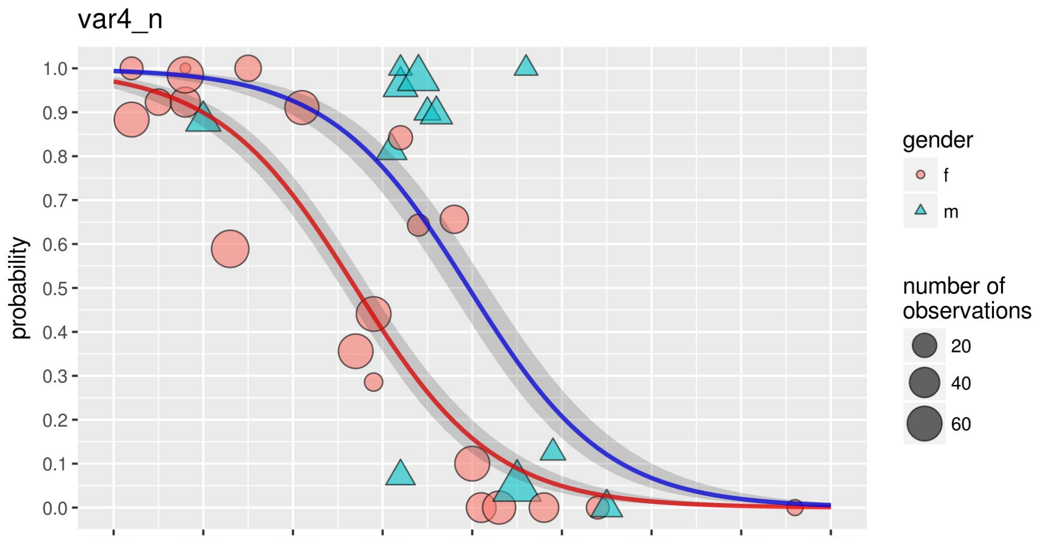


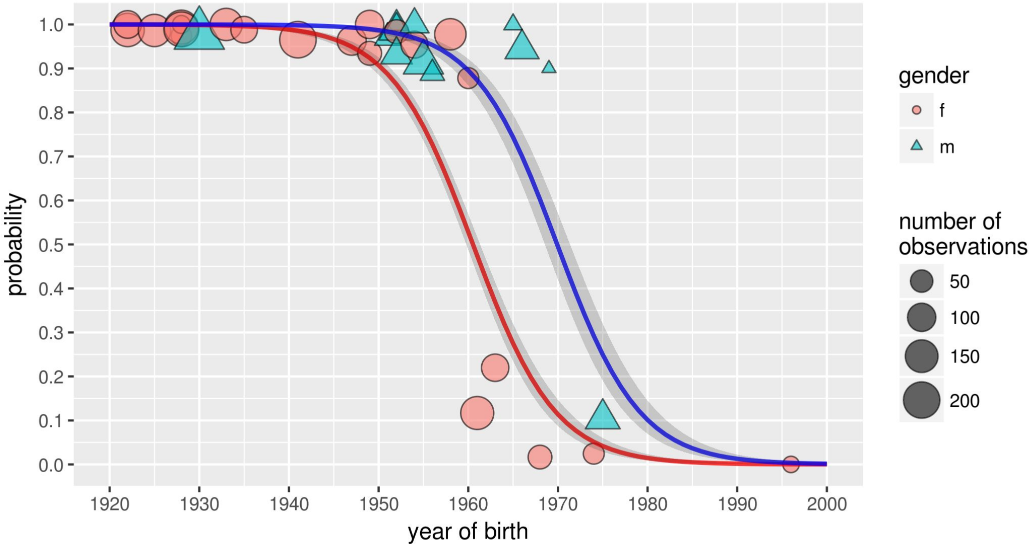
Does this chart shows that men are ahead or behind women?

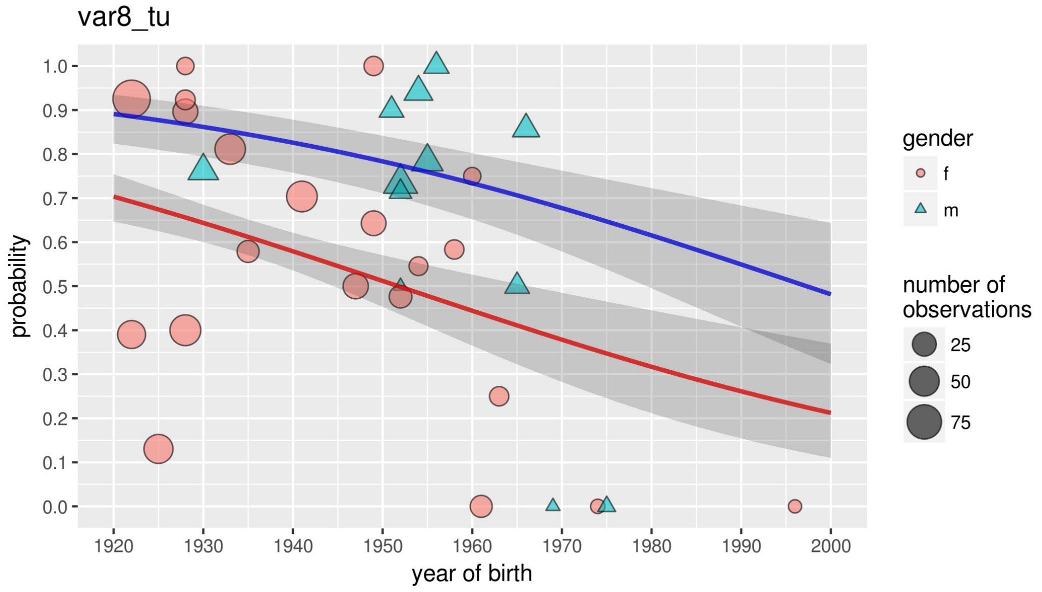
Logistic regression votes women

- Logistic regression shows that women are more innovative than men by many variables in our data

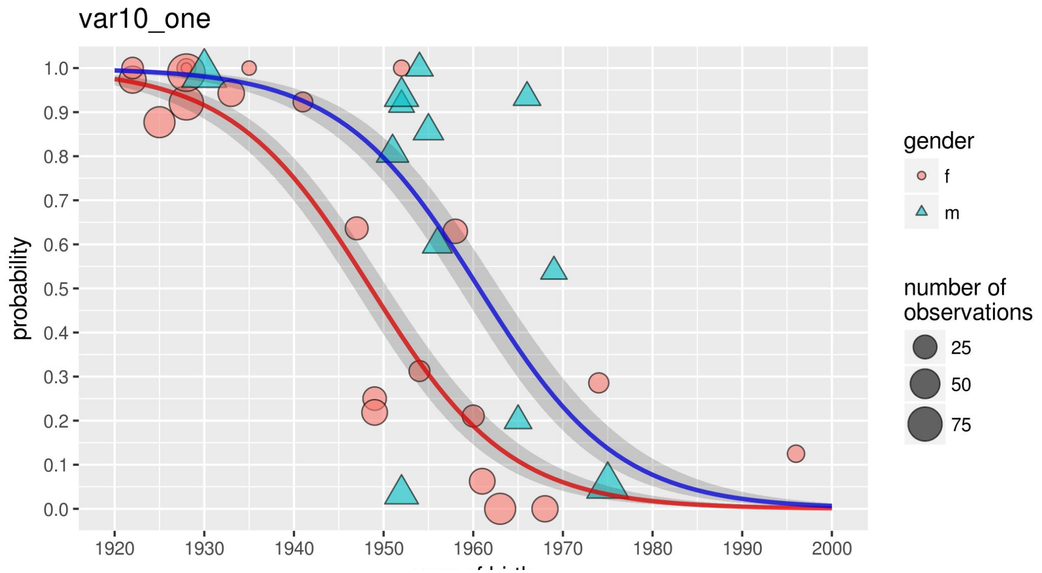

var6_sja_v

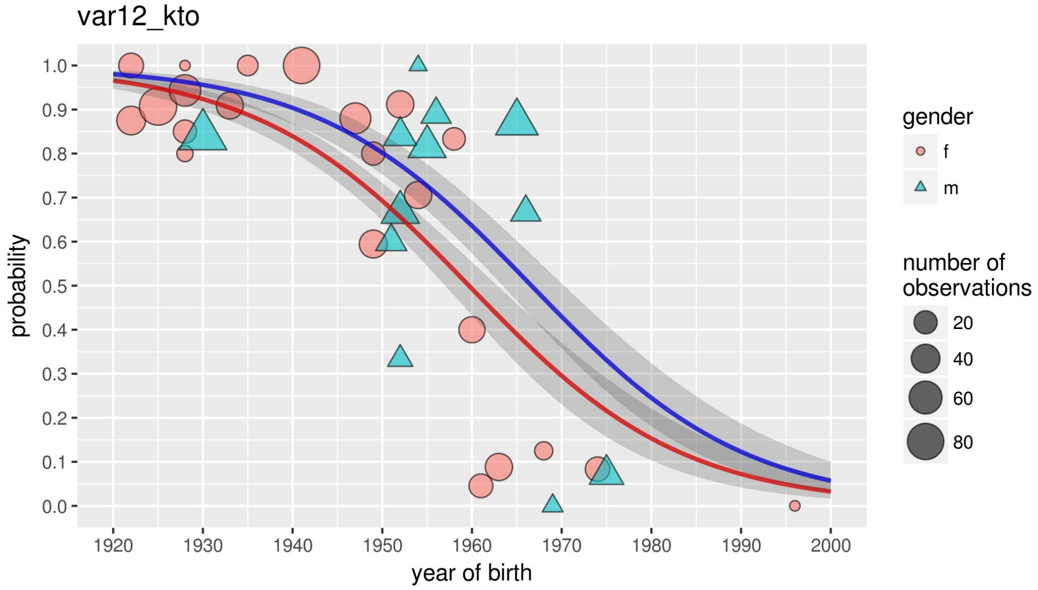

var4_n

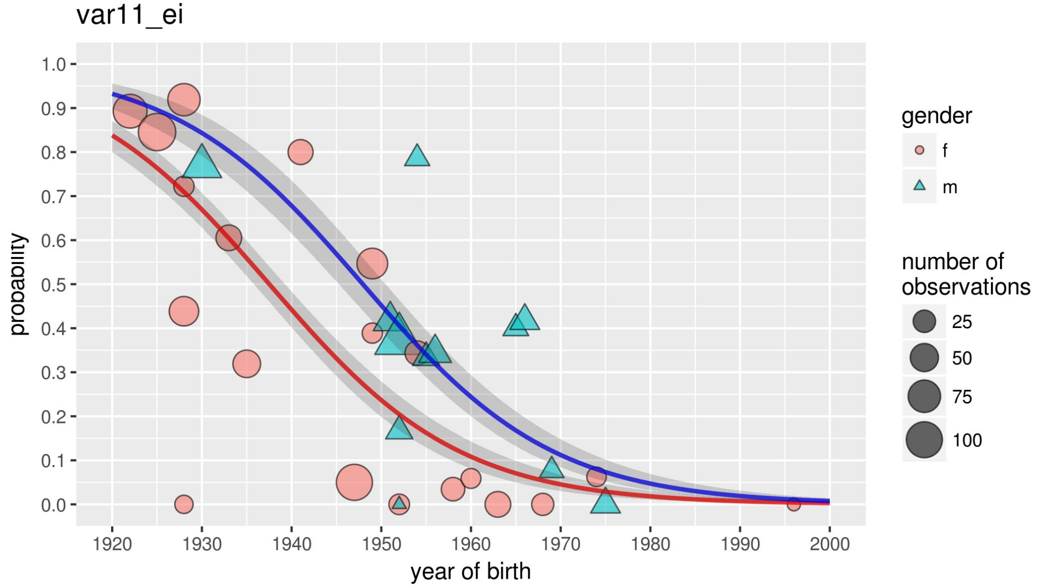



var8_tu

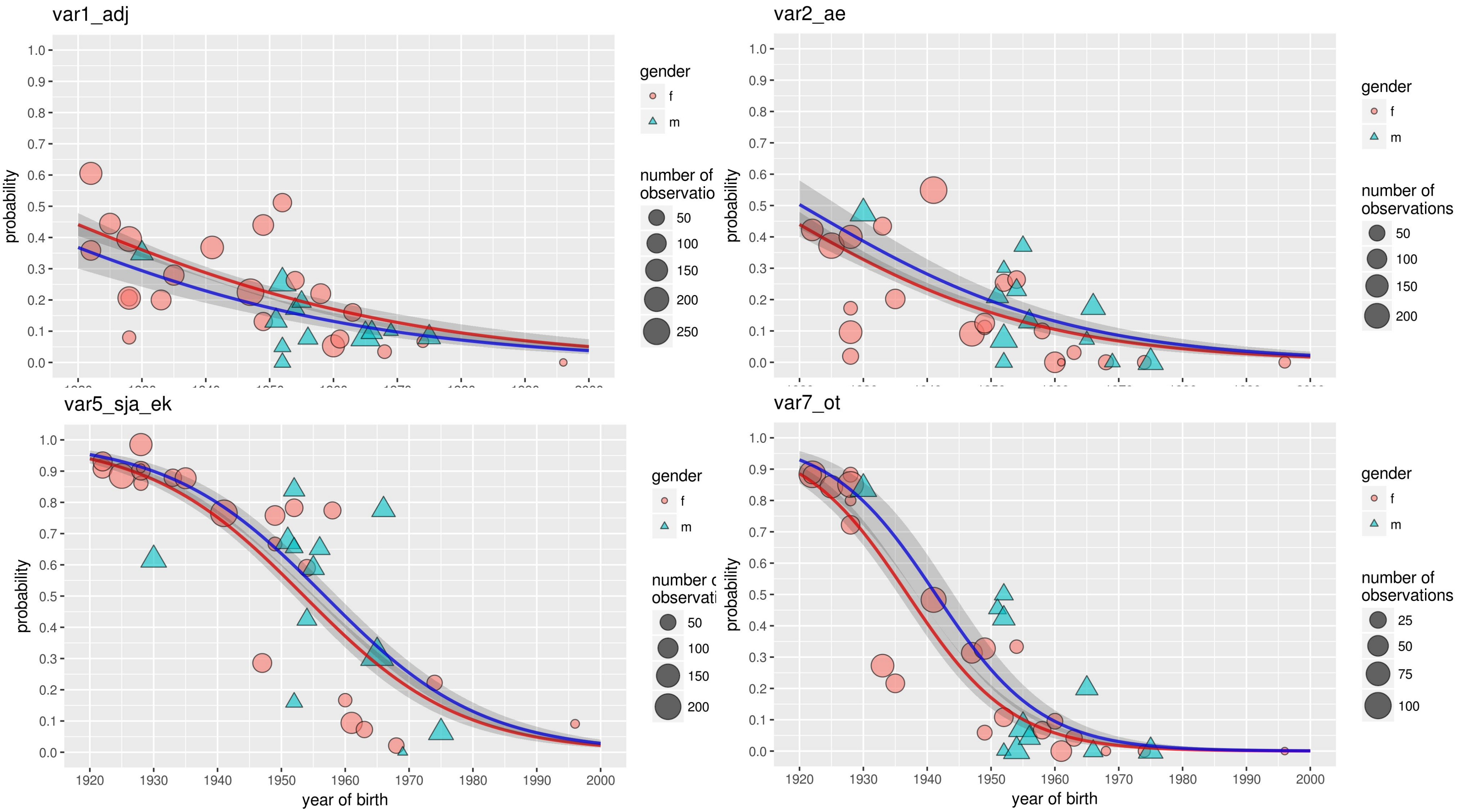
Logistic regression votes women







Logistic regression says nothing



Logistic regression models

- (Some technicalities: With additional predictor (gender), one of the logit coefficients is the same for both curves)

- Primarily: **Constant rate change assumption** (mutatis mutandis for logistic regression; symmetric re middle point 0,5)

Generalized additive regression model

- **Generalized additive models** are intended to overcome this by adding more (restricted) freedom to the curve. But they still contain assumptions (the set of curves) and at the same time make the data less easily interpretable (because of the multiple curves)
- I was looking for a method to try and look at the data with a minimum of assumptions.

Confidence interval

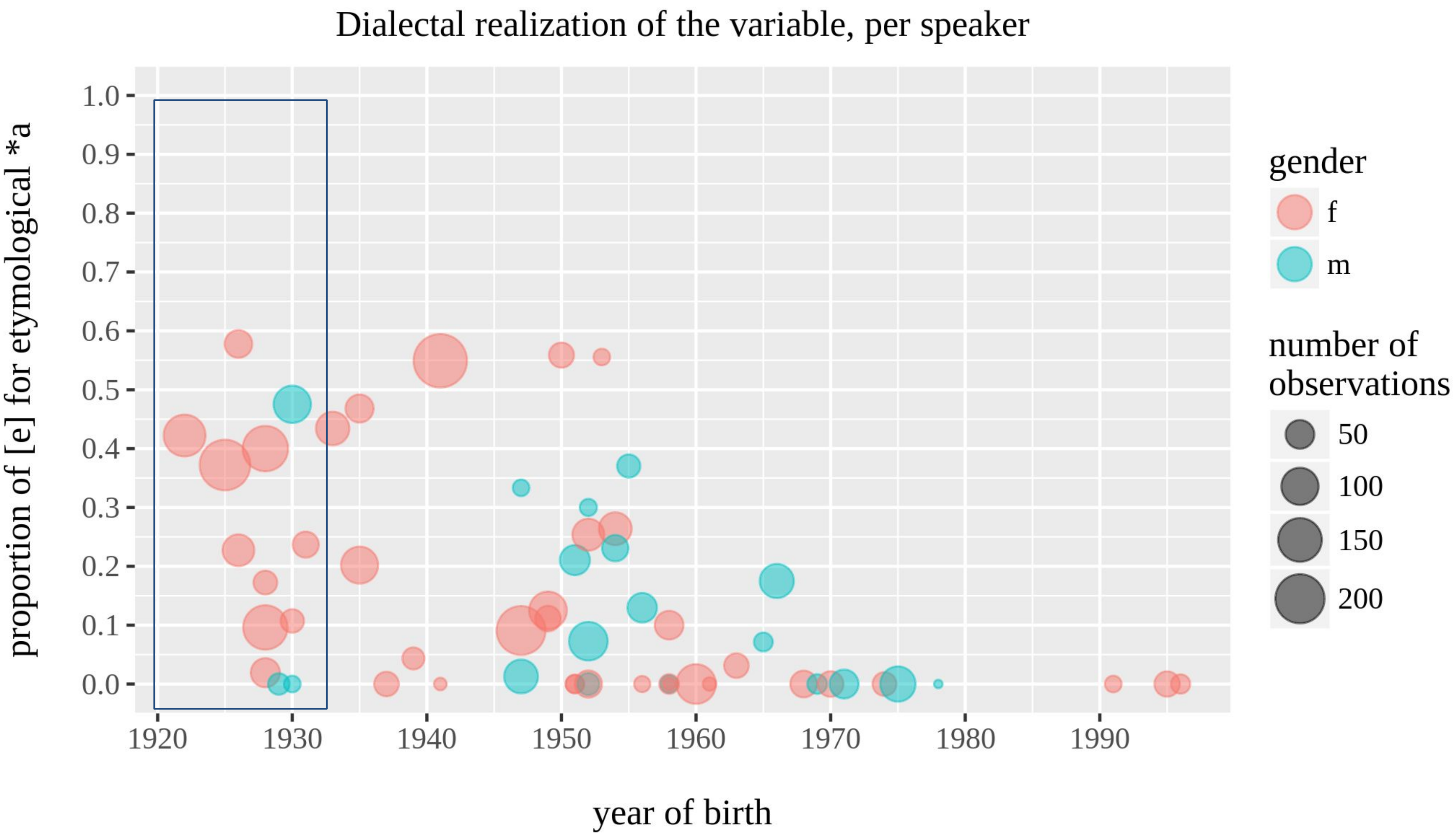
- Assume there is a fixed probability of conservative vs. innovative realization of a variable for a given speaker; this is a property of the speaker (how well does she preserve the variable)
- For a speaker, we have a number of conservative (CONS) vs. innovative (INS) observations
- Qualitatively: the higher is the number of observations, the closer CONS/(CONS+INS) is to this probability.

Confidence interval

- **CI**: Estimation of the expected range of the probability value based on the available observations.
- The size of the interval depends on the number of observations and, to a lesser degree, on the m/n value.

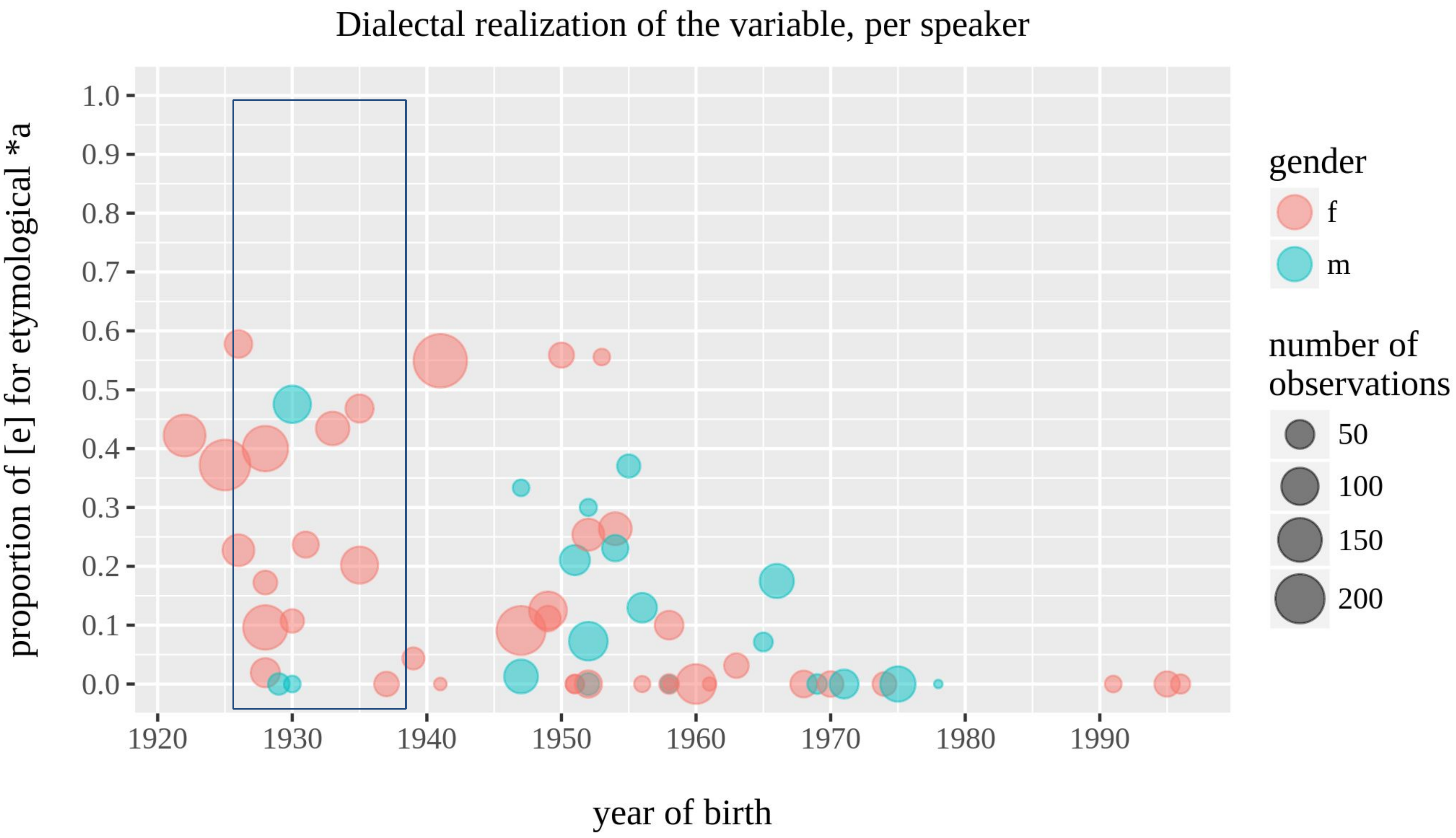
Qualitatively, the higher is the number of observations, the smaller is the confidence interval.

Confidence interval in a moving window



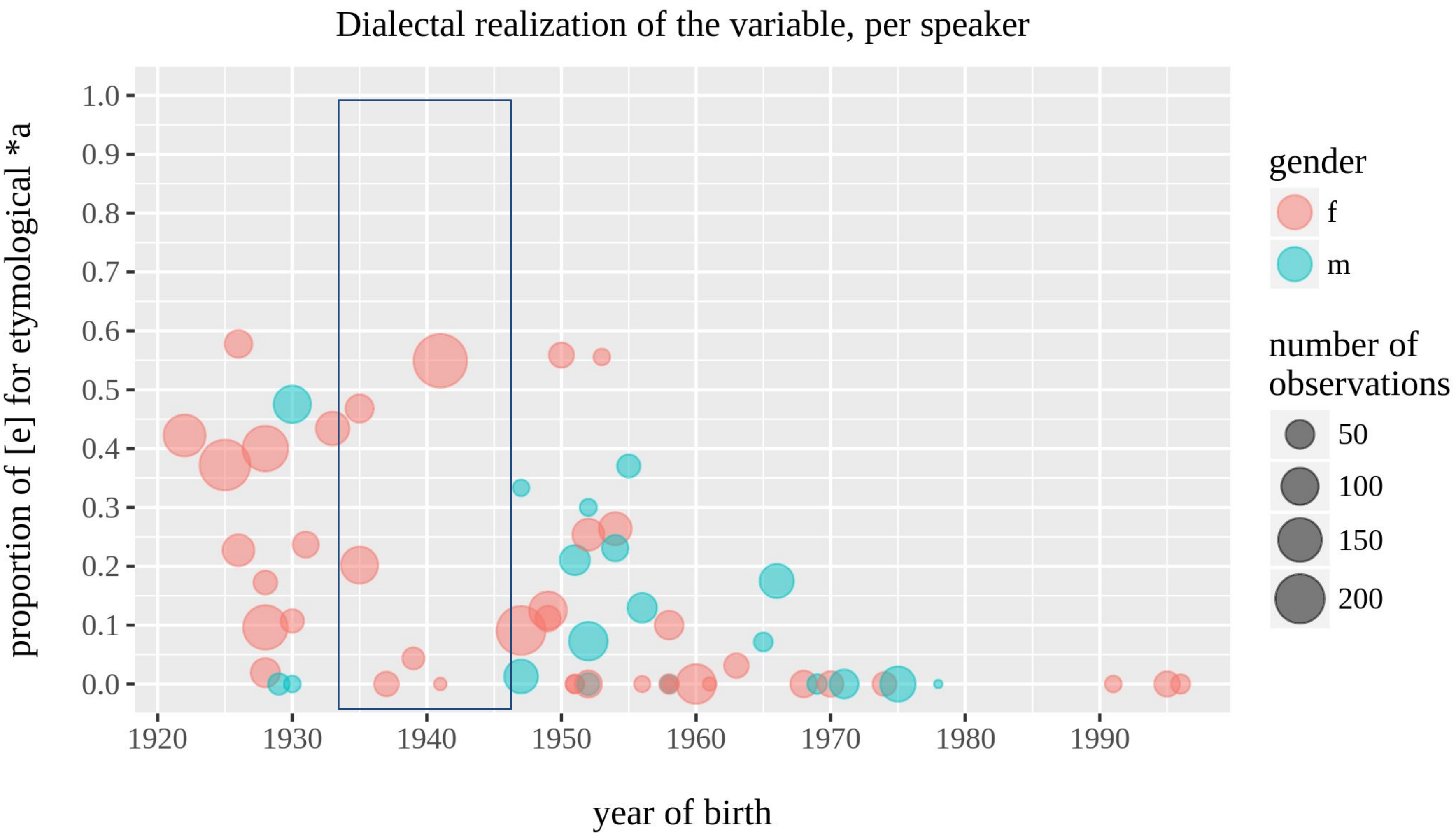
Moving window – a time interval which moves over the timeline and shows the change of a value calculated for this time interval.

Confidence interval in a moving window



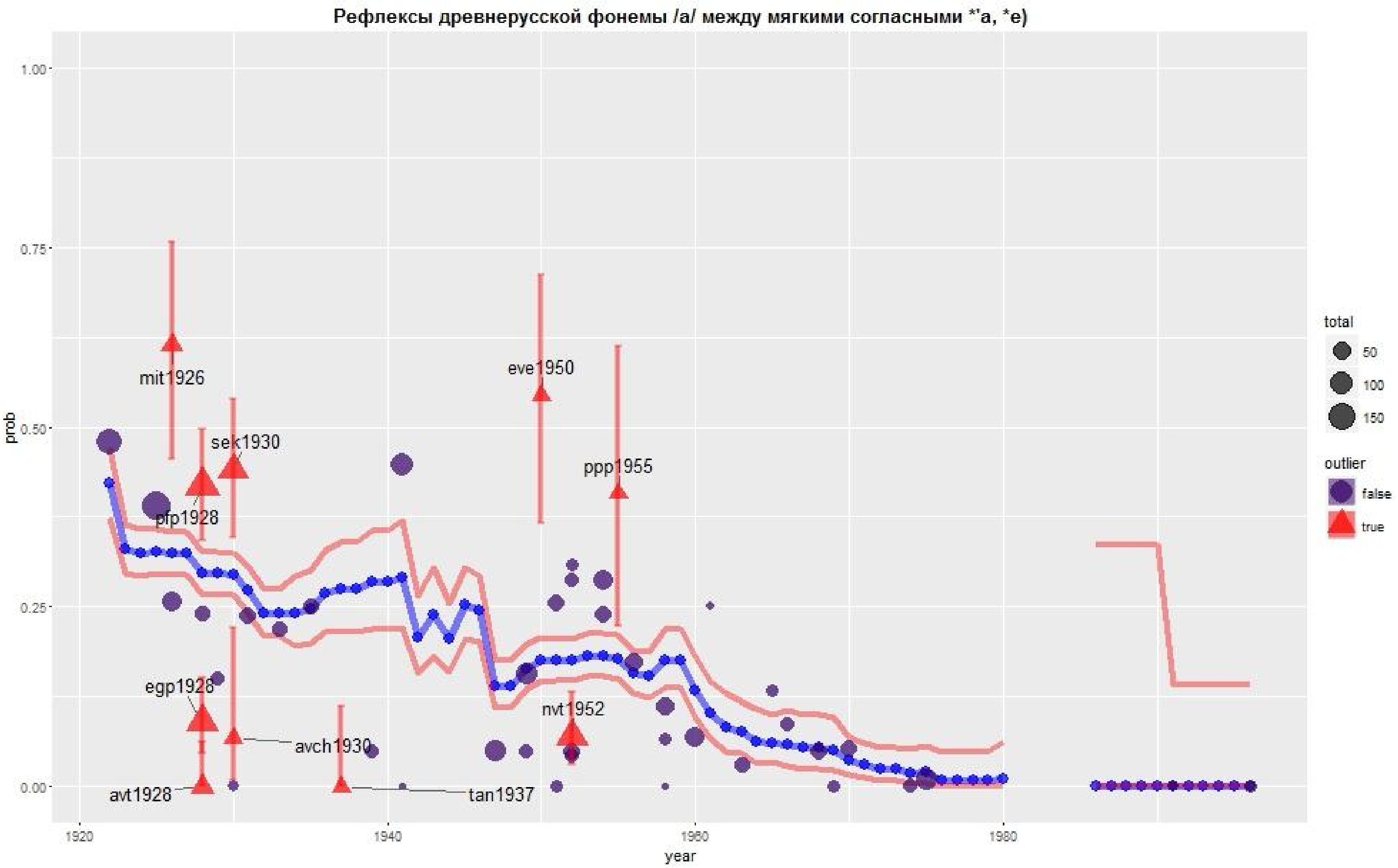
Add up all speakers in the window, build a confidence interval for the window.

Confidence interval in a moving window



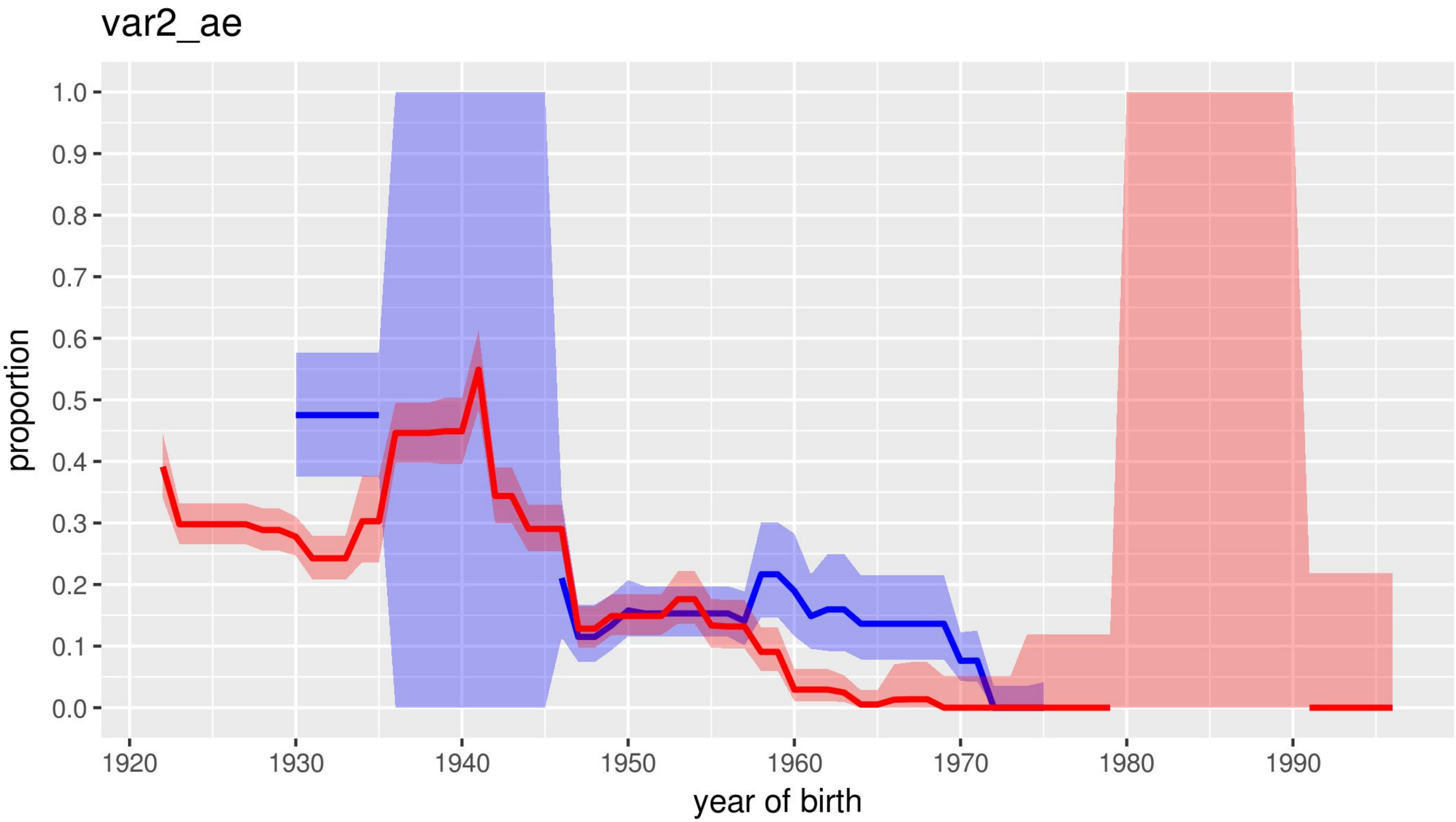
Move the window further and recalculate.

Outlier detection



Gender lag

- In this talk: separate women and men into two different moving CI charts, see the difference

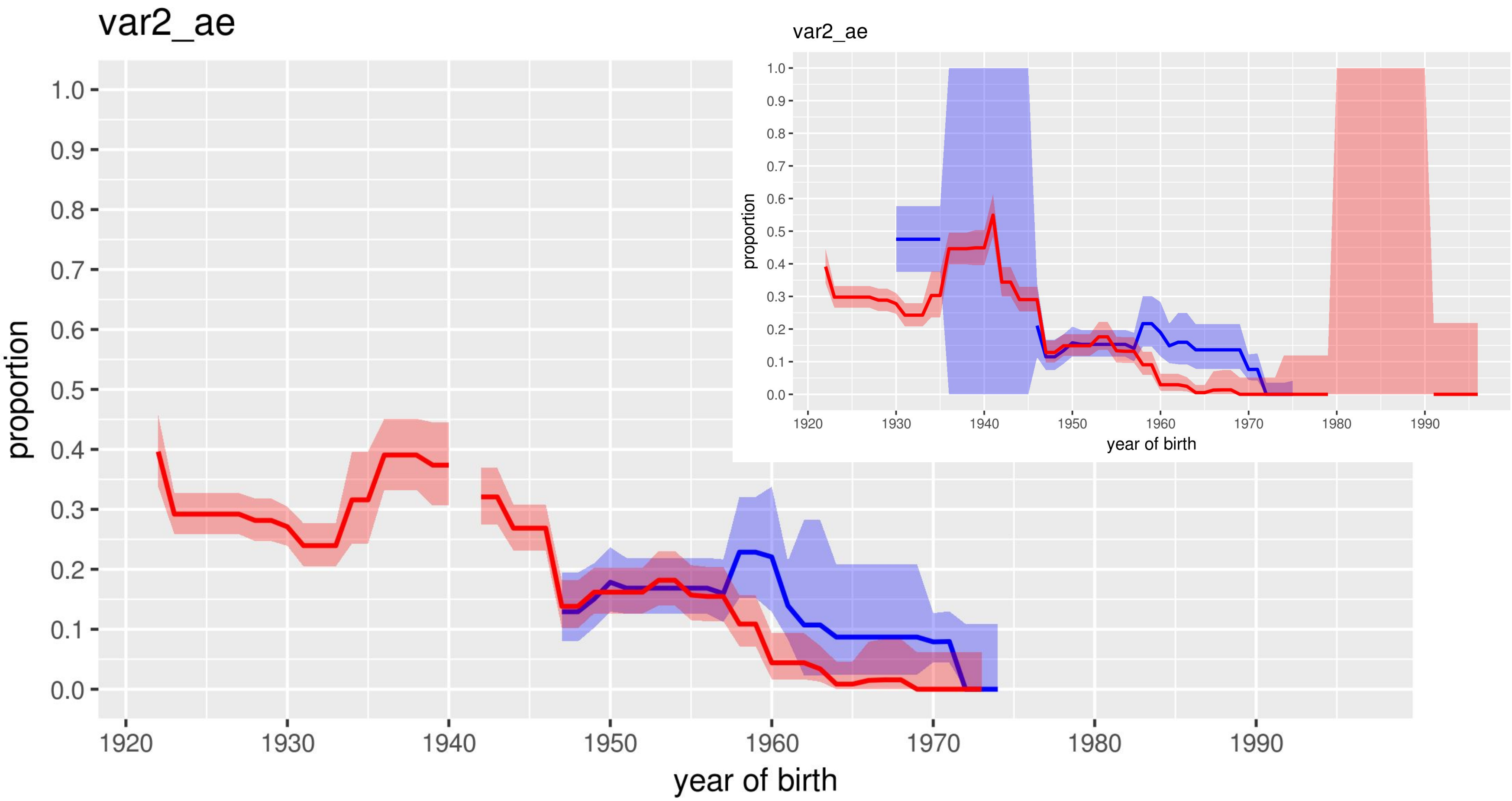


The Curse of the Pink Outlier

- In all windows she belongs to, a single but heavily represented outlier attracts the whole confidence interval to herself
- To compensate for this, reduce the n of observations to the second thickest outlier (or maybe to the third?) - outlier smoothing
- Also remove CI for windows where there is only one speaker

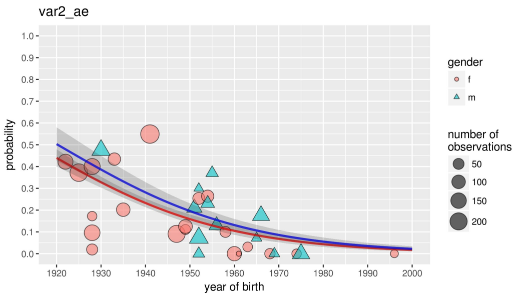
Moving window as change in time

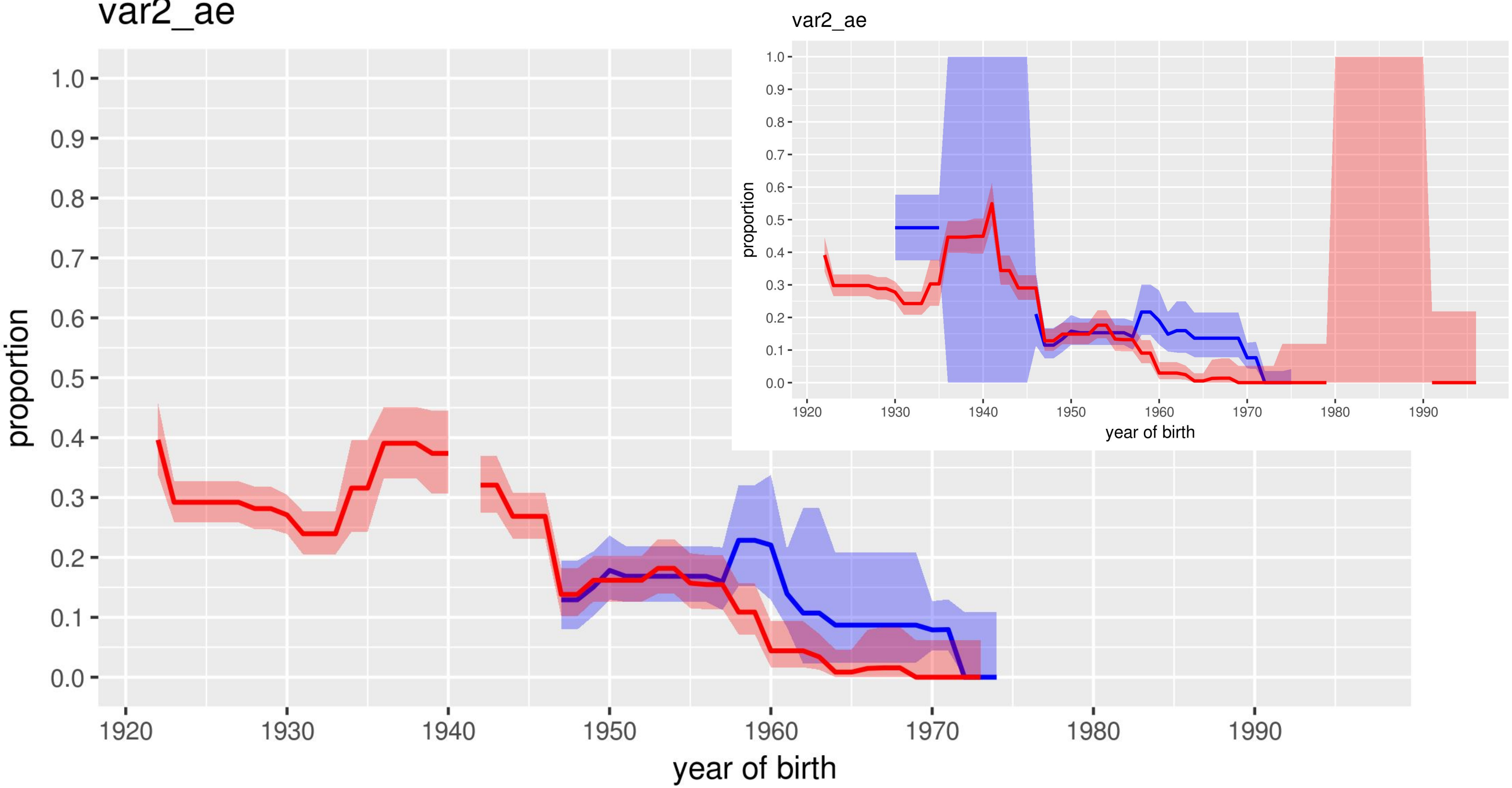
- After and before smoothing:



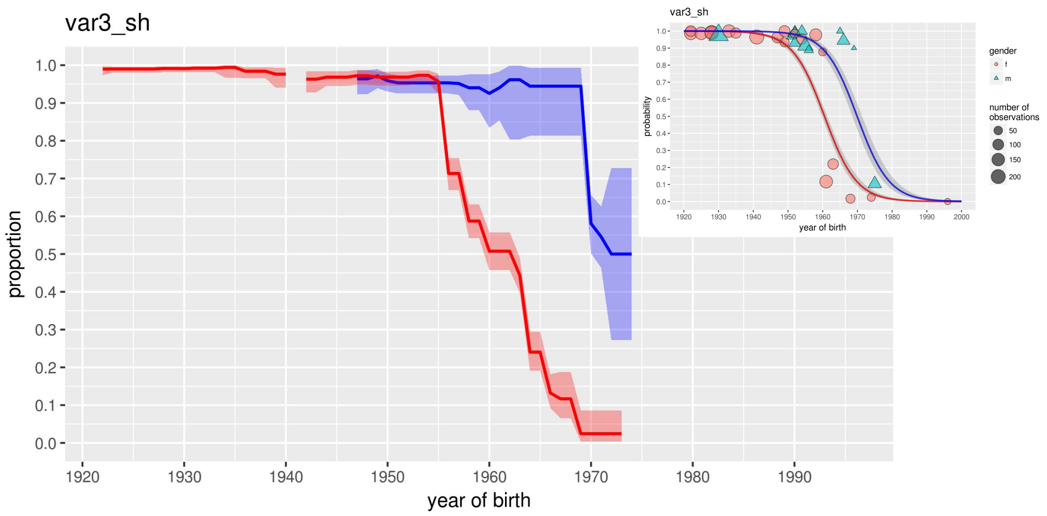
Moving window as change in time

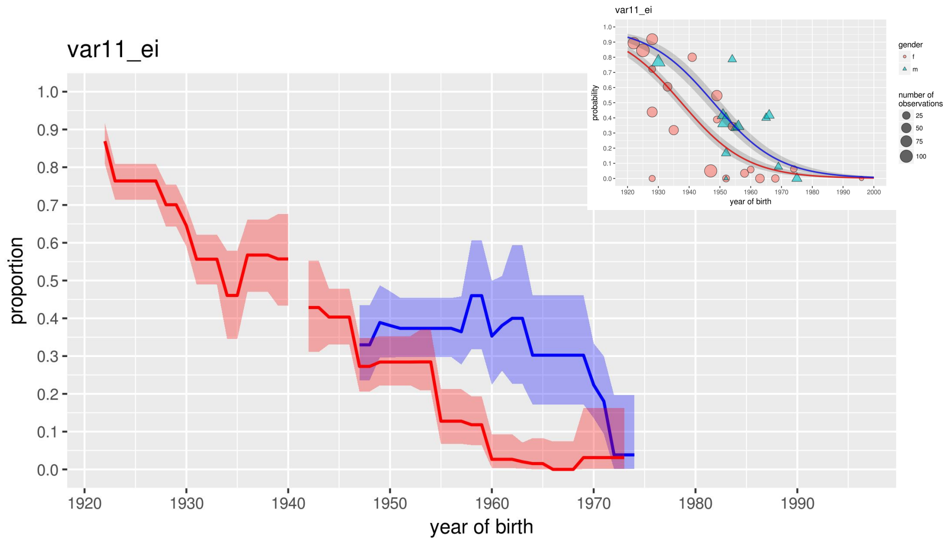
- After and before smoothing:

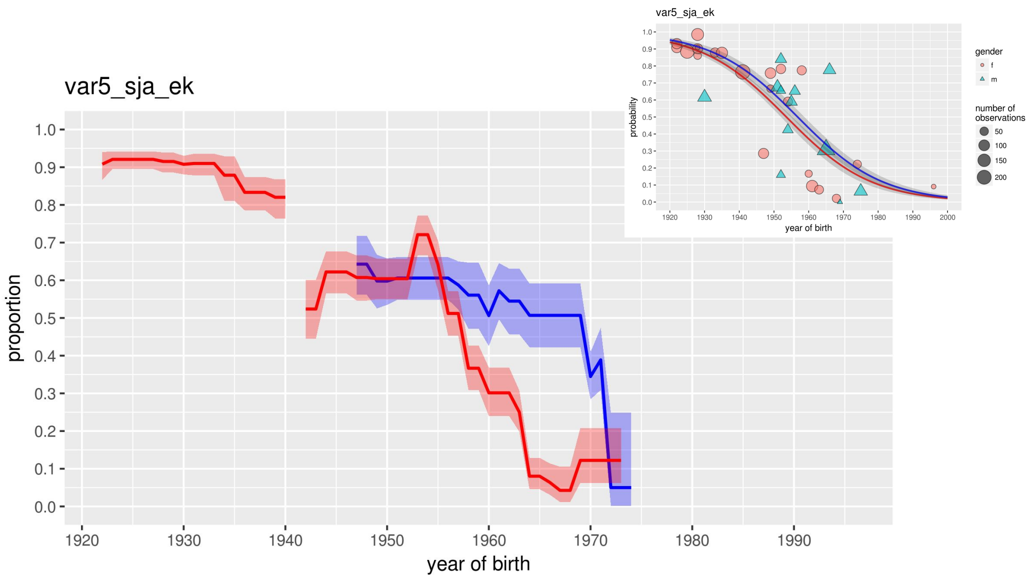


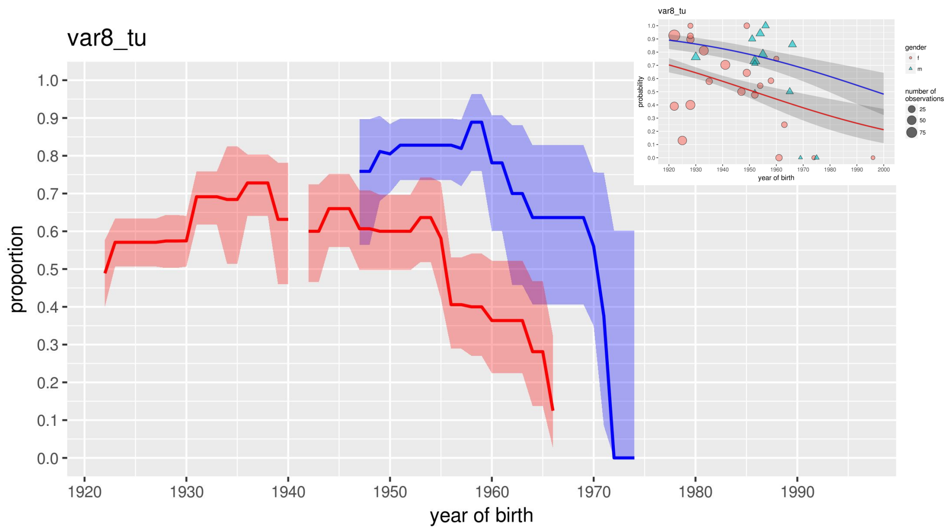


Gender lagged variables

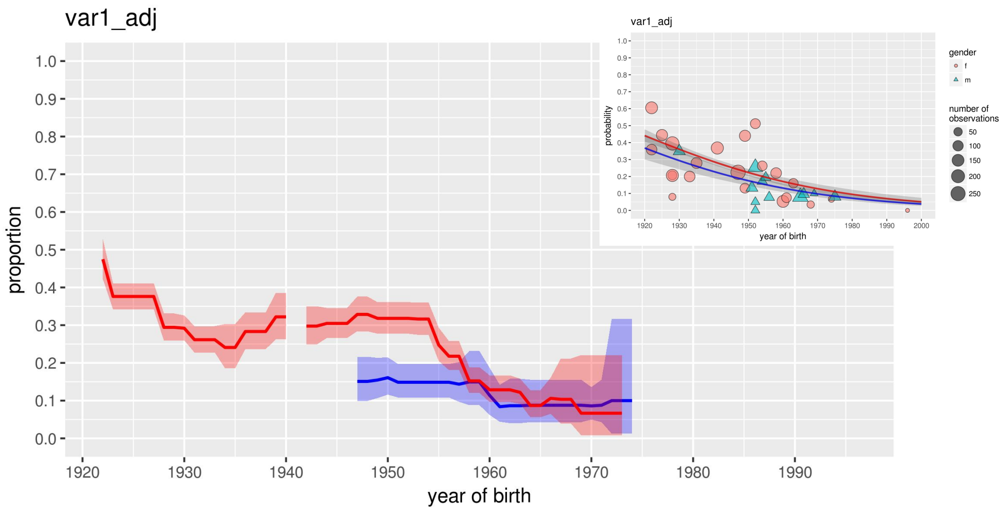


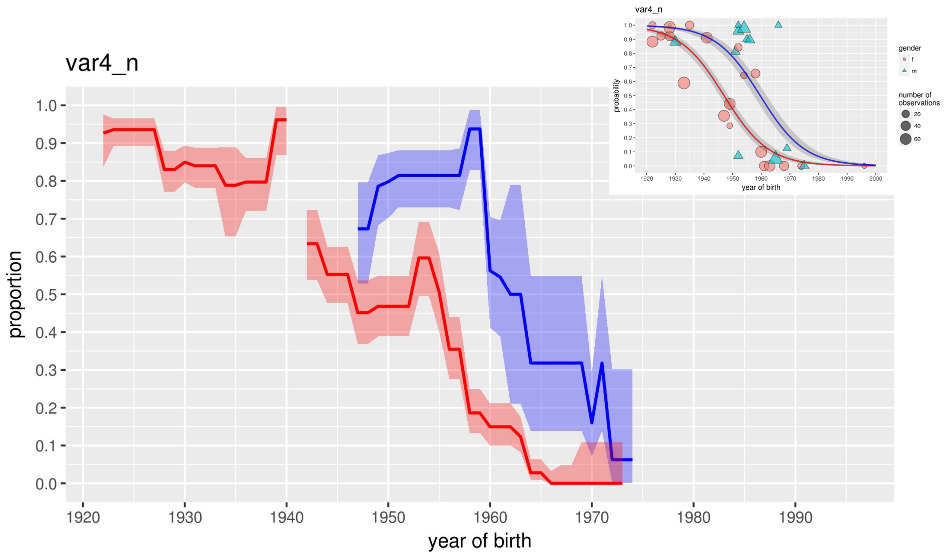


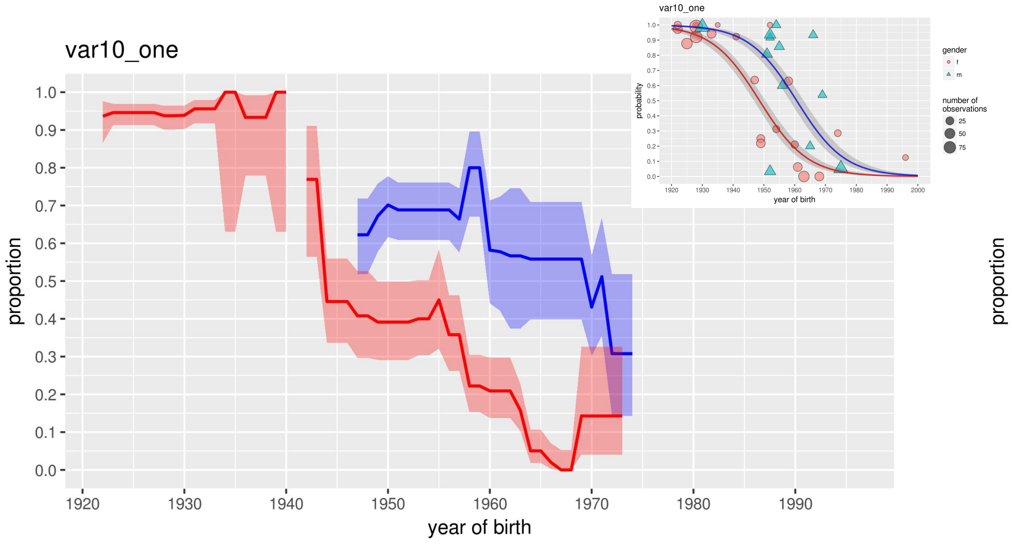


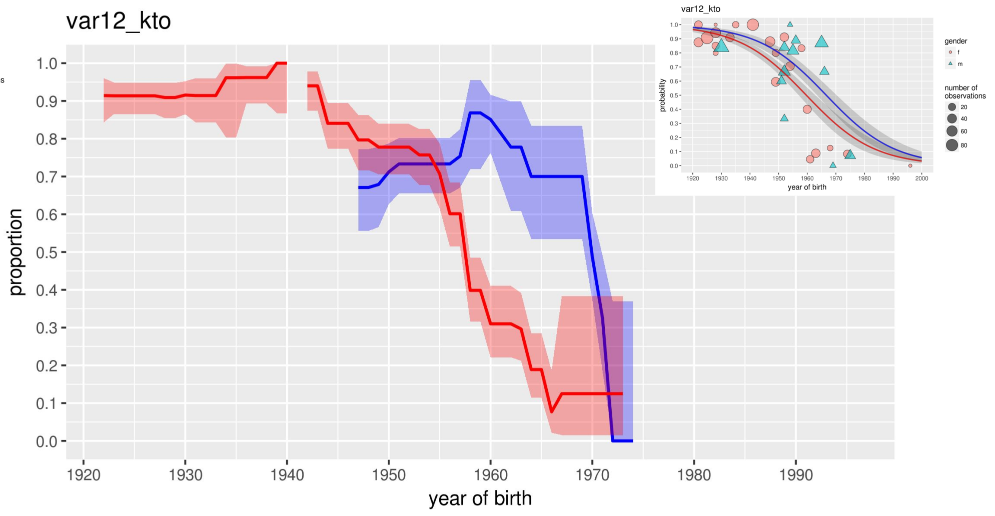


Gender lagged variables


var1_adj

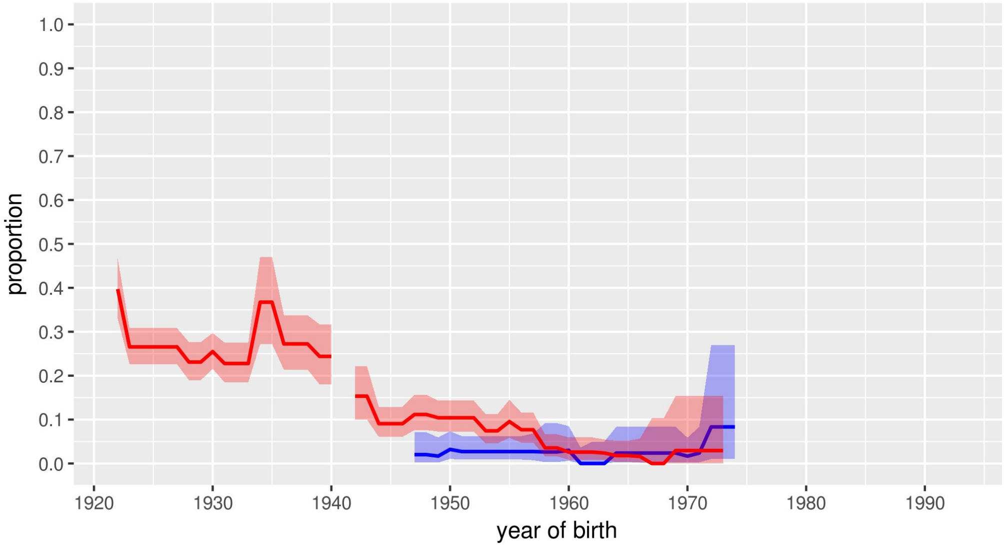

var4_n


var10_one

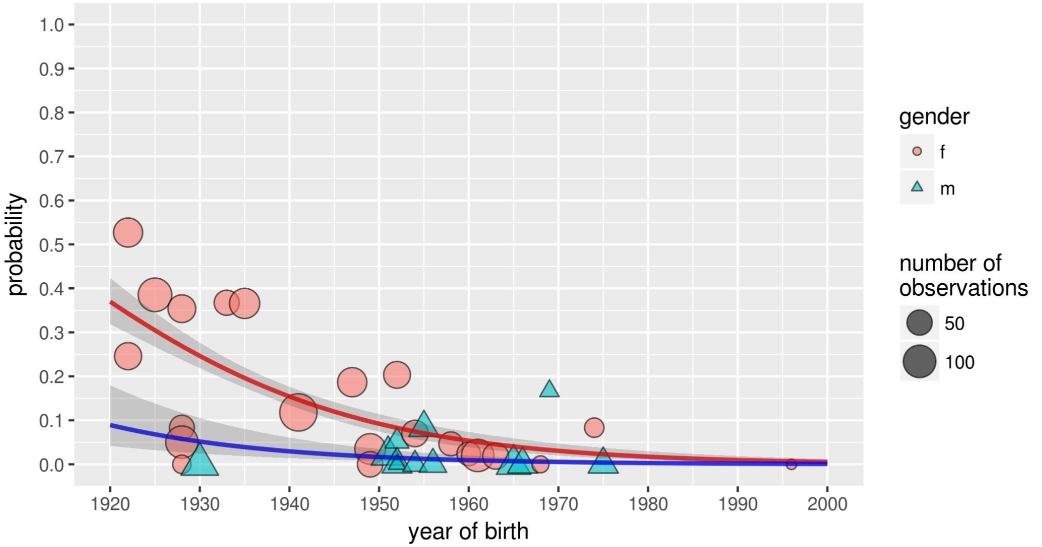

var12_kto

No lag

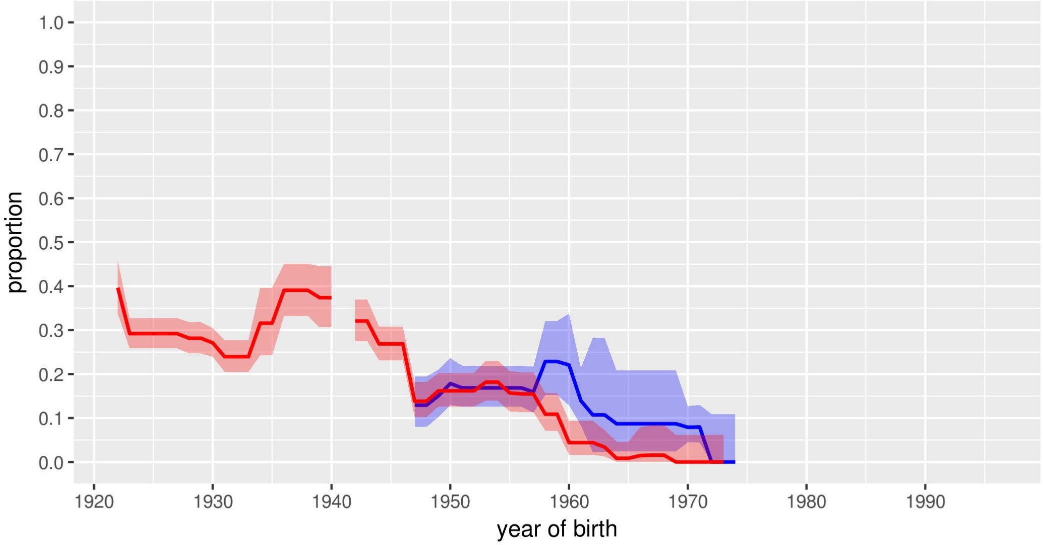
var6_sja_v



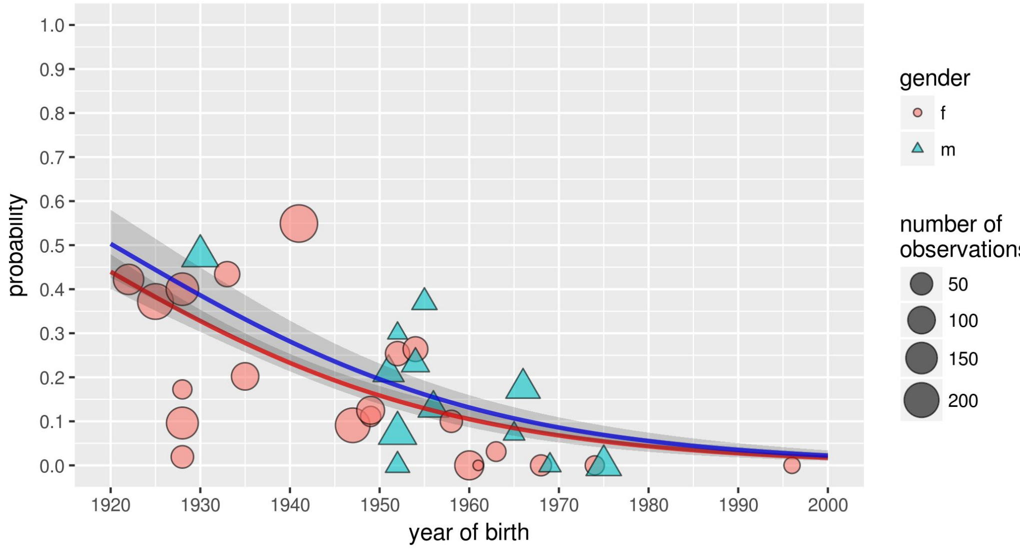
var6_sja_v



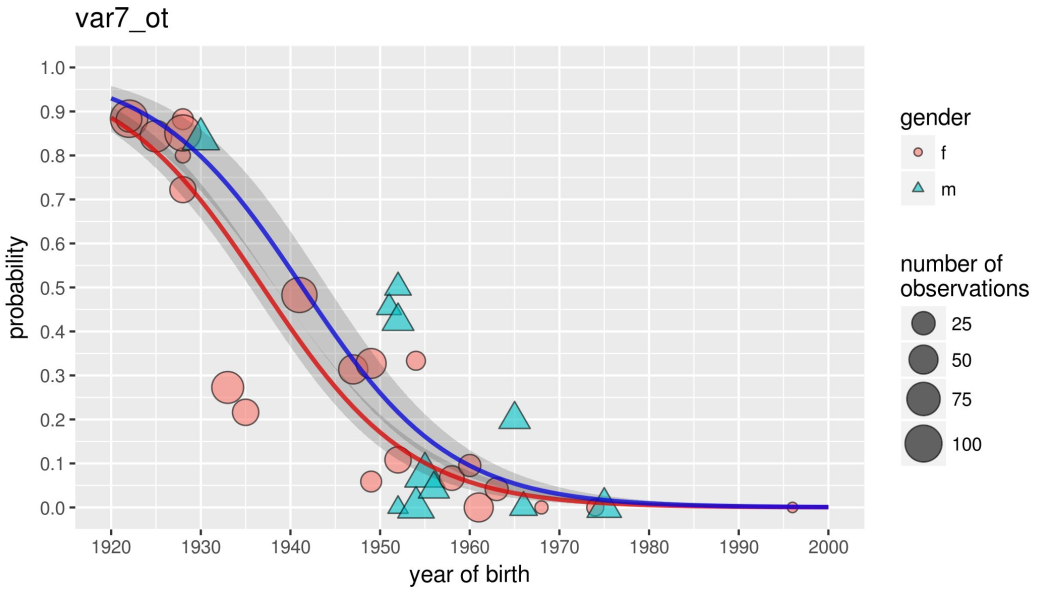
var2_ae

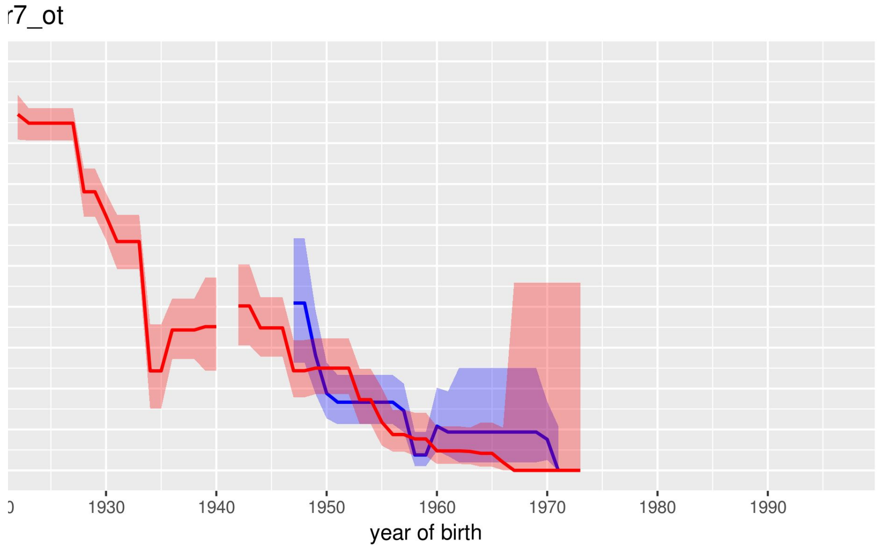


var2_ae



No lag


var7_ot
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
probability
1920
1930
1940
1950
1960
1970
1980
1990
2000
year of birth
gender
f
m
number of
observations
25
50
75
100


r7_ot
)
1930
1940
1950
1960
1970
1980
1990
year of birth

Interpretations

- Our data show that women lead the change (as in the Western SL), contrary to Bubrikh 1914 and other statements.

- (a) Russian dialectologists were wrong

- (b) Something has changed

- Women caught up with men and left them behind?

Interpretations

- Sovietization: villages becoming less traditional, more integrated; women's emancipation
- More upward mobility for women
- More women education
- More exposure to the standard

Interpretations (2)

- Why do some variables lag, and some do not? No idea, attitude / emblematicity perception studies are needed?

- LAGGING: adj, оне, хто, у ево
- LAGGING: щ, ять, ту, ся/се
- NOT LAGGING: ся/сь, а/е, то/от