Modelling linguistic and non-linguistic factors in sociosyntax

Dirk Speelman

QLVL, KU Leuven

Overview

Introduction

Complex or simple models?

Importance of factors and choice of tools

The example of the Dutch causatives

The was/were alternation in the York data

Questions

Conclusions

Overview

Introduction

Complex or simple models?

Importance of factors and choice of tools

The example of the Dutch causatives

The was/were alternation in the York data

Questions

Conclusions

Introduction

This presentation will be a plea for

- sufficiently complex models (including internal and external predictors, and random effects)
- much attention for model diagnostics and the inspection of the random effects (blups) in mixed models
- the confrontation of different tools, because they can help cancel out each other's 'blind spots' (we will confront mixed models with conditional inference trees)

Overview

Complex or Simple

Point of departure:

From the first draft of the program of the symposium

Which type of factor is more important with which type of variable (e.g. is it a general pattern that core syntactic variables are mainly linguistically determined, while discourse variables are both linguistically and socially determined? If so, what may explain this?) Can we focus on just social factors and exclude linguistic factors without compromising methodological standards?

Complex or Simple

From this I took the topics:

- Should we always try and analyze internal and external factors simultaneously?
- Can we always expect both these levels to be at work together (or at least take into account the possibility that this is the case)?

Complex or Simple

My 'belief system' regarding this topic:

- As soon as variation is possible (and therefore as soon as we can speak of a variable), there is at least possibility for socially determined variability (possibly next to or together with the working of internal factors)

Complex or Simple

My 'belief system' regarding this topic:

- As soon as variation is possible (and therefore as soon as we can speak of a variable), there is at least possibility for socially determined variability (possibly next to or together with the working of internal factors)
  - As soon as a certain function correlates with a certain external context (e.g. the mechanism of passive voice correlating with written language), it is possible that the form gets associated with this context.

Complex or Simple

My 'belief system' regarding this topic:

- As soon as variation is possible (and therefore as soon as we can speak of a variable), there is at least possibility for socially determined variability (possibly next to or together with the working of internal factors)
  - As soon as a certain function correlates with a certain external context (e.g. the mechanism of passive voice correlating with written language), it is possible that the form gets associated with this context.
  - I expect there to be a cline (across variables), showning varying importance of external versus internal factors (versus their interaction).

Complex or Simple

My 'belief system' regarding this topic:

- As soon as variation is possible (and therefore as soon as we can speak of a variable), there is at least possibility for socially determined variability (possibly next to or together with the working of internal factors)
  - As soon as a certain function correlates with a certain external context (e.g. the mechanism of passive voice correlating with written language), it is possible that the form gets associated with this context.
  - I expect there to be a cline (across variables), showning varying importance of external versus internal factors (versus their interaction).
  - I do believe it is worthwhile to examine the properties of this cline.

Complex or Simple

My 'belief system' regarding this topic:

- To my experience complex models typically are more realistic than simple models

Complex or Simple

My 'belief system' regarding this topic:

- To my experience complex models typically are more realistic than simple models
  - Alternation patterns that I have looked at in some detail typically show complex structure.

Complex or Simple

My 'belief system' regarding this topic:

- To my experience complex models typically are more realistic than simple models
  - Alternation patterns that I have looked at in some detail typically show complex structure.
  - Complex statistical models 'see more' (but we must also take care not to overfit; e.g. by using penalization techniques).

Complex or Simple

Example: word order variation in clause final verb clusters in Dutch

Internal factors:

- production pressure; semantic nuance (dynamism); rhythm/intonation patterns; syntactic weight of constituents; sentence final or not

External factors:

- country; dialect area; register/genre; mode

Complex or Simple

Example: inflectional variation in Dutch attributive adjectives

Internal factors:

- semantic specialization (official names); collocations

External factors:

- country; formality of register

Complex or Simple

Example: variation in presentative 'er'

Internal factors:

- adjunct type; verb class; predictability of subject ('er' as expectancy monitor)

External factors:

- country; register, genre

Complex or Simple

In these examples ...

- we see an internal 'function' (or a combination of functions) go together with external factors
- we sometimes see how internal and external factors interact, which would make it dangerous to only investigate one of the sources of variation

Complex or Simple

In these examples ...

- we see an internal 'function' (or a combination of functions) go together with external factors
- we sometimes see how internal and external factors interact, which would make it dangerous to only investigate one of the sources of variation
  - 'er': system of internal factors is simpler and stronger in NL

Complex or Simple

In these examples ...

- we see an internal 'function' (or a combination of functions) go together with external factors
- we sometimes see how internal and external factors interact, which would make it dangerous to only investigate one of the sources of variation
  - 'er': system of internal factors is simpler and stronger in NL
  - inflectional variation: different collocations in NL and BE

Overview

Introduction

Complex or simple models?

Importance of factors and choice of tools

The example of the Dutch causatives

The was/were alternation in the York data

Questions

Conclusions

Factors and tools

Point of departure:

From the first draft of the program of the symposium

How do we identify **which** factors determine the variation, and how do we assess the importance/weight/effect size of factors when we don‘t know all the determining factors? Which role can/should **random effects** (i.e. the use of mixed models analyses) play in the statistical modelling of variation?

Factors and tools

From this I extracted two topics:

- How to deal with imperfect knowledge about the variation at hand (cf. identifying the factors and assessing their importance)
- Which tools to use (cf. question about mixed models)

Factors and tools

How to deal with imperfect knowledge ...

Three procedures that can help:

- We can reduce the damage that our imperfect models can do (in the sense that they produce misleading results), by using random effects in regression models, and their counterparts in conditional inference trees.
- We can assess the quality of model and perform model diagnostics to look for indications that something (e.g. a predictor) is missing.
- We can gain further insight in our data by comparing results produced by competing tools for data analysis.

Factors and tools

Introducing random effects ...

- Unless you would end up with a really complex model (in which case there can be computational issues), I see no reason not to use random effects:
  - If there is no need to use random effects, and you do use them, then the output of the mixed model will contain indications that their usage was not necessary (small estimates of variance of random factor)
  - Conversely, a fixed effects only model will not necessarily contain indications that you should have introduced random effects.
- In conditional inference trees too the counterparts of what would be random factors in a regression model can play a similar role.

Introducing random effects ...

Which random factors?

- Preferably we capture the known potential sources of variation as directly as possible (verb, individual, register/genre, topic, ...)
- If this is not possible, then look for a 'rough' proxy for such sources of variation (e.g. document/text as a proxy for individual, or document source or set of documents as a proxy for register/genre, top document keyword as a proxy for topic, ...)

Introducing random effects ...

Again, why random factors? In general, they have different functions:

- to make the extimates for the fixed effects more robust (by reducing the damage that outliers can do)
- to make confidence intervals and p-values for the fixed effects more reliable (without them there is a risk of overestimating significance of fixed effects)
- a means to capture and show 'irregularities' and possibly 'patterns in such irregularities' (and insight in these patterns can help us better understand the variation at hand)

Factors and tools

Assessing model quality ...

Fixed effects logistic regression analysis offers several means to assess model quality.

- The goodness of fit of the model can be tested with the Hosmer-Lemeshov-Cessie test. When the test signals significance, this is a strong indication that e.g. a predictor is missing.
  - The converse is not true. Lack of significance is no guarantee that we're not missing anything.
- One can estimate how much overdispersion there is in the model. Overdispersion can be an indication that a predictor is missing.
- One can uses model quality measures such as C to test how well the model classifies. (This, however, cannot directly be used to test if predictors are missing.)
- The fixed effects in the model can be compared to a conditional inference tree.

Factors and tools

Assessing model quality ...

Mixed effects logistic regression analysis also offers several means to assess model quality.

- One can try and detect patterns in the random effects.
- The search for such patterns can be informed by how their counterparts behave in a conditional inference tree.
- Here too the fixed effects in the model can be compared to their counterparts in a conditional inference tree.

Factors and tools

Assessing model quality ...

In the case of conditional inference trees,

- one can compare the importance of the factors with the role of their counterparts in fixed effect and mixed effect regression models.

Factors and tools

Comparing results from different tools ...

Tools and models are there to be used, not believed.

- We must not expect there to be a perfect tool and a perfect model for our data.
- Statistical models are always simplifications of reality. Some of there are more acceptable than others. None are perfect.
- Sometimes different models are all acceptable.
- Sometimes different tools give different information on data.
- We can make use of this situation, by confronting the results that are given by different tools

Factors and tools

Comparing results from different tools ...

Today, I will only discuss regression models (both fixed effects and mixed effects) and conditional inference trees. Zooming in on these tools, we could say that:

- fixed effects in regression models were designed to look for global patterns
- random effects in regression models were designed to account for fine-grained patterns and (groups of) special cases
- from the design of conditional inference trees it follows that they have a strong capability of detecting interactions (i.e. of patterns that are somewhere between 'global' and 'local', in the sense that they hold true in a substantial number of cases, but don't apply in other contexts).

Overview

Dutch causatives

The alternation pattern:

Dutch causatives *doen* and *laten*

- Je mag hen niet **laten** geloven dat dit echt is.
  You must not **let** *them believe that this is real.*
- Je mag hen niet **doen** geloven dat dit echt is.
  You must not **make** *them believe that this is real.*

Dutch causatives

Internal predictors

- `inanim`: inanimateness of main clause subject (is expected to favor *doen*)
- `coref`: coreferentiality between main clause subject and subclause subject/object. (has been observed to favor *laten*)
- `col=lex`: lexical collocation between specific causative and subclause infinitive (has been observed to favor *doen*)
- `col=sem`: semantic collocation between subclause infinitive and causation (has been hypothesized to favor *doen*)

Dutch causatives

External predictors

- `country`: NL versus BE (BE is expected to favor *doen*)
- `spont`: spontaneous (spont=TRUE) versus prepared (spont=FALSE) speech
- `sex`: ...
- `age`: ...

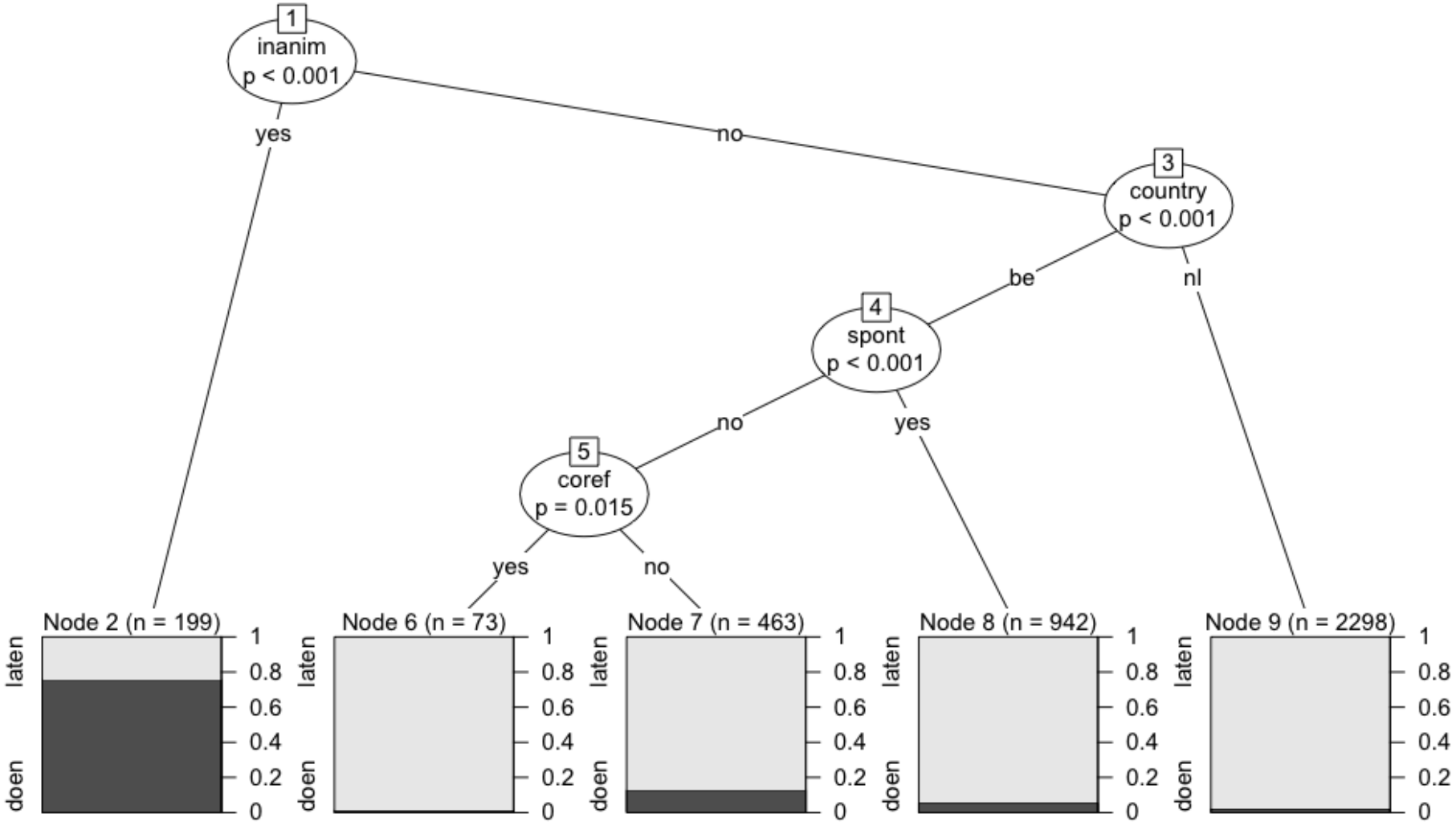
Dutch causatives

Random effects

- `inf`: the specific verb that appears in the infinitive slot in the subclause.
- `speaker`: (not discussed today)

Dutch causatives

Conditional inference tree (no random effects)


1
inanim
p < 0.001
yes
no
3
country
p < 0.001
be
nl
4
spont
p < 0.001
no
yes
5
coref
p = 0.015
yes
no
Node 2 (n = 199)
Node 6 (n = 73)
Node 7 (n = 463)
Node 8 (n = 942)
Node 9 (n = 2298)
laten
doen
1
0.8
0.6
0.4
0.2
0

Dutch causatives

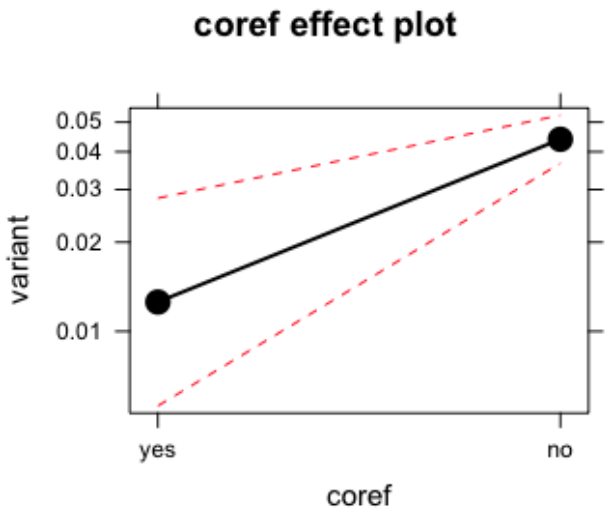
Conditional inference tree (no random effects)

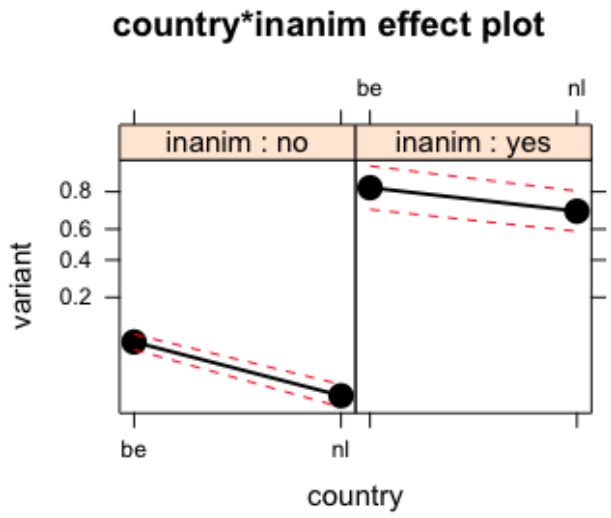
The conditional inference tree sees many interactions:

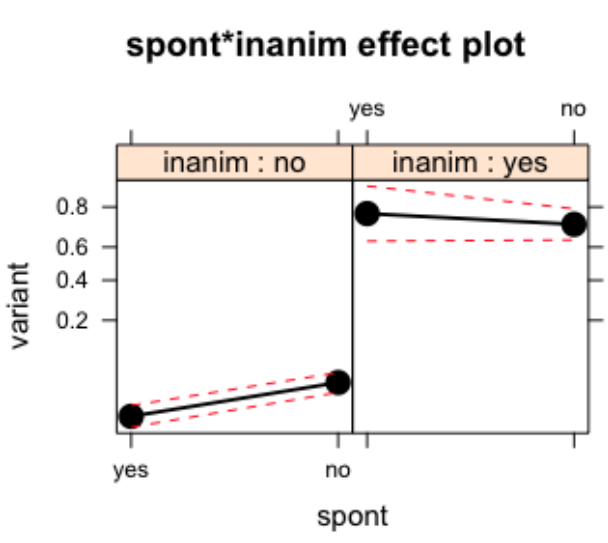
- inanim:country, inanim:spont, inanim:coref
- country:spont, coref:spont, coref:country

Dutch causatives

Fixed effects logistic regression (no random effects)


coref effect plot
variant
0.05
0.04
0.03
0.02
0.01
yes
no
coref


country*inanim effect plot
be
nl
inanim : no
inanim : yes
variant
0.8
0.6
0.4
0.2
be
nl
country


spont*inanim effect plot
yes
no
inanim : no
inanim : yes
variant
0.8
0.6
0.4
0.2
yes
no
spont

Dutch causatives

Fixed effects logistic regression (no random effects)

The regression model sees far fewer interactions:

- inanim:country, inanim:spont

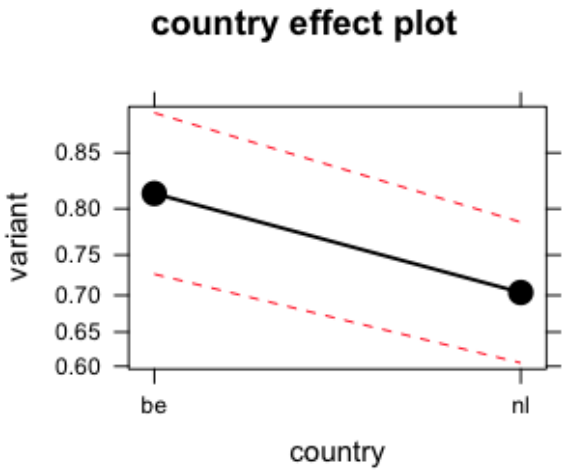
And it sees more global patterns:

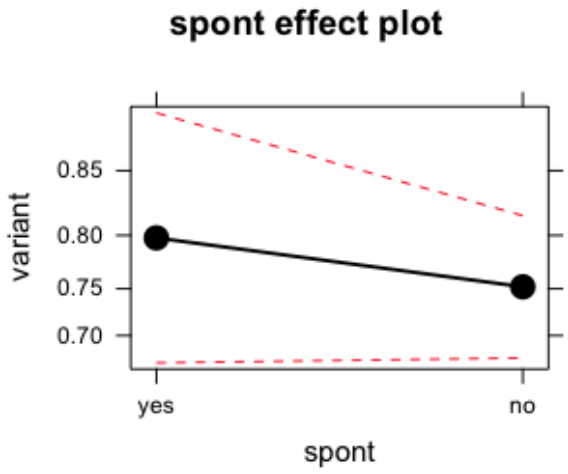
- e.g. coref

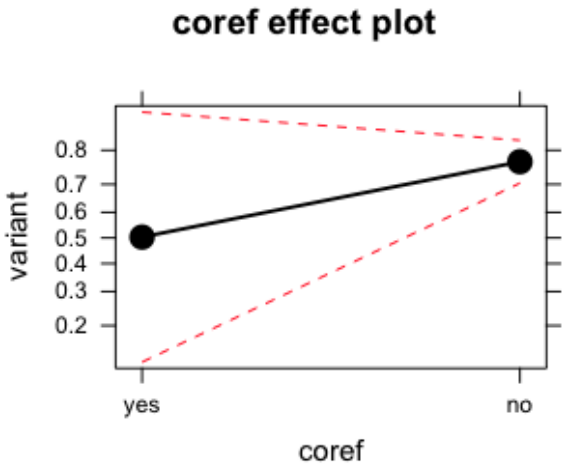
Dutch causatives

So which is correct ?

Let's run a regression on the subset inanim=yes. In this model only country is borderline significant, but let's look at the effects anyway.


country effect plot
variant
0.85
0.80
0.75
0.70
0.65
0.60
be
nl
country


spont effect plot
variant
0.85
0.80
0.75
0.70
yes
no
spont


coref effect plot
variant
0.8
0.7
0.6
0.5
0.4
0.3
0.2
yes
no
coref

Dutch causatives

So which is correct ?

It is hard to decide which model is more correct.

- the patterns we saw in the global model also show up here (e.g. coref)
- but they don't reach significance

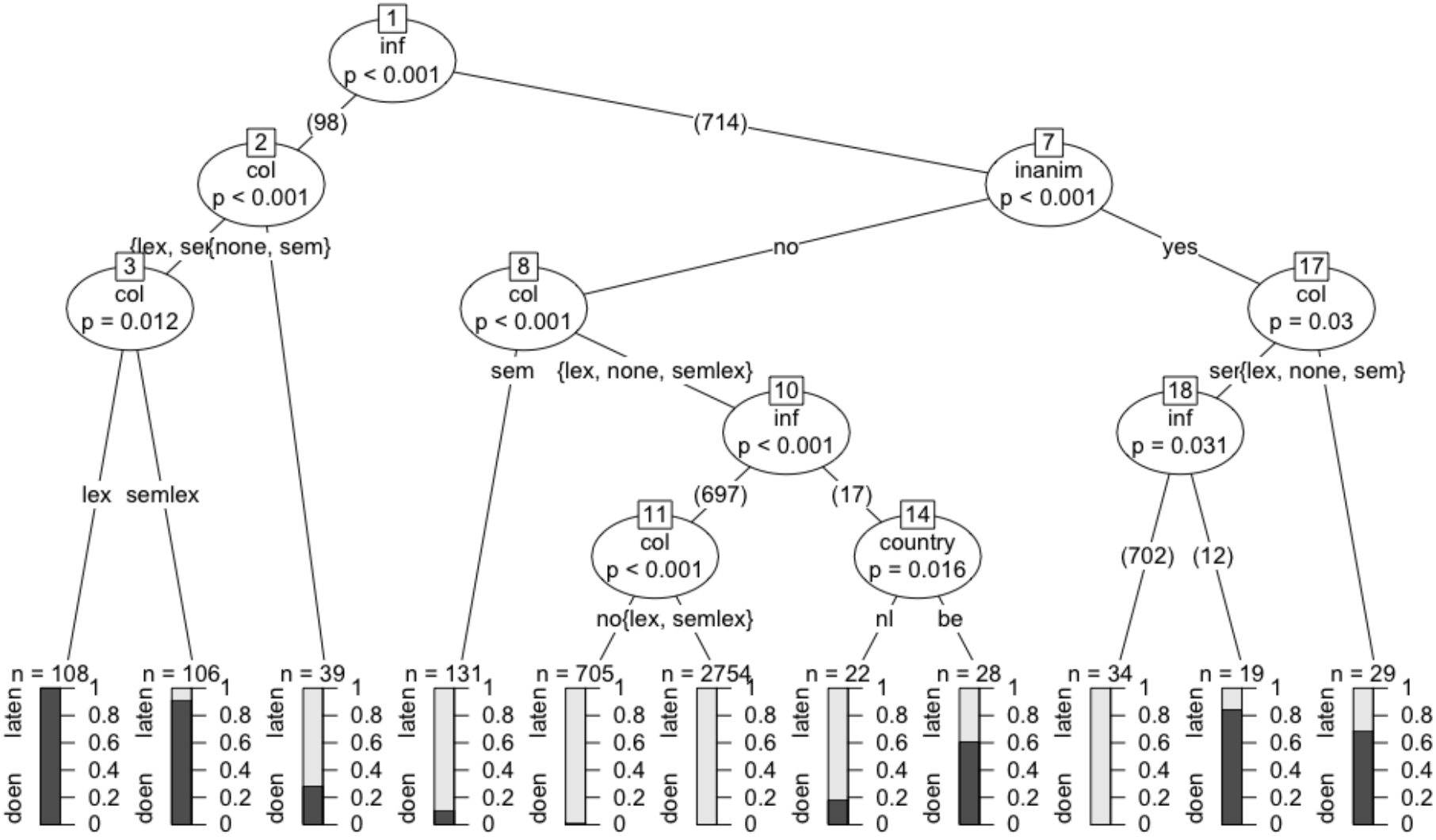
I would say both models are correct, or rather, both models 'have a point'. We learn

- that an effect such as coref can indeed be called global
- but that it is nevertheless somewhat less outspoken in the case of inanim=yes

Dutch causatives

Conditional inference tree (random factor inf)

Now let's try and use a random factor to catch irregularities of deficiencies in our naive fixed effects only model.


1
inf
p < 0.001
(98)
(714)
2
col
p < 0.001
7
inanim
p < 0.001
{lex, se
{none, sem}
no
yes
3
col
p = 0.012
8
col
p < 0.001
17
col
p = 0.03
lex
semlex
sem
{lex, none, semlex}
ser
{lex, none, sem}
10
inf
p < 0.001
18
inf
p = 0.031
(697)
(17)
11
col
p < 0.001
14
country
p = 0.016
(702)
(12)
no
{lex, semlex}
nl
be
n = 108
n = 106
n = 39
n = 131
n = 705
n = 2754
n = 22
n = 28
n = 34
n = 19
n = 29
laten
doen
1
0.8
0.6
0.4
0.2
0

Dutch causatives

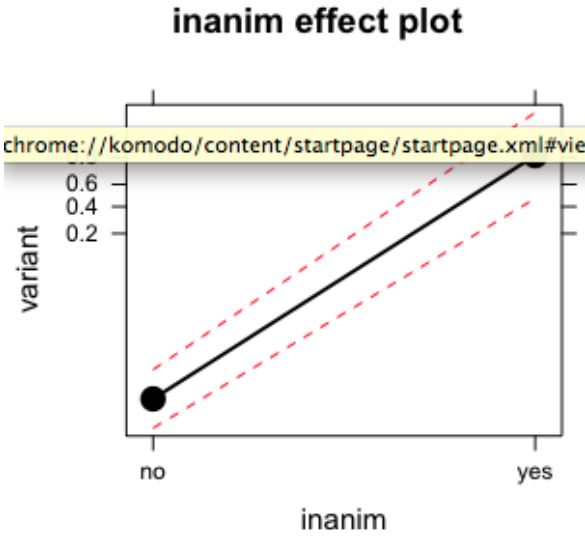
Conditional inference tree (random factor inf)

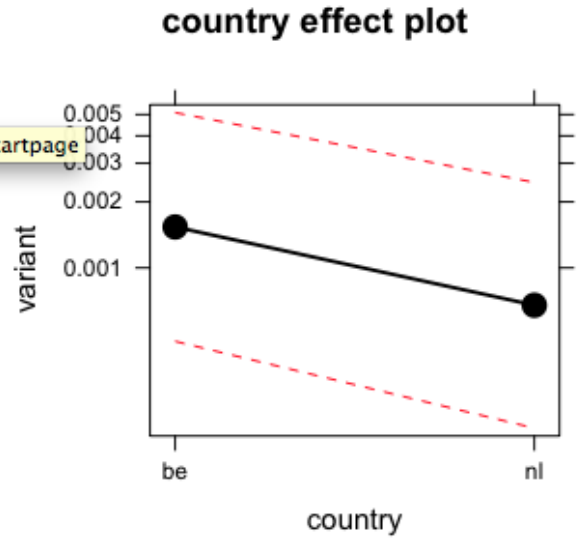
- Once again we see that the conditional inference tree succeeds in finding many interactions
- The predictor col enters the picture; perhaps it also to some extent is a proxy for 'verb class' (e.g. node 8)
- But inf clearly also plays a role, so there's structure in the data that isn't captured by the 'fixed effects'
- The factors spont and coref disappear, and country becomes less important

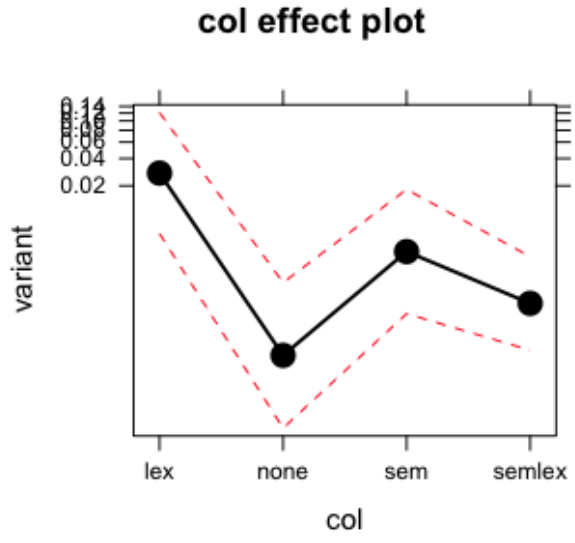
Dutch causatives

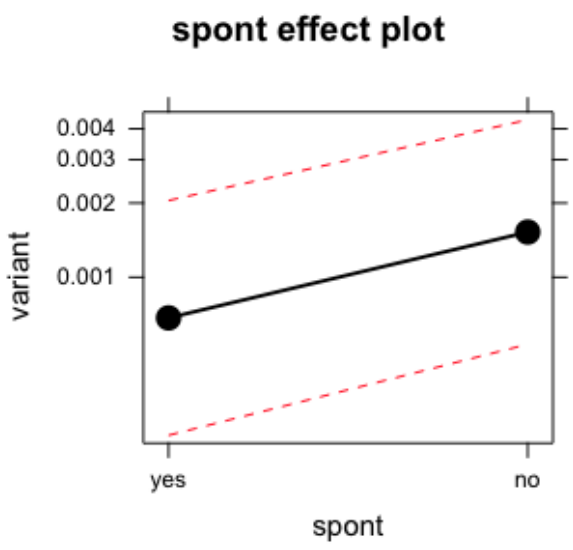
Mixed model (random factor inf)

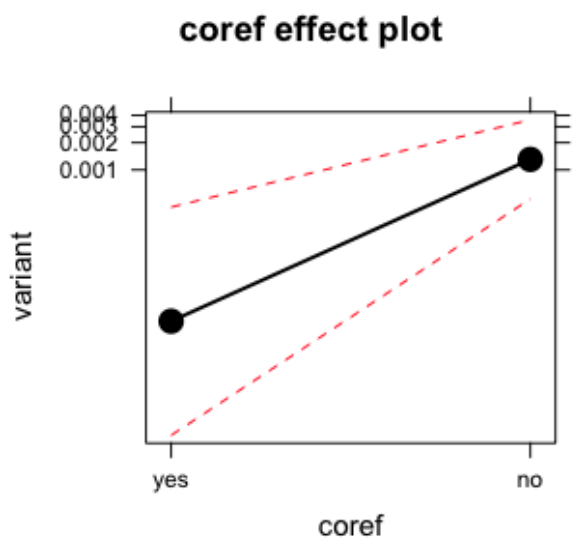
So what does a mixed model say? First, let's look at the fixed effects.


inanim effect plot
chrome://komodo/content/startpage/startpage.xml#view-startpage
0.6
0.4
0.2
variant
no
yes
inanim


country effect plot
0.005
0.004
0.003
0.002
0.001
variant
be
nl
country


col effect plot
0.14
0.12
0.1
0.08
0.06
0.04
0.02
variant
lex
none
sem
semlex
col


spont effect plot
0.004
0.003
0.002
0.001
variant
yes
no
spont


coref effect plot
0.004
0.003
0.002
0.001
variant
yes
no
coref

Dutch causatives

Mixed model (random factor inf)

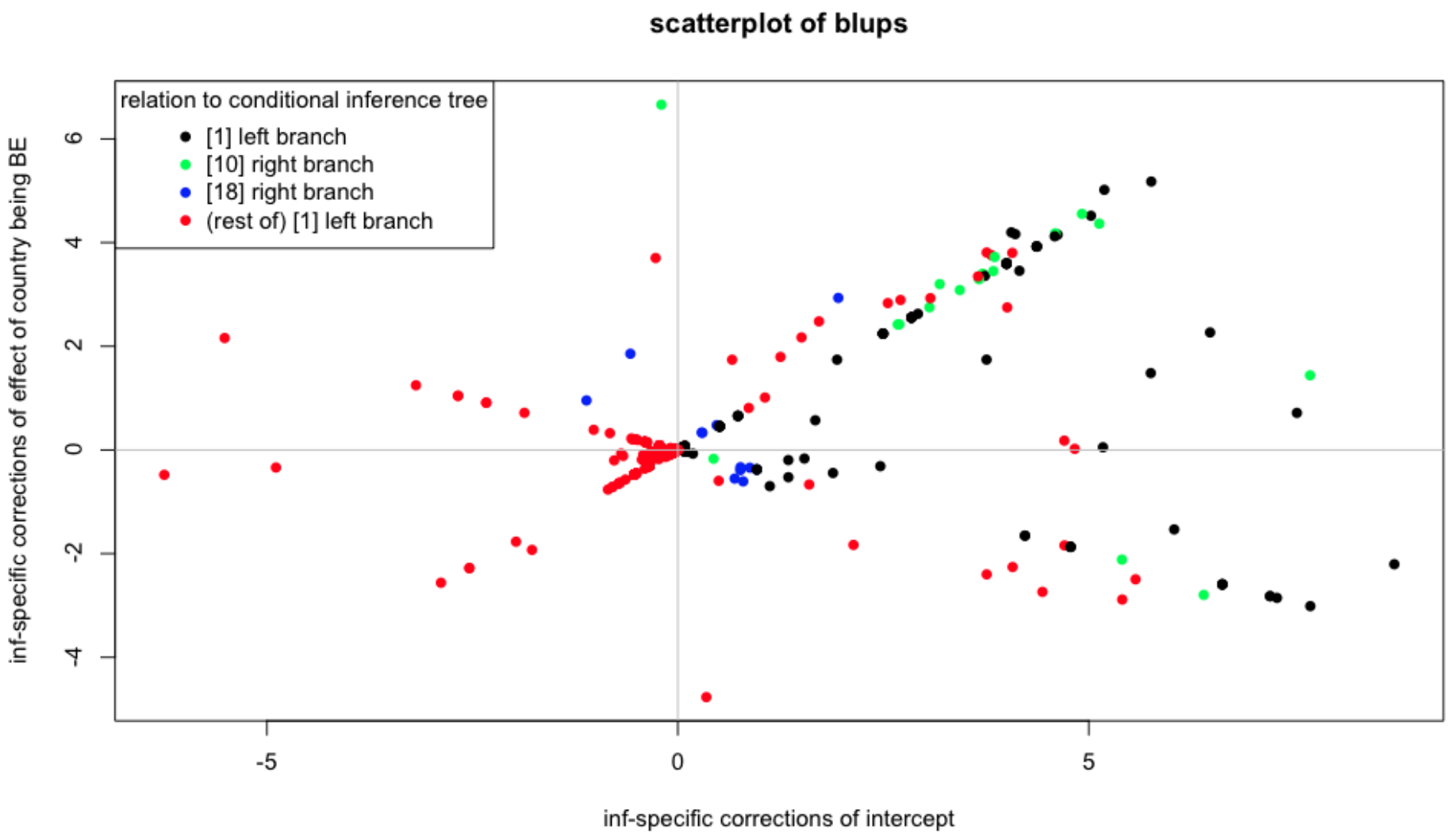
- the factors spont and coref stay in the picture
- here too inf is important (substantial variance)
- we clearly see a different 'perspective' between mixed models and conditional inference trees (no interactions in the mixed model; nothing but interactions in the conditional inference tree) [here too it would be a good exercise to examine in detail how and to which extent the actual state of affairs in the data is to be found somewhere in between both perspectives]

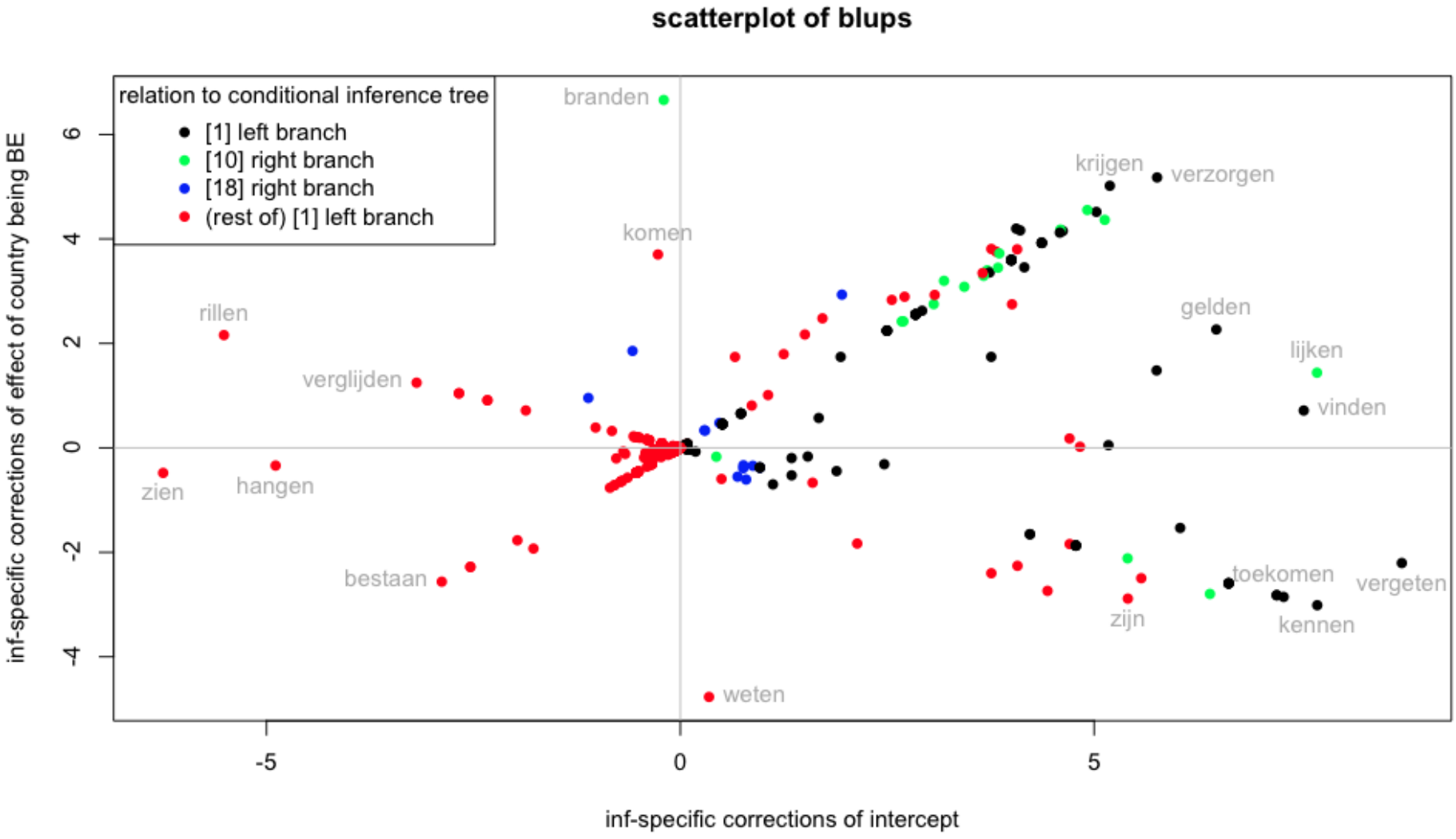
Dutch causatives

Mixed model (random factor inf)

Let's turn to the random effects. How then can we find structure in them? We'll use three strategies:

- We'll try and locate the infinitives from the branches in the conditional inference tree in the blups (= random effects), to see whether they form clusters in the blups
- We'll manually identify individual infinitives in the scatterplots of the blups, on the basis of their position
- We'll compile and examine lists of infinitives the blups of which have extreme values. We might see interesting verb classes of collocation patterns in there, that may inspire us to come up with a new fixed effect (for follow-up research).


scatterplot of blups
relation to conditional inference tree
[1] left branch
[10] right branch
[18] right branch
(rest of) [1] left branch
inf-specific corrections of effect of country being BE
6
4
2
0
-2
-4
-5
0
5
inf-specific corrections of intercept


scatterplot of blups
relation to conditional inference tree
[1] left branch
[10] right branch
[18] right branch
(rest of) [1] left branch
branden
krijgen
verzorgen
komen
rillen
gelden
lijken
verglijden
vinden
zien
hangen
bestaan
toekomen
vergeten
zijn
kennen
weten
inf-specific corrections of effect of country being BE
inf-specific corrections of intercept
6
4
2
0
-2
-4
-5
0
5

Dutch causatives

Mixed model (random factor inf)

- **least likely to be used with doen**: "zien" "rillen" "hangen" "verglijden" "bestaan" "rijpen" "communiceren" "binnentreden" "commanderen" "combineren" "lekken" "opgaan" "nemen" "beseffen" "slaan" "leggen" "genieten" "sturen" "starten" "verstaan" "liggen" "invullen" "uitleggen" "voorzien" "studeren" "horen" "maken" "bestellen" "verdwijnen" "modelleren" "vallen" "doorwerken" "fuseren" "grazen" "kennismaken" "klonen" "nadoen" "ontsmetten" "overdoen" "renoveren" "slachten" "verblinden" "verenigen" "afbreken" "verzekeren" "beschrijven" "wassen" "herstellen" "uitspreken" "vermoorden"

Dutch causatives

Mixed model (random factor inf)

- **most likely to be used with doen**: "vergeten" "kennen" "lijken" "vinden" "toekomen" "instromen" "reactiveren" "beven" "opbiechten" "voortbestaan" "wedergeboren" "gelden" "overtuigen" "geloven" "verzorgen" "denken" "geven" "zijn" "instellen" "krijgen" "aanvaarden" "stoppen" "keren" "verkopen" "gaan" "voorkomen" "belanden" "melken" "uitproberen" "slagen" "draaien" "kruipen" "leren" "sneuvelen" "uitvoeren" "herleven" "foerieren" "heropleven" "imploderen" "ondertekenen" "ondervinden" "huiveren" "rijzen" "plaatsvinden" "praten" "omslaan" "veranderen" "dalen" "opspringen" "kabbelen"

Dutch causatives

Mixed model (random factor inf)

- **strongest negative correction of "BE" effect**: "weten" "kennen" "zijn" "toekomen" "instromen" "reactiveren" "overtuigen" "uitvoeren" "beven" "opbiechten" "voortbestaan" "wedergeboren" "bestaan" "geven" "schijnen" "binnentreden" "commanderen" "plaatsvinden" "vergeten" "instellen" "beseffen" "voorkomen" "belanden" "melken" "uitproberen" "slagen" "staan" "opgaan" "ondertekenen" "ondervinden" "geloven" "genieten" "starten" "verstaan" "leven" "ontstaan" "invullen" "uitleggen" "voorzien" "studeren" "zweten" "trekken" "bestellen" "schrikken" "opwaaien" "zien" "doorwerken" "fuseren" "grazen" "kennismaken"

Dutch causatives

Mixed model (random factor inf)

- **strongest negative correction of "BE" effect**: "branden" "verzorgen" "krijgen" "verkopen" "keren" "stoppen" "omslaan" "leren" "rijzen" "kruipen" "sneuvelen" "herleven" "foerieren" "heropleven" "imploderen" "samenwerken" "praten" "produceren" "vollopen" "komen" "dalen" "opspringen" "kabbelen" "kelderen" "kokhalzen" "opbloeien" "opwakkeren" "plengen" "verstommen" "welslagen" "huiveren" "tekenen" "lachen" "ontluiken" "winnen" "eten" "opmerken" "ontploffen" "springen" "lopen" "zeggen" "overkomen" "zitten" "leeglopen" "veranderen" "inzien" "afdwingen" "afwijken" "behoren" "tennissen"

Overview

Was/were

The alternation pattern:

was/were

- There **was** one or two killed in that area.
- There **were** one or two killed in that area.

Was/were

Internal predictors

- `polarity`: negation or no negation of to be
- `DPconstituency`: determiner phrase constituency; type of determination (bare NP, ...)
- `adjacency`: and other measures of proximity between verb and referent

Was/were

External predictors

- `education`: ...
- `sex`: ...
- `age`: ...

Dutch causatives

Random effects

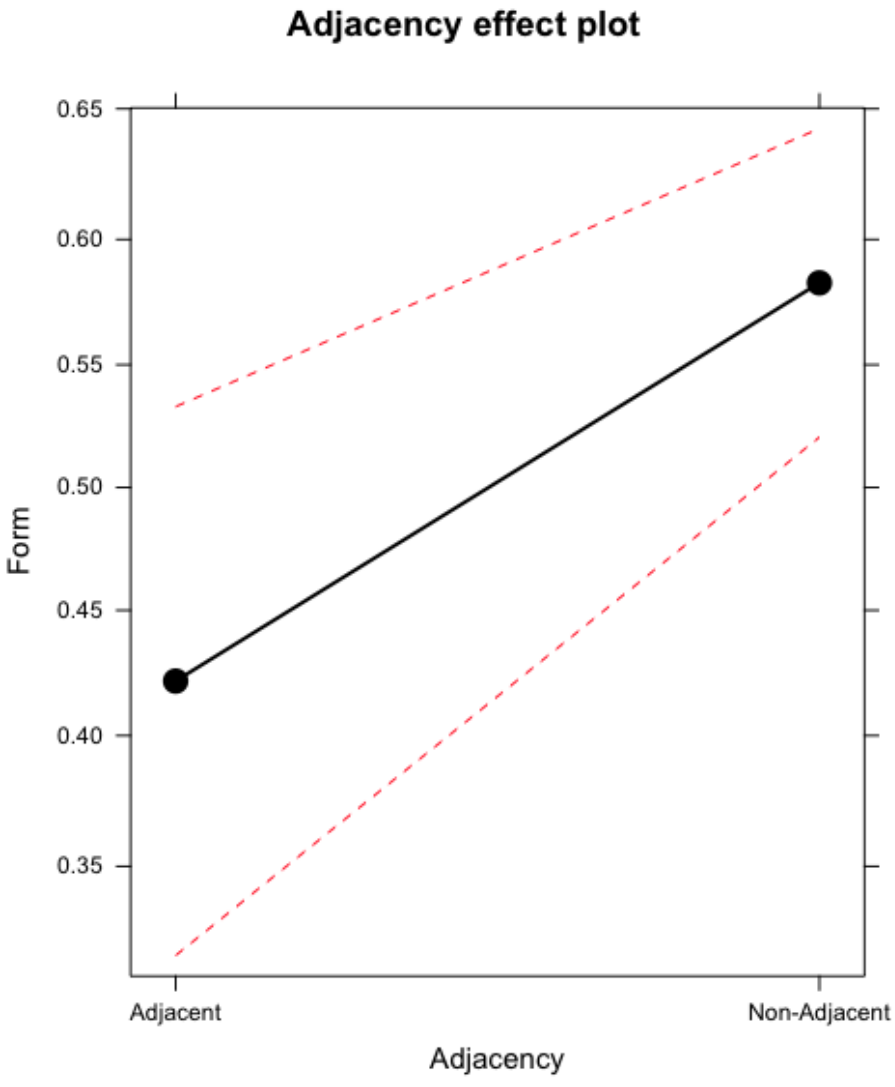
- `individual`: ...

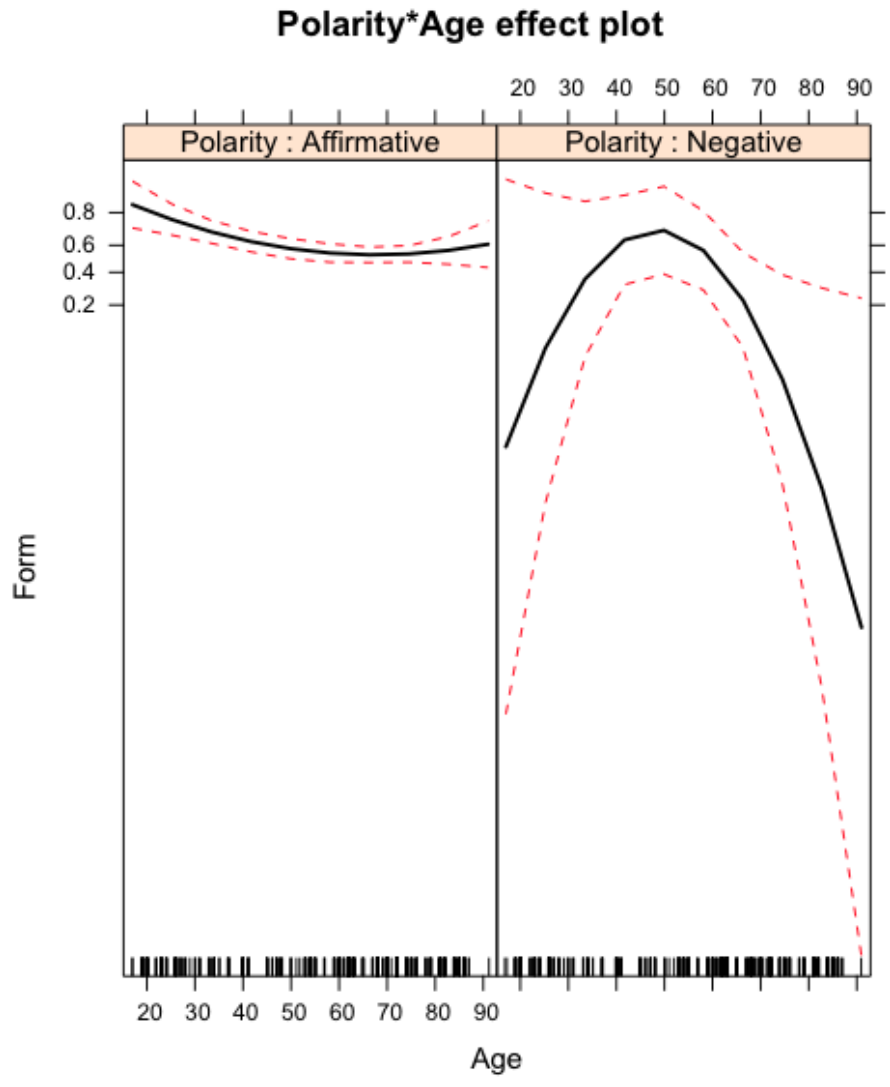
Was/were

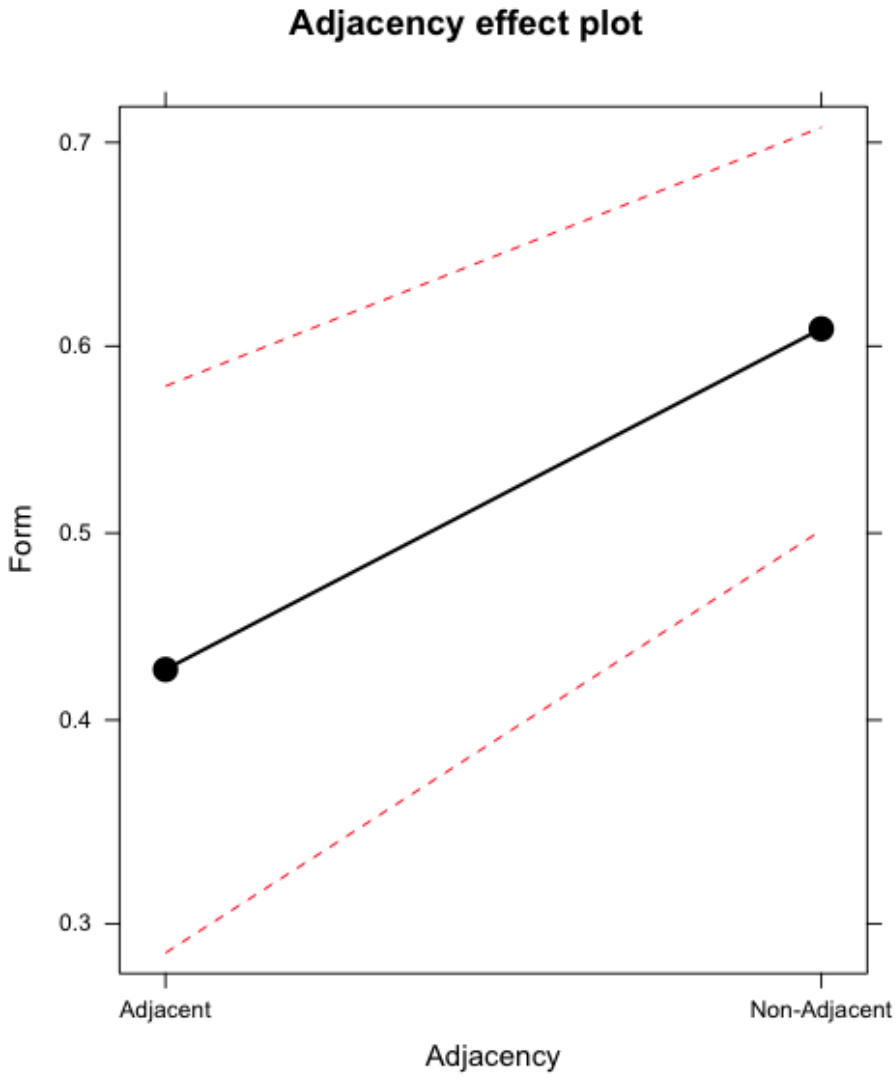
Data and analyses were mainly taken from Tagliamonte and Baayen (2012), so let's start from their argumentation. In a nutshell:

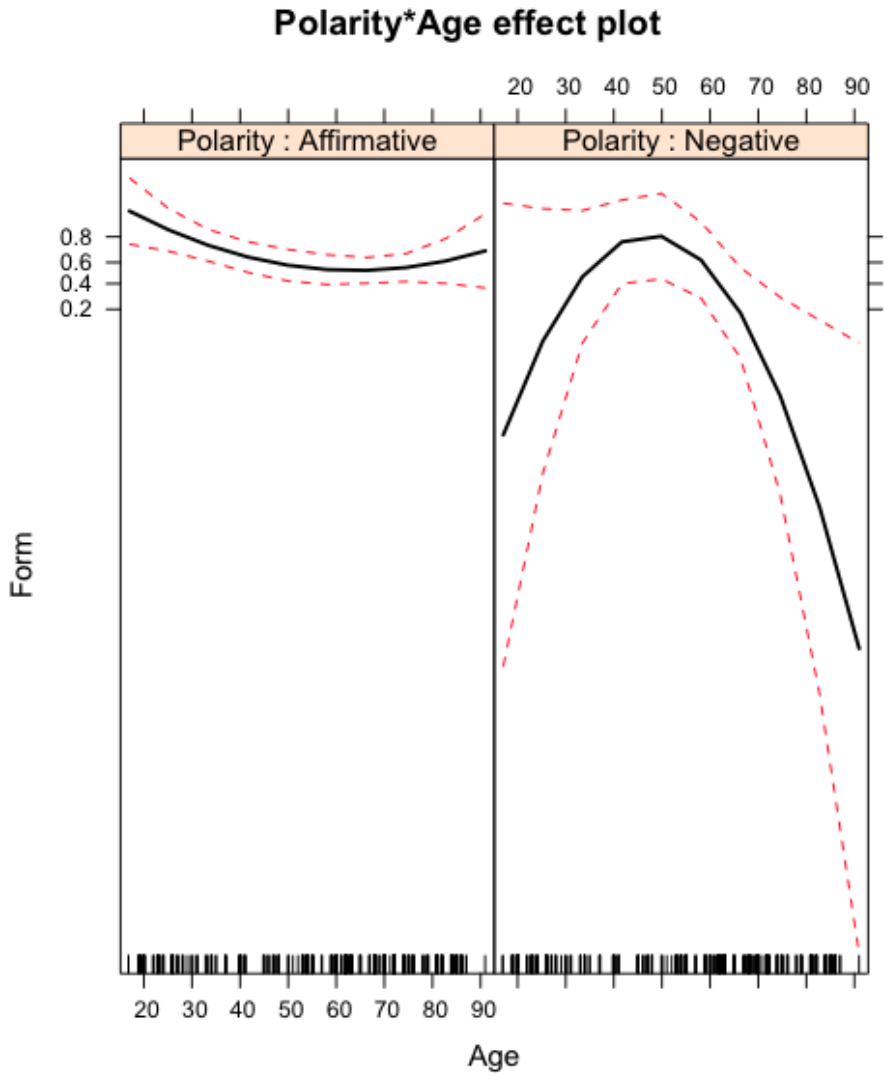
- fixed effect models cannot handle sources of variation such as individual variation in a satisfactory way
- mixed models can, but they in turn cannot deal with predictors that correlate too much
- random forests can deal with individual variation and can deal with highly correlated predictors

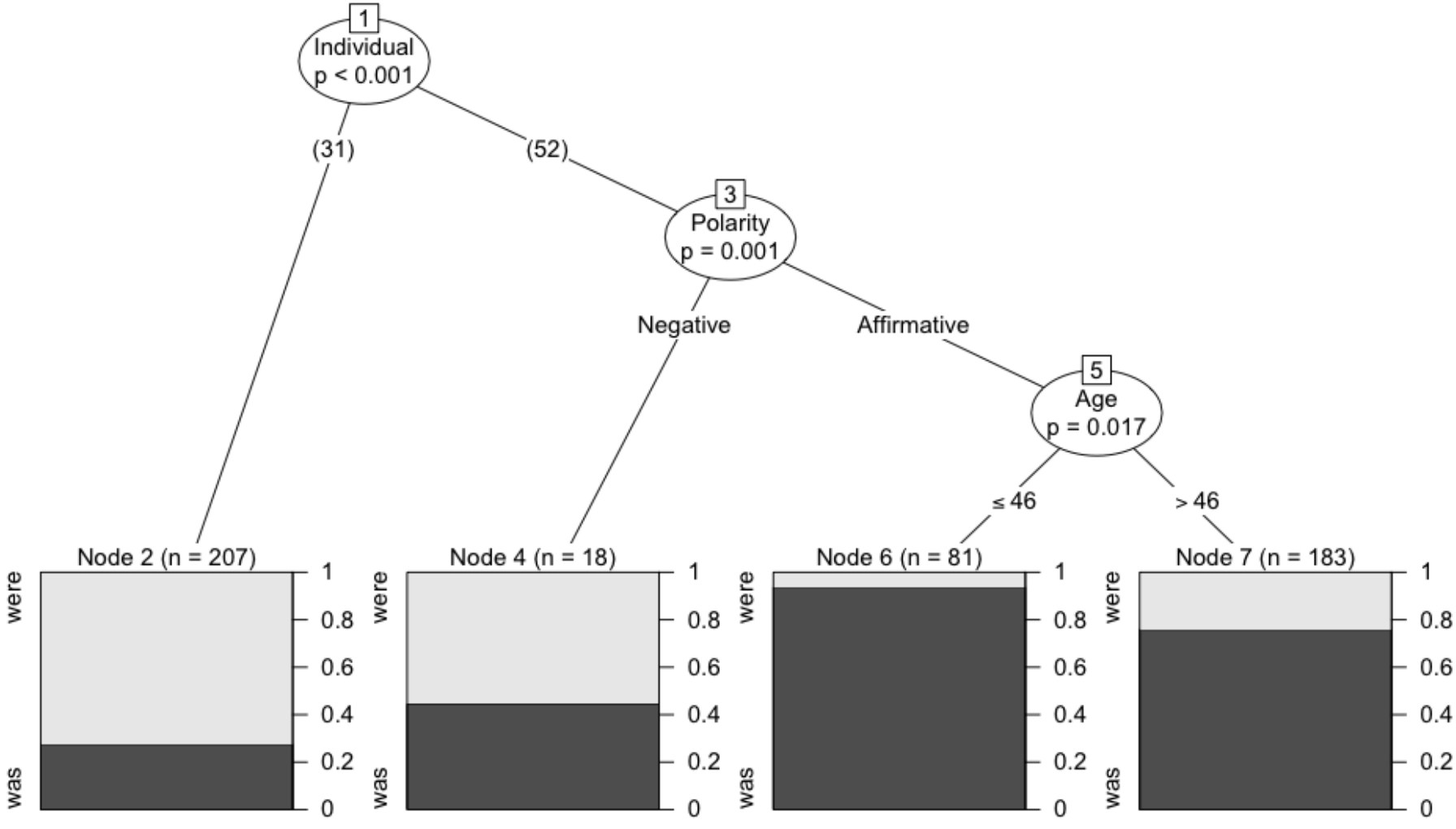
Conclusion: we should add these tools to our toolset, next to fixed effects logistic regression analysis.


Adjacency effect plot
0.65
0.60
0.55
0.50
0.45
0.40
0.35
Form
Adjacent
Non-Adjacent
Adjacency


Polarity*Age effect plot
20 30 40 50 60 70 80 90
Polarity : Affirmative
Polarity : Negative
0.8
0.6
0.4
0.2
Form
20 30 40 50 60 70 80 90
Age


Adjacency effect plot
0.7
0.6
0.5
0.4
0.3
Form
Adjacent
Non-Adjacent
Adjacency


Polarity*Age effect plot
20 30 40 50 60 70 80 90
Polarity : Affirmative
Polarity : Negative
0.8
0.6
0.4
0.2
Form
20 30 40 50 60 70 80 90
Age


1
Individual
p < 0.001
(31)
(52)
3
Polarity
p = 0.001
Negative
Affirmative
5
Age
p = 0.017
≤ 46
> 46
Node 2 (n = 207)
Node 4 (n = 18)
Node 6 (n = 81)
Node 7 (n = 183)
were
was
1
0.8
0.6
0.4
0.2
0

Was/were

We will take these analyses as a starting point, and apply the same strategy of tool comparison that we applied to doen/laten.

- Notice how in the was/were data again the conditional inference tree differs from the regression models: other factors, fewer global patterns, more interactions
- We will zoom in on the random effects

Was/were

We will start from two observations:

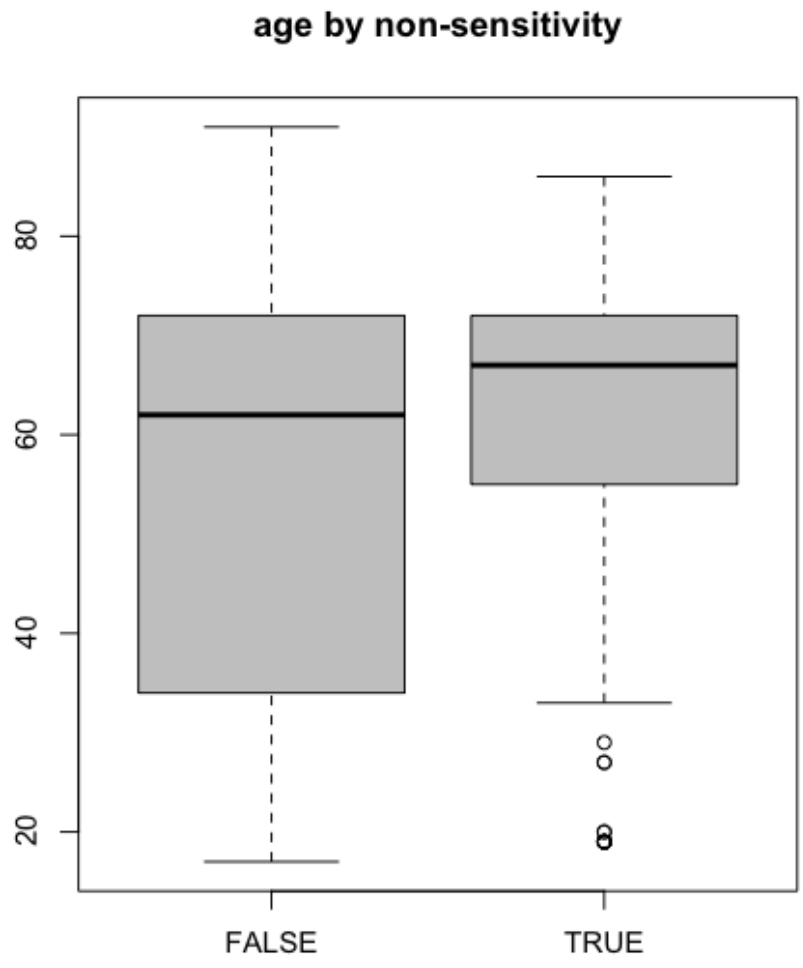
- Tagliamonte and Baayen (2012) observe that non-variable individuals behave differently from variable individuals
- In the conditional inference tree we just saw (top left branch), it is suggested that 31 individuals are not sensitive to the overall patterns of age and polarity (or any other patterns); We'll call these individuals non-sensitives

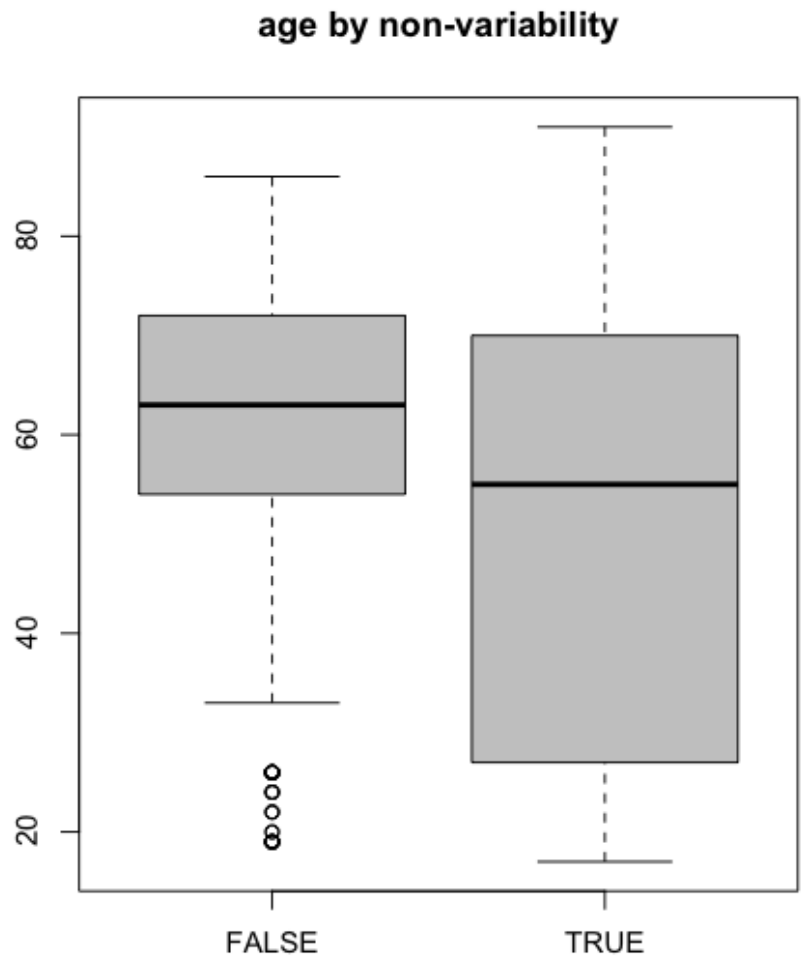
We will examine in more detail what the data and the tool comparison tell us about these two groups (non-variables, non-sensitives).

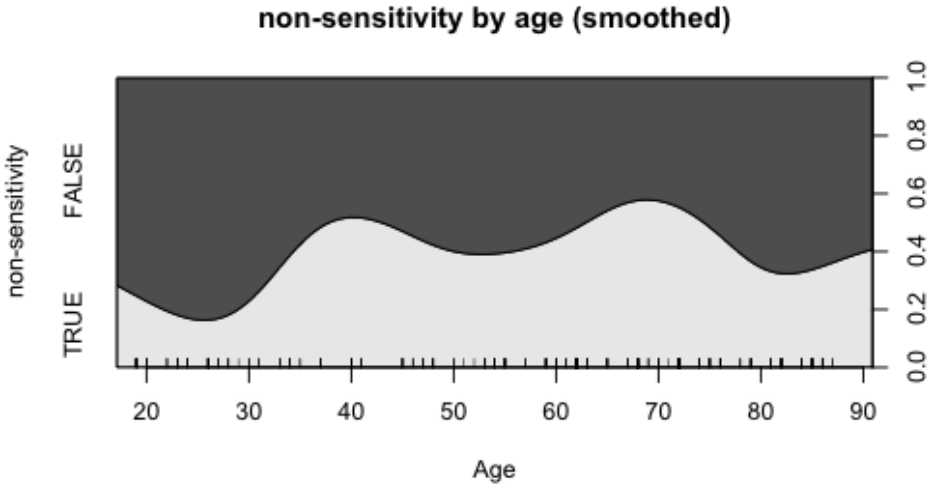
Was/were

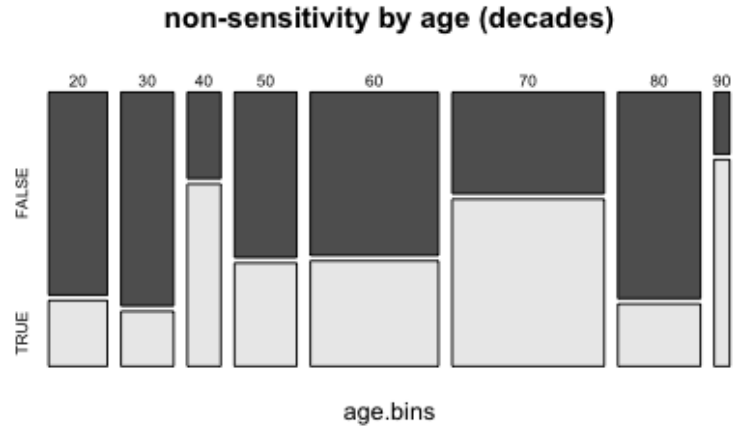
Simple descriptive mosaic plots (and some other plots) suggest that:

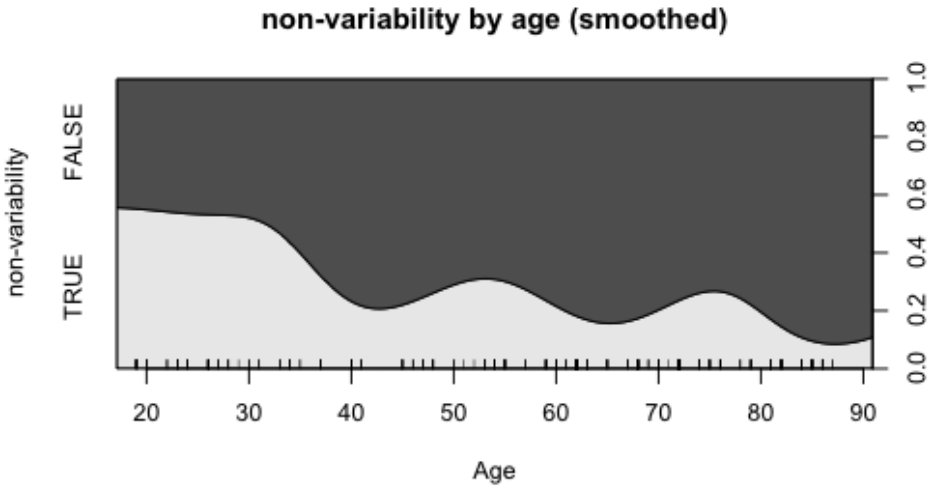
- non-sensitives tend to be older, and tend to be were users
- non-variables tend to be younger, and in that case tend to be was users (but the older non-variables are were users)

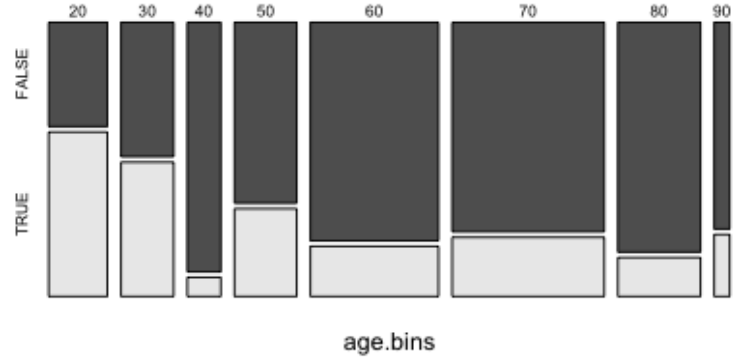

age by non-sensitivity
80
60
40
20
FALSE
TRUE

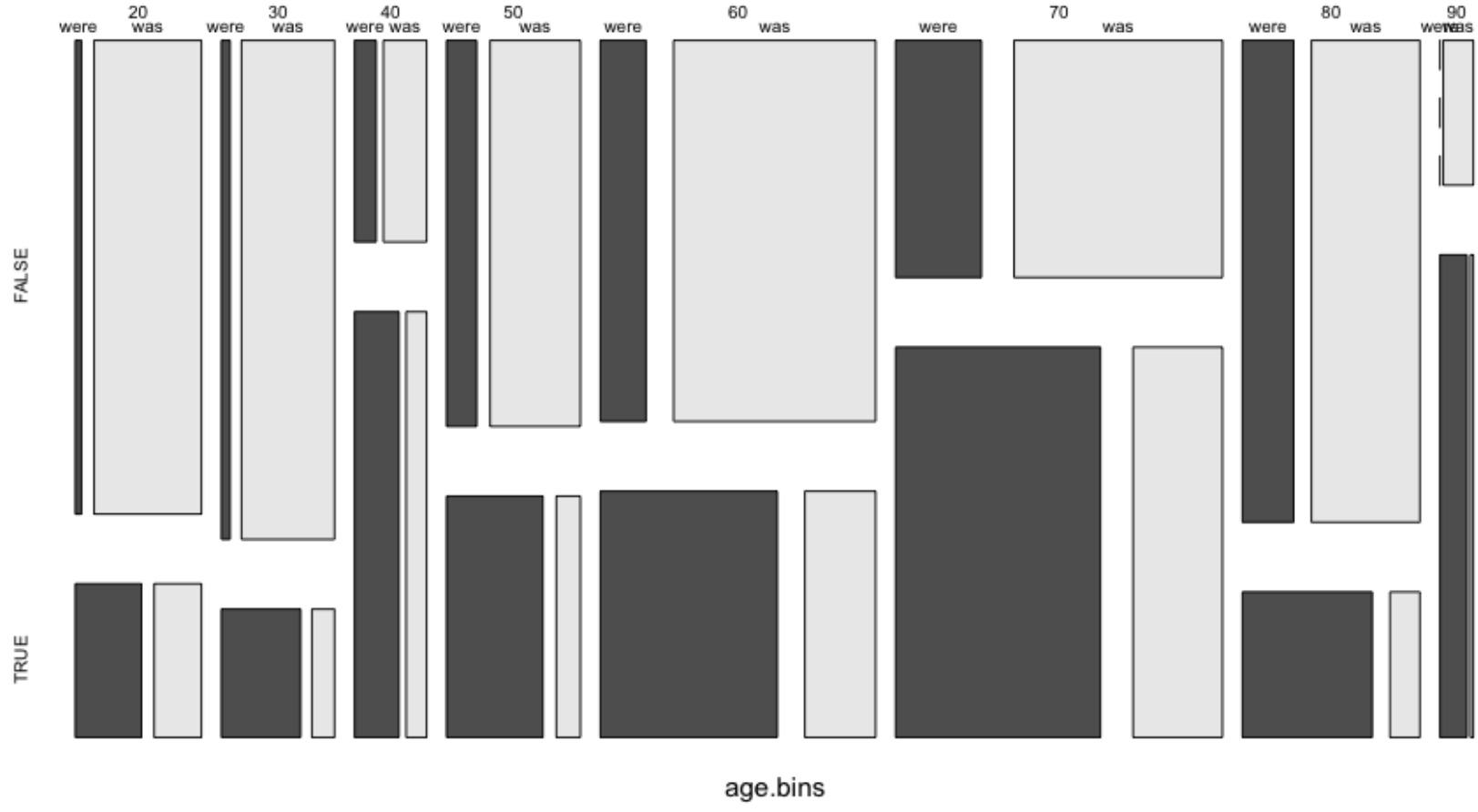

age by non-variability
80
60
40
20
FALSE
TRUE

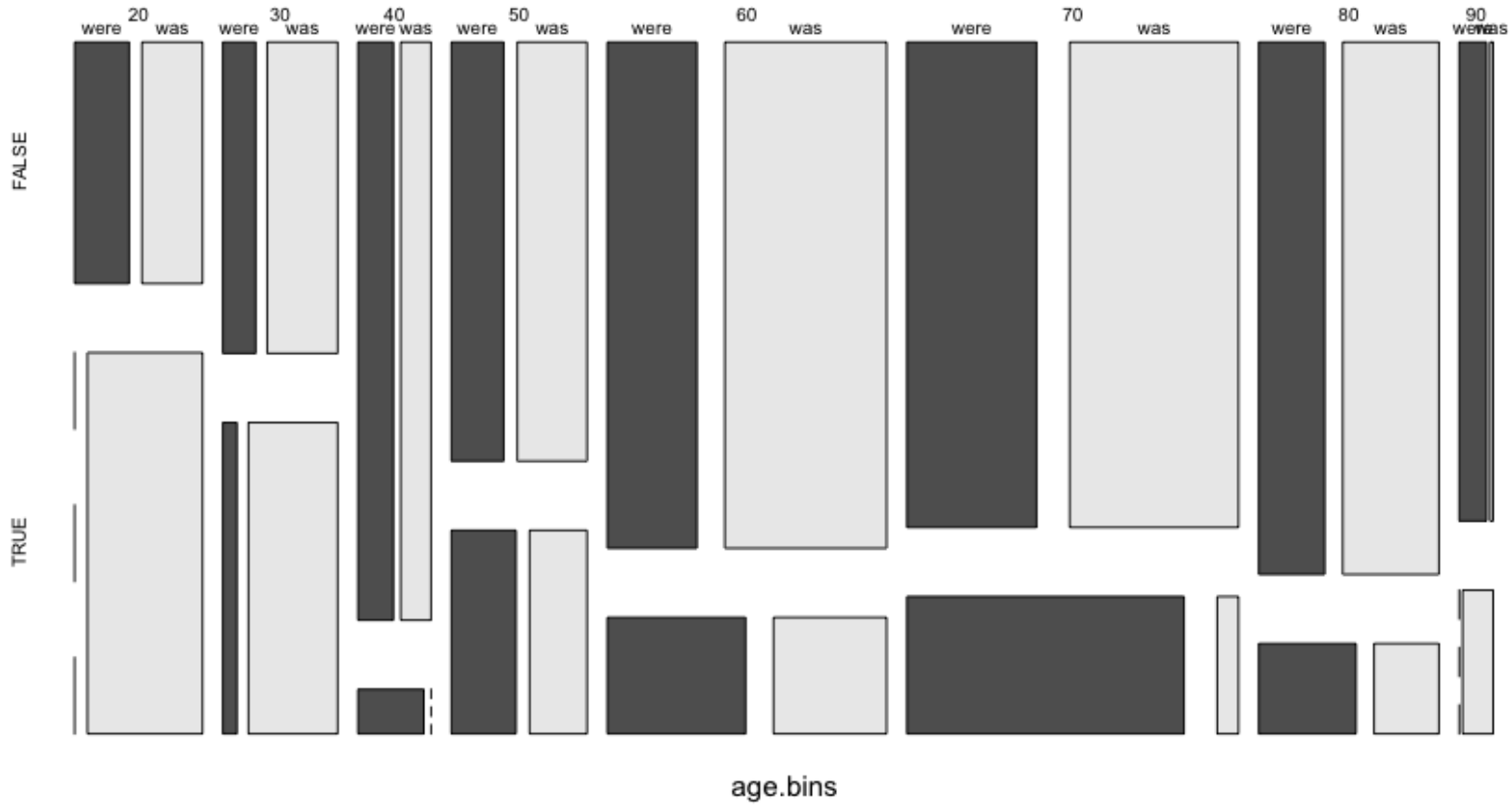

non-sensitivity by age (smoothed)
non-sensitivity
FALSE
TRUE
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
70
80
90
Age


non-sensitivity by age (decades)
20
30
40
50
60
70
80
90
FALSE
TRUE
age.bins


non-variability by age (smoothed)
non-variability
FALSE
TRUE
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
70
80
90
Age

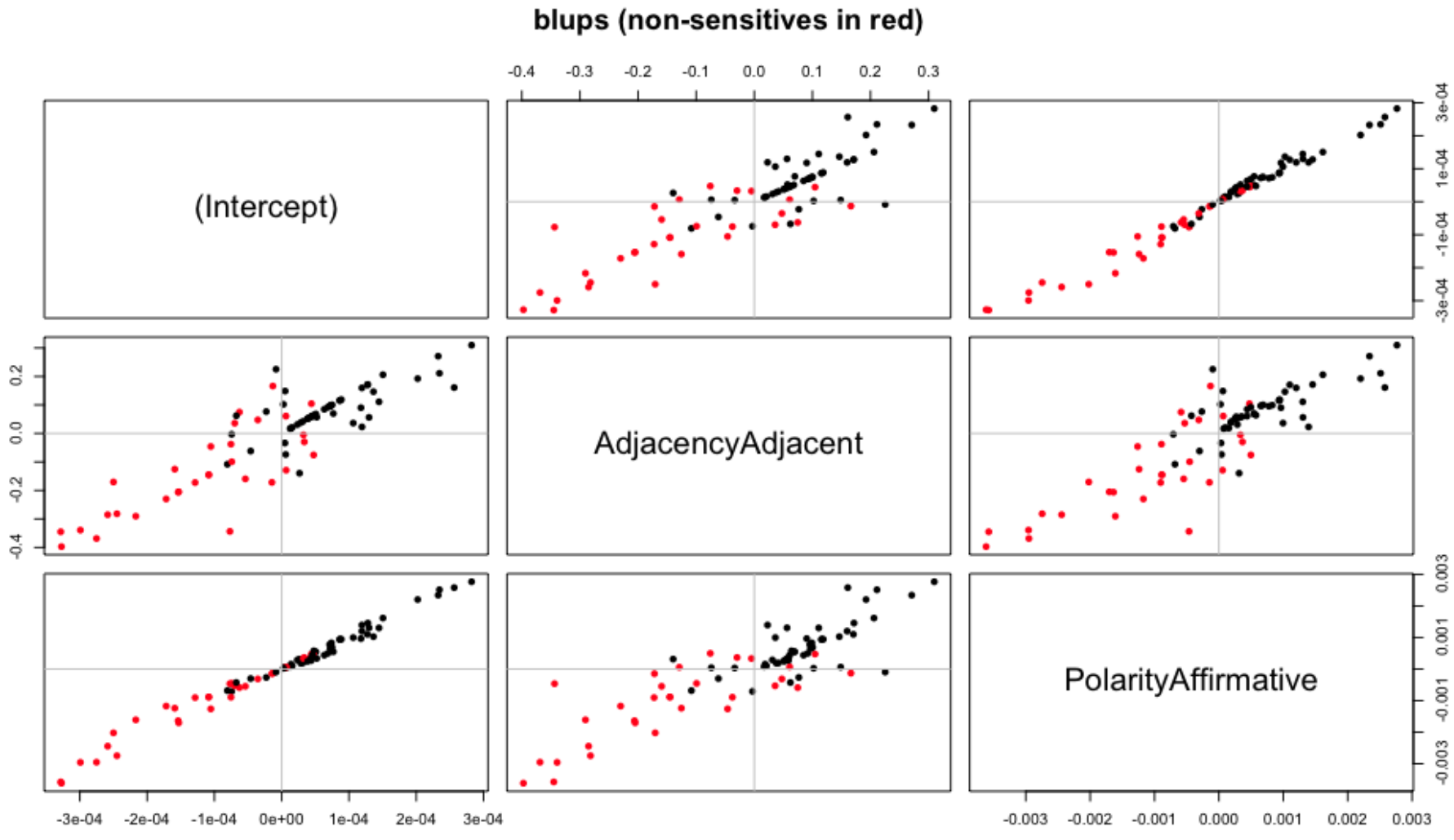

non-variability by age (decades)
20
30
40
50
60
70
80
90
FALSE
TRUE
age.bins

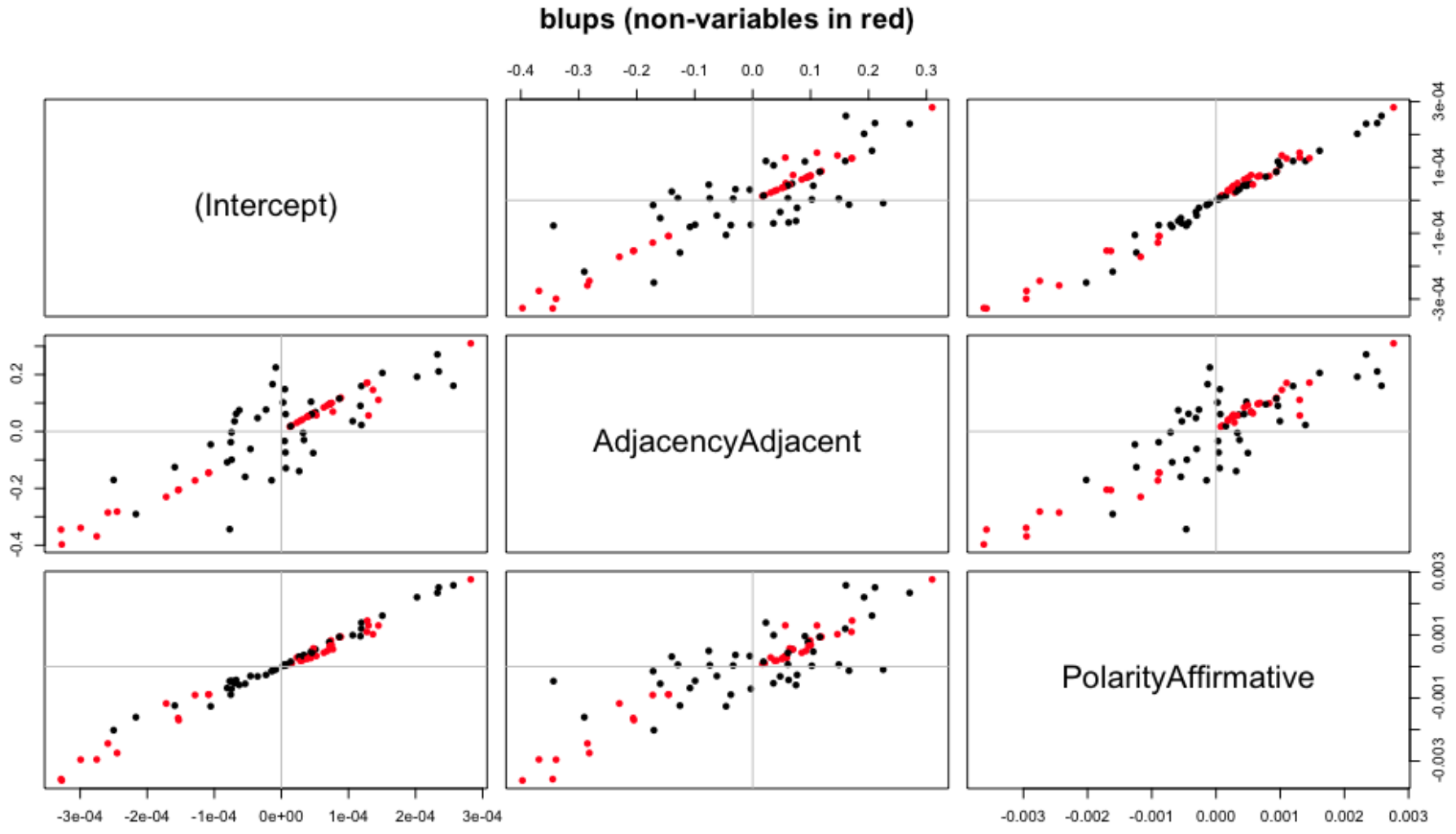

Age by non-sensitivity by form
20
30
40
50
60
70
80
90
were
was
FALSE
TRUE
age.bins


Age by non-variability by form
20
30
40
50
60
70
80
90
were
was
FALSE
TRUE
age.bins

Was/were

Can we locate the non-sensitives and the non-variables in the blups?


blups (non-sensitives in red)
(Intercept)
AdjacencyAdjacent
PolarityAffirmative
-0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3
-3e-04 -1e-04 1e-04 3e-04
-0.4 -0.2 0.0 0.2
-0.003 -0.001 0.001 0.003
-3e-04 -2e-04 -1e-04 0e+00 1e-04 2e-04 3e-04
-0.003 -0.002 -0.001 0.000 0.001 0.002 0.003

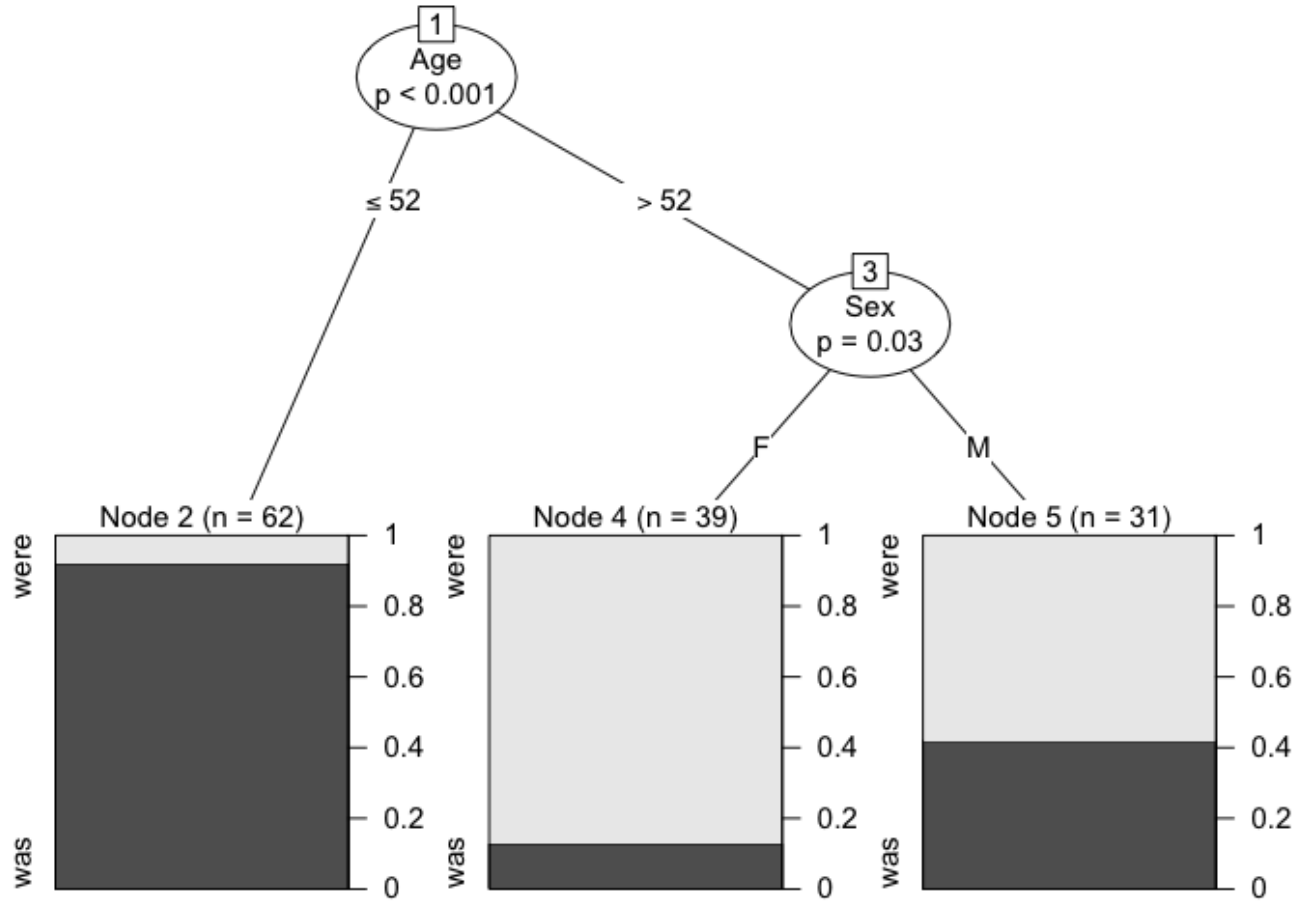

blups (non-variables in red)
(Intercept)
AdjacencyAdjacent
PolarityAffirmative
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
-3e-04
-1e-04
1e-04
3e-04
-0.4
-0.2
0.0
0.2
-0.003
-0.001
0.001
0.003
-3e-04
-2e-04
-1e-04
0e+00
1e-04
2e-04
3e-04
-0.003
-0.002
-0.001
0.000
0.001
0.002
0.003

Was/were

Can we locate the non-sensitives and the non-variables in the blups?

- Both the conditional inference tree and the mixed model agree upon the existence of a group of individuals that are indeed non-sensitives
- However, the non-variables don't behave that special in the blups (apart from the fact that they tend to avoid the neutral center in blups space)

So how then do non-variables behave differently? A conditional inference tree for just them (without 'individual' as factor, which would 'explain' everything), looks as follows.


1
Age
p < 0.001
≤ 52
> 52
3
Sex
p = 0.03
F
M
Node 2 (n = 62)
Node 4 (n = 39)
Node 5 (n = 31)
were
was
1
0.8
0.6
0.4
0.2
0

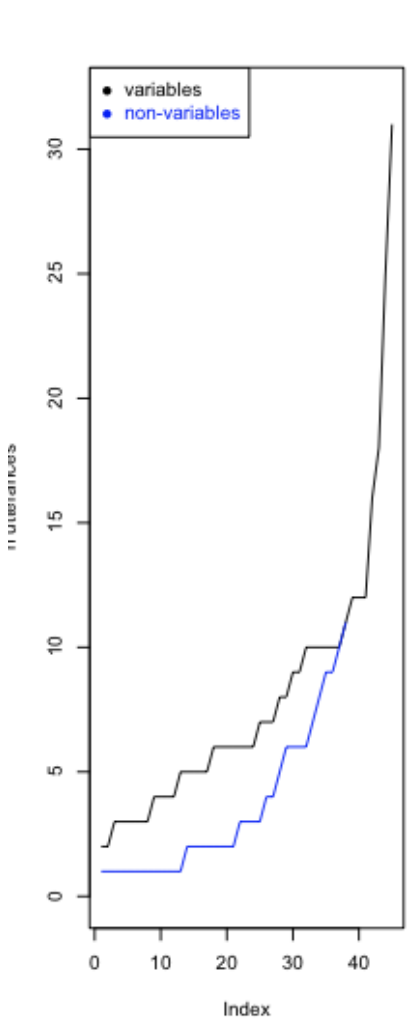
Was/were

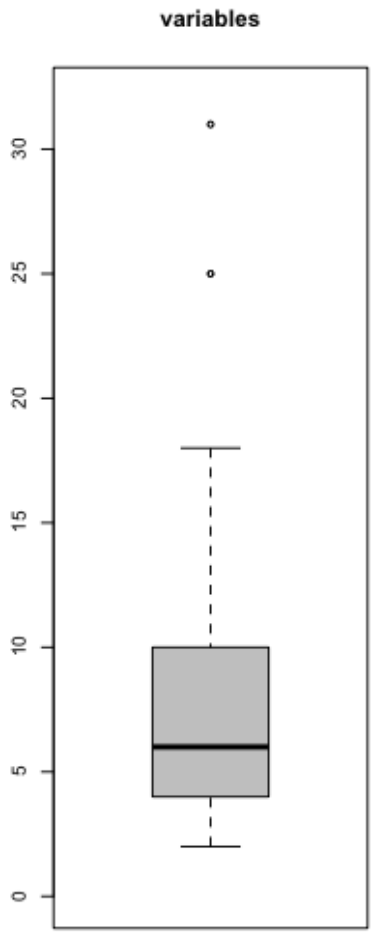
- The most important reason for the difference seems to be the failure to have a significant effect of polarity. So let's look at the relation between the group of non-variables and polarity.

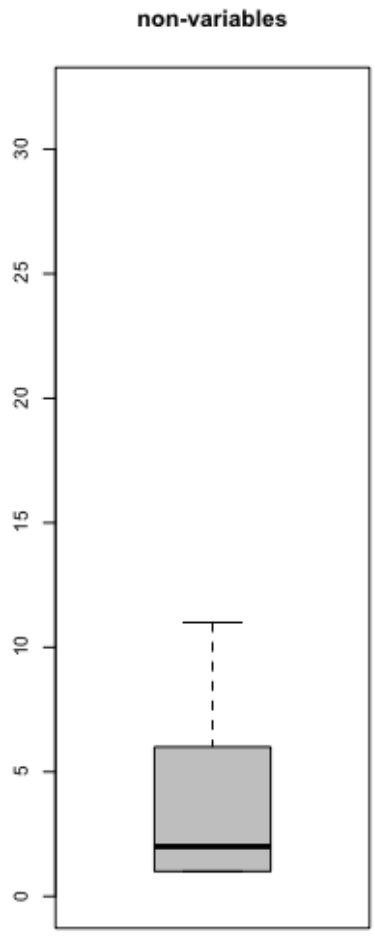

polarity by (non-)variability
FALSE
TRUE
Affirmative
Negative

Was/were

- We find a remarkable low frequency of negative polarity among non-variables (two percent, versus almost nine percent for the variables). This low frequency can explain why the factor fails to reach significance.
- But what explains the low frequency of negative polarity in this group? It turns out that the non-variables are a group that produces few utterances in the dataset (3.5 on average, versus 8 on averages for the variables), and that negative polarity is the realm of individuals who produce many utterances (perhaps because one typically needs to produce some utterances before negative polarity pops up).


variables
non-variables
0
5
10
15
20
25
30
0
10
20
30
40
Index


variables
0
5
10
15
20
25
30


non-variables
0
5
10
15
20
25
30

Overview

Questions

Is what is proposed here 'data fishing'?

Does what is proposed here guaranteed deeper insight in your data?

Questions

Is what is proposed here 'data fishing'?

- no, we don't select between models in search for the 'best one'; we stick to a single model (I prefer to make mixed models my primary tool), and then do many post hoc tests to try and better understand this model.

Does what is proposed here guaranteed deeper insight in your data?

Questions

Is what is proposed here 'data fishing'?

- no, we don't select between models in search for the 'best one'; we stick to a single model (I prefer to make mixed models my primary tool), and then do many post hoc tests to try and better understand this model.

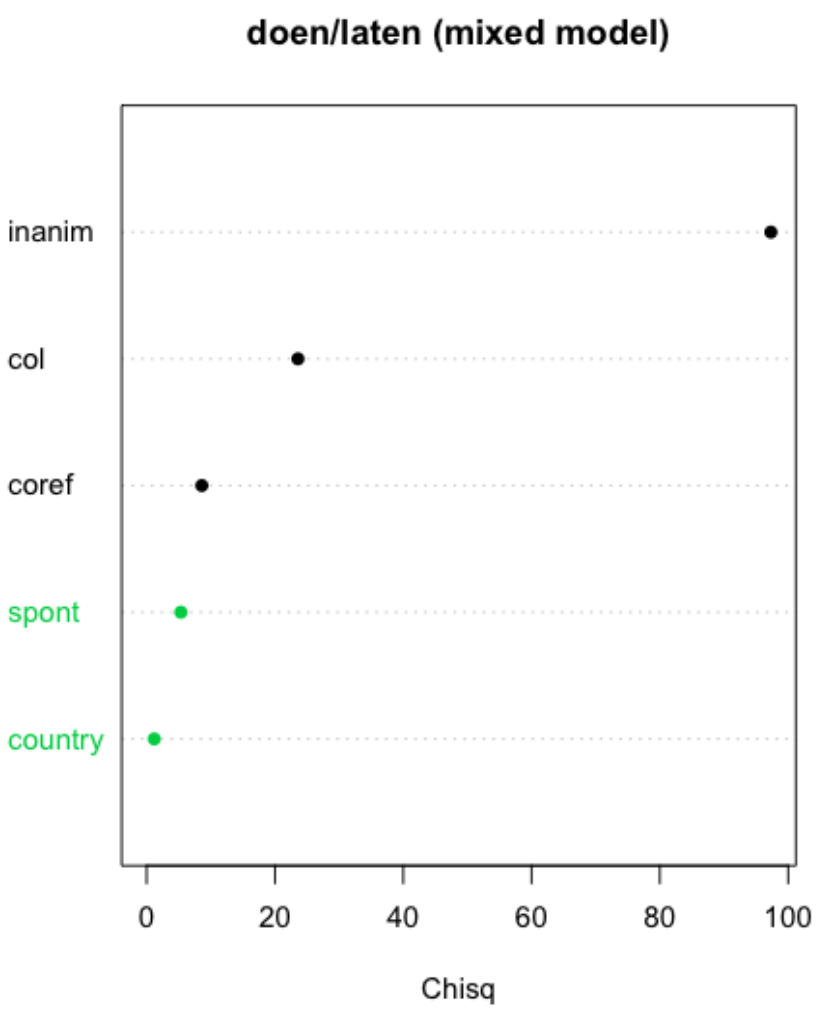
Does what is proposed here guaranteed deeper insight in your data?

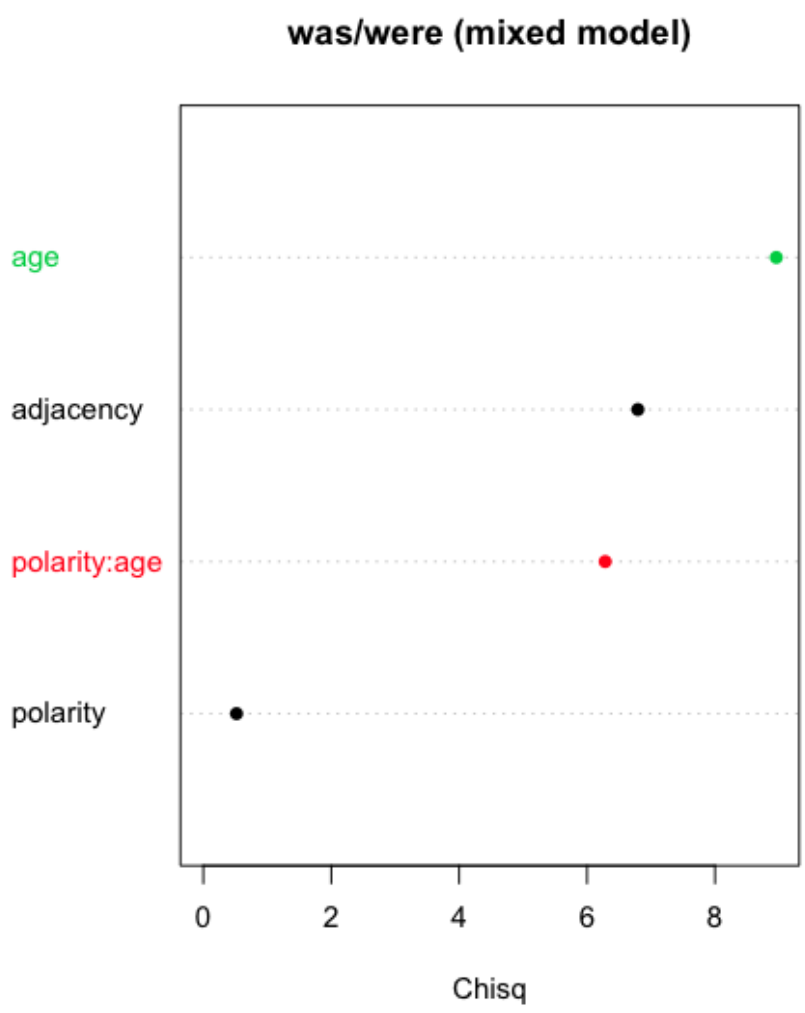
- no, there is no guarantee, but my experience is that you typically learn a lot

Overview

Conclusions

- This has been a plea to always simultaneously investigate internal and external factors.
- This way we can eventually obtain a bird's-eye view of their relative importance across (types of) variables


doen/laten (mixed model)
inanim
col
coref
spont
country
0
20
40
60
80
100
Chisq


was/were (mixed model)
age
adjacency
polarity:age
polarity
0
2
4
6
8
Chisq

Conclusions

- Methodologically it has been a plea for the use of complex models
- I have tried to illustrate that tool comparison and inspection of random effects can help ...

Conclusions

- Methodologically it has been a plea for the use of complex models
- I have tried to illustrate that tool comparison and inspection of random effects can help ...
  - gain insight in what are truly global effects and what are strong interactions

Conclusions

- Methodologically it has been a plea for the use of complex models
- I have tried to illustrate that tool comparison and inspection of random effects can help ...
  - gain insight in what are truly global effects and what are strong interactions
  - see which local patterns exist in (and complicate) our models [e.g. effects that are specific to certain verb classes, or specific to certain individuals (e.g. non-variables; non-sensitives].

Thank you!

For more information:

dirk.speelman@arts.kuleuven.be