Methods for detection of word usage over time

Ondřej Herman

FI MUNI

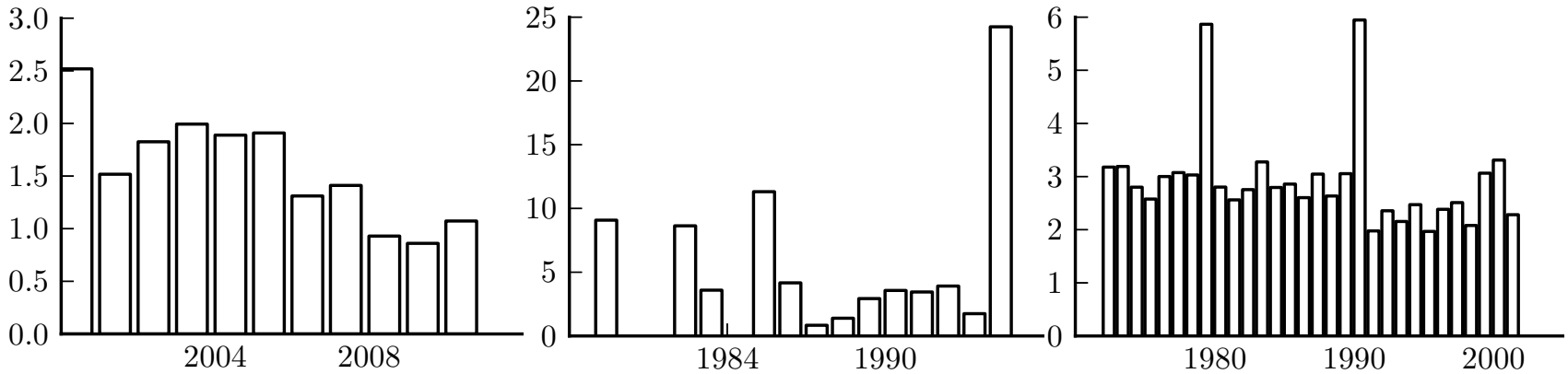
7. 12. 2013

Motivation

- natural language is not a static object
- word usage changes over time

Motivation

- natural language is not a static object
- word usage changes over time
- natural language corpora provide relevant data

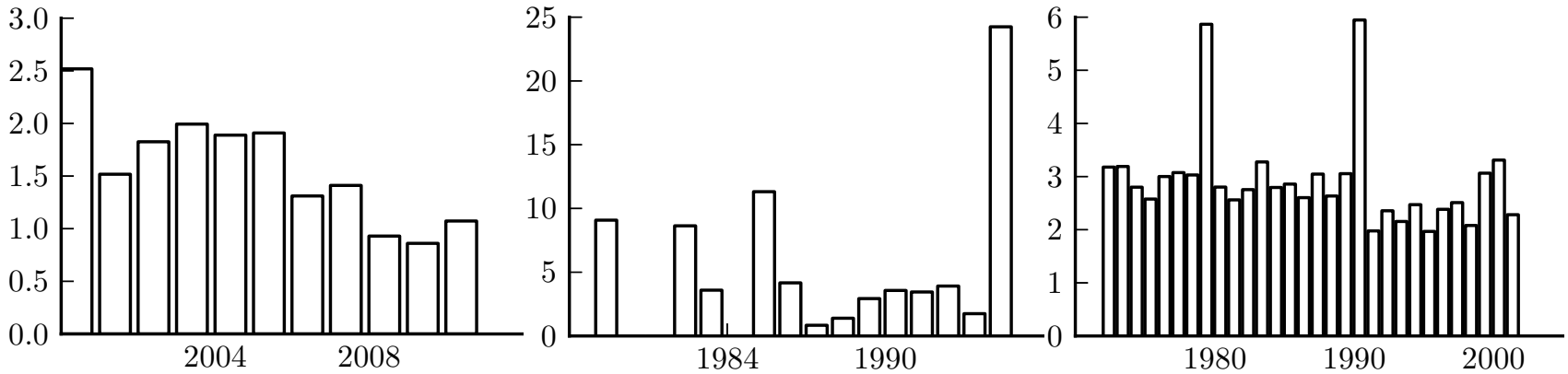


(a) OEC (b) BNC (c) Google ngrams

yearly occurences of the word 'ant'

Motivation

- natural language is not a static object
- word usage changes over time
- natural language corpora provide relevant data



(a) OEC (b) BNC (c) Google ngrams

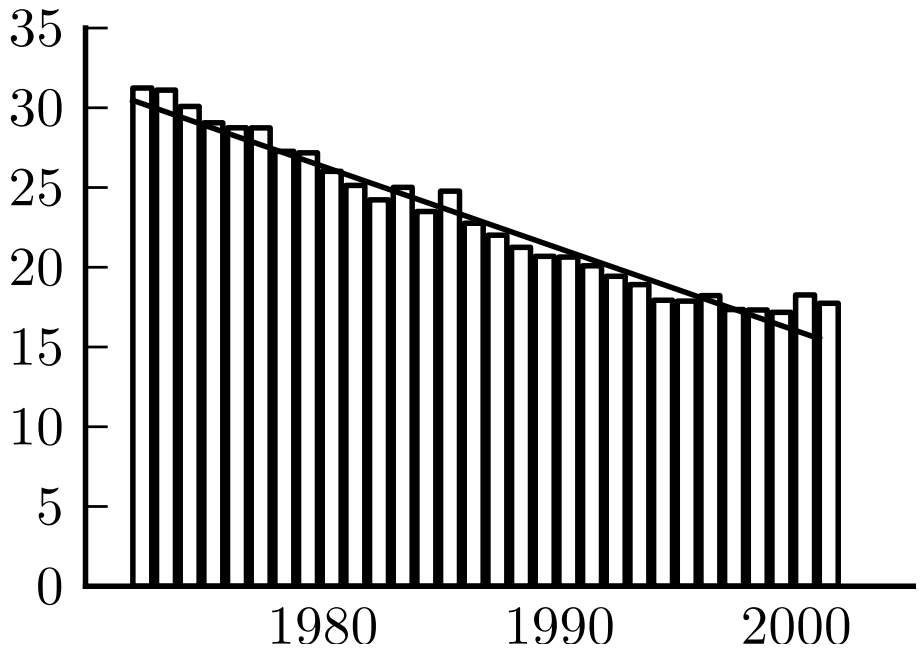
yearly occurences of the word 'ant'

- difficult to interpret

Overview

- classical least-squares regression analysis
- robust regression methods

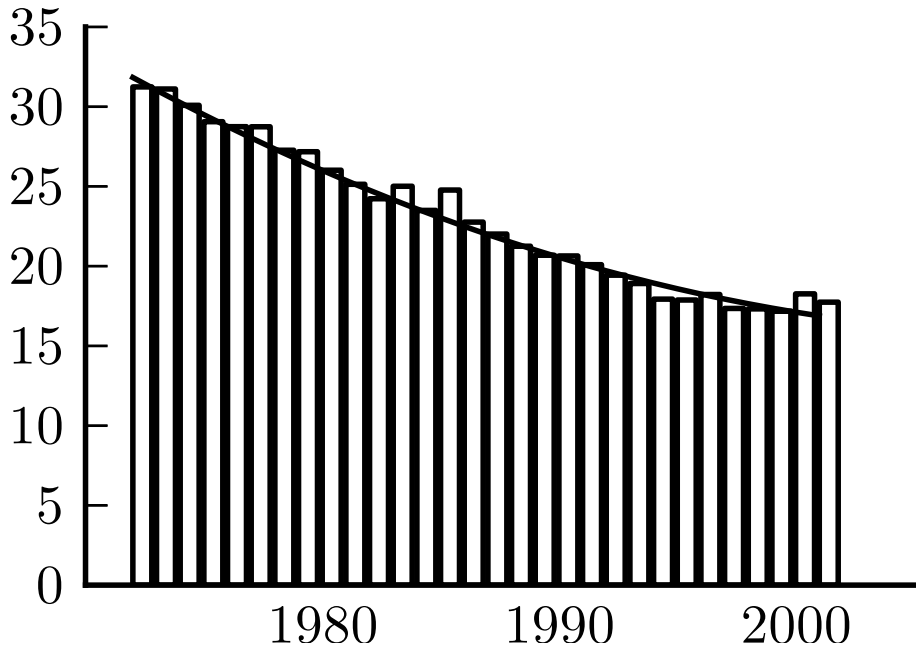
Linear regression



'slight' - Google ngrams

- simple linear model: $y = a + bx + \epsilon$
- regression line calculated using the least-squares method, that is, by minimizing the value of $e = \sum_{i=1}^{n} \epsilon_i^2$

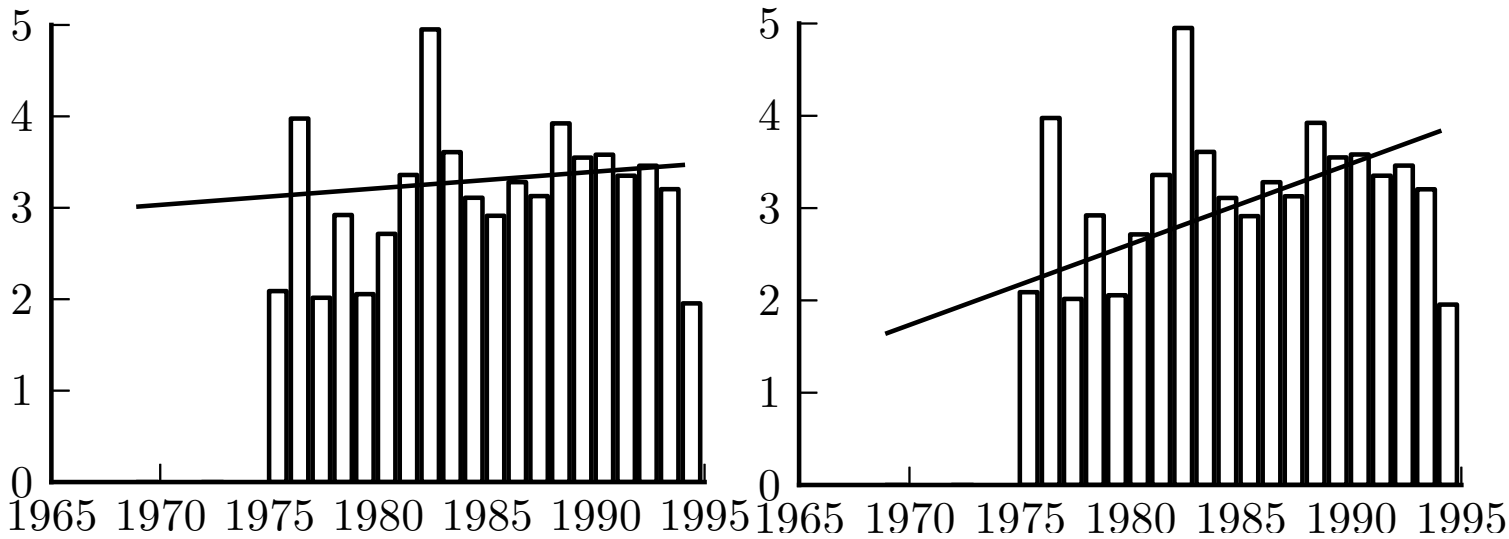
Linear regression



'slight' - Google ngrams

- polynomial model
- coefficient of determination (R^2)
- adjusted R^2

Weighted linear regression

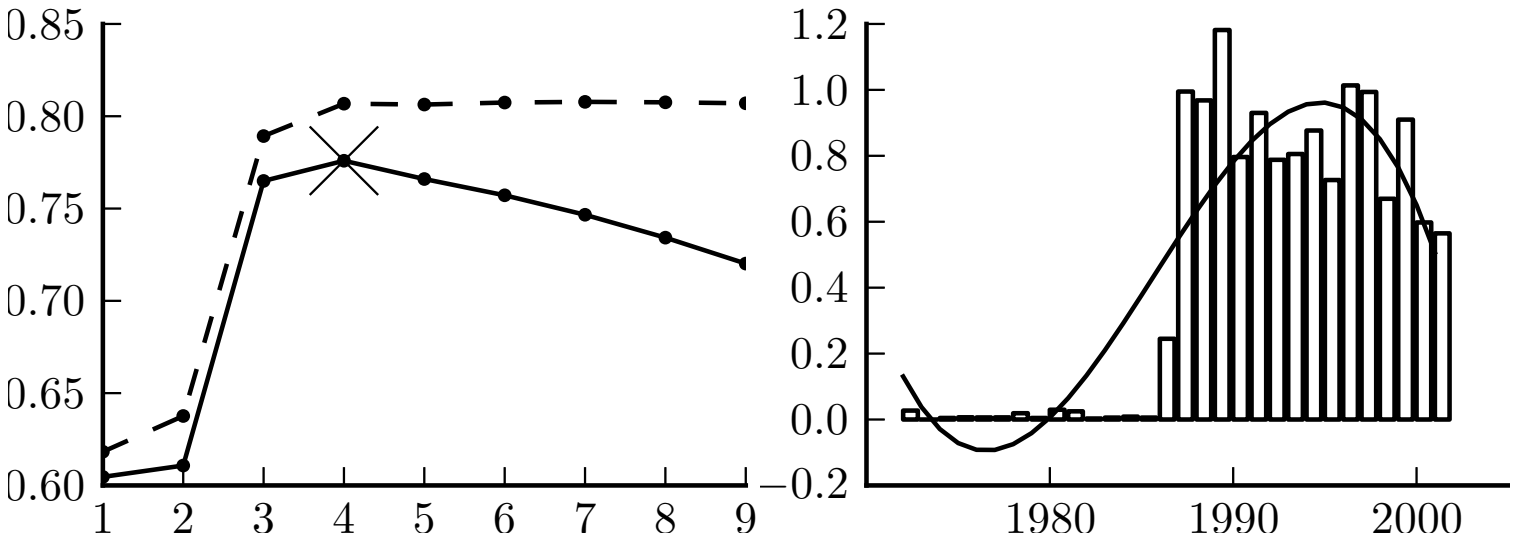


(a) W = total counts (b) W = log(total counts)

'evil' - British National Corpus

- linear model
- directly using the total counts as the weights skews the results

Weighted linear regression

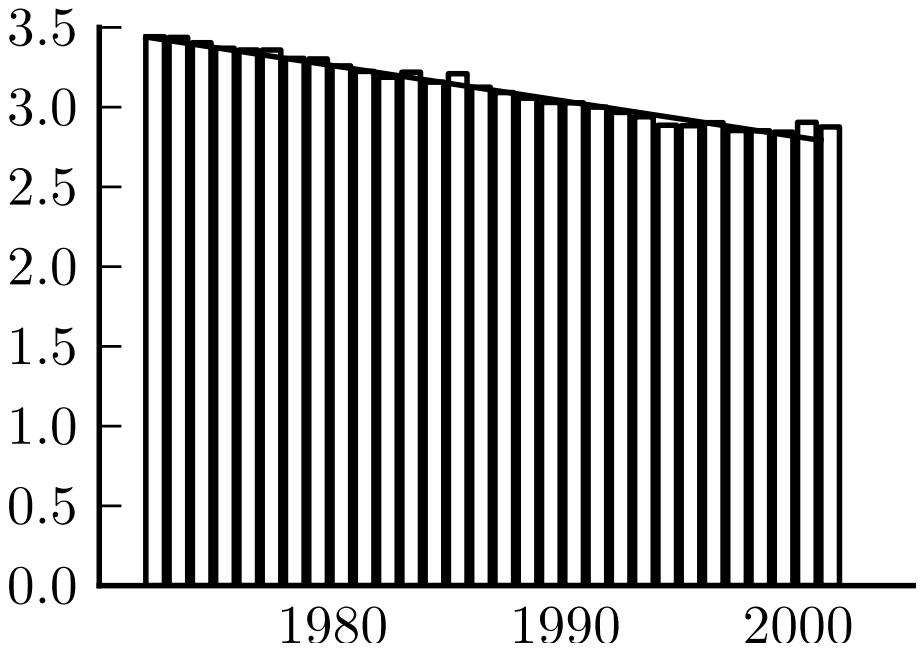


(a) adjusted R^2 (b) model with maximal R^2_{adj}

'Chernobyl' - Google ngrams

- R^2, the coefficient of determination, is the fraction of variance explained by the regression model
- R^2 increases with the degree of the regression model
- kitchen sink regression

Weighted linear regression



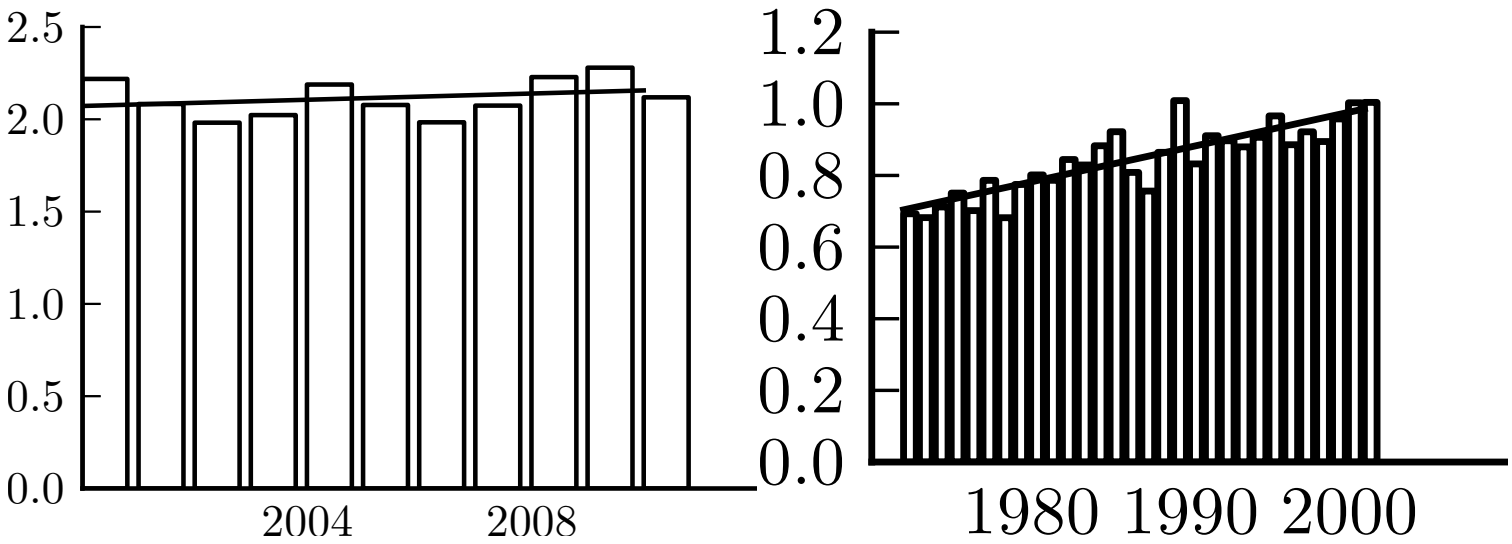
'slight' - Google ngrams

- linear model
- logarithmic transformation of frequencies

Linear regression - significance testing

- t-test
  - tests the significance of a single regression coefficient
- F-test
  - tests the significance of the whole model

Linear regression - significance testing



(a) 'steep' Oxford English Corpus, $p = 0.414$

(b) 'carrot' from Google ngrams, $p = 4.3 \times 10^{-10}$

example F-test p-values

H_0 : the mean predicts the behavior of the series well

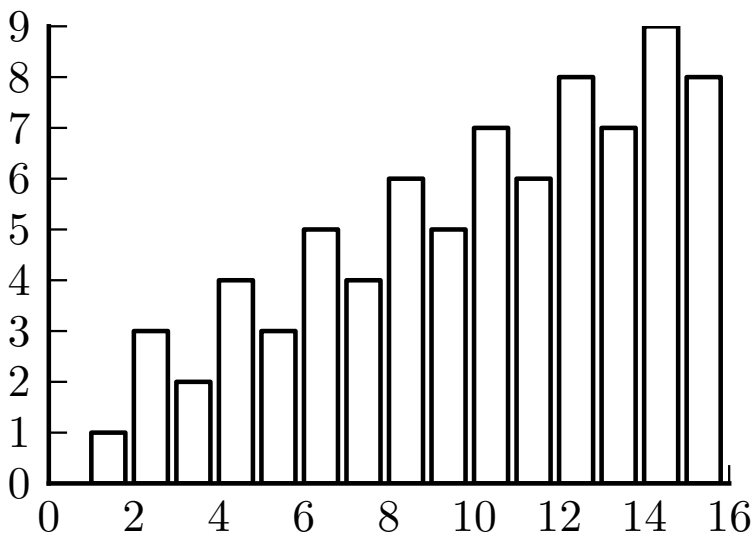
H_1 : the given regression model predicts the behavior well

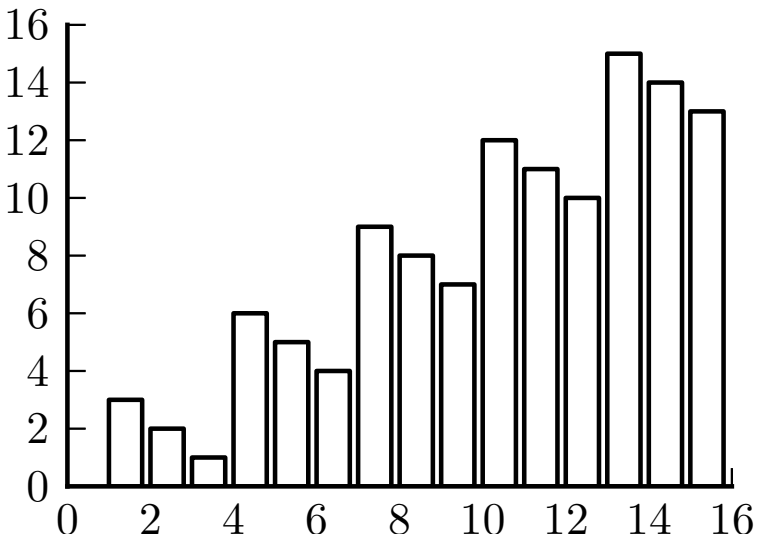
Robust regression

- Moore-Wallis test
- Mann-Kendall test
- Spearman's ρ
- Theil-Sen metod

Moore-Wallis test

- also known as the sign-difference test



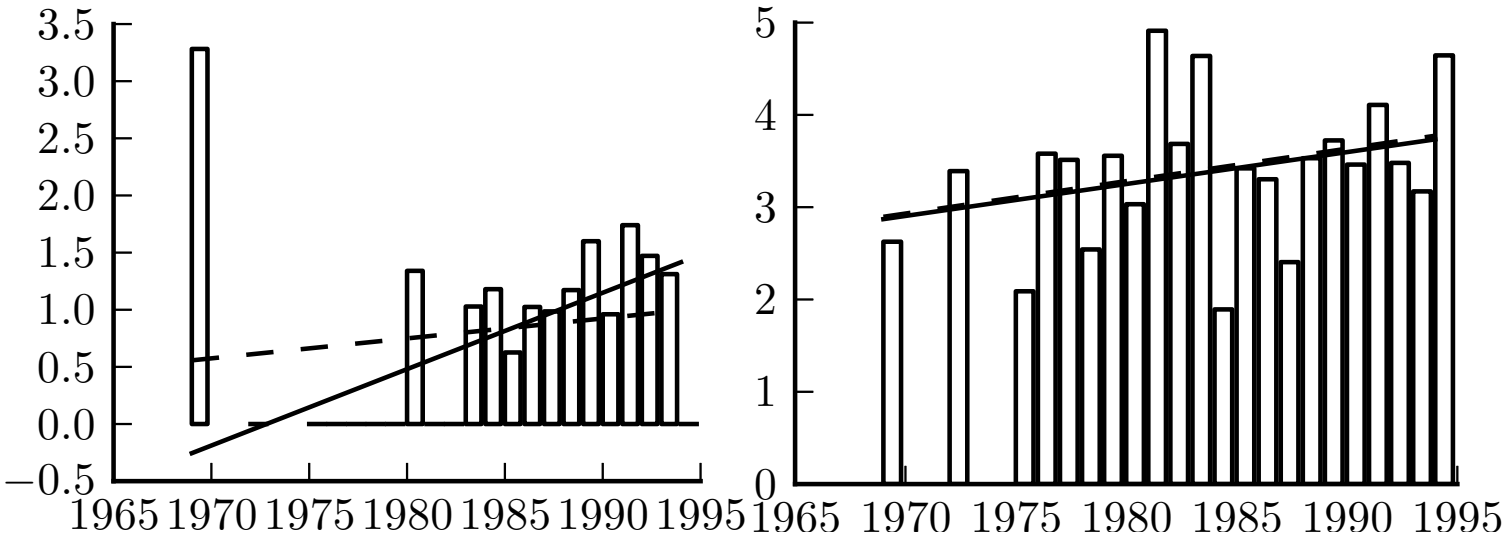


- no trend is detected in the first series, a downward trend is detected in the second series
- asymtotically optimal
- on short series the power of the test is low

Theil-Sen estimator

- defined as the median of the pairwise slopes of the samples:

$$b' = \operatorname{med} \frac{y_i - y_j}{x_i - x_j}, \quad i \neq j$$



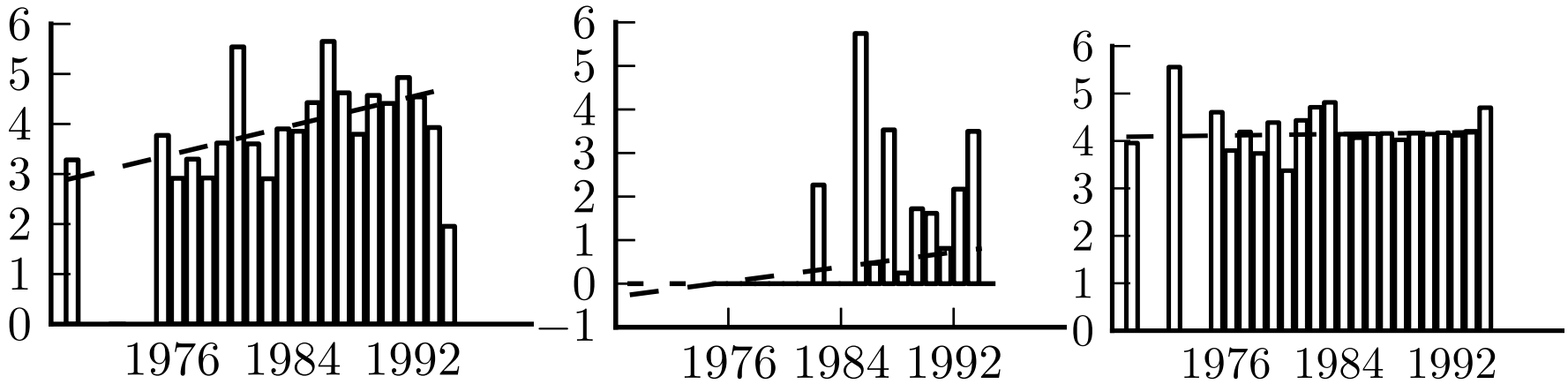
(a) 'spice' (b) 'snow'

Behavior of the Theil-Sen estimator for words encountered in the British National Corpus

Mann-Kendall test

- used to test the significance of a regression model fitted using the Theil-Sen estimator

$$S = \sum_{i=1}^{n} \sum_{j=1}^{i} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j)$$



(a) 'oil', $p = 0.021$ (b) 'disk', $p = 0.009$ (c) 'slow', $p = 0.821$

Words from the British National Corpus tested using the Mann-Kendall test with the trend line fitted using the Theil-Sen estimator

Spearman's ρ

- calculated as the correlation coefficient of a linear model obtained by using the rank of the observations instead of the actual value
- yields almost the same results as the Mann-Kendall test
- the distribution of the test scores is more difficult to calculate

Slope normalization

- the slope estimates are not directly comparable, they need to be normalized

$$d = \frac{b'}{\bar{y}}$$

where $\hat{b}$ is the estimated slope and $\bar{y}$ is the mean of y, the observed frequencies.

On the next slide: the slopes obtained from Google ngrams of the 50 most common words from the Oxford English Corpus ordered by the slope relative to the mean d

word	d	$\hat{b}$
which	−1.256	−36.231
been	−0.862	−13.977
his	−0.836	−23.698
he	−0.804	−20.531
It	−0.744	−9.081
were	−0.713	−16.541
be	−0.69	−33.798
by	−0.669	−31.319
there	−0.645	−7.4
was	−0.601	−31.296
has	−0.572	−9.966
of	−0.527	−171.39
had	−0.512	−12.061
would	−0.5	−7.105
all	−0.496	−8.423
but	−0.451	−8.853
one	−0.427	−7.862
not	−0.422	−14.476
the	−0.4	−194.423
it	−0.381	−15.137
will	−0.359	−4.771
is	−0.346	−30.998
at	−0.338	−10.113
The	−0.326	−20.124
this	−0.308	−8.888

word	d	$\hat{b}$
in	−0.302	−50.594
have	−0.249	−6.764
who	−0.209	−2.853
are	−0.206	−8.2
more	−0.2	−3.146
from	−0.178	−6.051
to	0.0	0.0
and	0.0	0.0
a	0.0	0.0
for	0.0	0.0
on	0.0	0.0
with	0.0	0.0
as	0.0	0.0
an	0.0	0.0
or	0.0	0.0
they	0.0	0.0
we	0.0	0.0
their	0.0	0.0
said	0.0	0.0
that	0.098	7.687
up	0.279	2.488
I	0.488	15.796
about	0.504	5.583
can	0.678	10.541
you	1.523	23.465

Future work

- anomaly detection
- piecewise linear model

Conclusion

- Mann-Kendall test together with the Theil-Sen estimator give the best results
- standard linear regression model gives satisfactory results most of the time