A compositional approach to sociophonetic variation

Daniel Ezra Johnson, Lancaster University
Carmen Llamas, University of York
Dominic Watt, University of York

Sociophonetic analysis

- Data not inherently compositional
- Many observations (tokens) of a variable
- Regression analysis
    - Which predictors affect the response?
    - What is their effect?

Sociophonetic responses

- Continuous
    - vowel shifts, VOT, suprasegmentals
- Binary
    - most common: (ing), (td)
- More than 2 Categories
    - less common: Scottish (r), Spanish (s)
    - but also avoided due to stats

Sociophonetic predictors

- categorical and continuous (not comp.!)
- "social factors"
    - many grouped by speaker
- "linguistic factors"
    - many grouped by word

Regression models

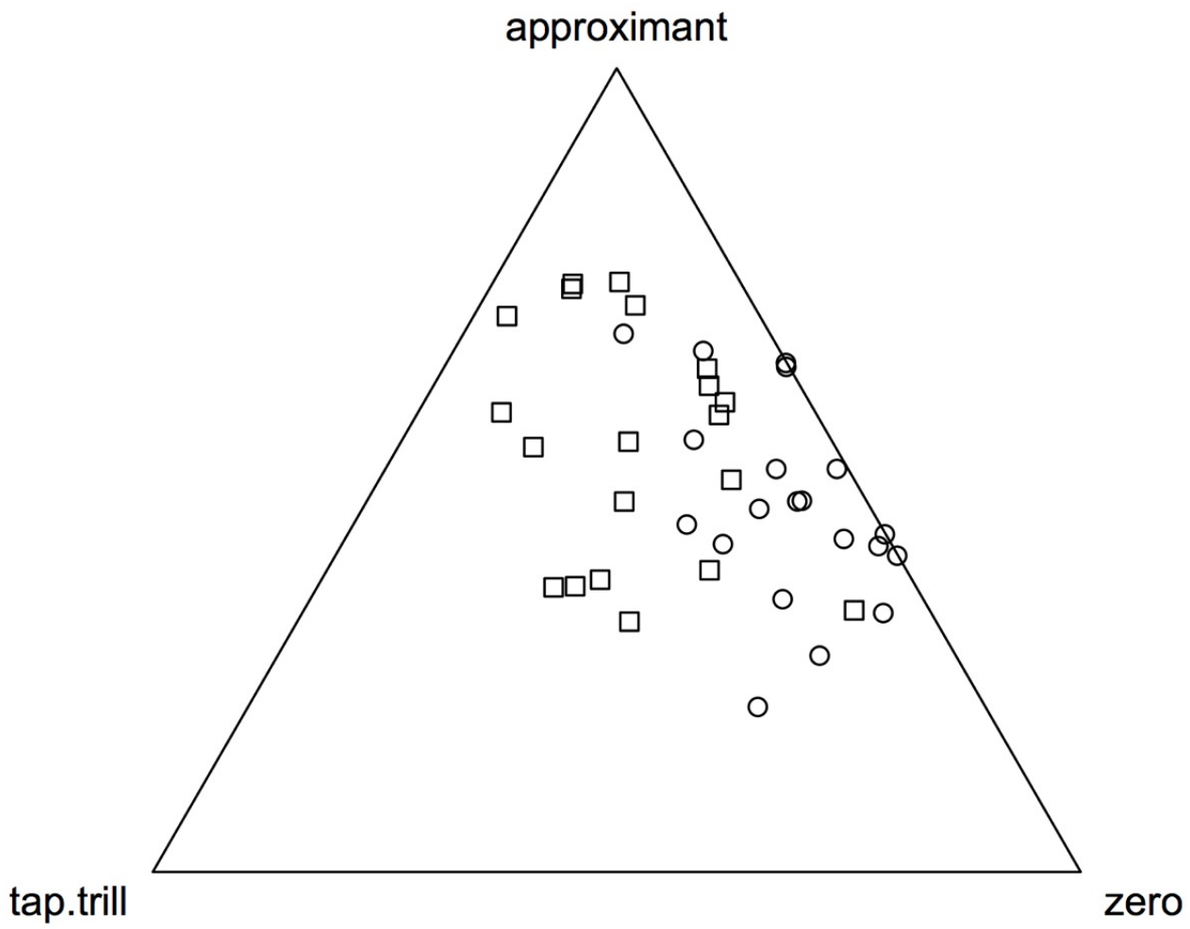
- Logistic regression
- Mixed-effects logistic regression
    - Random effect(s) for speaker
- *Multinomial logistic regression*
- *Mixed-effects multinomial logistic regression?*
- *CoDa linear regression: by-speaker composition?*

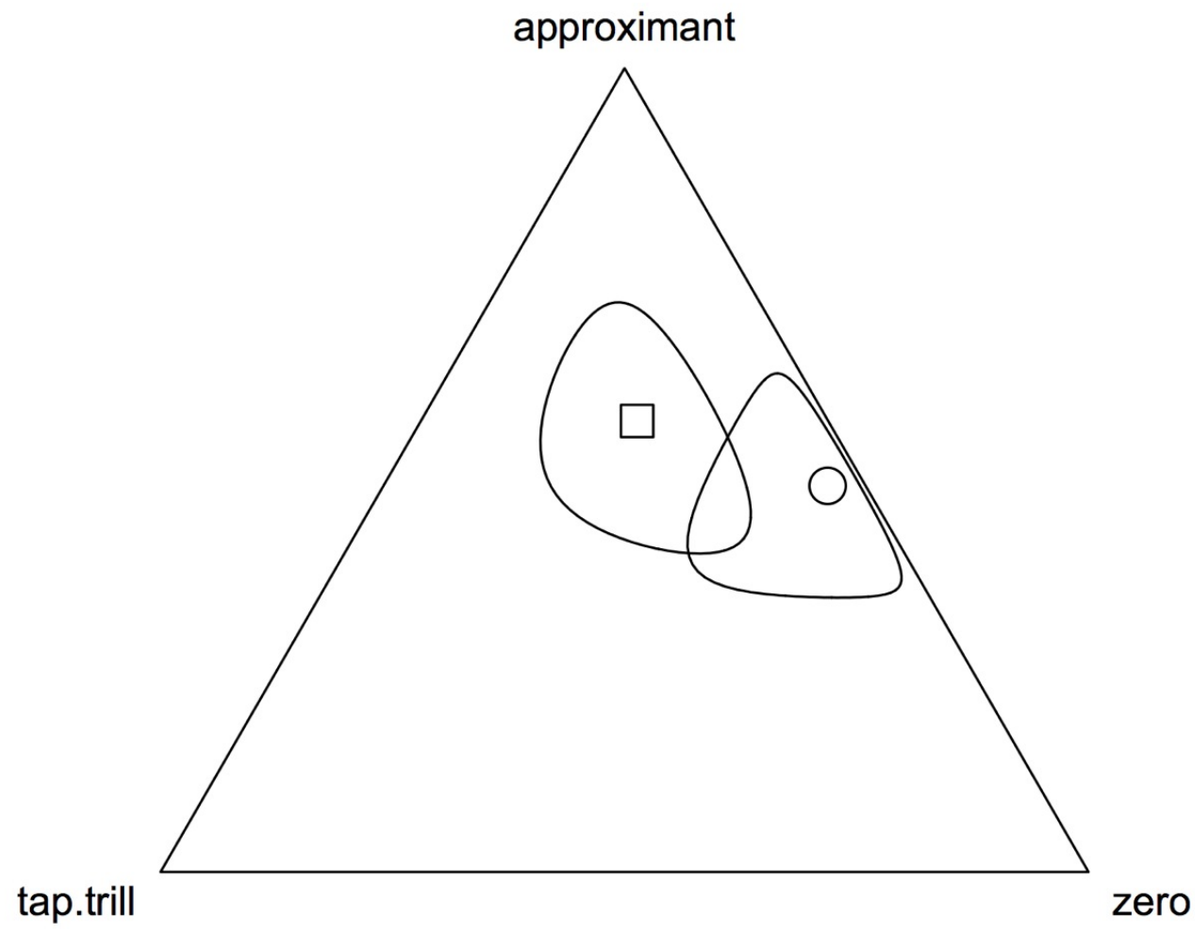
A sociophonetic data set

- AISEB: how identity relates to accent
- 160 speakers = 4 localities x 40 speakers
- 2 age groups x 2 classes x 2 genders
- 5 speakers / cell
- ~350 tokens of (r) per speaker

40 speakers from Gretna

((tap > approximant) > zero)

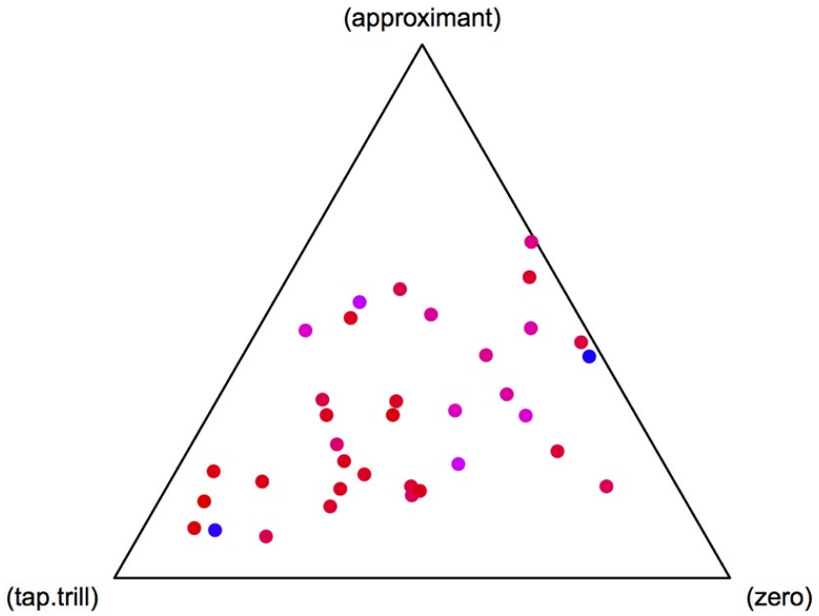




squares: ages 15-27; circles: ages 57-82

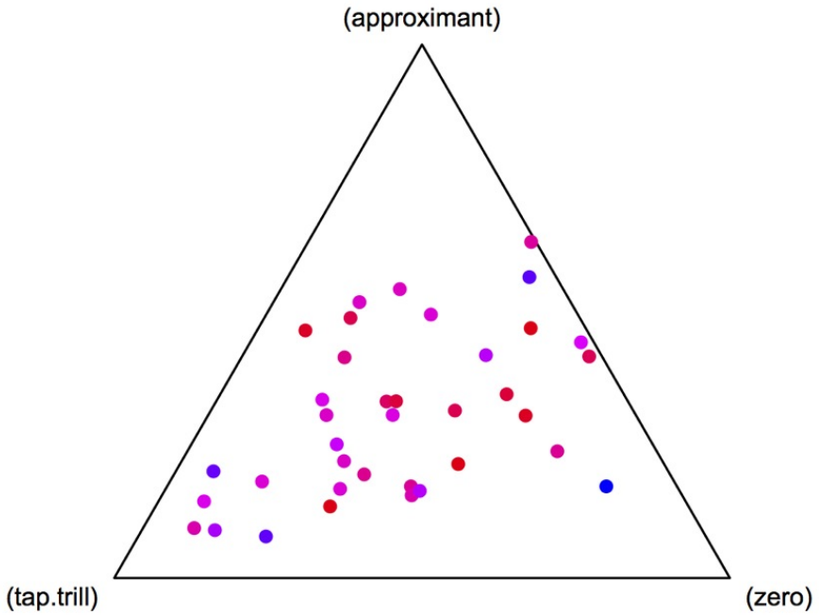
CoDa linear regression

residuals from r ~ age * class * gender
residuals ~ identity score



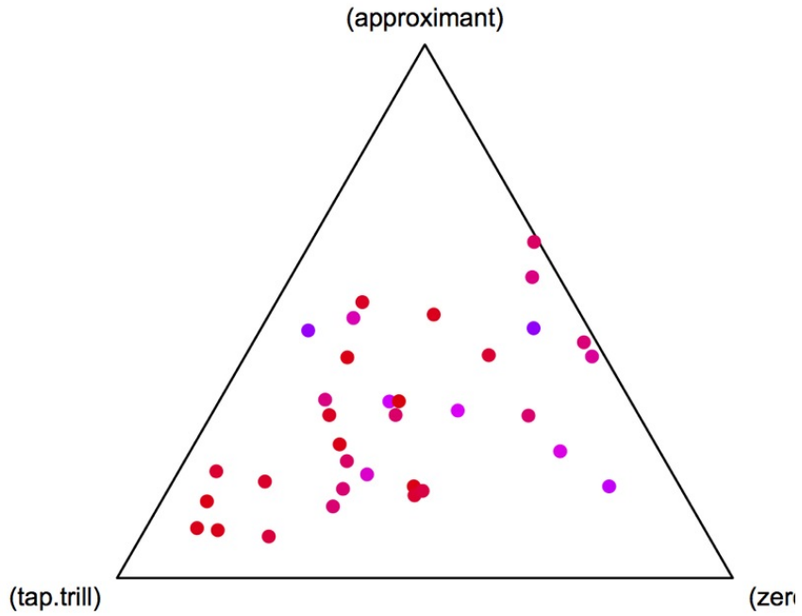
Scottish

not Scottish



British

not British

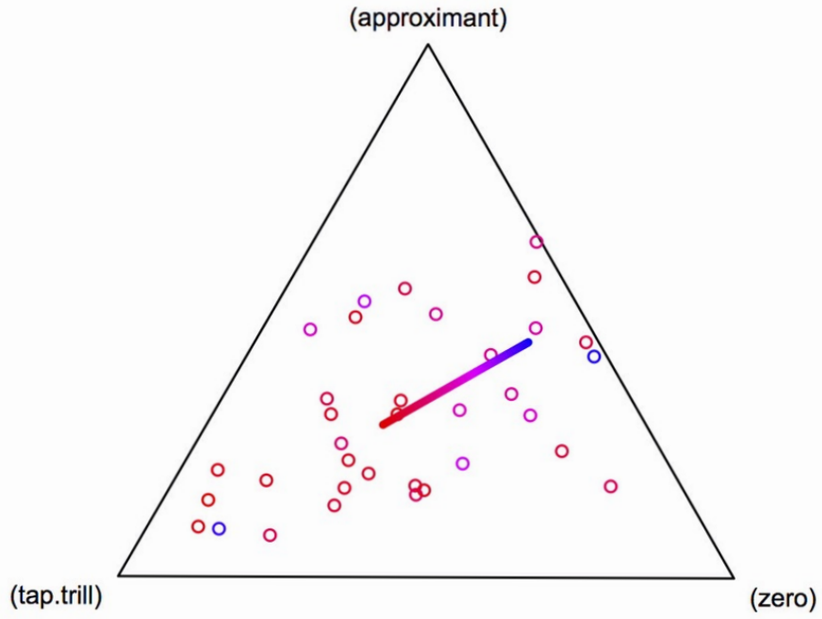
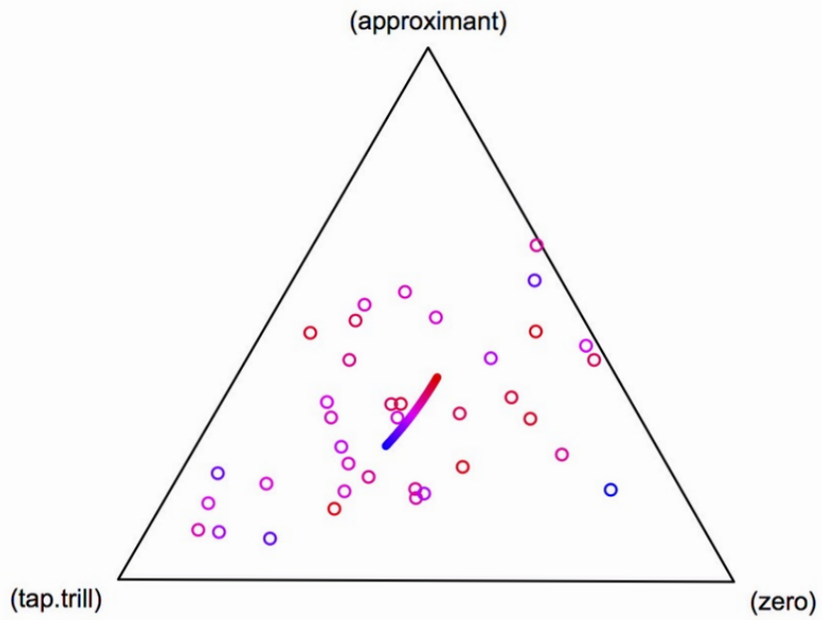
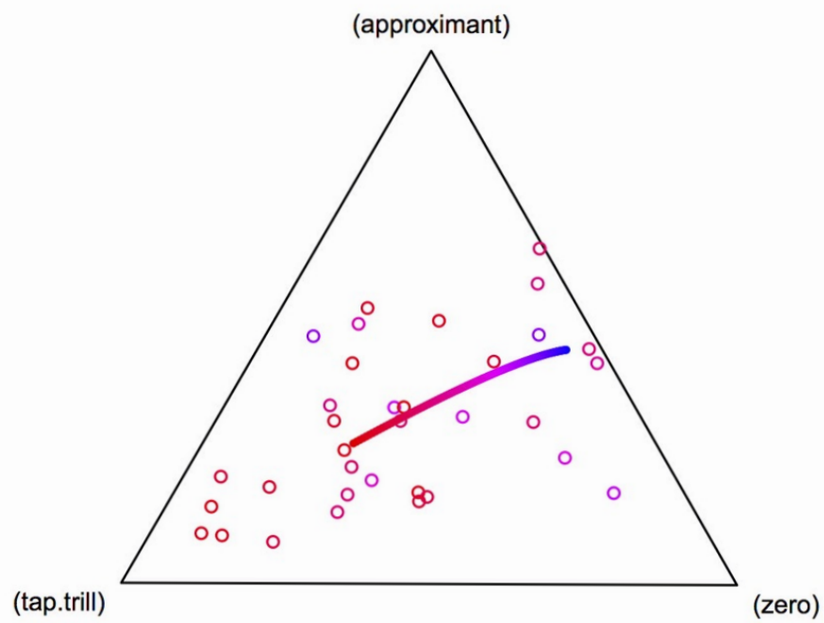


Gretna

not Gretna

CoDa linear regression

residuals ~ identity score

Scottish	British	Gretna
		
b=(-0.72, -1.25)	(0.05, 0.59)	(-1.26, -1.89)
R^2 = .071	.017	.106
p = .29	.45	.15

However...

- Averaging by speaker loses data
- We are also interested in (or need to control for) within-speaker variables of several types
- Easier? Style divides each speaker's data
- Harder? Syllable-position, preceding vowel divide each speaker but also grouped by word

How to proceed

- Keep using by-speaker compositions
    - Incorporate within-speaker variables?
    - Incorporate by-word grouping?
- Go back to a case-wise approach
    - Incorporate SBP partition ideas into (mixed) multinomial logistic regression?