Regional variation in 140 characters:

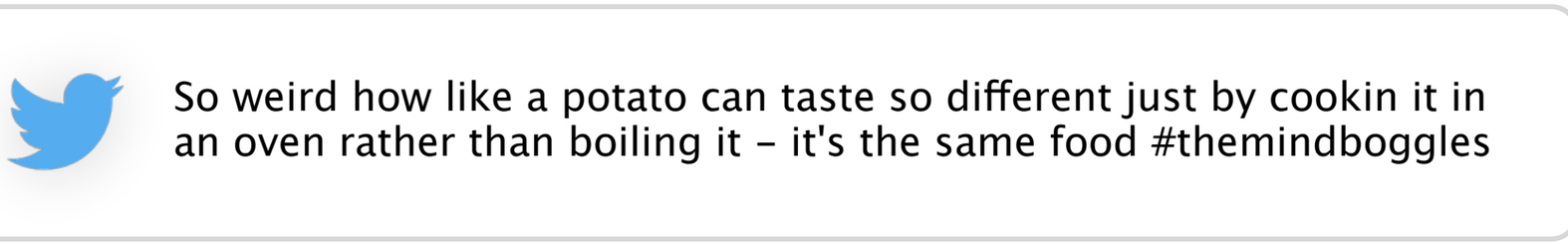
Mapping geospatial tweets

George Bailey
University of Manchester
@grbails

Using Twitter for Linguistic Research - University of Kent, 31st May 2016

What is Twitter?

- Social media platform where users post short, 140-character messages called ‘Tweets’



- In most cases, Tweets are visible to anyone

- You can choose to ‘follow’ other users, so that their Tweets populate your timeline

- Noted for viral content, internet memes, early adoption (innovation?) of online slang terms, hashtags etc.

Why study it?

- Great source of natural language data
    - more than **500 million** tweets sent each day (Twitter 2015)
        - as of 2013, seven years after its founding, over **170 billion** tweets had been sent (Leetaru et al. 2013)
- Easy to collect, just leave the script running!
- Informal style, which leads to lots of variation and creative use of language
- **Lots** of metadata…

Metadata

{"created_at":"Mon May 09 07:59:23 +0000 2016","id":729581517257740288,"id_str":"729581517257740288","text":"**Dumb idea. Won't achieve anything**. https:\/\/t.co\/EGOno9VXNw","source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":28327397,"id_str":"28327397","name":"Amy\u270c\ufe0f","screen_name":"amycfc90","location":"Birmingham, England","url":"http:\/\/fiskwilson.tumblr.com\/","description":"Gamer, geek, avid film and TV watcher. @ChelseaFC, @BCFC, @packers. Whovian. Marvel. DC. Work in retail. Rock, punk, metal, screamo, metalcore, EDM.","protected":false,"verified":false,"followers_count":2664,"friends_count":2736,"listed_count":89,"favourites_count":1110,"statuses_count":76062,"created_at":"Thu Apr 02 12:25:38 +0000 2009","utc_offset":3600,"time_zone":"London","geo_enabled":true,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"1A1B1F","profile_background_image_url":"http:\/\/pbs.twimg.com\/profile_background_images\/560595733749825538\/O_A_CGtm.jpeg","profile_background_image_url_https":"https:\/\/pbs.twimg.com\/profile_background_images\/560595733749825538\/O_A_CGtm.jpeg","profile_background_tile":false,"profile_link_color":"3B94D9","profile_sidebar_border_color":"FFFFFF","profile_sidebar_fill_color":"252429","profile_text_color":"666666","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/727275189772423169\/MN80I6sN_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/727275189772423169\/MN80I6sN_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/28327397\/1462736107","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":{"id":"53b67b1d1cc81a51","url":"https:\/\/api.twitter.com\/1.1\/geo\/id\/53b67b1d1cc81a51.json","place_type":"city","name":"Birmingham","full_name":"Birmingham, England","country_code":"GB","country":"United Kingdom","bounding_box":{"type":"Polygon","coordinates":[[[-2.033651,52.381063],[-2.033651,52.606870],[-1.747630,52.606870],[-1.747630,52.381063]]]},"attributes":{}},"contributors":null,"quoted_status_id":729408069961191425,"quoted_status_id_str":"729408069961191425","quoted_status":{"created_at":"Sun May 08 20:30:10 +0000 2016","id":729408069961191425,"id_str":"729408069961191425","text":"Fans beginning to plan a 26th minute walk out during our last 2 #BPL games for John Terry.. A good idea? #CFC https:\/\/t.co\/VY77JPBZAj","source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":3302879767,"id_str":"3302879767","name":"BreatheChelsea","screen_name":"Breathechels","location":"WorldWide","url":"http:\/\/www.breathechelsea.com","description":"Breathe Chelsea's official Twitter! Check out our website, Instagram, and our verified Facebook page for the full Breathe Chelsea experience! #CFC","protected":false,"verified":false,"followers_count":560,"friends_count":67,"listed_count":10,"favourites_count":133,"statuses_count":4124,"created_at":"Fri Jul 31 23:44:18 +0000 2015","utc_offset":null,"time_zone":null,"geo_enabled":false,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"C0DEED","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_link_color":"0084B4","profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/718320511911071744\/qWH-Kvwf_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/718320511911071744\/qWH-Kvwf_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/3302879767\/1460098189","default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_status":false,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[{"text":"BPL","indices":[64,68]},{"text":"CFC","indices":[105,109]}],"urls":[],"user_mentions":[],"symbols":[],"media":[{"id":729408062449168384,"id_str":"729408062449168384","indices":[110,133],"media_url":"http:\/\/pbs.twimg.com\/media\/Ch9gxKqWsAAJkw6.jpg","media_url_https":"https:\/\/pbs.twimg.com\/media\/Ch9gxKqWsAAJkw6.jpg","url":"https:\/\/t.co\/VY77JPBZAj","display_url":"pic.twitter.com\/VY77JPBZAj","expanded_url":"http:\/\/twitter.com\/Breathechels\/status\/729408069961191425\/photo\/1","type":"photo","sizes":{"medium":{"w":600,"h":602,"resize":"fit"},"thumb":{"w":150,"h":150,"resize":"crop"},"small":{"w":340,"h":341,"resize":"fit"},"large":{"w":639,"h":641,"resize":"fit"}}}]},"extended_entities":{"media":[{"id":729408062449168384,"id_str":"729408062449168384","indices":[110,133],"media_url":"http:\/\/pbs.twimg.com\/media\/Ch9gxKqWsAAJkw6.jpg","media_url_https":"https:\/\/pbs.twimg.com\/media\/Ch9gxKqWsAAJkw6.jpg","url":"https:\/\/t.co\/VY77JPBZAj","display_url":"pic.twitter.com\/VY77JPBZAj","expanded_url":"http:\/\/twitter.com\/Breathechels\/status\/729408069961191425\/photo\/1","type":"photo","sizes":{"medium":{"w":600,"h":602,"resize":"fit"},"thumb":{"w":150,"h":150,"resize":"crop"},"small":{"w":340,"h":341,"resize":"fit"},"large":{"w":639,"h":641,"resize":"fit"}}}]},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low","lang":"en"},"is_quote_status":true,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[{"url":"https:\/\/t.co\/EGOno9VXNw","expanded_url":"https:\/\/twitter.com\/breathechels\/status\/729408069961191425","display_url":"twitter.com\/breathechels\/s\u2026","indices":[35,58]}],"user_mentions":[],"symbols":[]},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low","lang":"en","timestamp_ms":"1462780763340"}

Overview of this talk

- Big methodological aspect
    - Mining tweets
    - Cleaning up data
    - Geotagging
    - Mapping techniques
- But also some results!
    - Lexical variation, i.e. dialect maps
    - ~~Phonological variation~~ ~~Orthographic variation?~~

      Orthographic reflections of a phonological variable

Methodology

Data collection

- Python script to access Twitter streaming API
    - Grants you free access to a sample of all tweets sent in real-time (between 1% to 40%)
- Further restricted to tweets with geographic metadata (around 2% of all tweets), using a bounding box of the UK
- Extracts a number of fields from the metadata:
    - time/date
    - handle (e.g. @grbails) and full name (e.g. George Bailey)
    - latitude / longitude
- Rest of metadata saved to a separate file, for two reasons:
    - Might need it at a later date
    - R crashes otherwise

Methodology

Cleaning up and POS-tagging

- Cleaning up data
    - Removing ‘prolific’ tweeters for fears of imbalance in the corpus
    - Removing automated tweets (making up almost 25% of all data)

ThurrockWeather –
0100hrs Forecast: Fine weather. Temp: 14.7C. UV:0. Baro:1012.0hpa. Steady WindGust:4.9mph. Rain2Day:0.0mm.

Weather forecasts

OxfordSolarLive –
Solar Realtime Event: 66 watts.

Solar activity

TrafficStAlbans –
Area A1 southbound within the A5135 junction | Southbound | Congestion: Location : The A1 #stalbans #harpenden

Traffic alerts

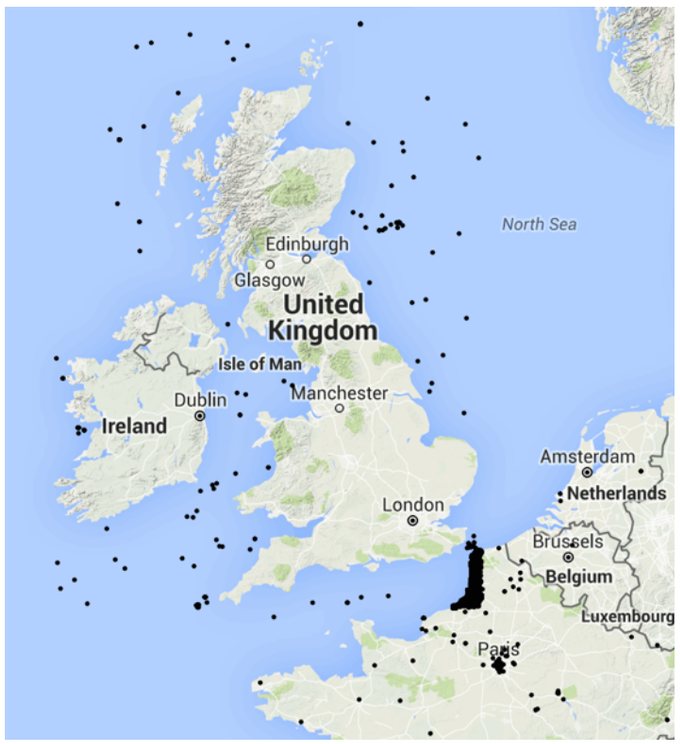
- POS-tagging
    - twitie-tagger (Derczynski et al. 2013) uses the Penn Treebank tagset and has an accuracy rate of 91%
    - Errors are largely systematic, e.g. identification of proper nouns too heavily influenced by initial-grapheme capitalisation:
        - e.g. “Can’t. Stop. **Eating**.”
    - Can deal with Twitter-specific peculiarities like hashtags (#lol_**HT**), usernames (@morrissey_**USR**), and hyperlinks (http://www.google.co.uk/_**URL**)

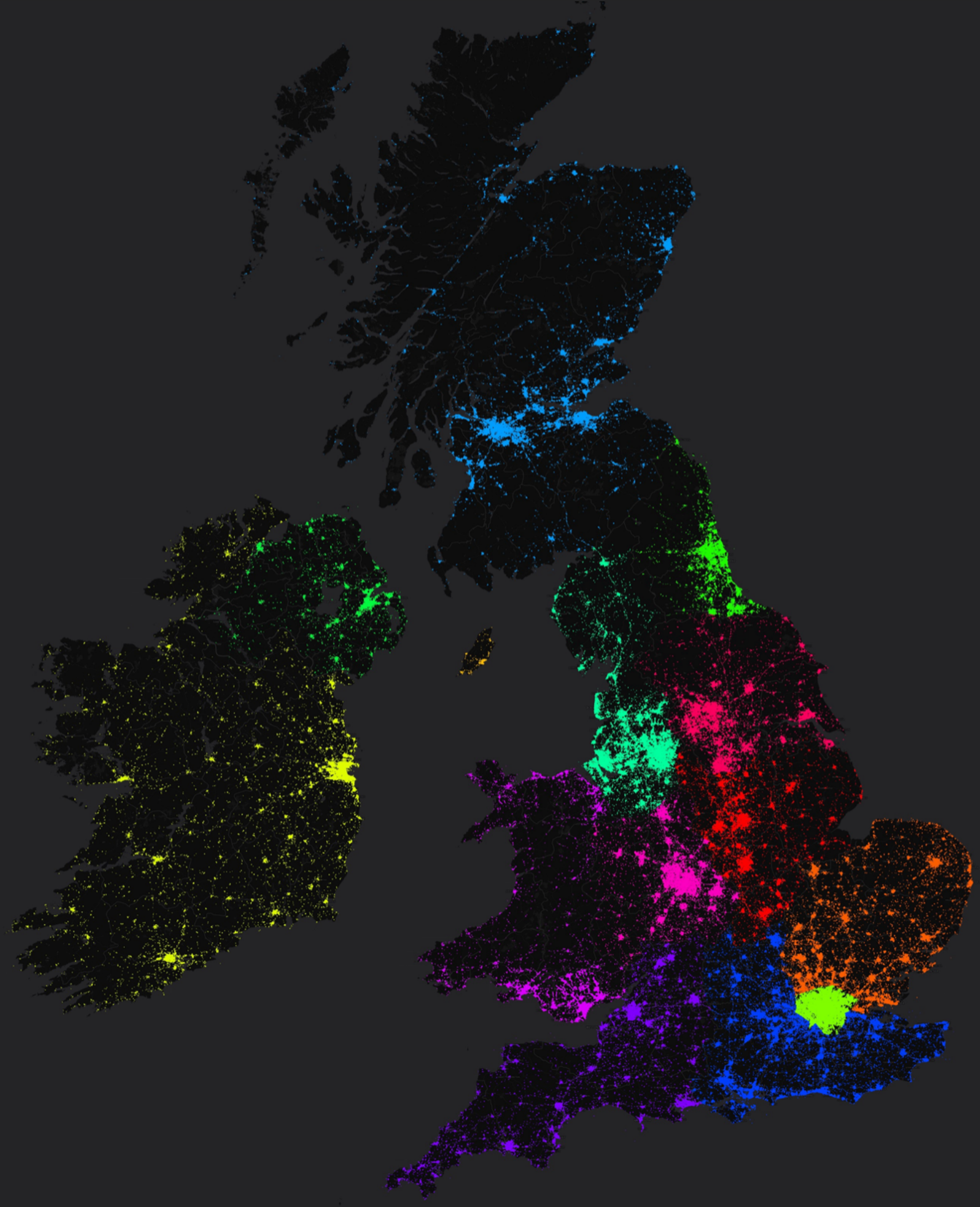
Methodology

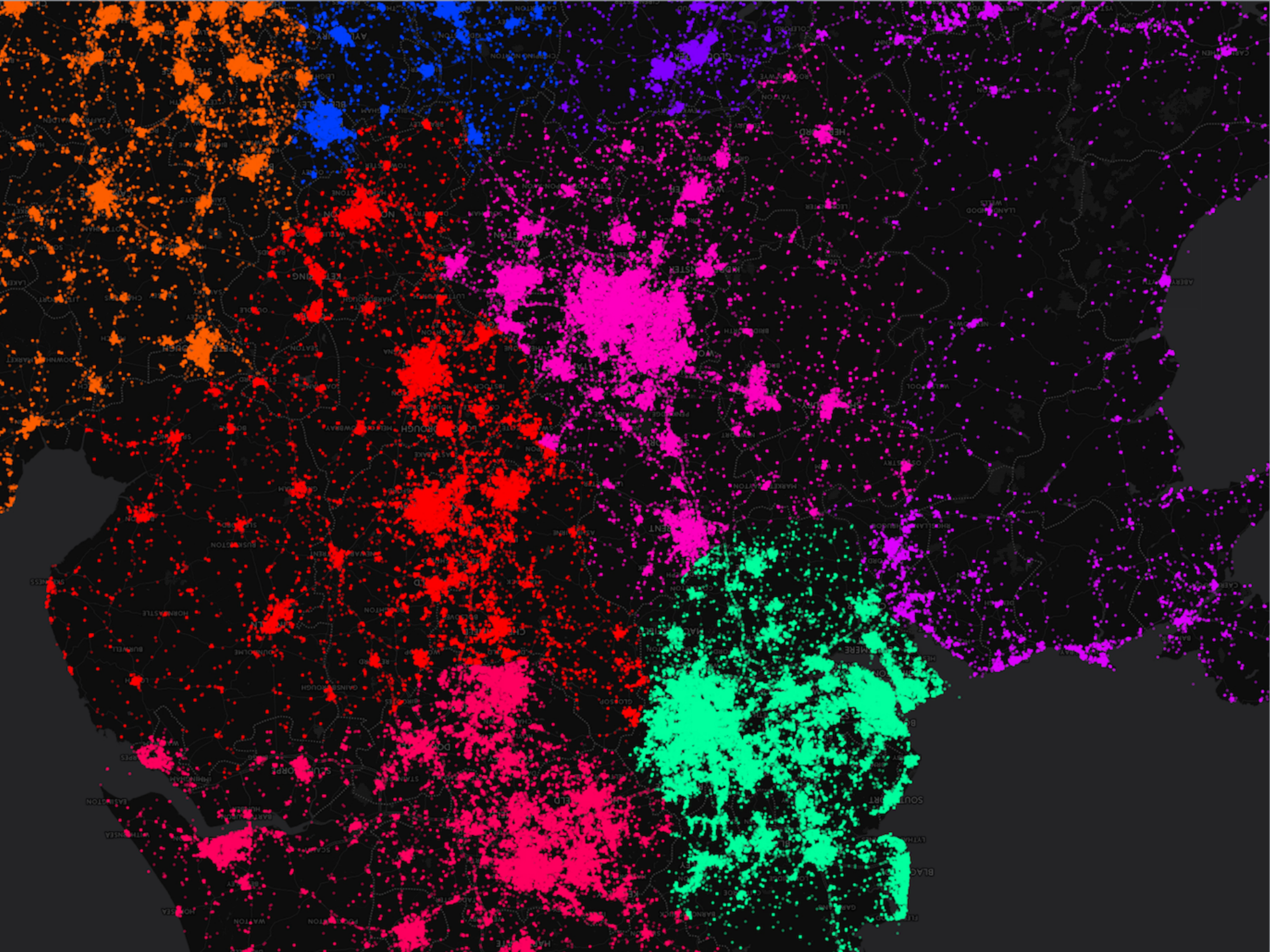
Geotagging

- Two types of geospatial metadata:
  1. specific latitude/longitude points
  2. bounding box demarcating a more general area
     - around 75% of geotagged tweets sent with this less precise geospatial metadata
     - these areas vary wildly in size, but can be as large as the UK itself!
     - solution: automatically generate a latitude/longitude point within the bounding box, but discard tweets where the box's area exceeds some arbitrary limit

- Python script to discretise these latitude/longitude points into broad regional categories
  - uses shapefiles of UK regions and a point-in-polygon function to determine which regional polygon a tweet falls into (or is a reasonable distance from)
  - tweets that don't overlap with any regional polygon are discarded







Mapping techniques

- Many ways of plotting geospatial data in R

a) Individual points using **`ggmap`**

  - serious issue of overplotting
  - extremely slow to plot

b) Individual points using shapefiles

  - same issues, but *can* conduct spatial aggregation by plotting points over regional polygon shapefiles, allowing you to create…

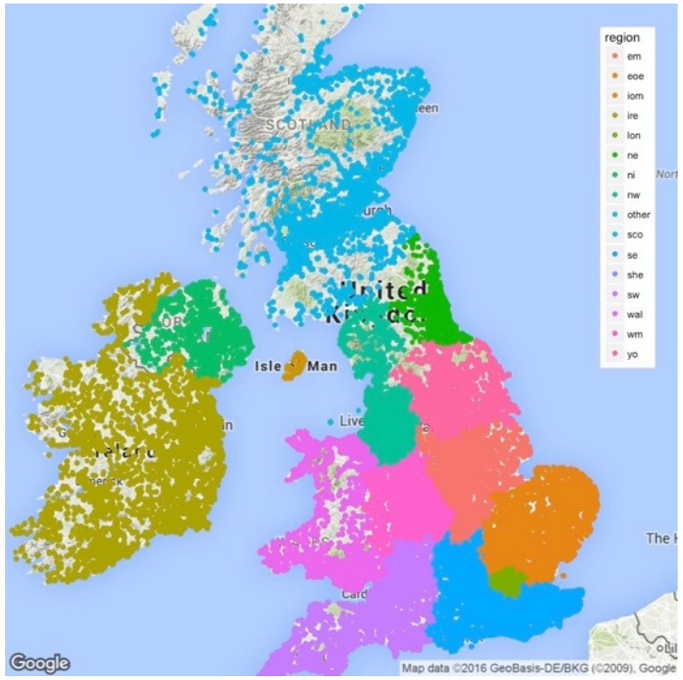
c) Choropleth maps using shapefiles

  - regional polygons colour-coded by level of variable, e.g. employment rate

d) Interactive maps using **`leaflet`**

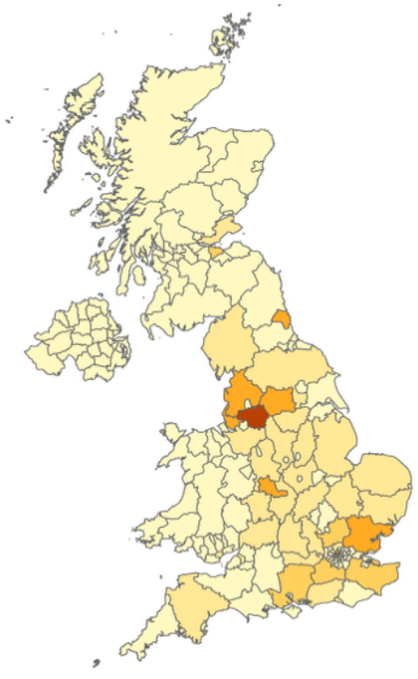
  - can zoom in and pan around the maps, filter data, and include pop-up windows

(a)

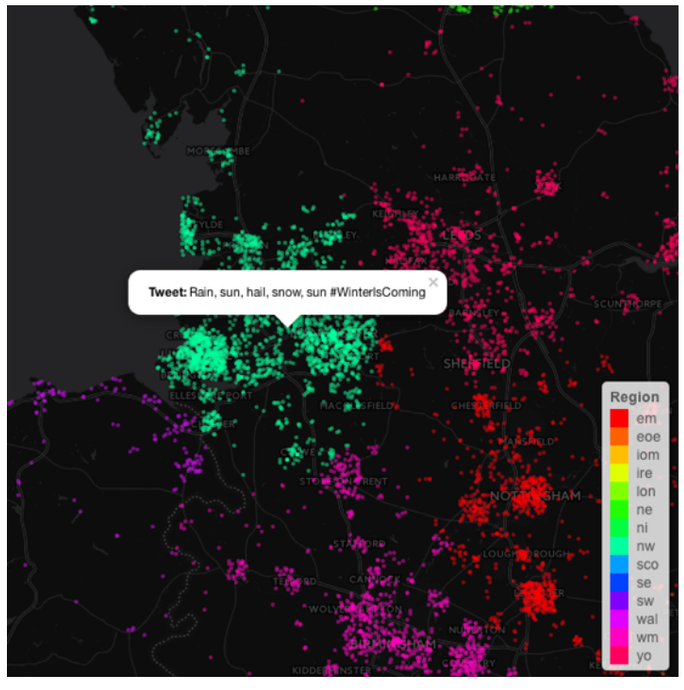


(b)

(c)



(d)

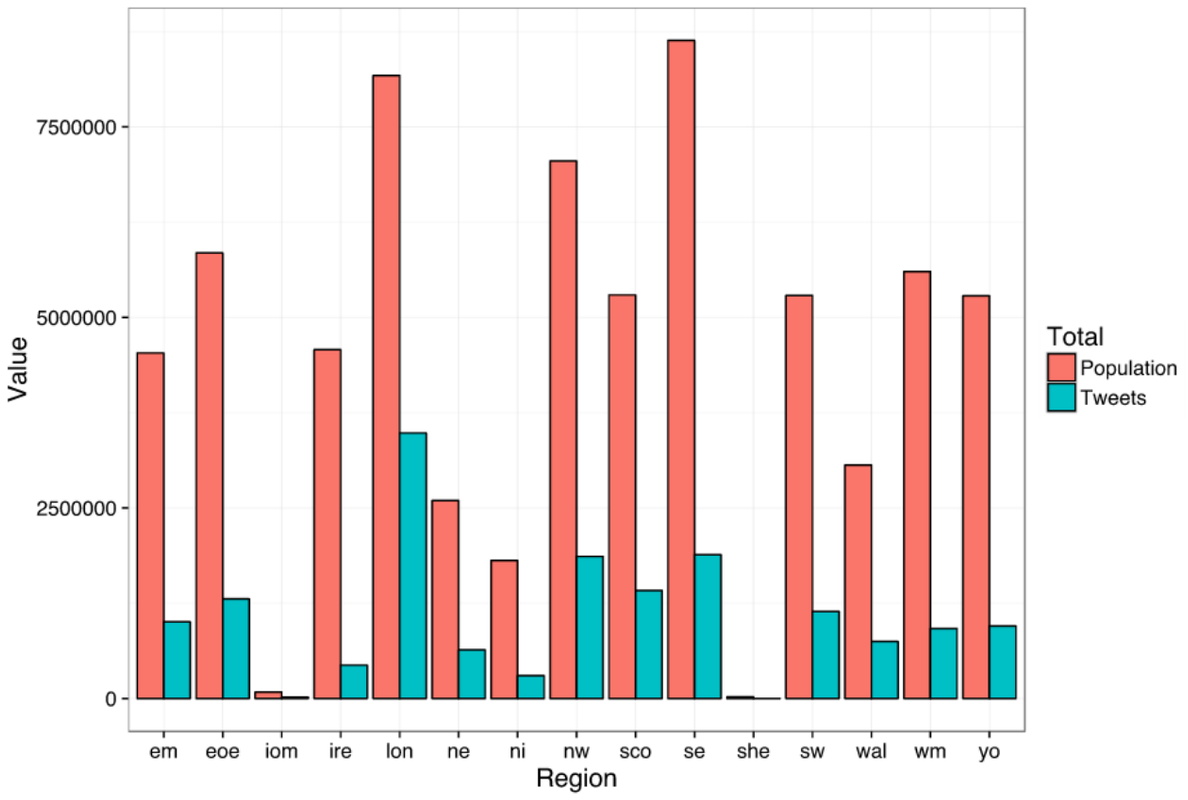


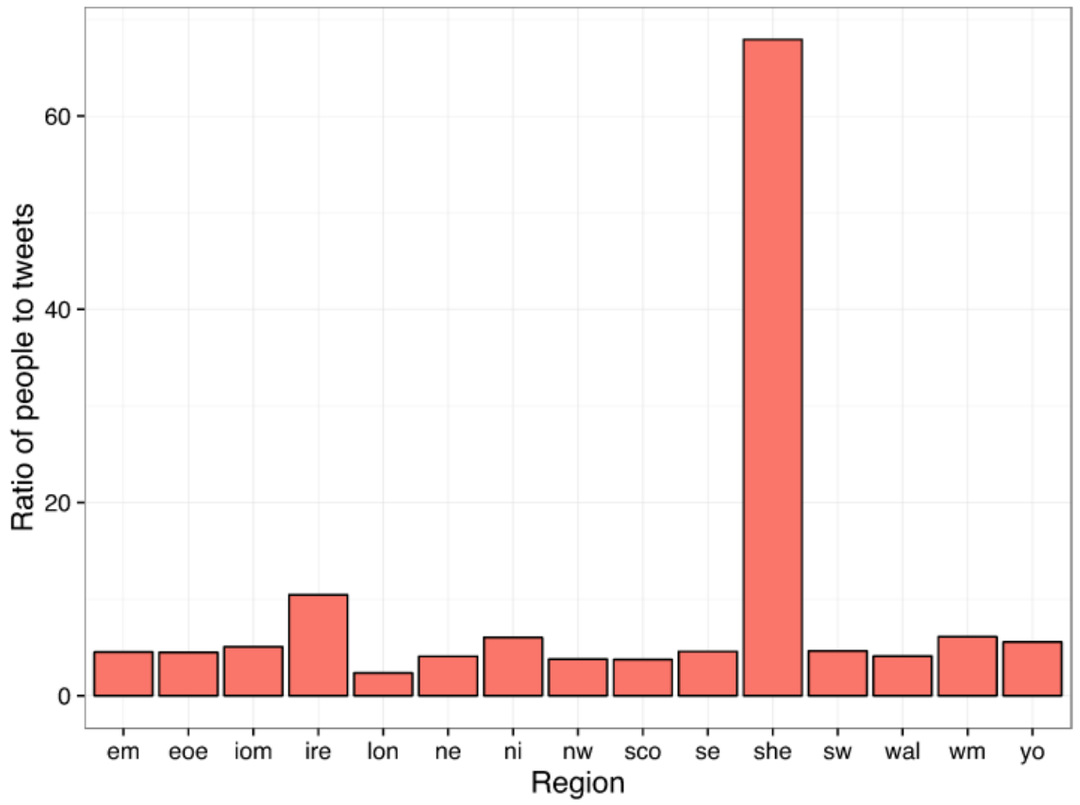
Methodology

The corpus

- Results presented here based on a Twitter corpus of over 16 million tweets (around 175 million words), collected mostly in a 4-month period between January and April 2016

- Good regional distribution overall, where tweet density in each region corresponds perfectly with population density





Results

Lexical variation

- Regular expressions using **`grep`** in R to extract tweets containing words of interest
  - e.g. *mortal,* 'drunk' (821 tokens)

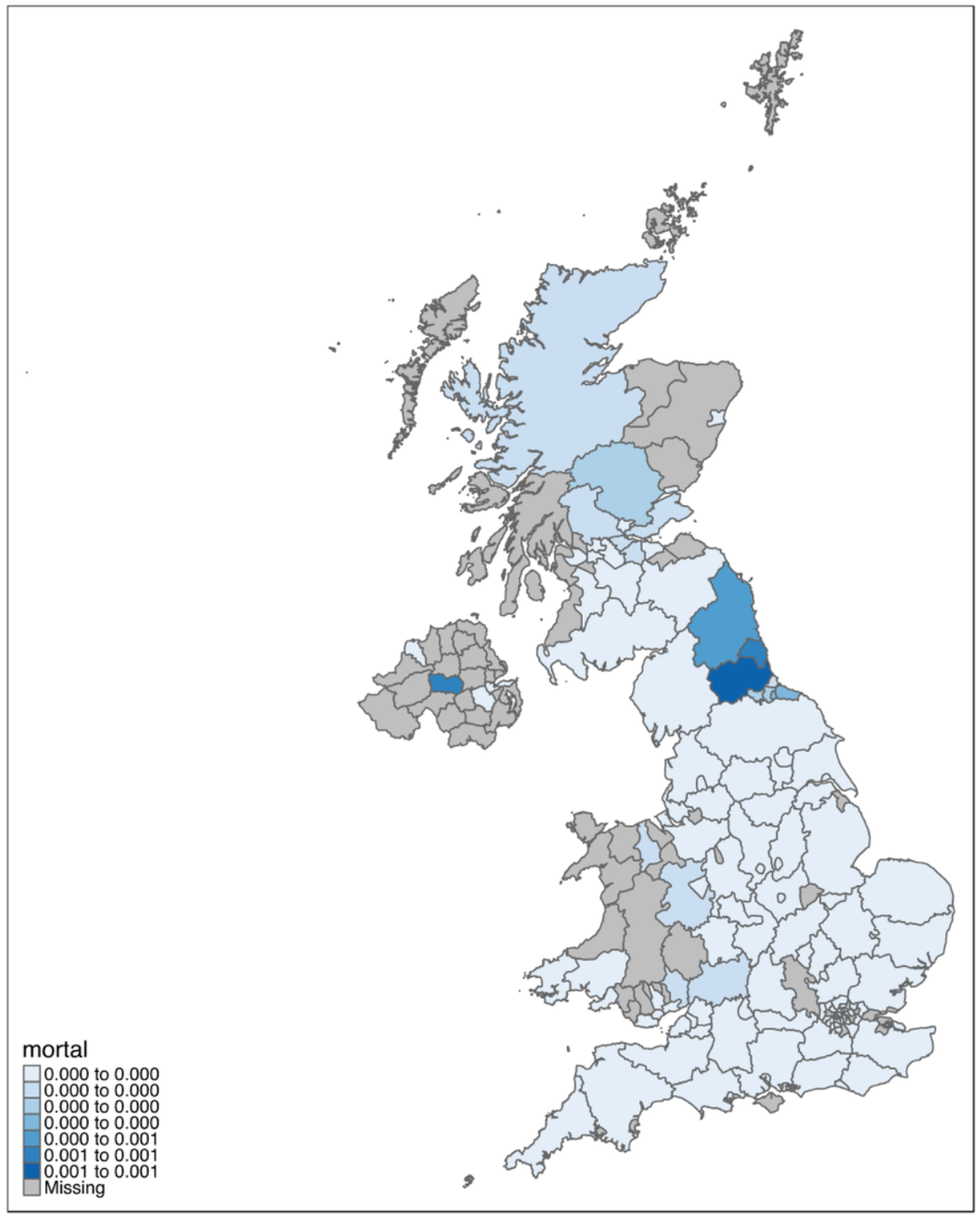
> Half of me wants to stay in and watch a film and the other half of me wants to get **mortal**

> My whole life just consists of me being **mortal**. Yay

↑

could be existential crisis

- wider issue of only capturing the desired sense of the word
  - manual inspection ideal, but time-consuming



Results

Lexical variation

- Regular expressions using **`grep`** in R to extract tweets containing words of interest
  - e.g. *fleek*, 'looking good' (752)

> When you don't wanna take your make up off cus your eyebrows are too on **fleek**

> Dads bbq game on **fleek** #bbq #barbecue #chicken

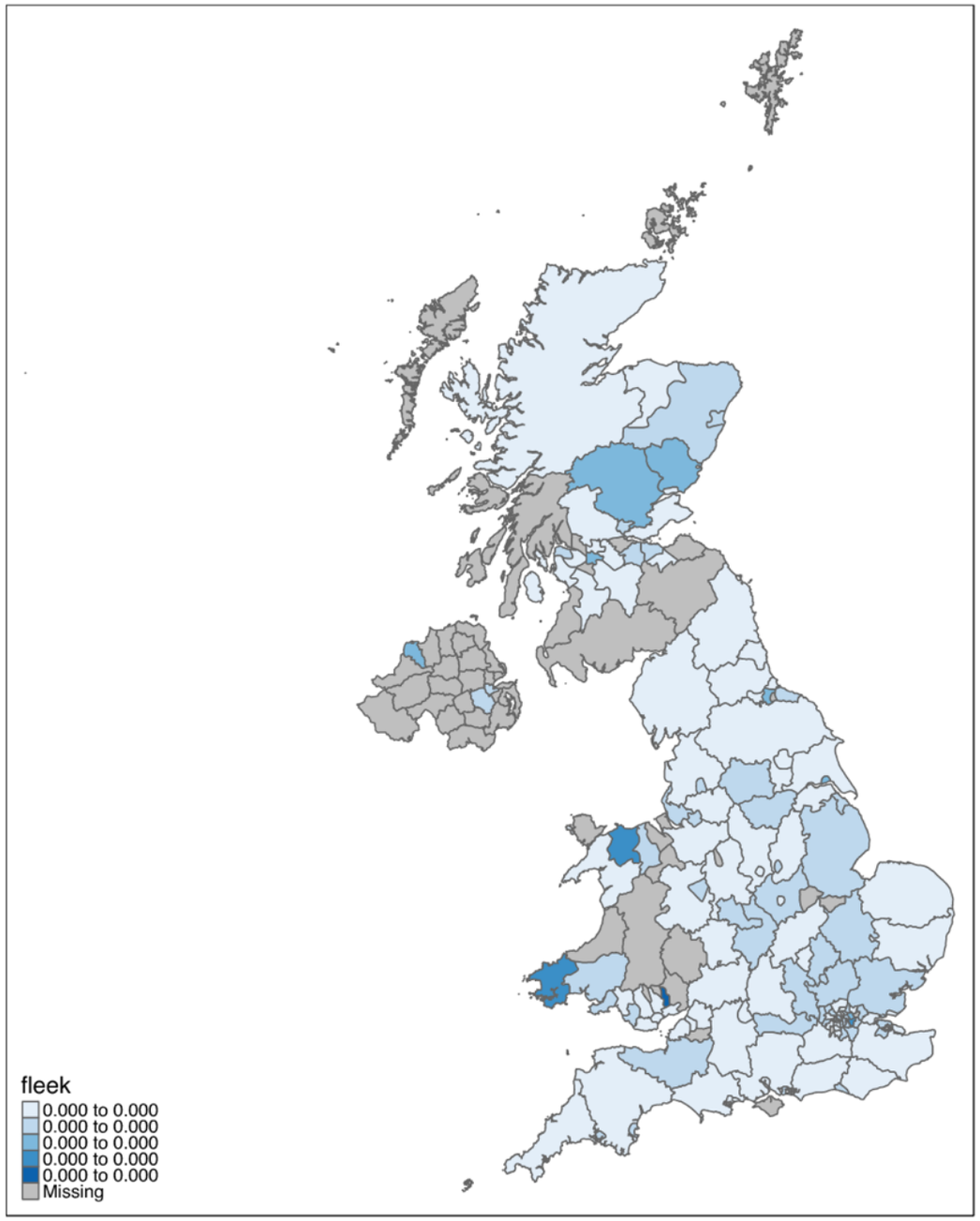
↑

meaning generalised?

> Men who use the word **fleek** make me die inside. Don't ever come near me

↑

do we really want to include this?



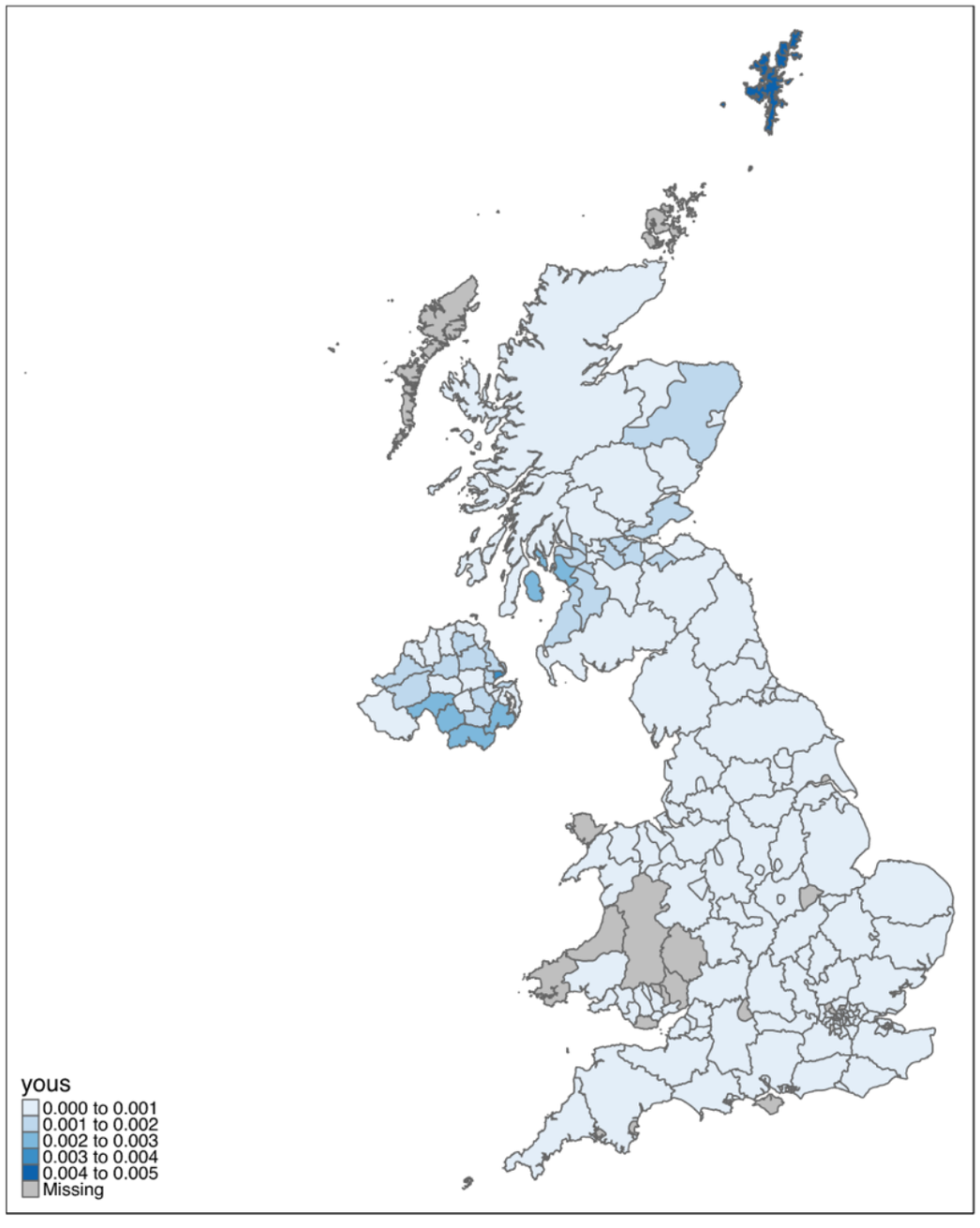
Results

Lexical variation

- Regular expressions using **`grep`** in R to extract tweets containing words of interest
  - e.g. *yous,* you.PL (3419)

> sorry to disappoint **youse**, no stunts from me tonight, was on my road bike tonight.

> Do you know what the only issue with going out out with your BF is that when one of **yous** needs a wee the other one is on your own #loner

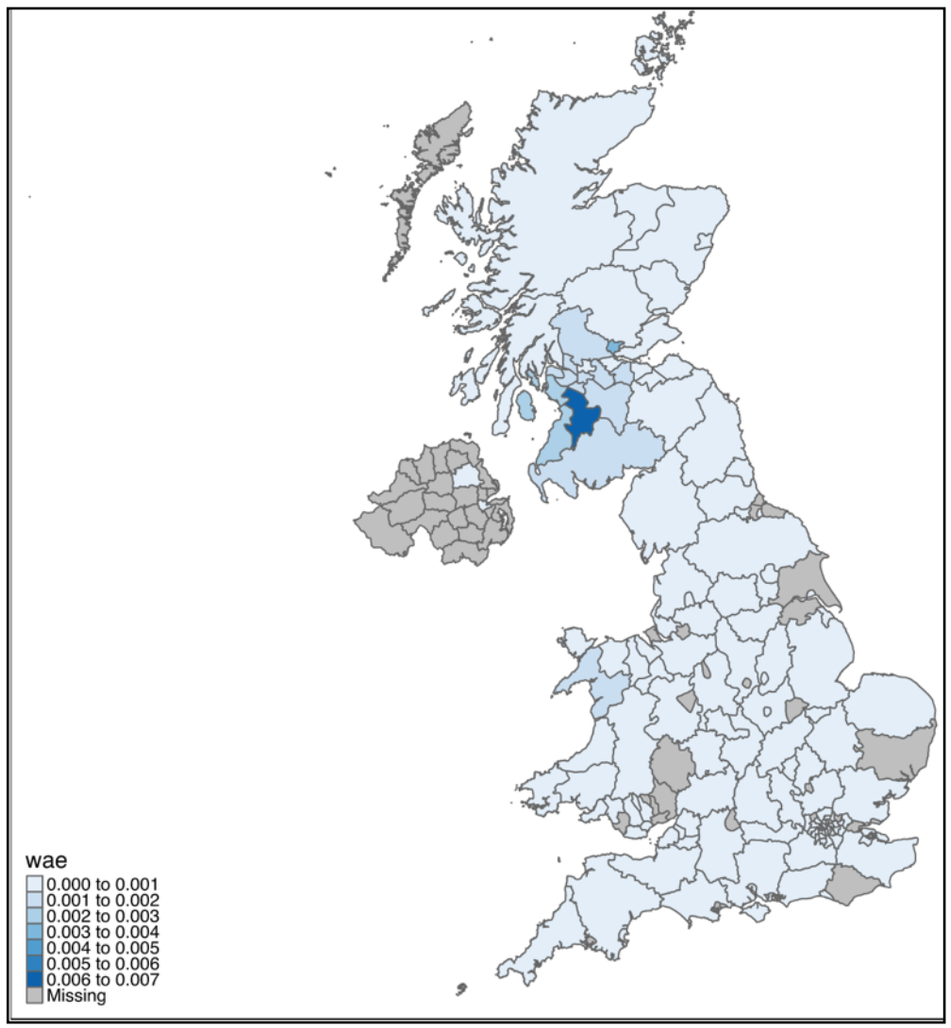


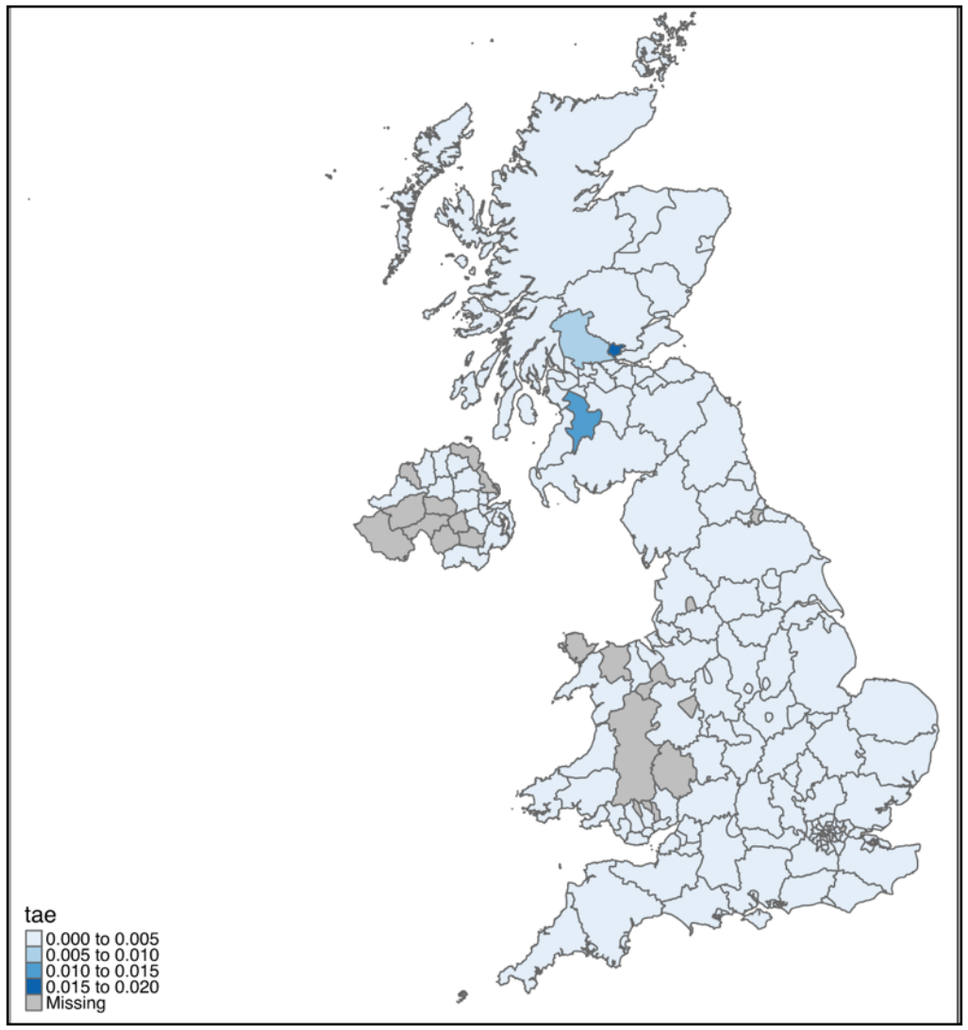
Results

Lexical variation

- Regular expressions using **`grep`** in R to extract tweets containing words of interest
  - e.g. *tae,* 'to' and *wae,* 'with' (7350)

> yous should be in dumfries **wae** me jst aboot **tae** see the hostiles again

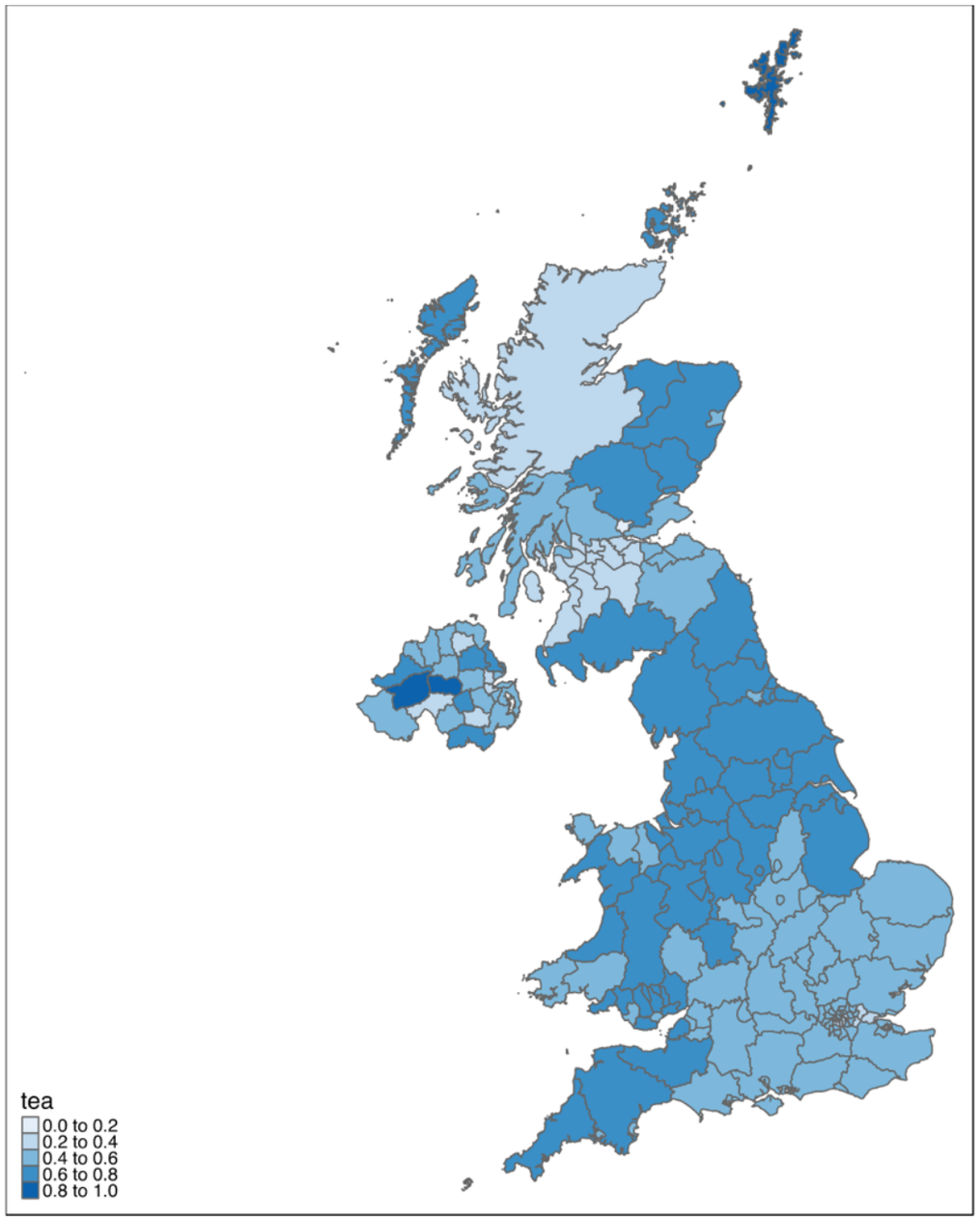


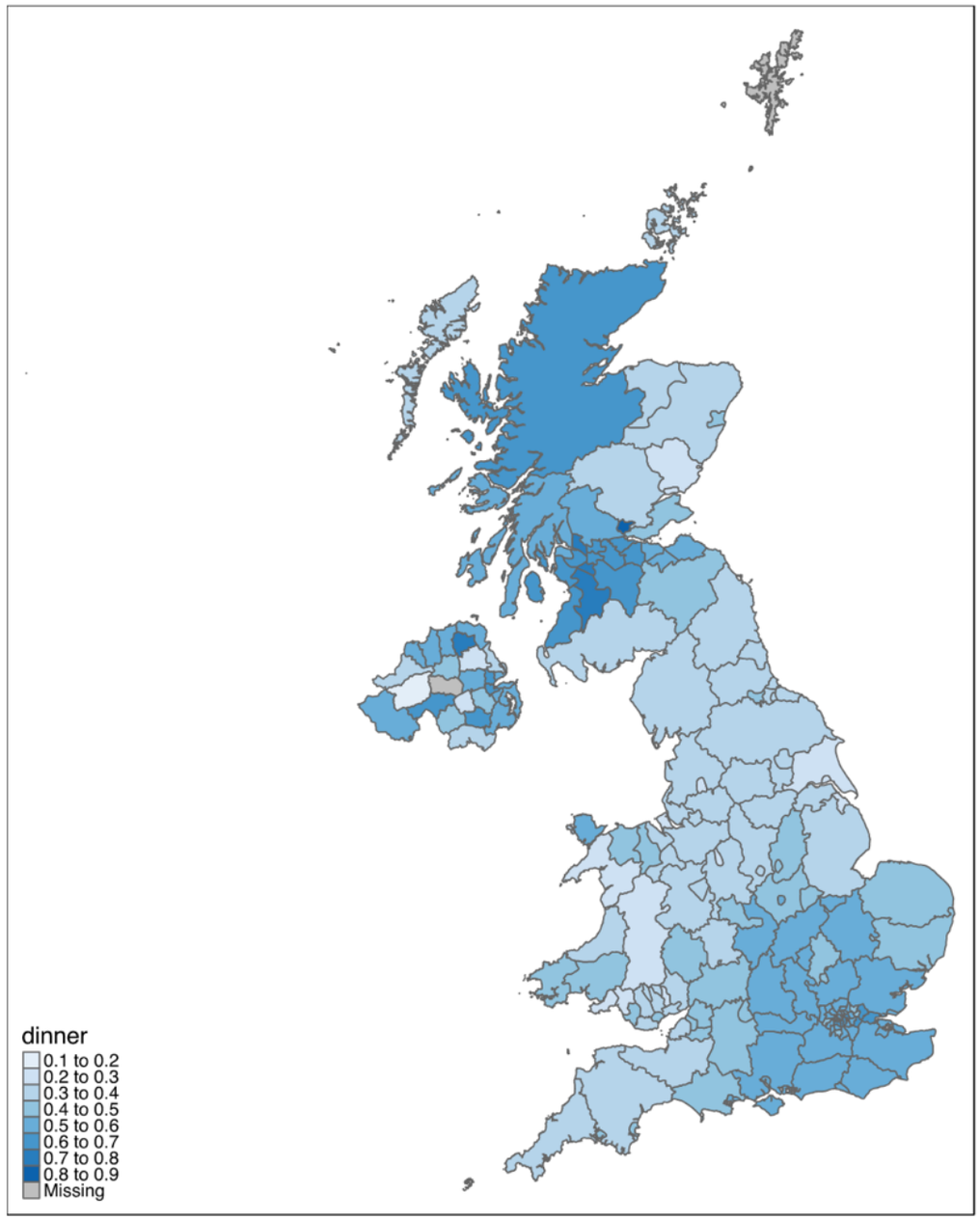


Results

Lexical variation

tea ~ dinner (85275)

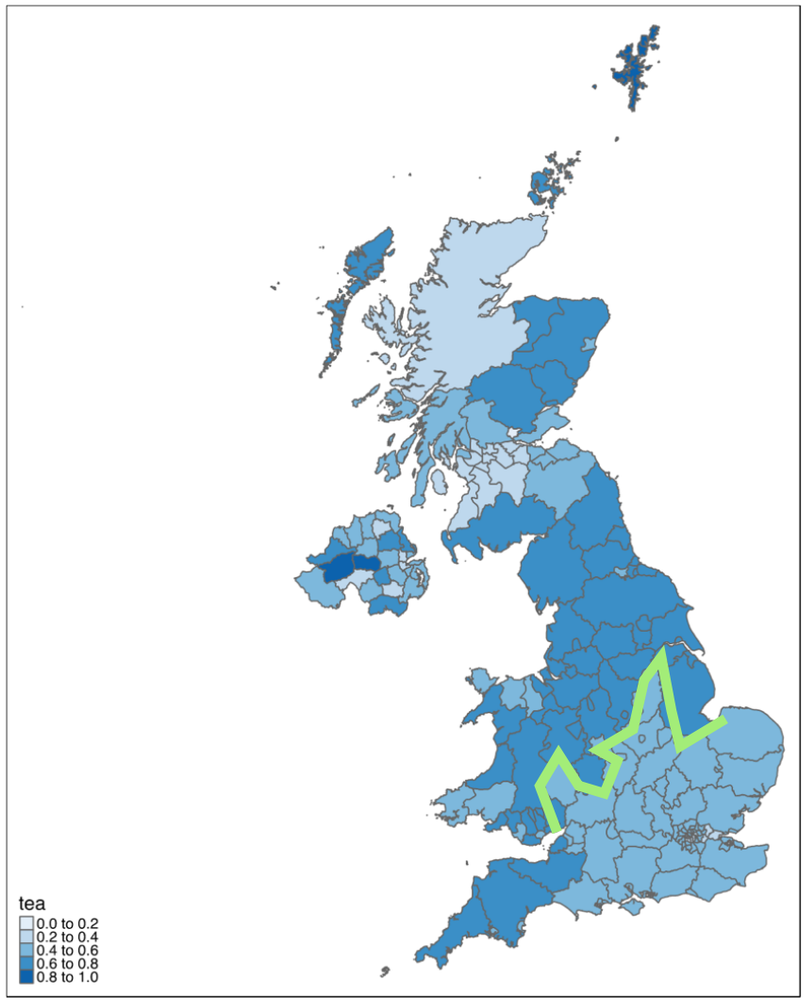


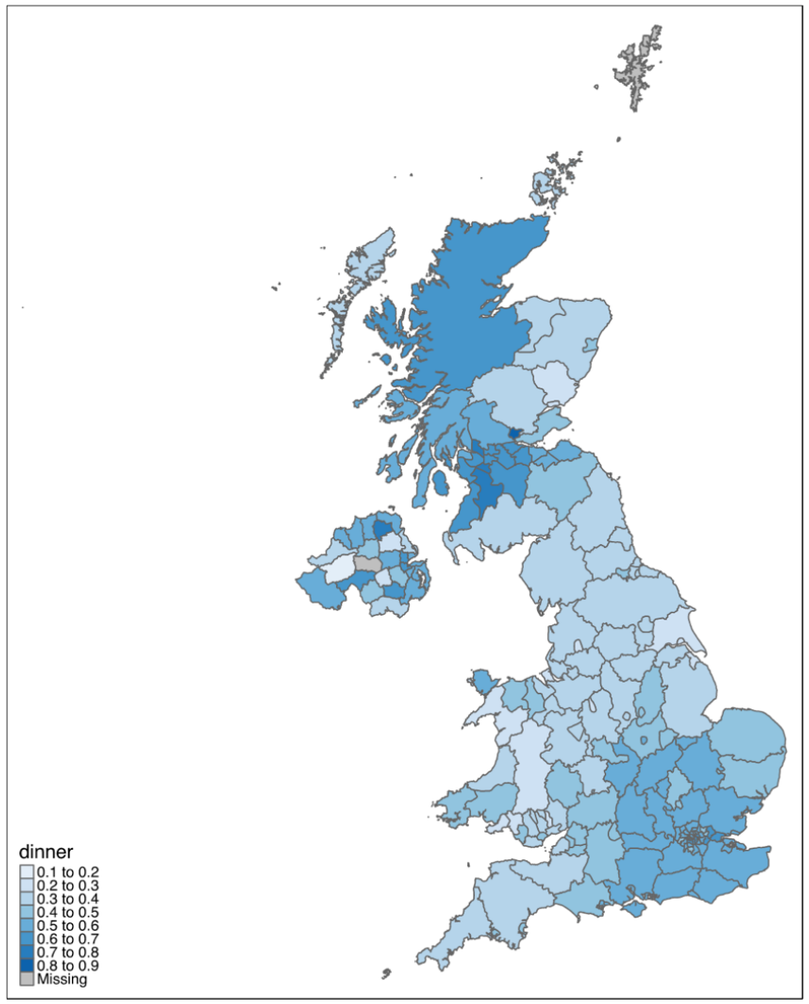


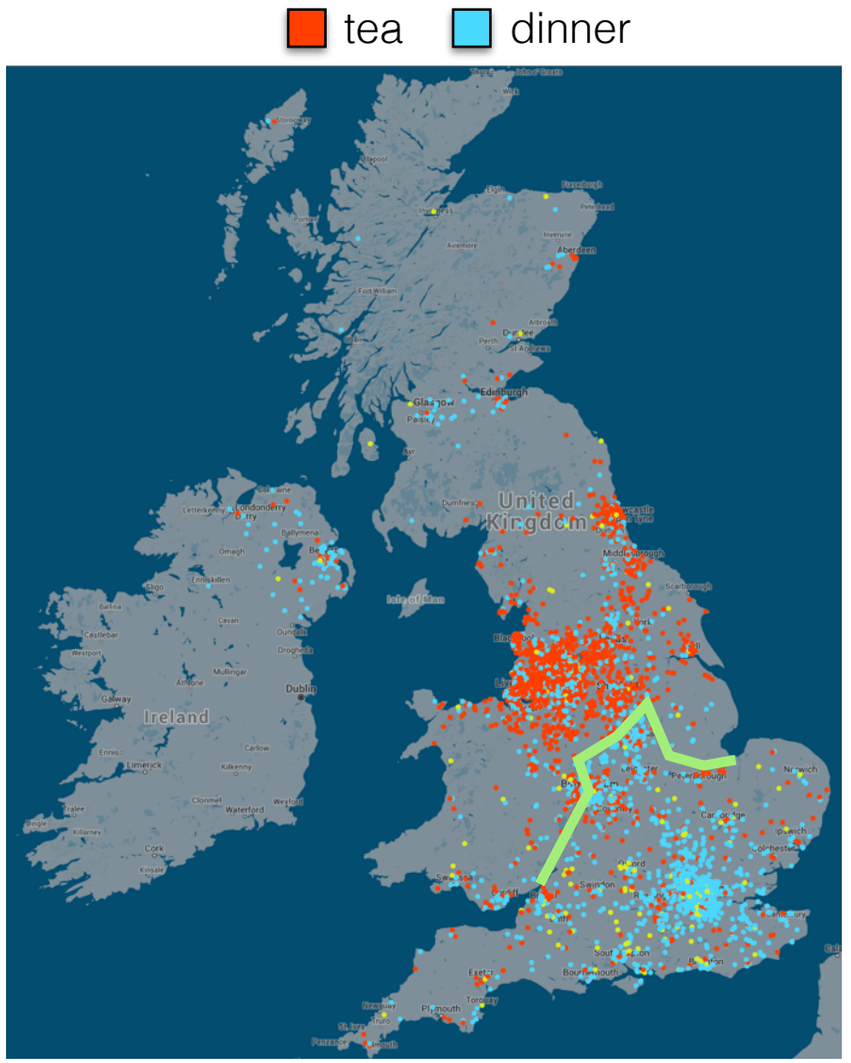
Results

Lexical variation

tea ~ dinner (85275)







from *Our dialects* (MacKenzie et al. 2015)

Results

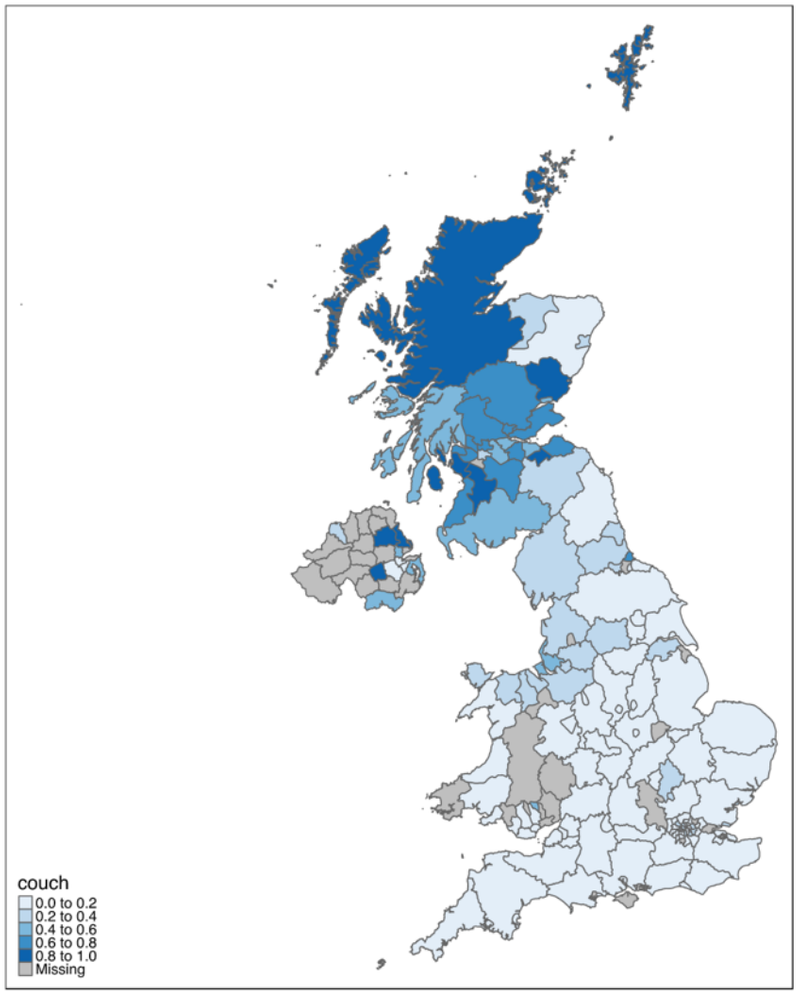
Lexical variation

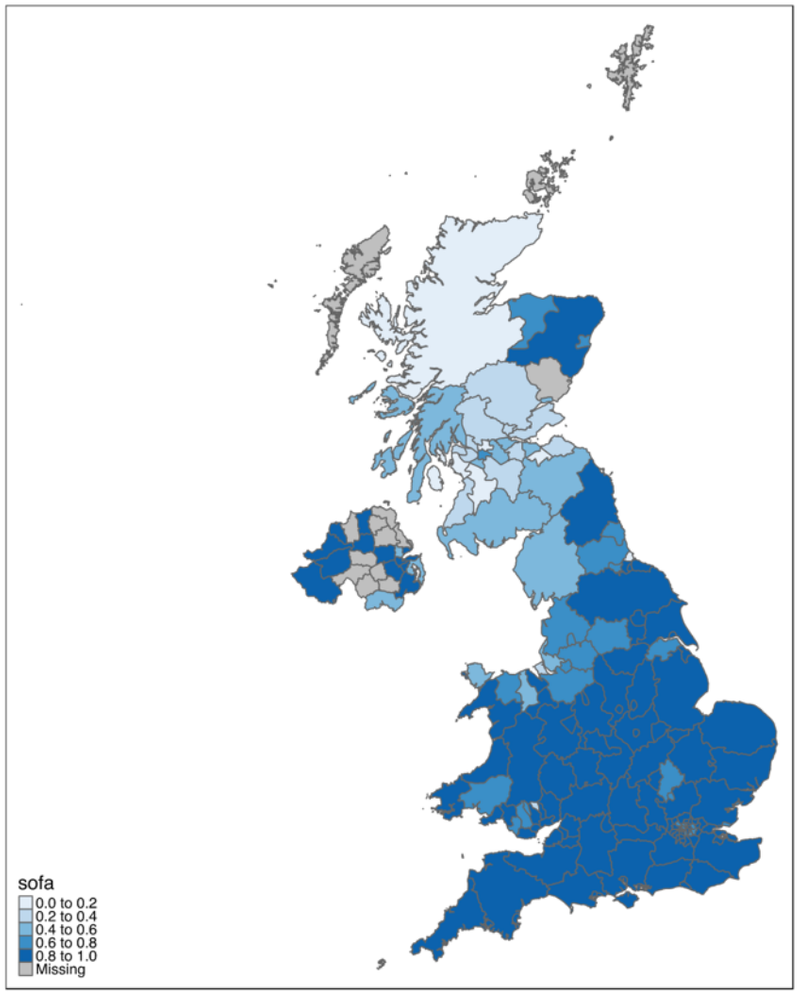
couch ~ sofa ~ settee (7076)

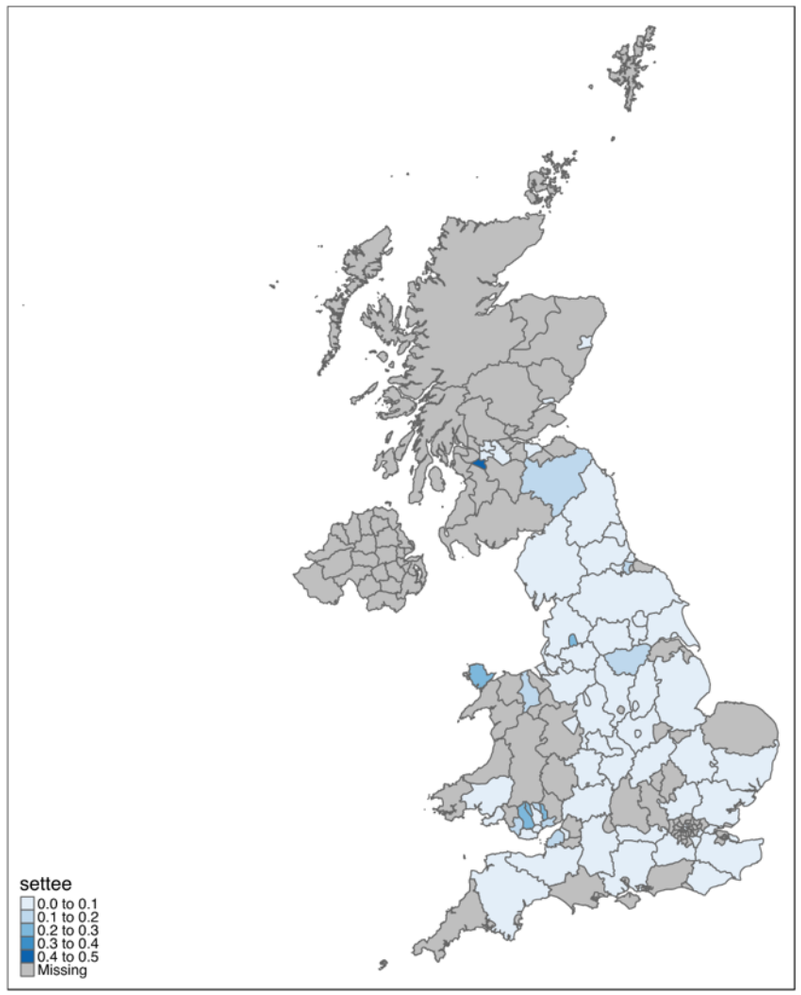
Results

Lexical variation

couch ~ sofa ~ settee (7076)

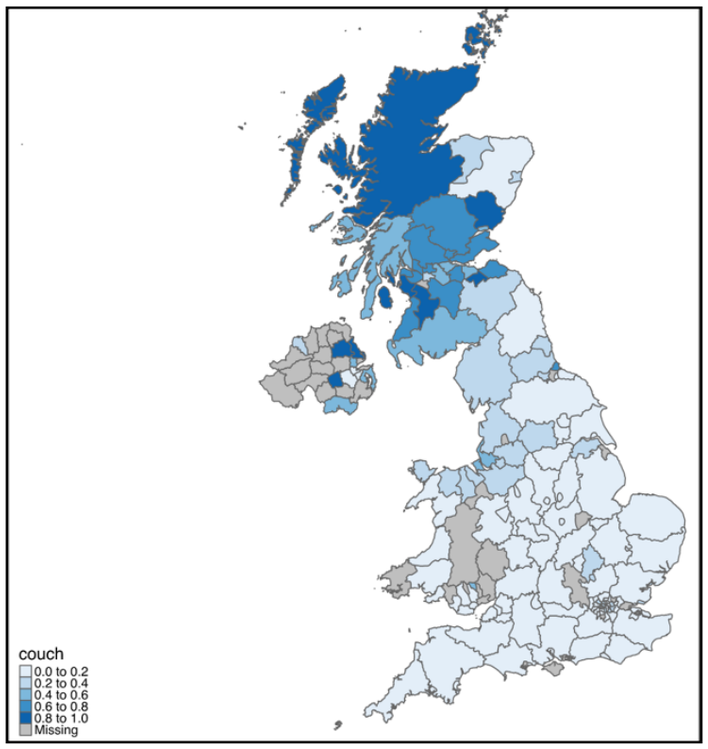


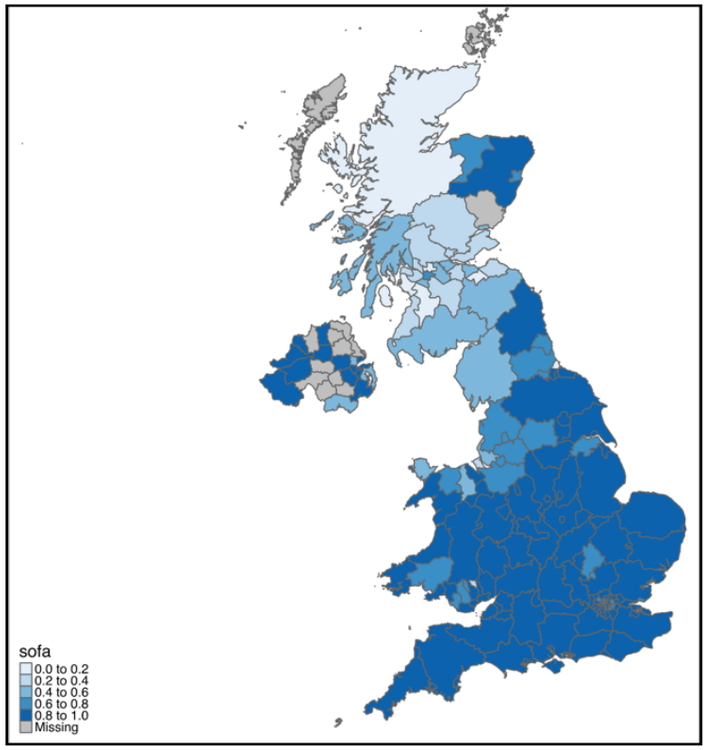


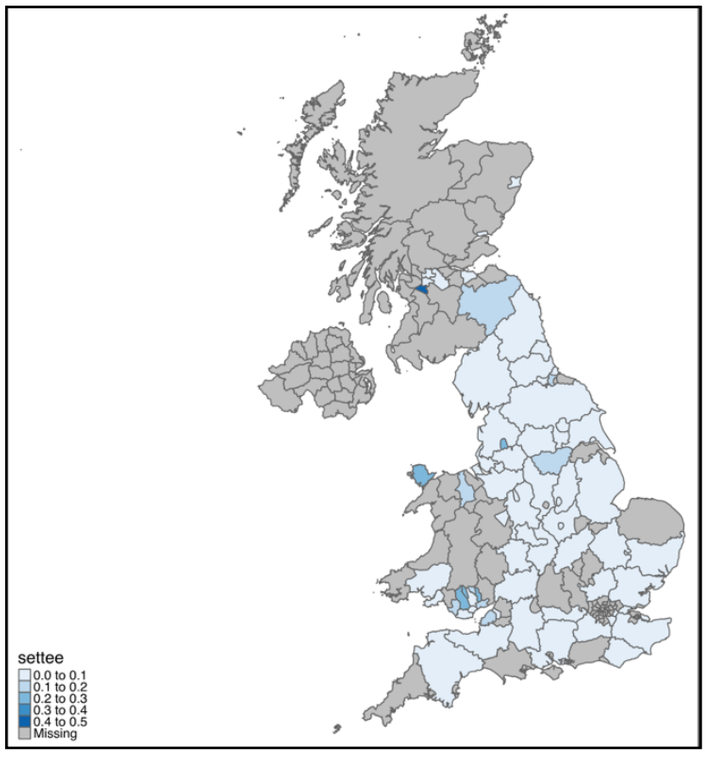


Results

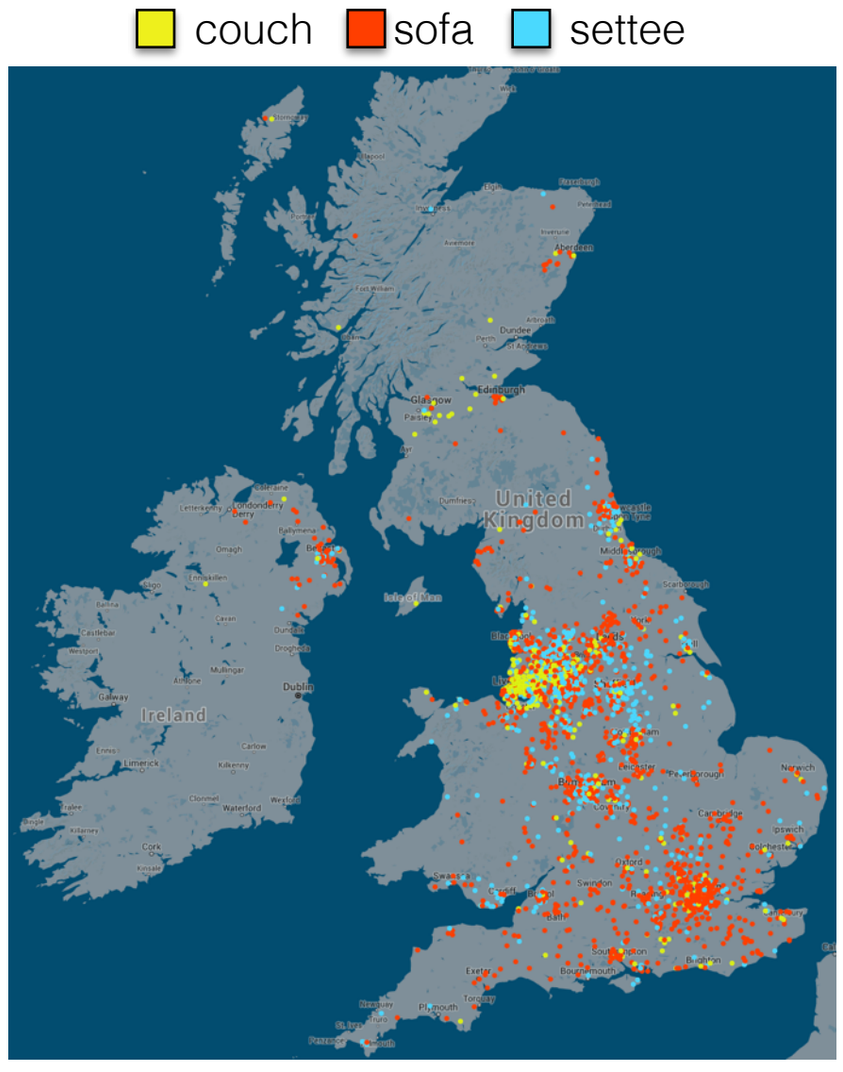
Lexical variation







couch ~ sofa ~ settee (7076)



from *Our dialects* (MacKenzie et al. 2015)

Results

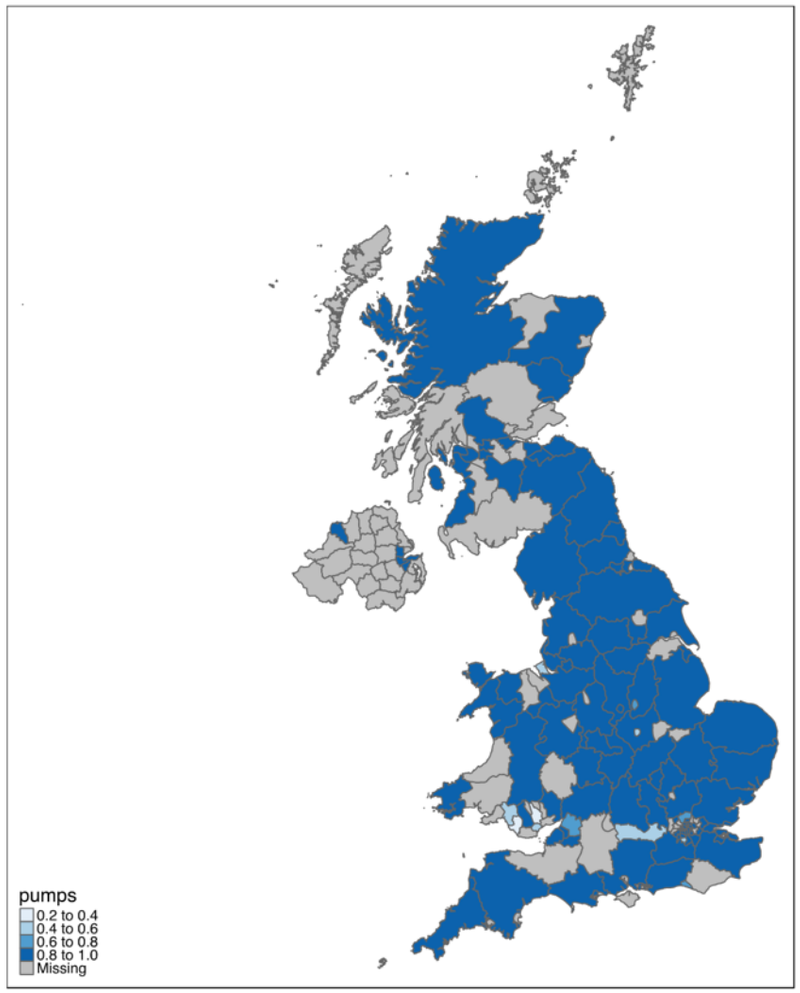
Lexical variation

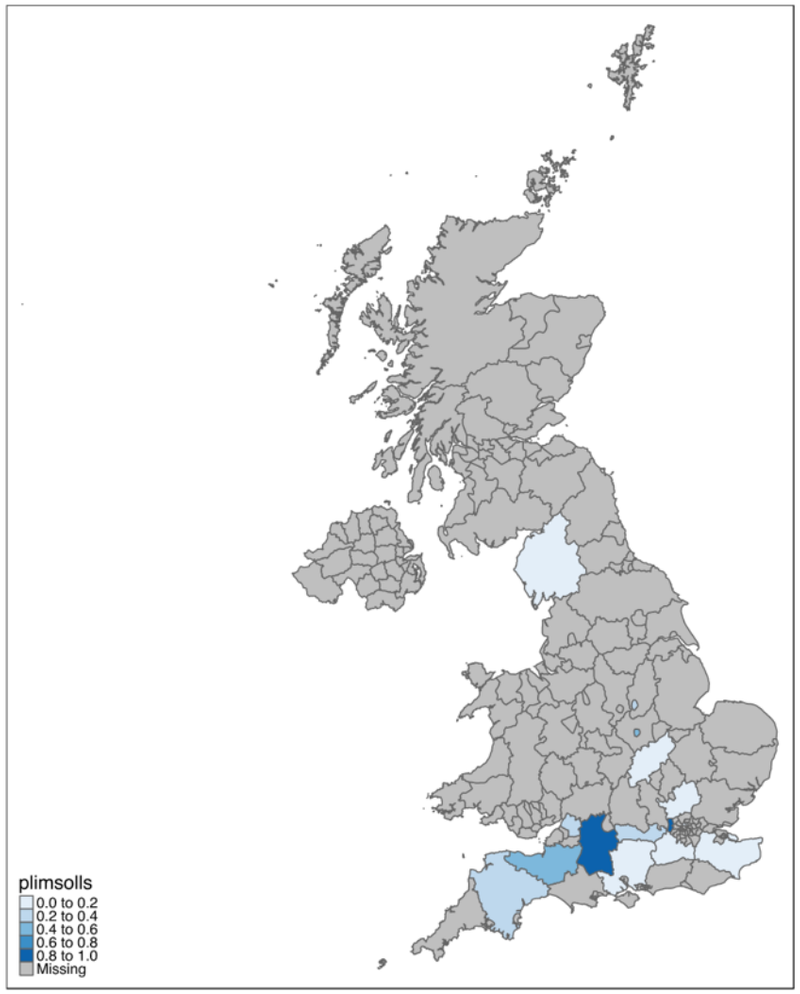
pumps ~ plimsolls ~ daps (552)

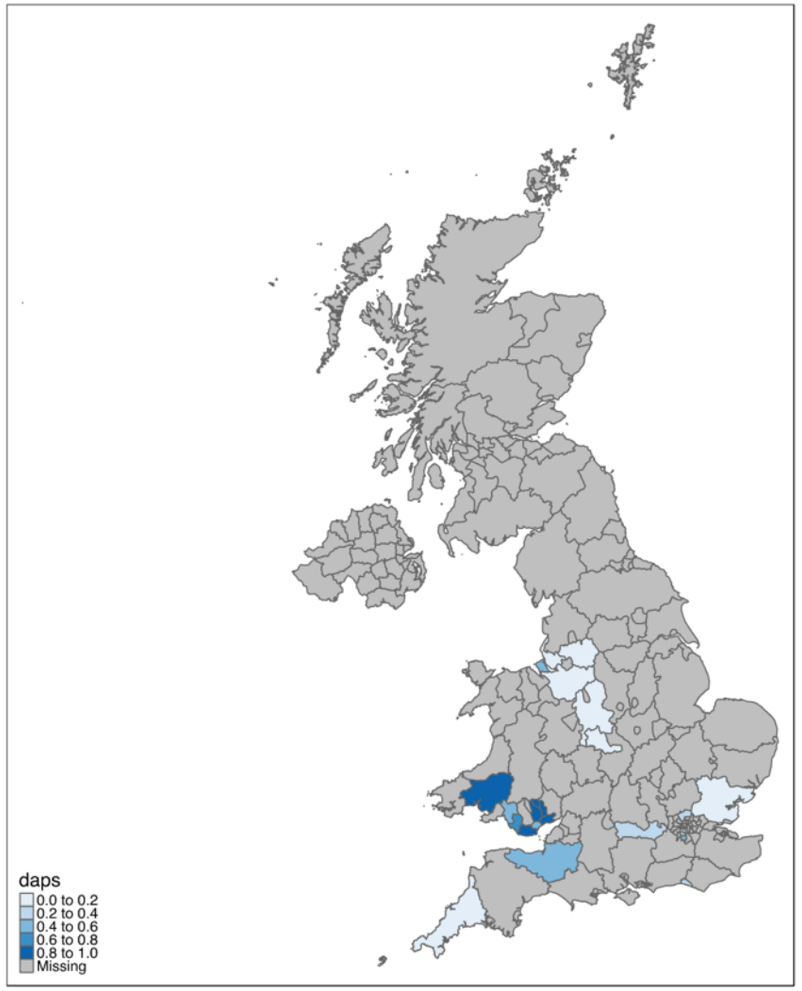
Results

Lexical variation

pumps ~ plimsolls ~ daps (552)



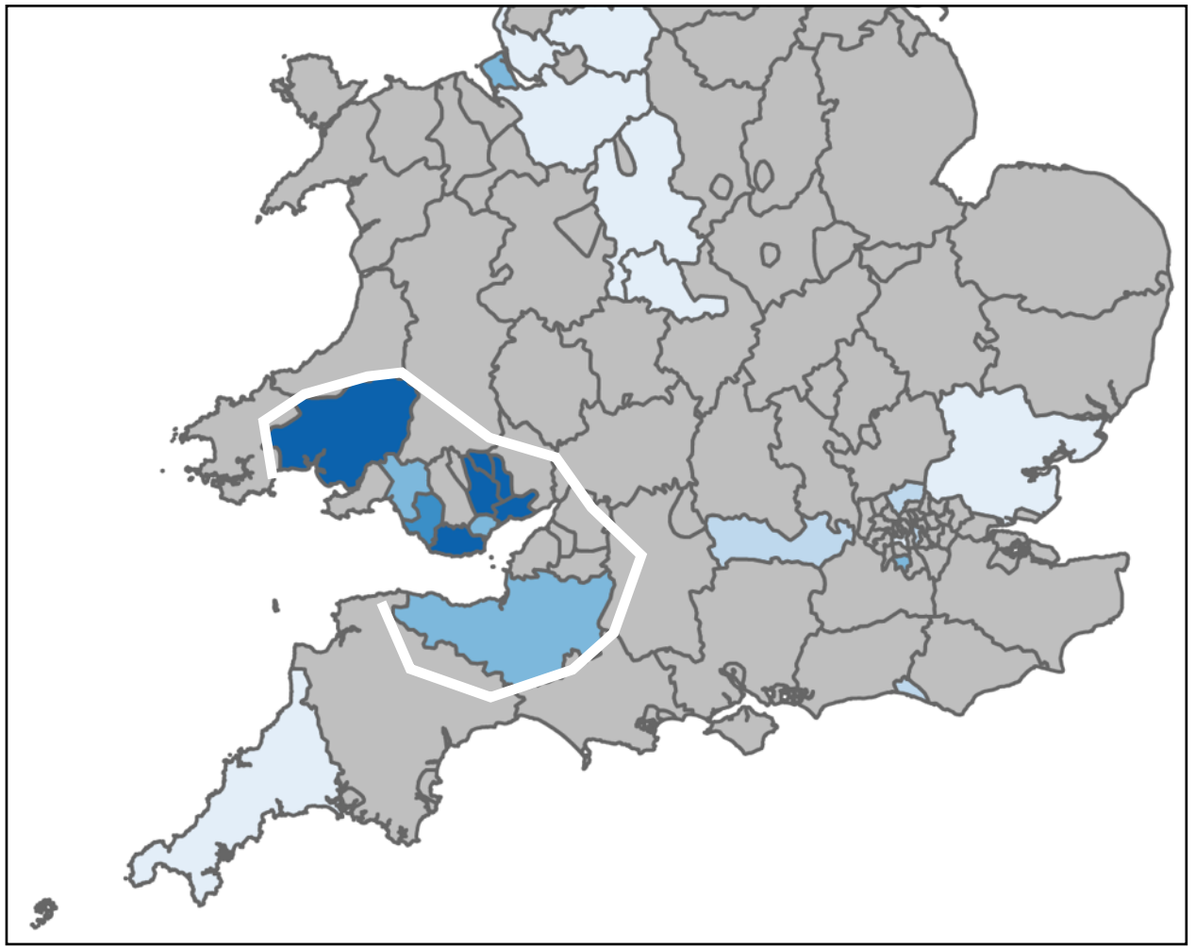


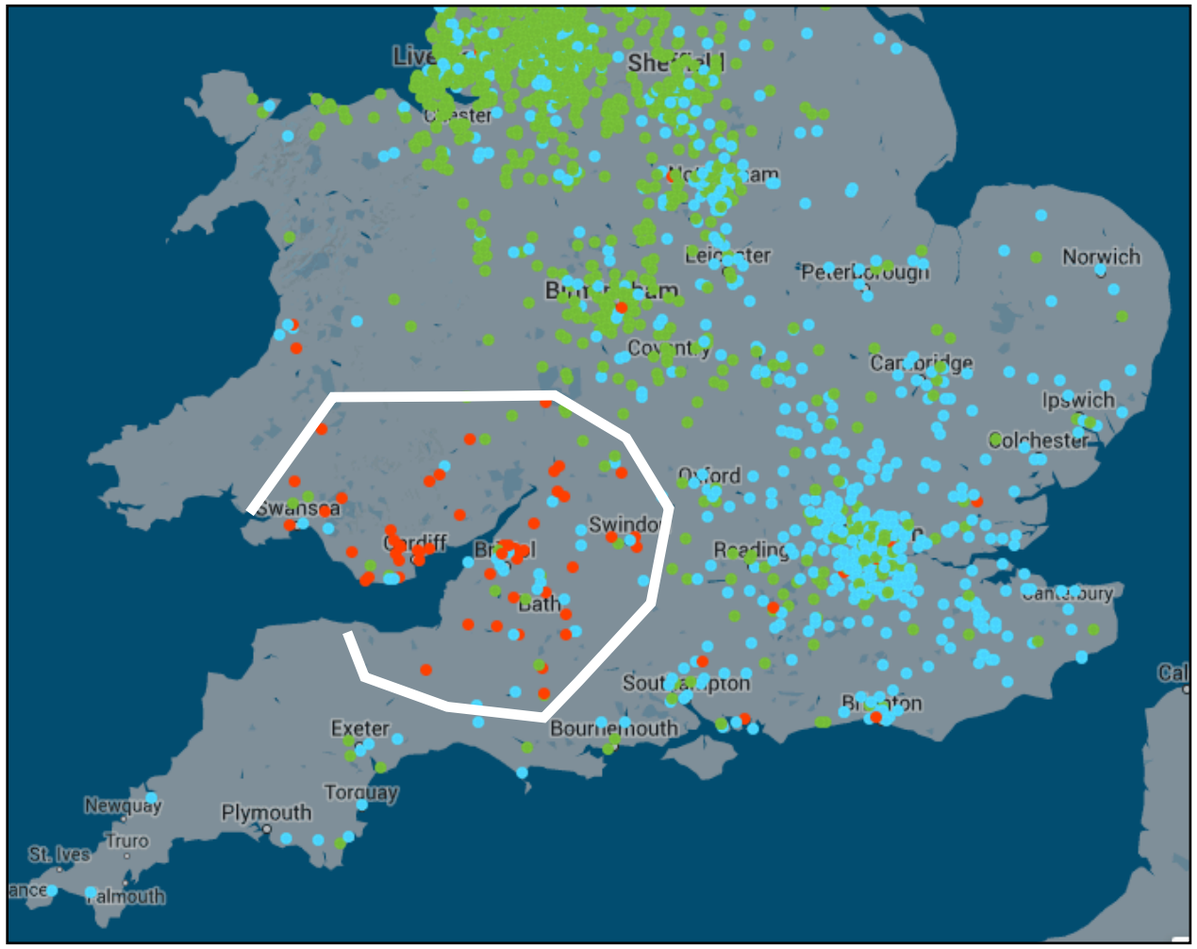


Results

Lexical variation

daps (41)





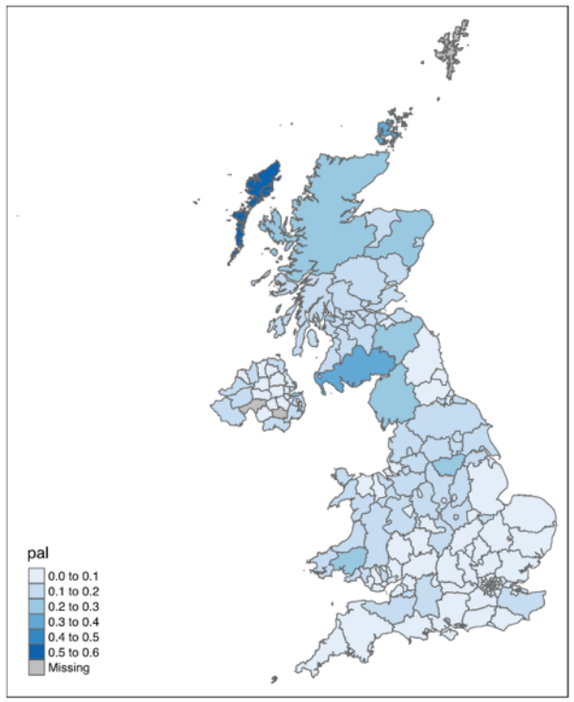
from *Our dialects* (MacKenzie et al. 2015)

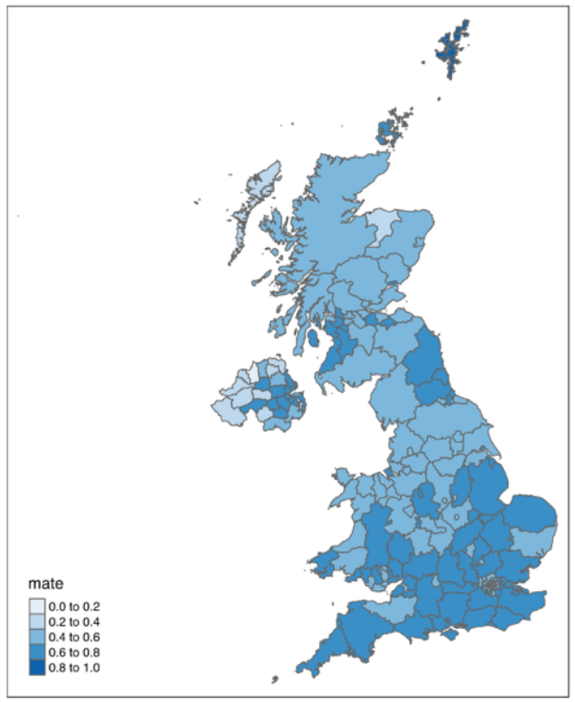
(**D**unlop **A**thletic **P**limsolls) - factory based in Bristol

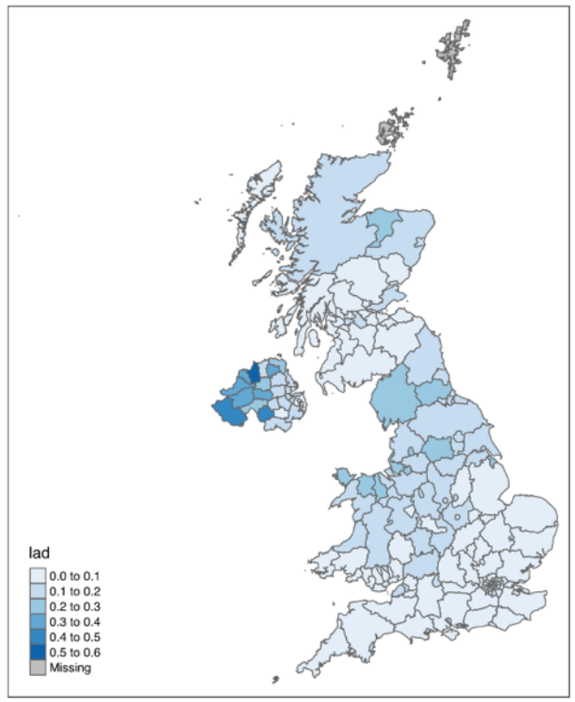
Results

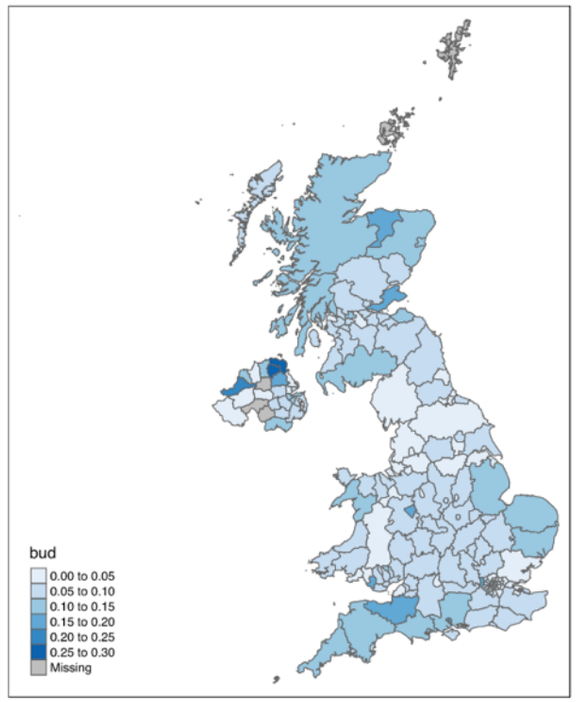
Lexical variation

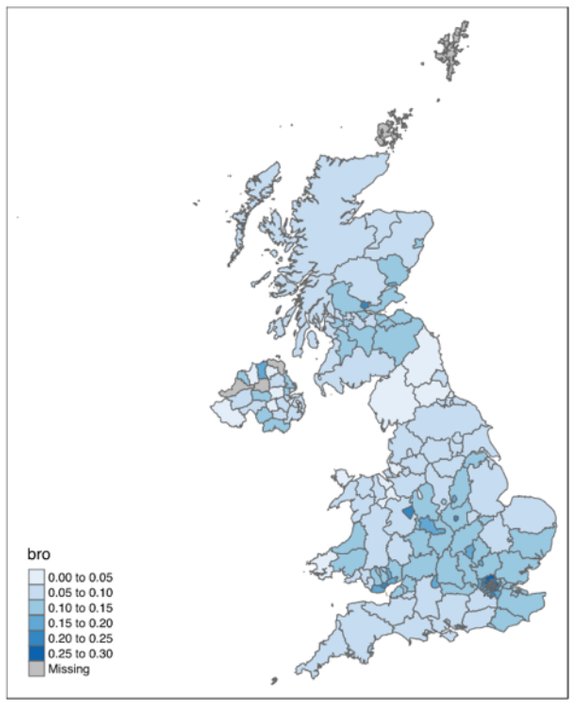
pal ~ mate ~ lad ~ bud ~ bro ~ bruv ~ blud

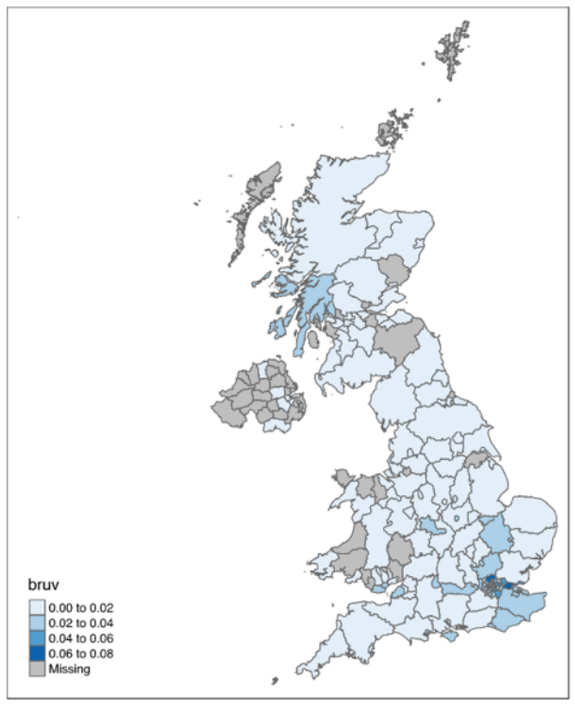


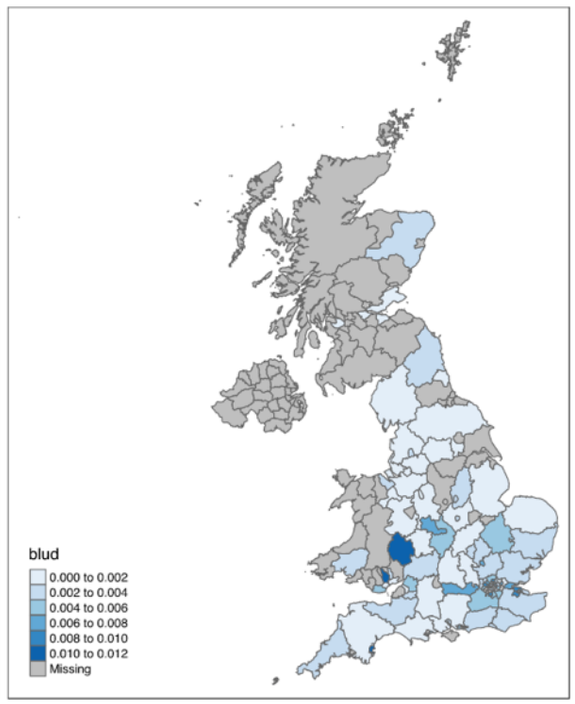










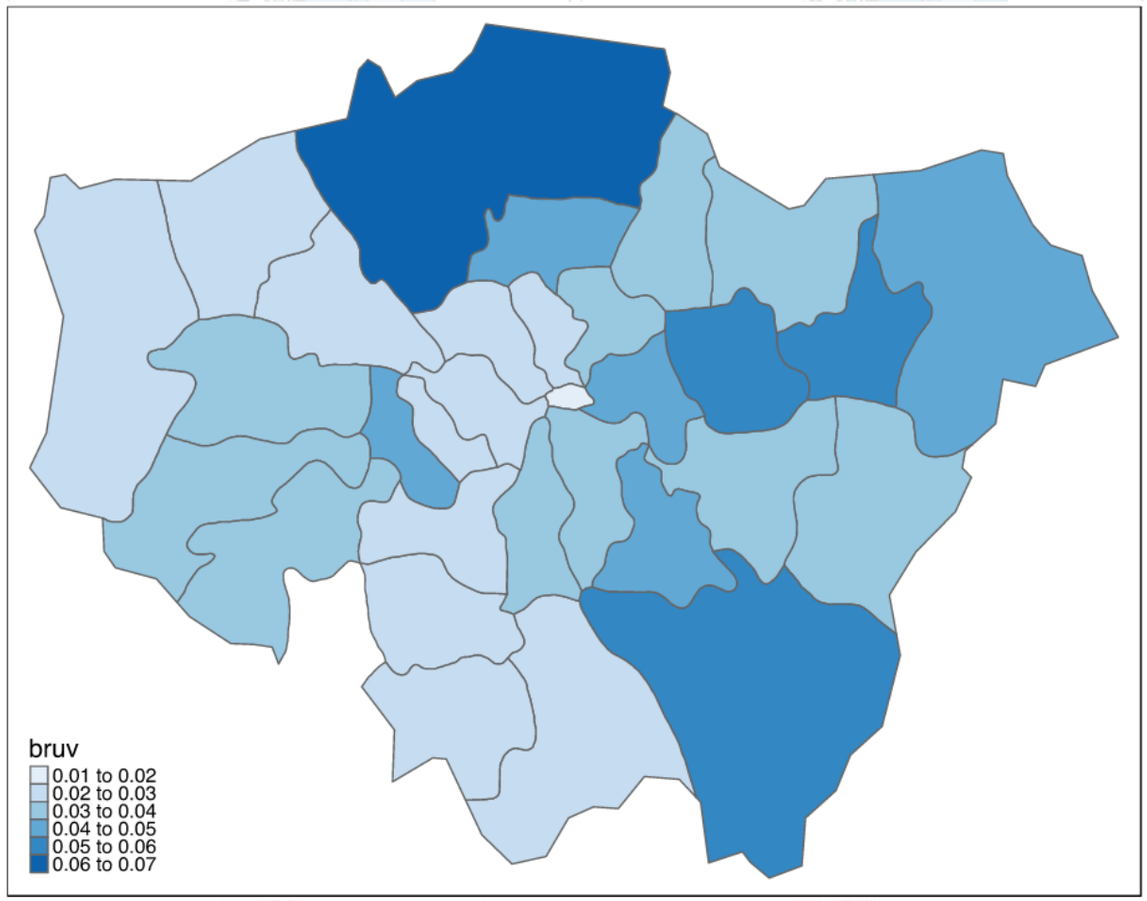


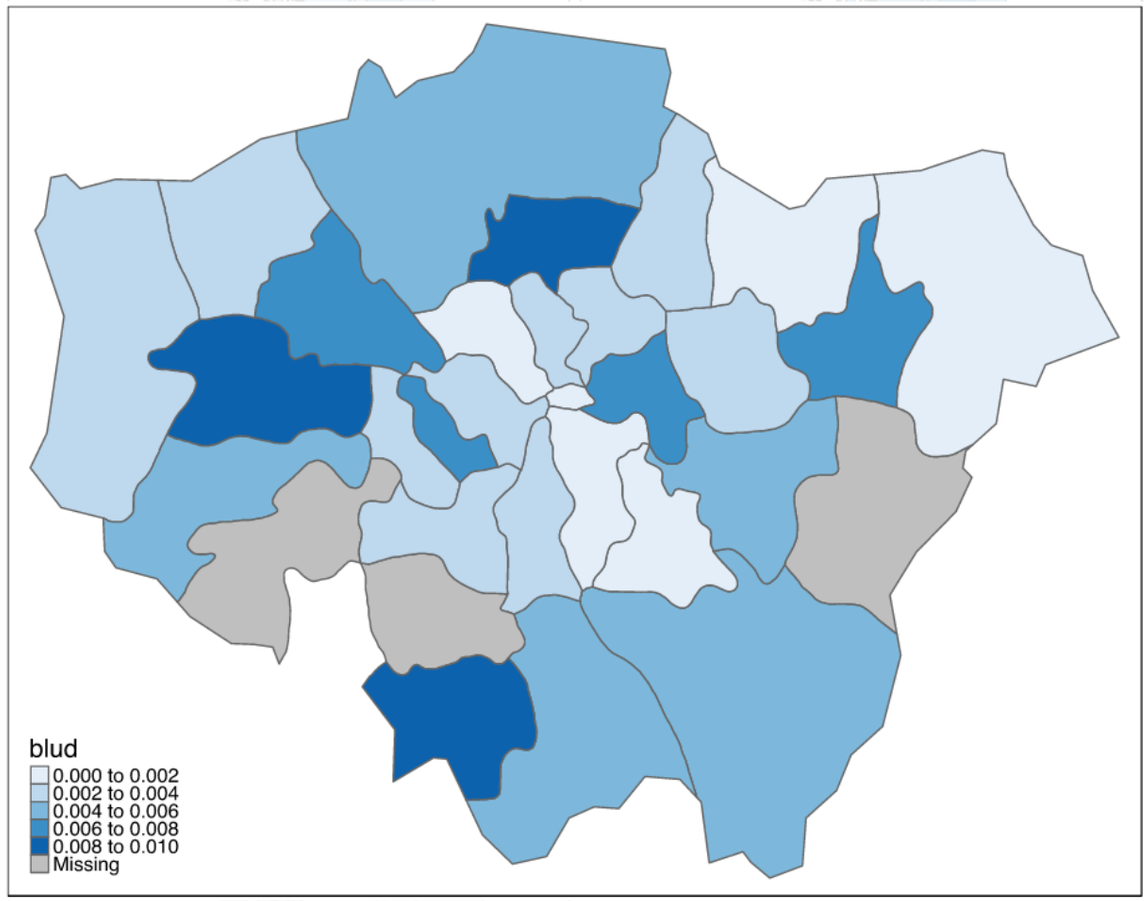
Results

Lexical variation

bruv ~ blud

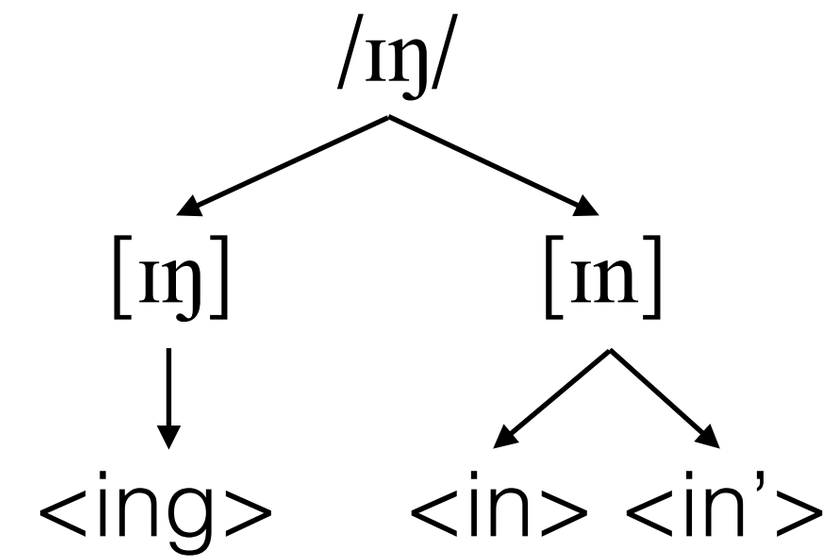
London Boroughs





(ing)

- (ing) a rare example of phonological variation being reflected in orthography



- To what extent do the factors influencing phonological (ing) variation also play a role in orthographic (ing) variation?

(ing)

- Very well-studied in speech, in both American and British varieties of English (e.g. Labov 2001 in Philadelphia; Tagliamonte 2004 in York; Bailey 2015 in Manchester and Blackburn)

- Major conditioning factors:

    - **style**, and speaker **age**, **sex**, and **social class**: all the typical behaviour of a stable sociolinguistic variable (see Hazen 2006 for an overview)

- **grammatical category**: nominal-verbal continuum, reflecting the historical origin of the (ing) alternation by more verb-like words favouring *-in* and more noun-like words favouring *-ing* (see Houston 1985)

- **region**: rates of *-in* much higher in northern regions of the UK, including Scotland (Labov 2001: 90)

    + **word frequency**: not attested in US varieties (see Abramowicz 2007), but small effect in northern English dialects where rates of *-in* highest in the most frequent words

Methodology
(ing)

- Envelope of variation: word-final <ing> cluster…
  - …but only when this represents an unstressed cluster in speech (e.g. *walk**ing*** /ˈwɔː.kɪŋ/, cf. *s**ing*** /sɪŋ/
  - how do you distinguish these?
    - by referring to a pronouncing dictionary, e.g. Carnegie Mellon University dictionary (CMUdict)
    - look up the orthographic word in this dictionary - does the phonemic transcription end in **IH0 NG** (Arpabet equivalent of /ɪŋ/)?
- Word frequency counts from SUBTLEX-UK, measured along the Zipf-scale (van Heuven et al. 2014)

(very) Big data

Bailey (2015) data on (ing):
3,700 tokens
1.2MB

This talk on (ing):
4,706,425 tokens
2.15GB

Bailey (2015) data on (ing):
3,700 tokens
1.2MB

Results

(ing) - Part of speech

- Part of speech effect seems to be absent (Figure 1)
- …but there are major outliers in the adjectival category that show unusually high g-dropping rates e.g. *fricking, frigging, freaking, fucking* etc. (Figure 2)
- Removing these is justified on two accounts:
    - their extremely informal style likely contributes to this effect
    - strictly speaking, they don't often function as adjectives, despite occupying that syntactic position
- Excluding these, there is an extremely small (but significant, $p < 0.001$) trend where verbs show more g-dropping (Figure 3)

Figure 1

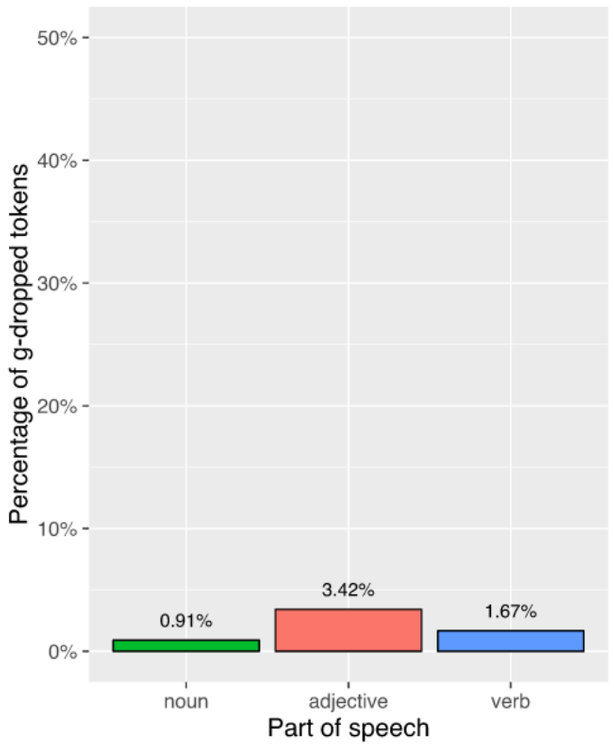


Figure 2

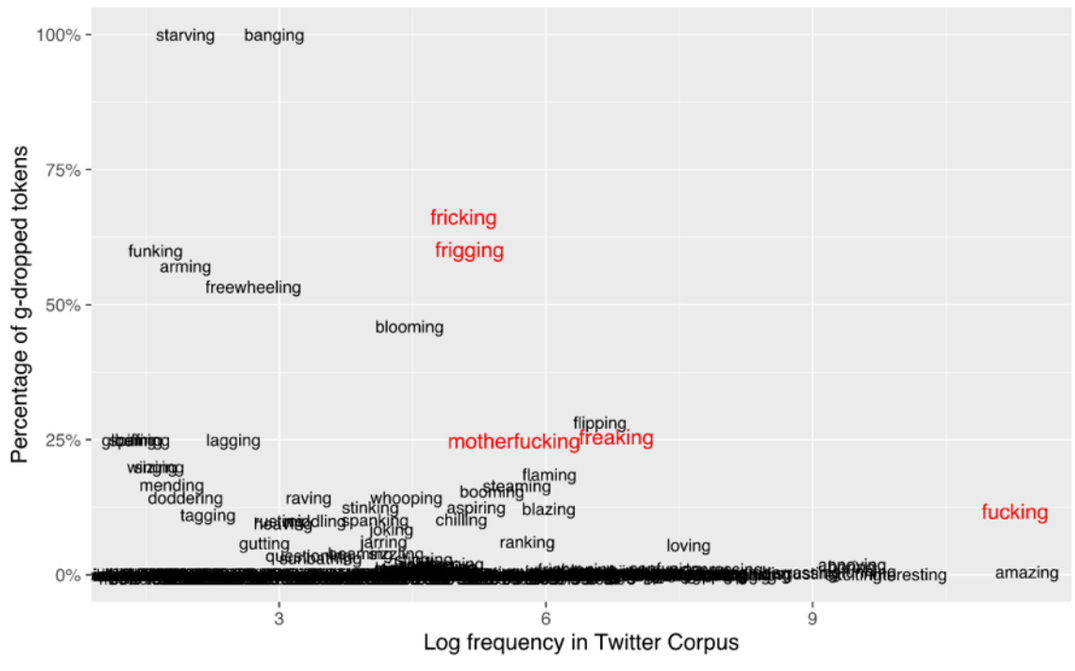
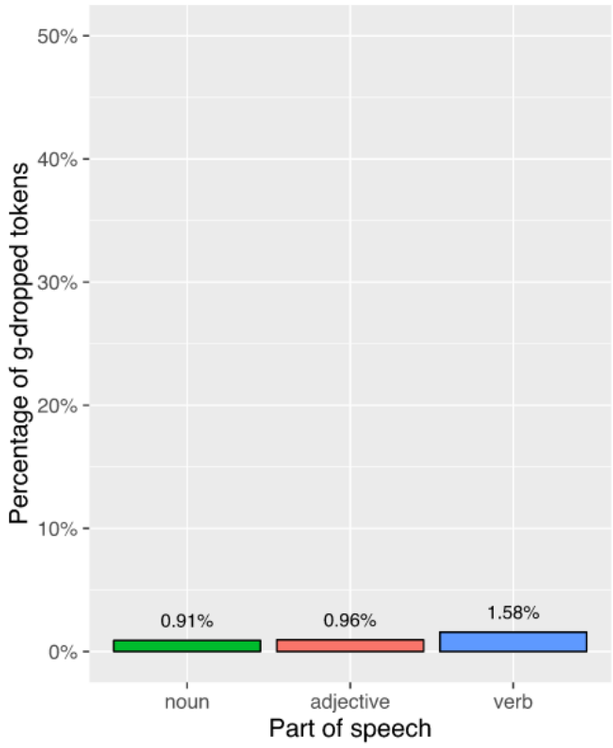


Figure 3



Results

(ing) - Part of speech

- POS effect stronger in the US (Figure 2), but based on a much smaller dataset (around 250,000 tokens, from 1 million tweets)
- But the POS effect is absent in parts of England anyway (Bailey 2015), so it's not surprising that it doesn't show up as strongly in UK Twitter data

Figure 1 - UK Twitter data

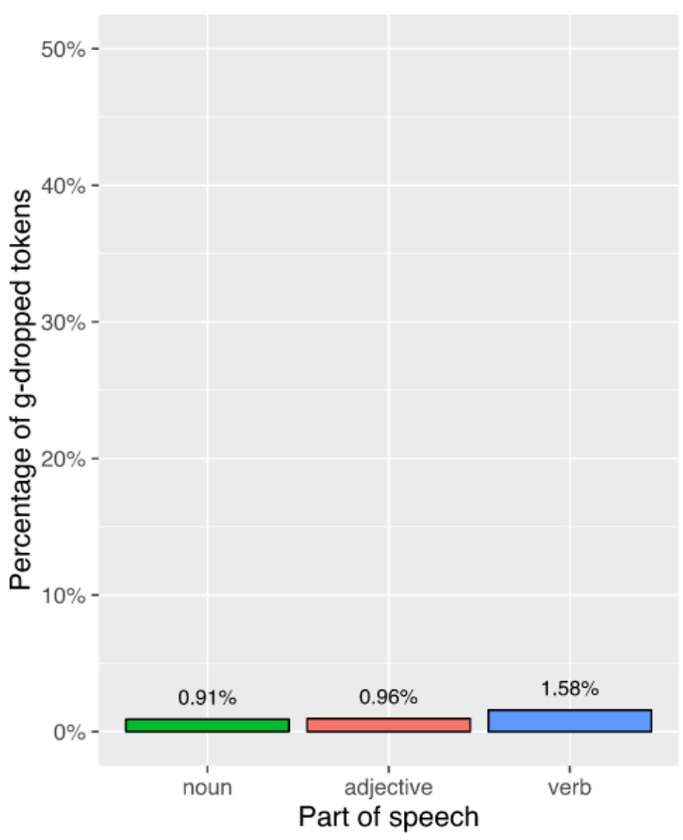


Figure 2 - US Twitter data

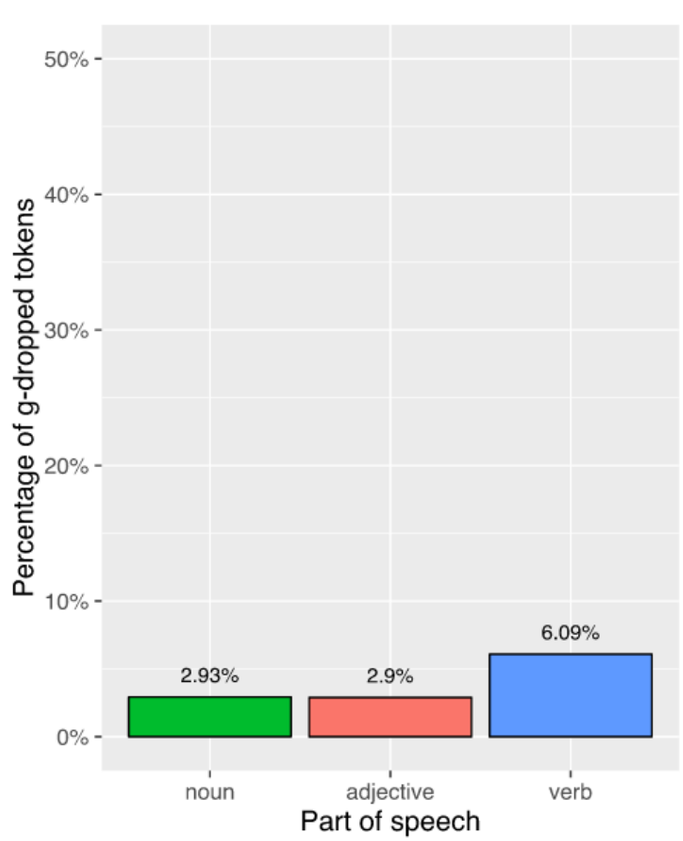
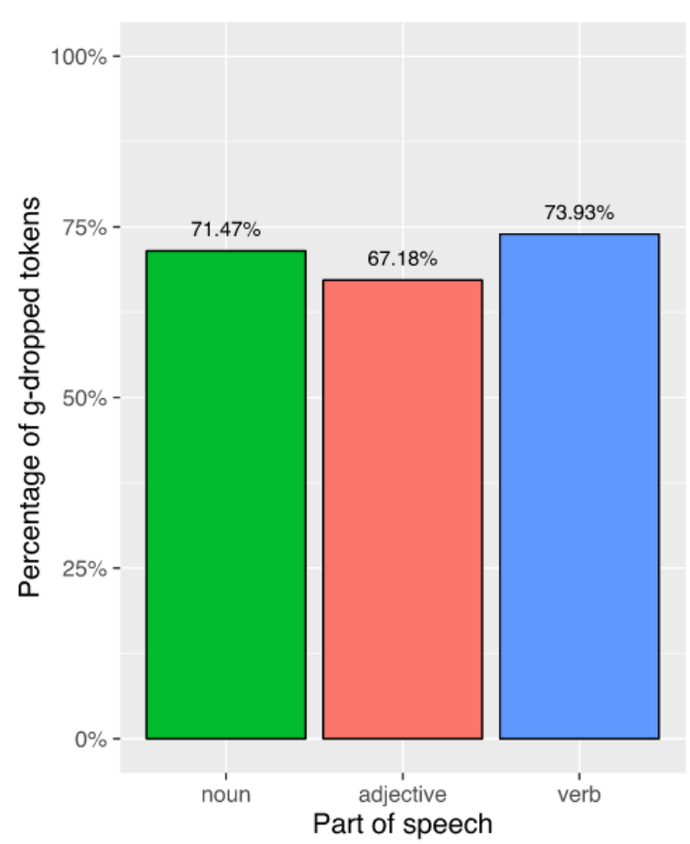
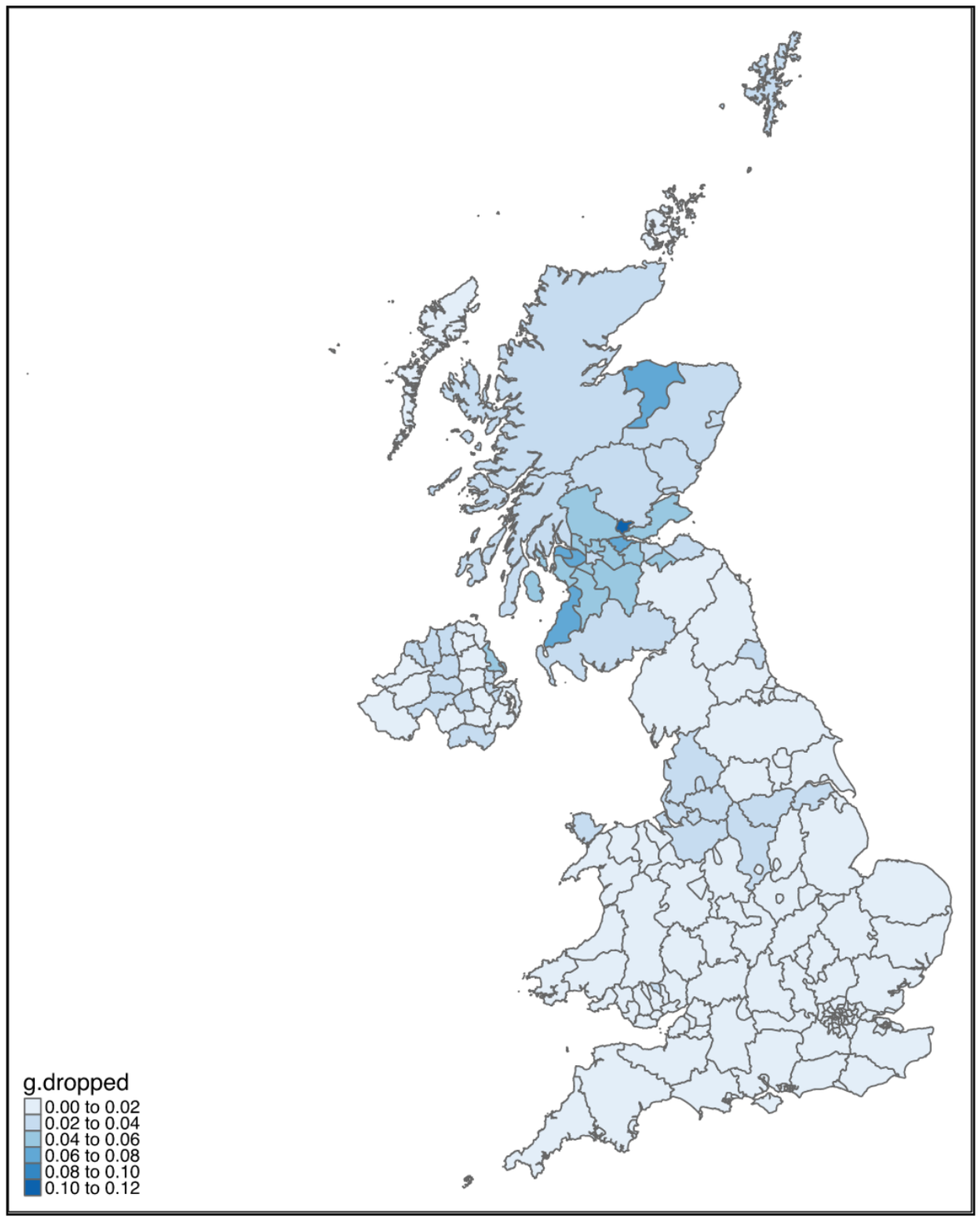


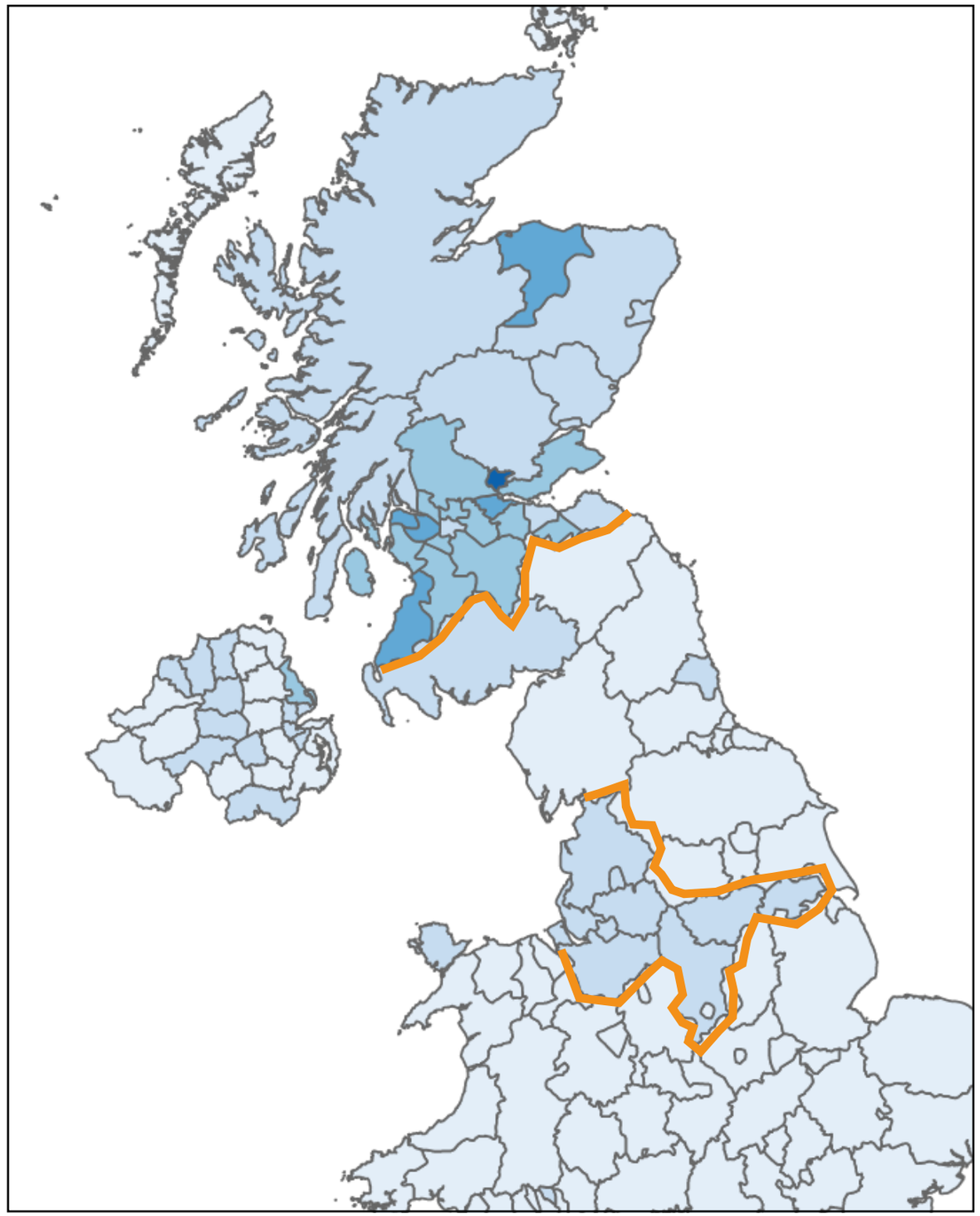
Figure 3 - UK spoken data



Results

(ing) - Region

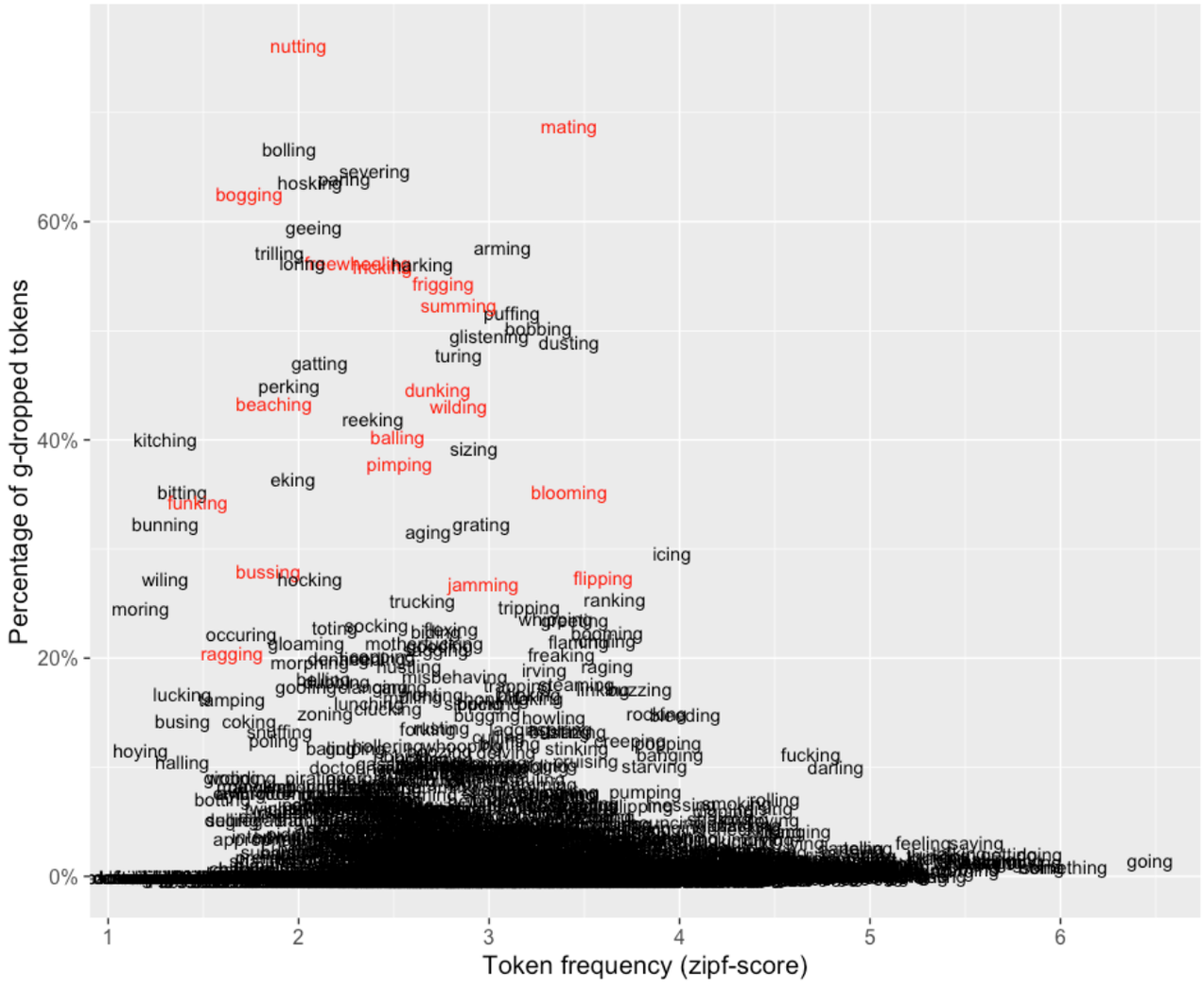




Results

(ing) - Word frequency

- In speech, the most frequent words show even higher rates of *-in,* though the effect is small (Bailey 2015)
- Here, there seems to be an opposite trend
- online slang terms unsurprisingly show the highest rates of g-dropping, despite being infrequent words by normal measures
- e.g. *nuttin* ‘nothing’; *summin* ‘something’; *pimpin, ballin, jammin*



Summary

Methodological:

- Twitter gives you *lots* of data…
- … but not necessarily *good* data
- Geospatial metadata one of its biggest selling points
    - reassuring that patterns found here largely correspond to earlier findings gathered using more traditional methods

Theoretical:

- Interesting grapheme-phoneme parallels between /ɪŋ/ and <ing>
    - not completely unrelated phenomena: phonetically-motivated spelling of (ing) is sensitive to the same social/regional and grammatical factors as its spoken counterpart
    - that said, the behaviour of these low frequency (ing) words is a reminder that stylistic properties of online social media are quite distinct from those in sociolinguistic interviews

#thanks

References

Abramowicz, Ł. 2007. Sociolinguistics meets exemplar theory: frequency and recency effects in (ing). *University of Pennsylvania Working Papers in Linguistics* 13(2): 27-37.

Bailey, G. 2015. *Social and internal constraints on (ing) in northern Englishes*. MA dissertation, University of Manchester.

Derczynski, L., A. Ritter, S. Clarke, & K. Bontcheva. 2013. Twitter part-of-speech tagging for all: overcoming sparse and noisy Data. In Angelova, G., K. Bontcheva, & R. Mitkov (eds.), *Proceedings of Recent Advances in Natural Language Processing*, 198-206. Shoumen, Bulgaria: INCOMA Ltd.

Grieve, J. 2015. Dialect variation. In Biber, D. & R. Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 362-380. Cambridge: Cambridge University Press.

Hazen, K. 2006. The IN/ING variable. In K. Brown (ed.), *Encyclopedia of Language and Linguistics*, 581-584. St. Louis: Elsevier.

Houston, A. 1985. *Continuity and change in English morphology: the variable (ING)*. Ph.D dissertation, University of Pennsylvania.

Labov, W. 1989. The child as linguistic historian. *Language Variation and Change* 1: 85-97.

Labov, W. 2001. *Principles of linguistic change vol. 2: social factors*. Malden, MA: Blackwell.

MacKenzie, L., G. Bailey & D. Turton. 2015. Our Dialects: Mapping variation in English in the UK [Website]. <http://tiny.cc/OurDialects>

Pak, A. & P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.

Schnoebelen, T. 2012. Do You Smile with Your Nose? Stylistic Variation in Twitter Emoticons. *University of Pennsylvania Working Papers in Linguistics* 18(2): 117-125.

Tagliamonte, S. 2004. Somethi[ŋ]'s goi[n] on! Variable (ing) at ground zero. In Gunnarsson, B., L. Bergström, G. Eklund, S. Fridell, L. Hansen, A. Karstadt, B. Nordberg, E. Sundgrenand & M. Thelander (eds.), *Language Variation in Europe: Papers from the Second International Conference on Language Variation in Europe (ICLaVE 2)*, 390-403. Uppsala: University Press.

Van Heuven, W. J. B., P. Mandera, E. Keuleers, & M. Brysbaert. 2014. Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology* 67: 1176-1190.