Assessing the Use of Social Media for Mapping Lexical Variation in British English

Jack Grieve, Aston University (@JWGrieve)

Chris Montgomery, University of Sheffield (@montgomerychris)

Andrea Nini, University of Manchester (@and_nini)

Diansheng Guo, University of South Carolina (@Diansheng_Guo)

6 June 2017

ICLAVE 9

Malaga, Spain

Twitter Dialect Studies

Corpus-based dialect studies are increasingly common.

Very large geocoded Twitter corpora in particular have been studied in detail.

But social media is often seen as being too unusual to tell us anything about linguistic variation more generally.

Spanish:
Gonçalves & Sánchez (2014)

American English:
Eisenstein et al. (2014)
Huang et al. (2016)

British English:
Bailey (2016)

Traditional Dialect Studies

Dialect data is traditionally collected by eliciting language directly from individual informants either through surveys or interviews.

Surveys allow for rare forms to be studied, but the forms must be pre-selected.

Interviews allow for language use to be directly observed, but only frequent forms.

Both approaches suffer from the Observer's Paradox.

Twitter Dialect Studies: Pros and Cons

Pros:

Natural language

No need to pre-select features

Infrequent features (incl. Lexis)

High spatial resolution/precision

Large numbers of informants

Relatively easy to implement

Cons:

Bad Tweets: RTs/Spam/Ads/Bots

Phonetics/Phonology

Control over informants

Control over features (Polysemy)

Situational/Social generalisability

Do Twitter Maps Represent General Patterns?

Twitter is an important variety of language and is worth studying in its own right.

But it is also the only current source of big data for geolinguistics and so it is important to asses whether Twitter maps represent general patterns.

In this study, we therefore map lexical alternations in a corpus of UK Tweets and compare our maps to the maps from the BBC Voices survey (Wieling et al. 2014).

UK Twitter Corpus

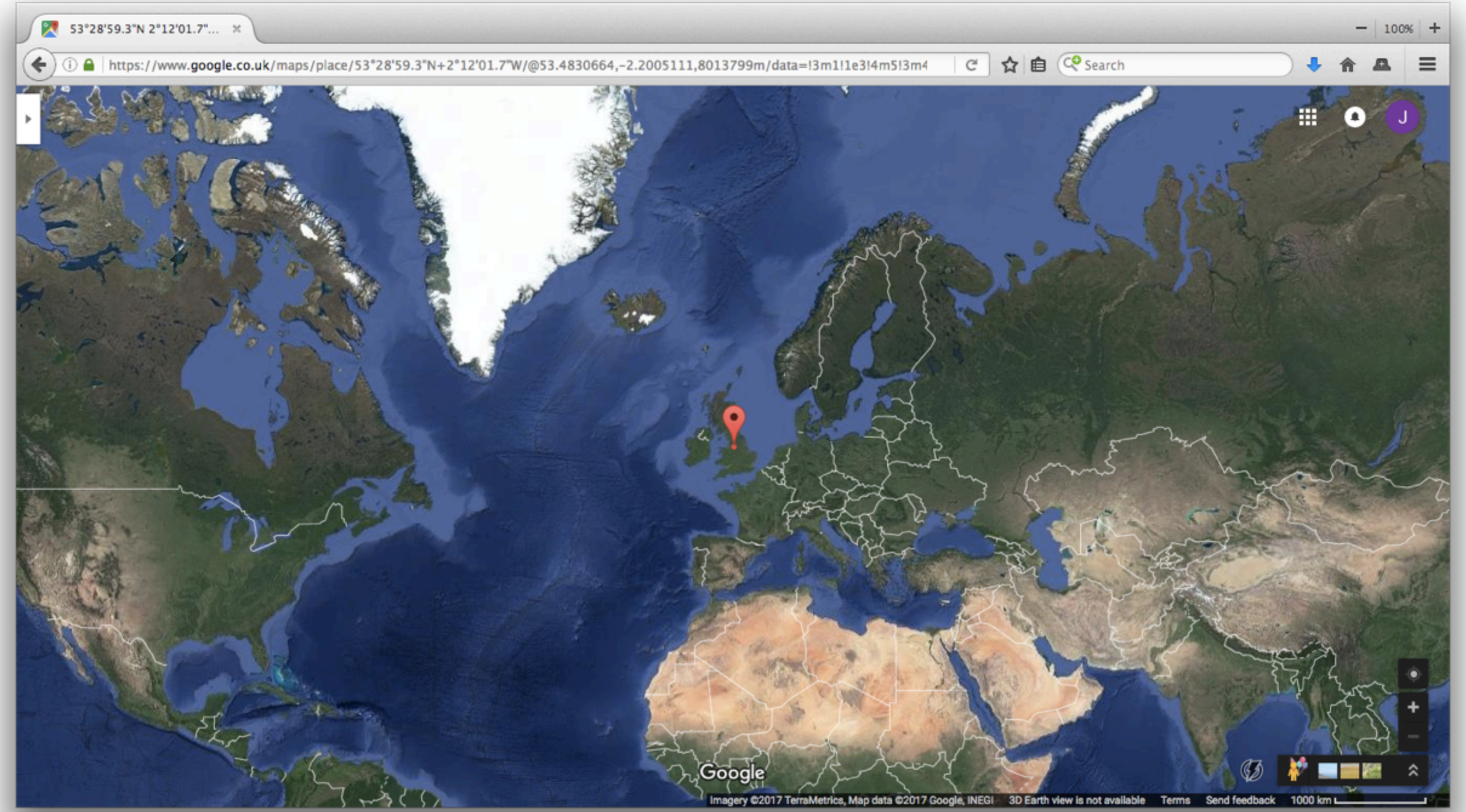
The corpus only contains Tweets that are geocoded with the longitude and latitude of the user at the time of posting

Most Tweets are not geocoded, but we still get enough to compile a very large corpus.

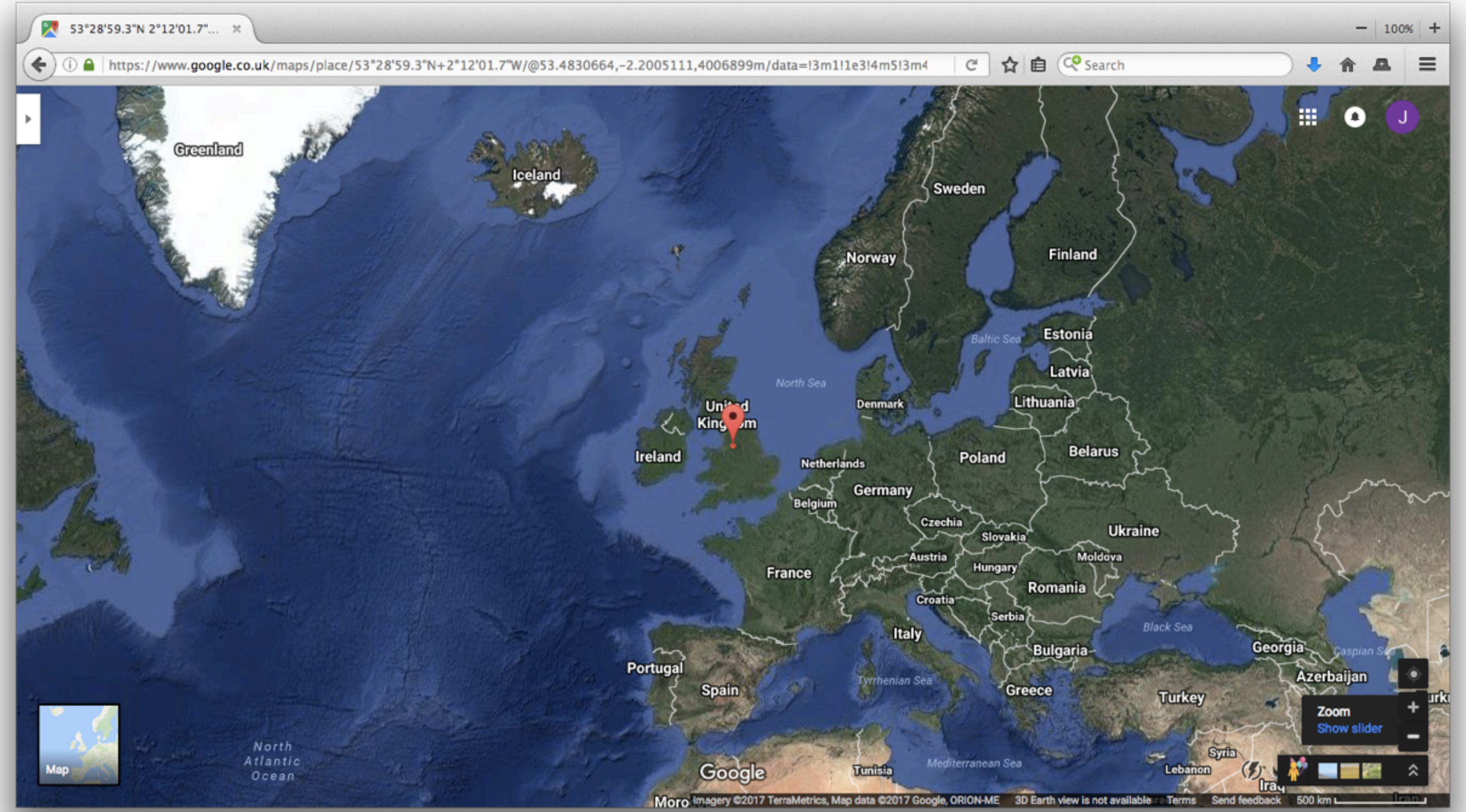
We downloaded the UK Twitter Corpus between 1/1/2014 and 31/12/2014 using the Twitter API.

In total our corpus contains approximately 2 billion words and 180 million Tweets, written by 1 million users.

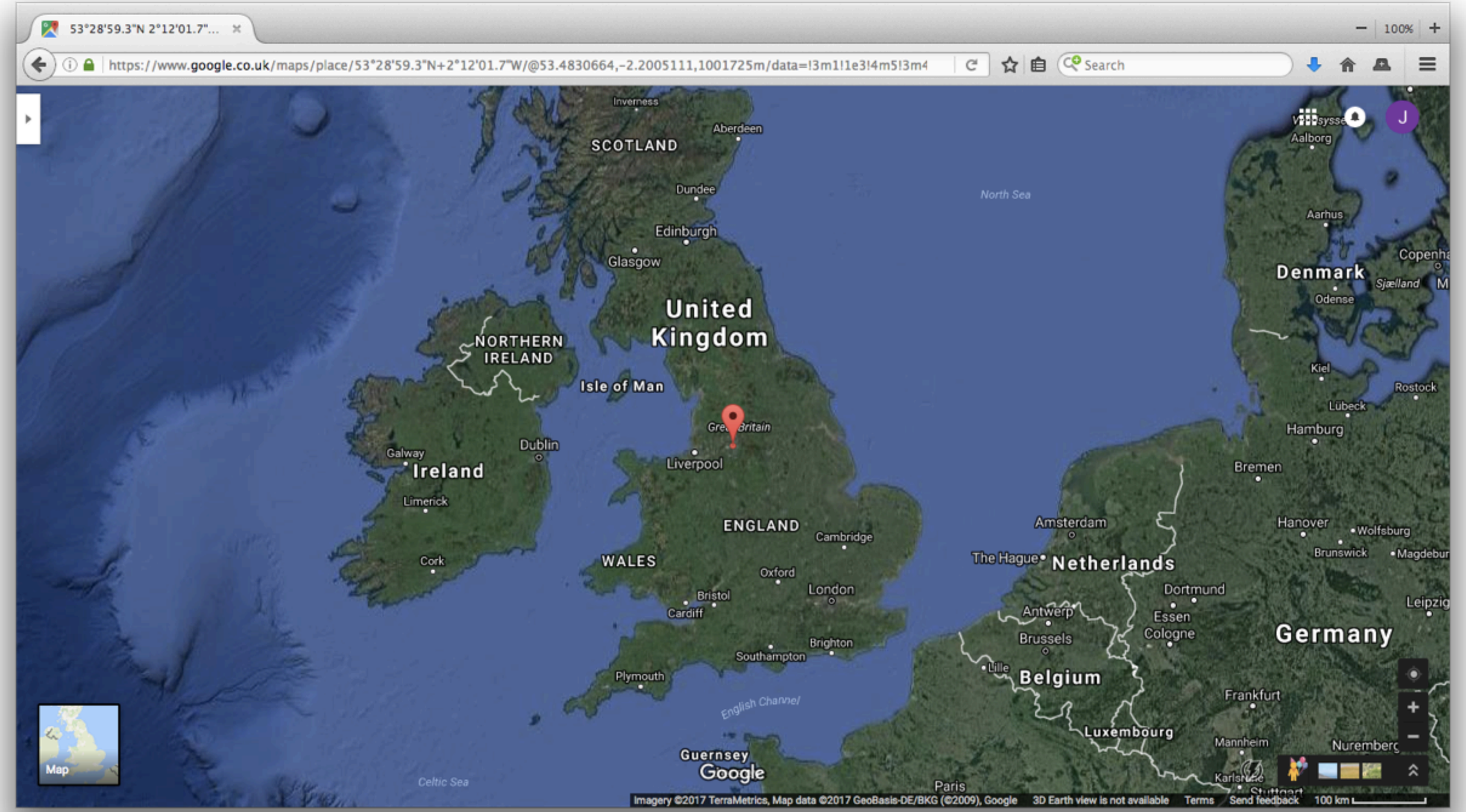
53.483143, -2.2004628



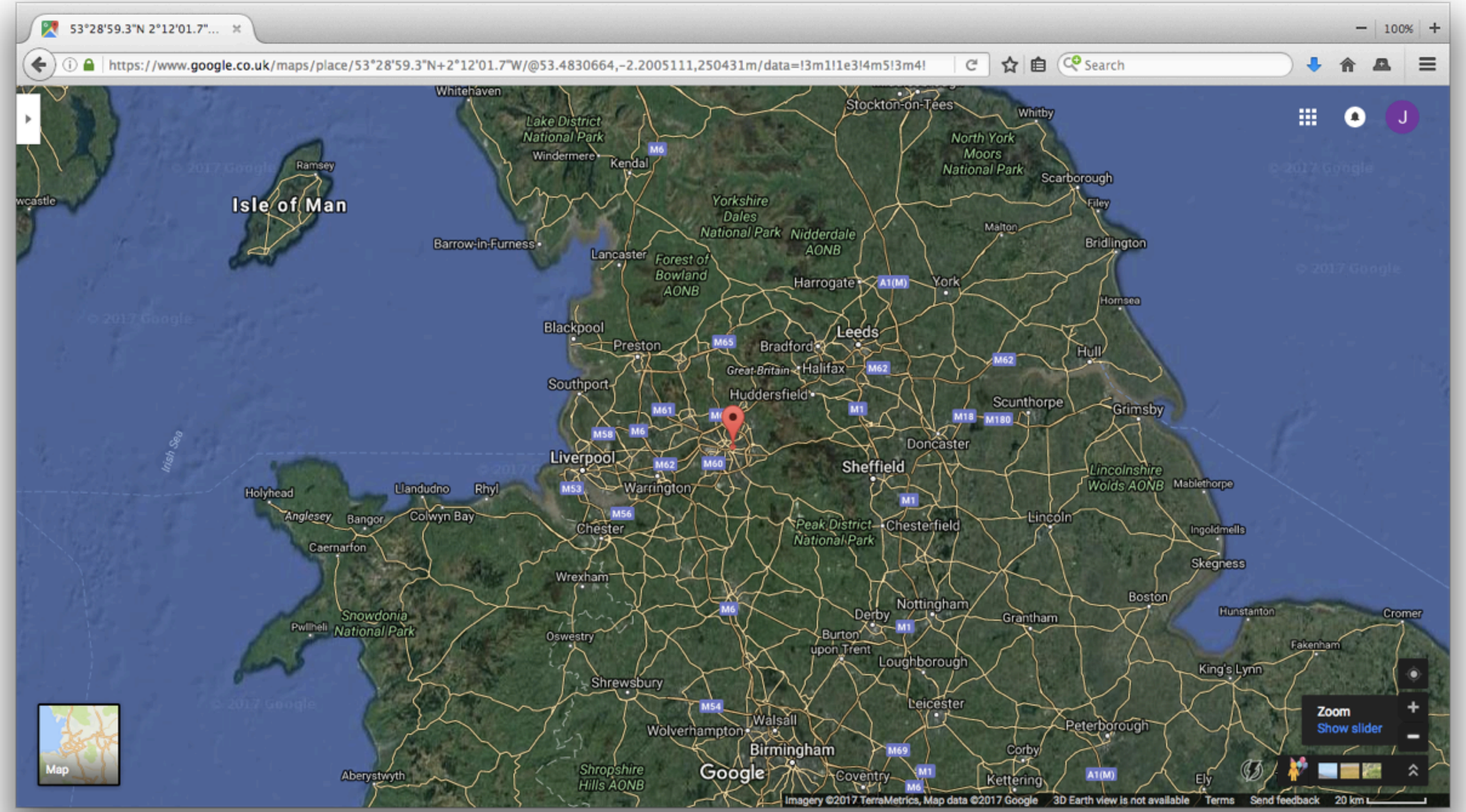
53.483143, -2.2004628, Patience Paid Off! @ Etihad Stadium



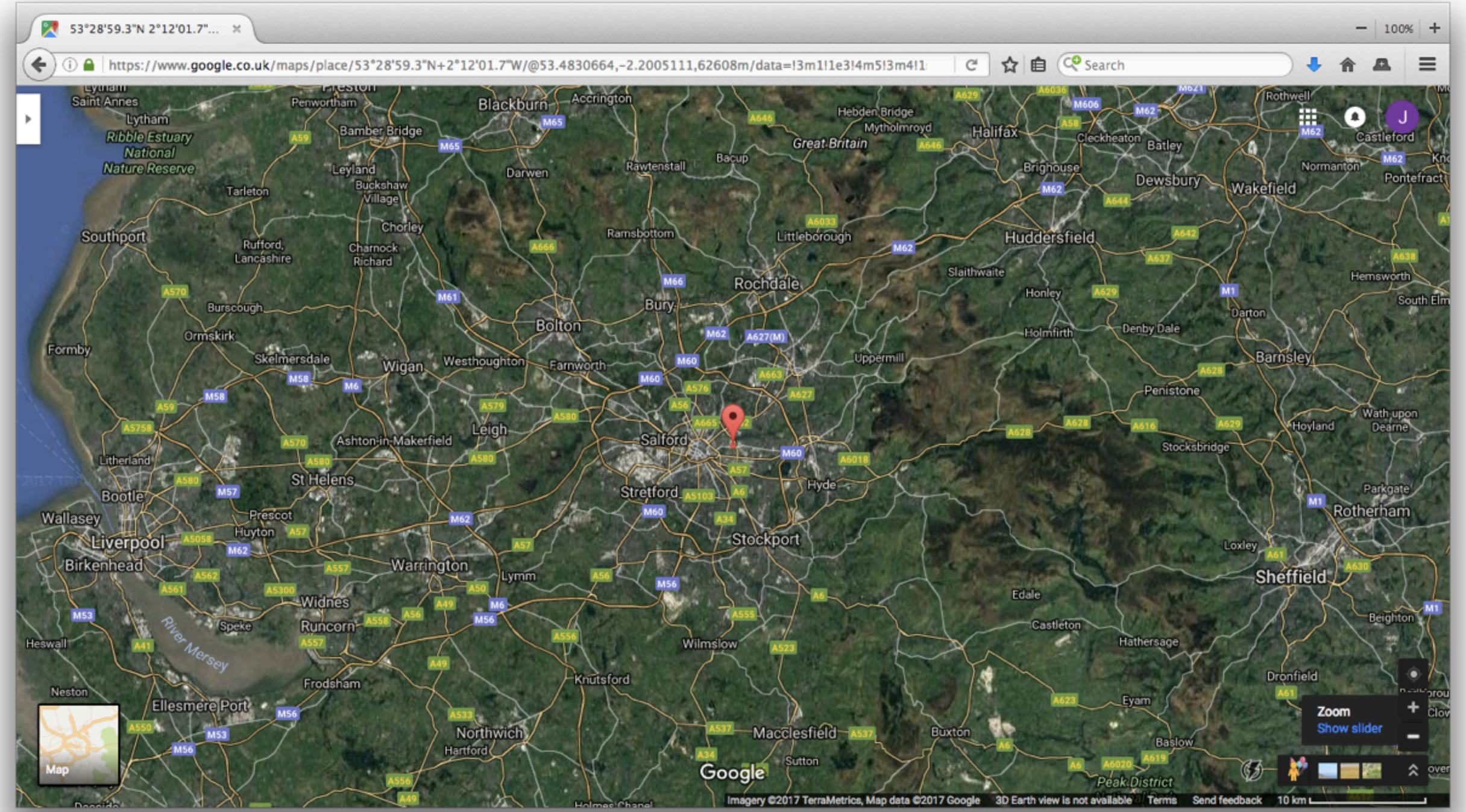
53.483143, -2.2004628, Patience Paid Off! @ Etihad Stadium



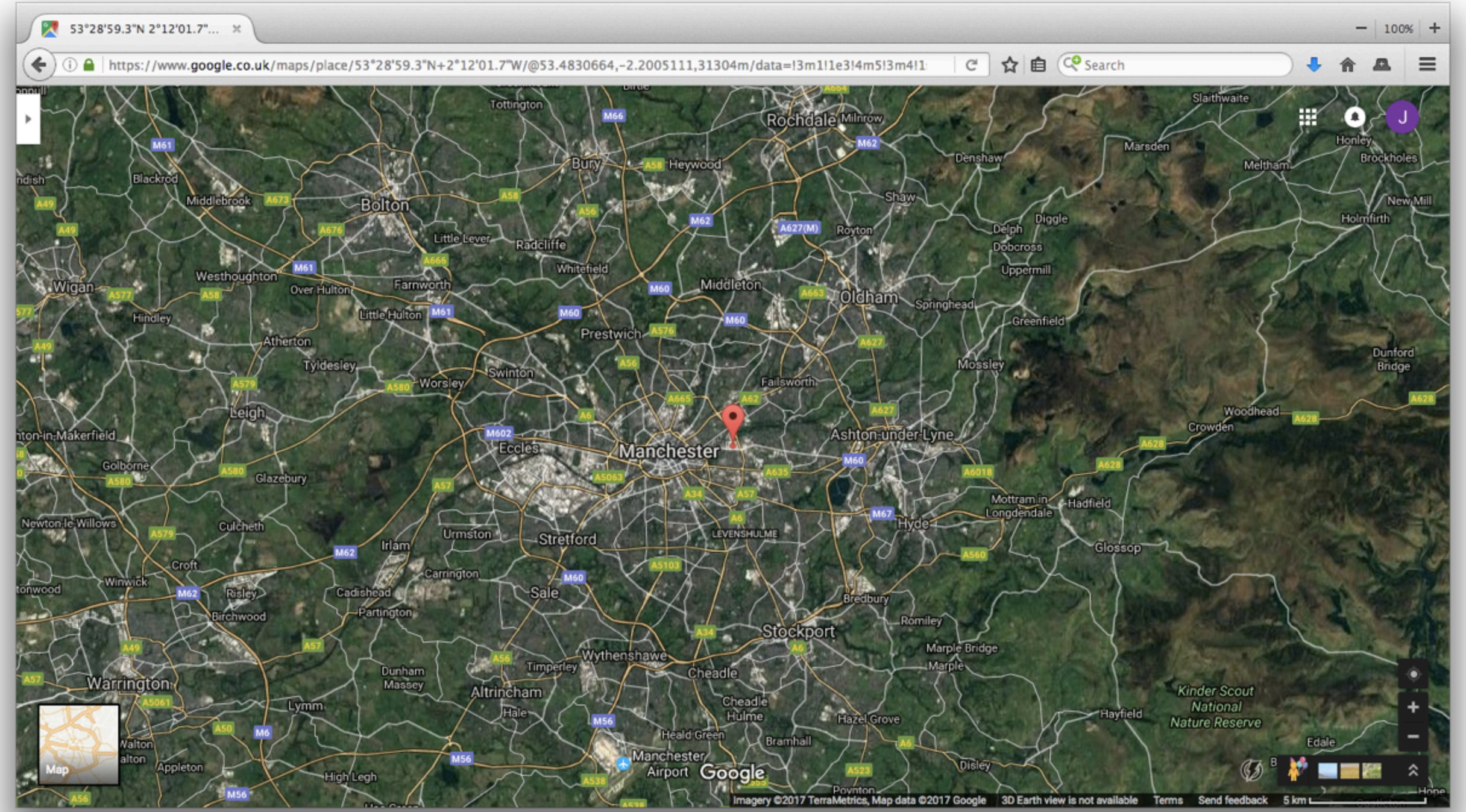
53.483143, -2.2004628, Patience Paid Off! @ Etihad Stadium



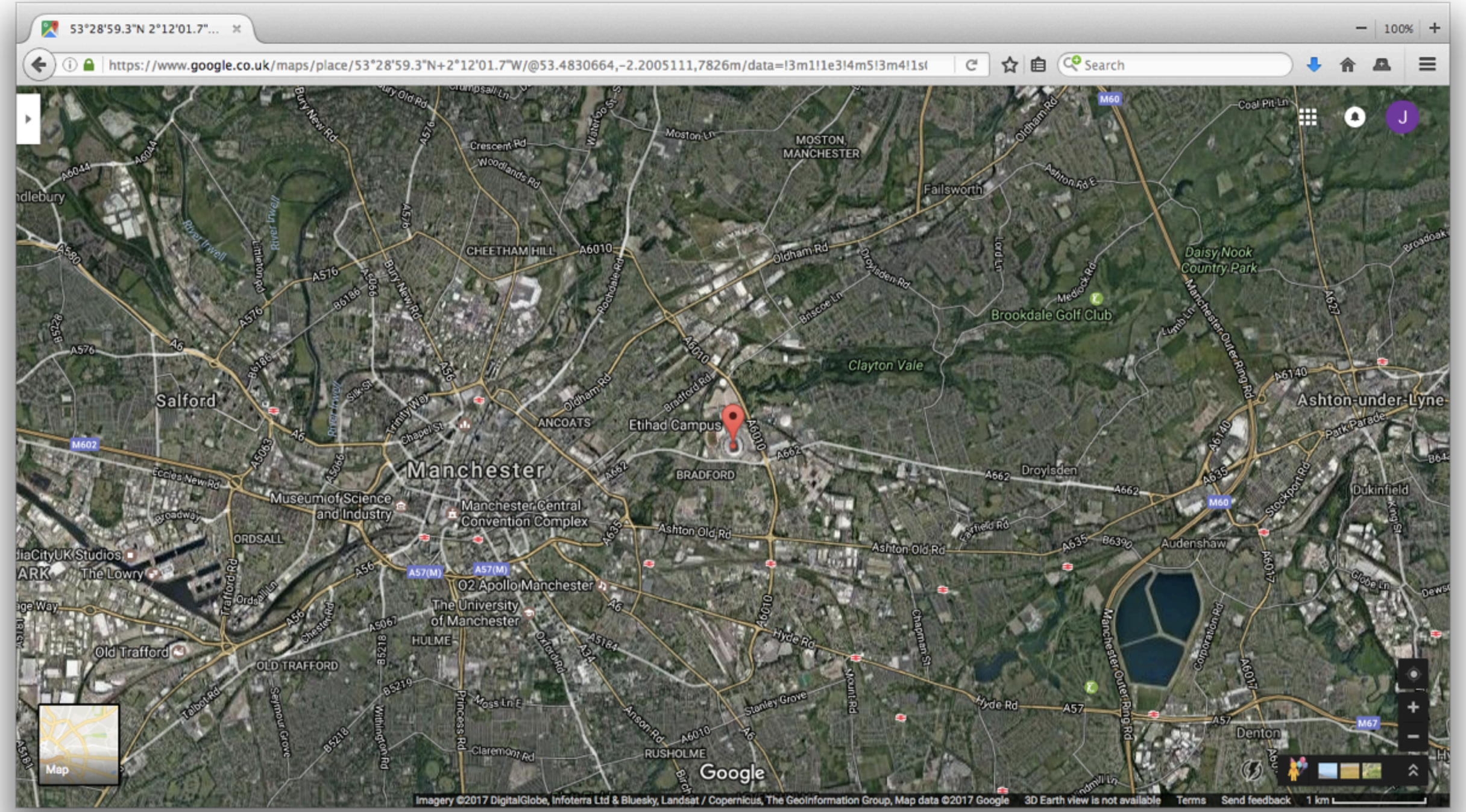
53.483143, -2.2004628, Patience Paid Off! @ Etihad Stadium



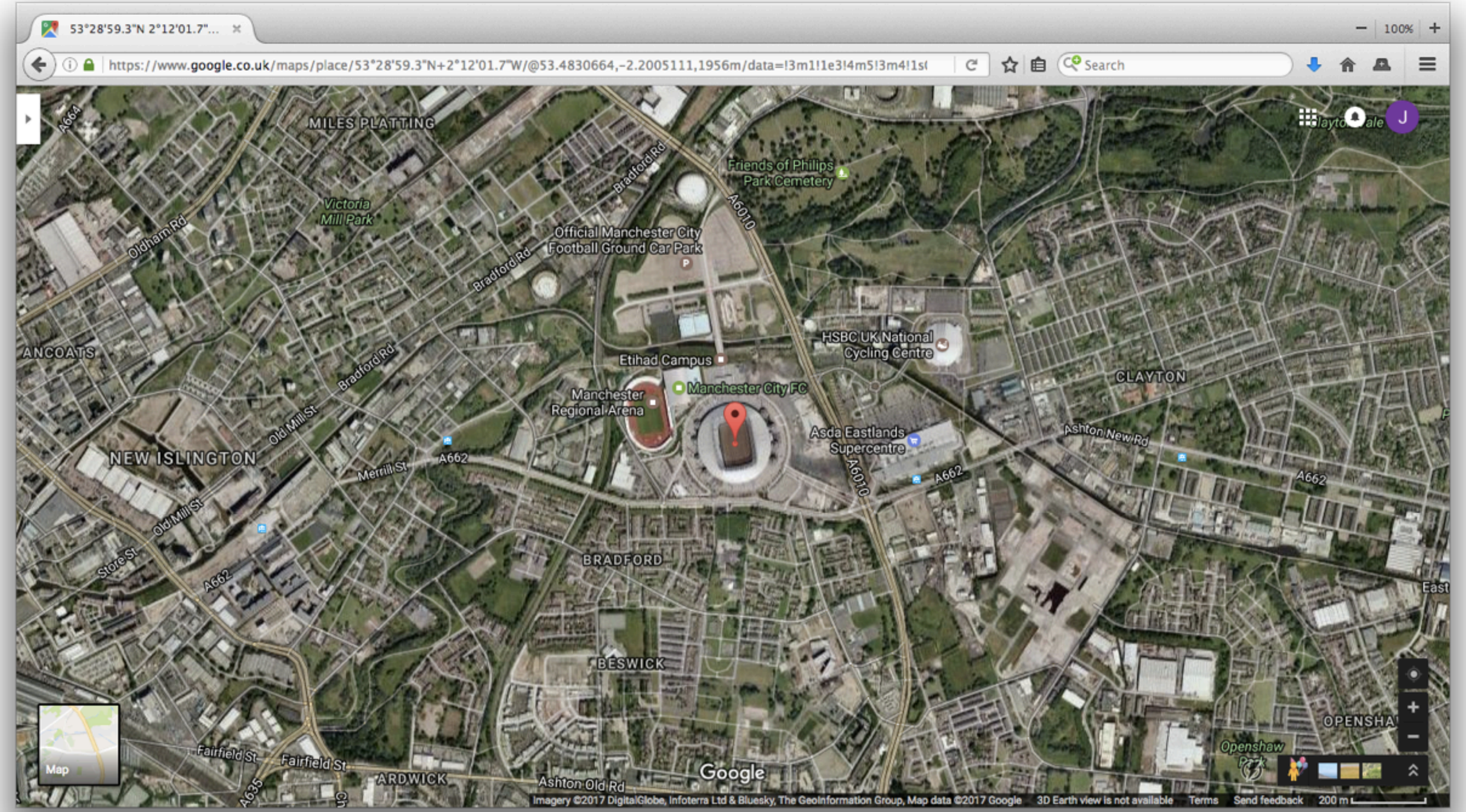
53.483143, -2.2004628, Patience Paid Off! @ Etihad Stadium



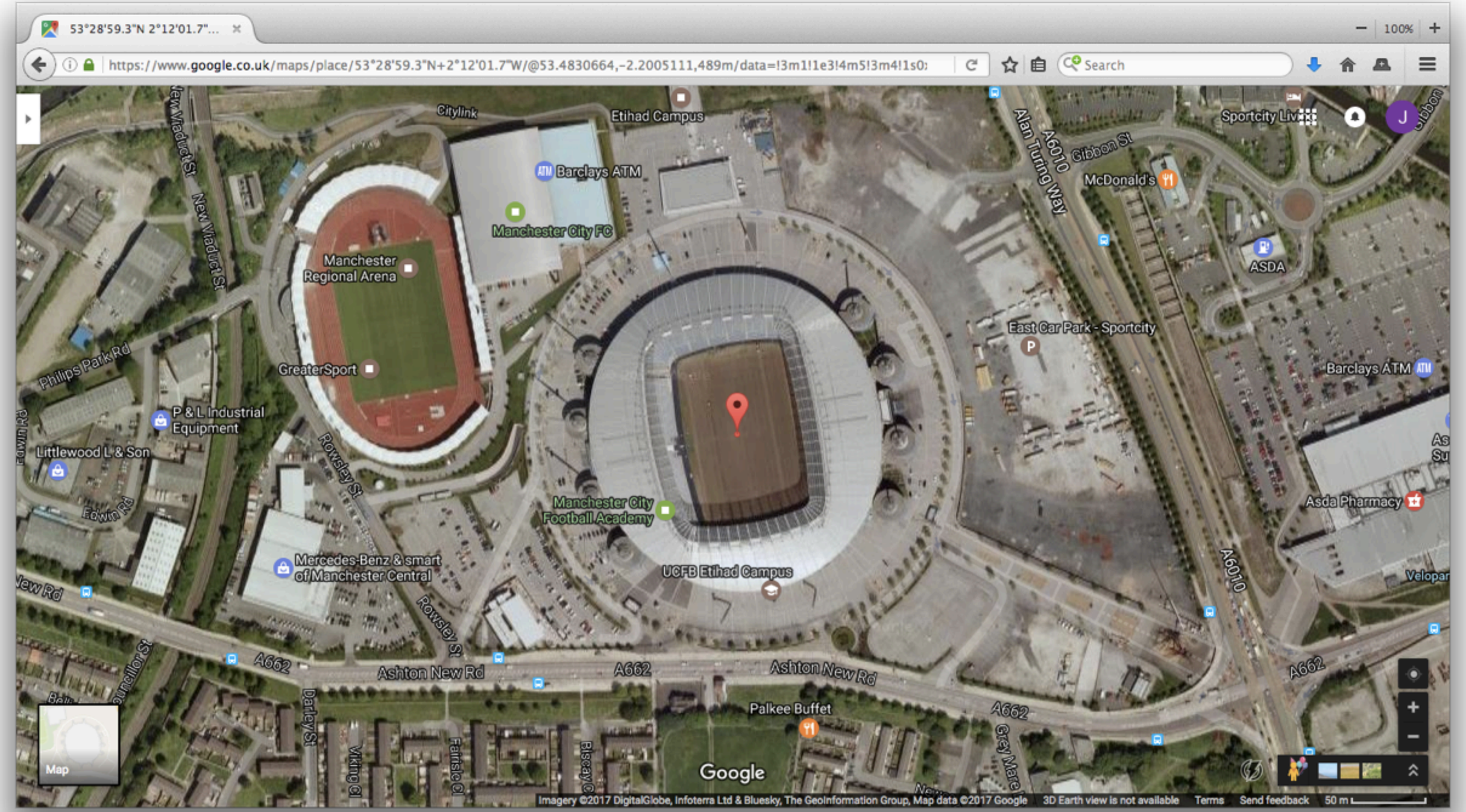
53.483143, -2.2004628, Patience Paid Off! @ Etihad Stadium



53.483143, -2.2004628, Patience Paid Off! @ Etihad Stadium



53.483143, -2.2004628, Patience Paid Off! @ Etihad Stadium



53.483143, -2.2004628, Patience Paid Off! @ Etihad Stadium

Analysis

We grouped tweets by postal code regions to facilitate comparison with BBC Voices.

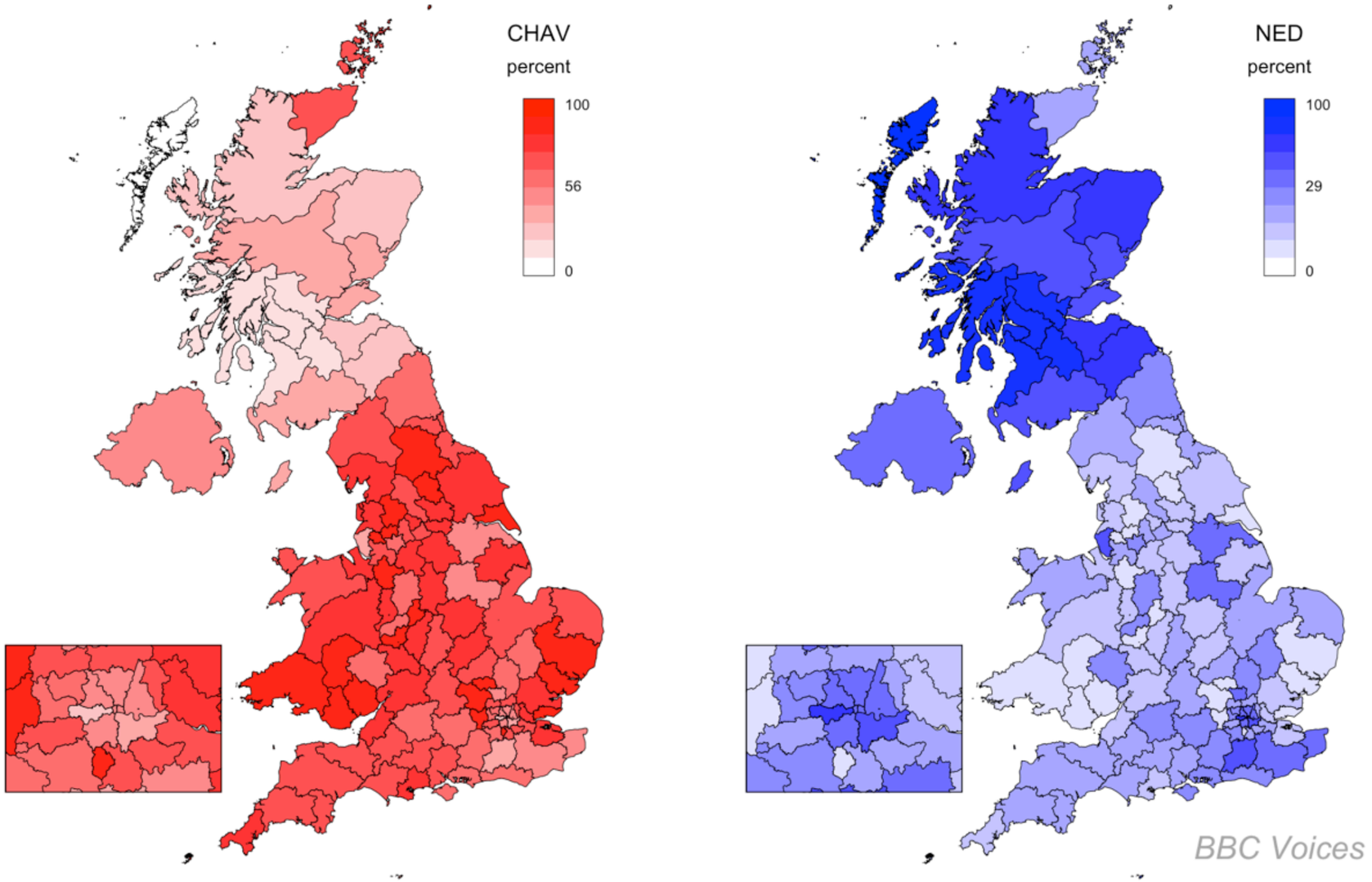
We took 35 lexical alternations and their 115 variants that were returned by at least 5% of the BBC informants and that occurred at least 1,000 times in our corpus.

For each variant we calculated its proportional use vs. all the variants of that alternation in each postal code area.

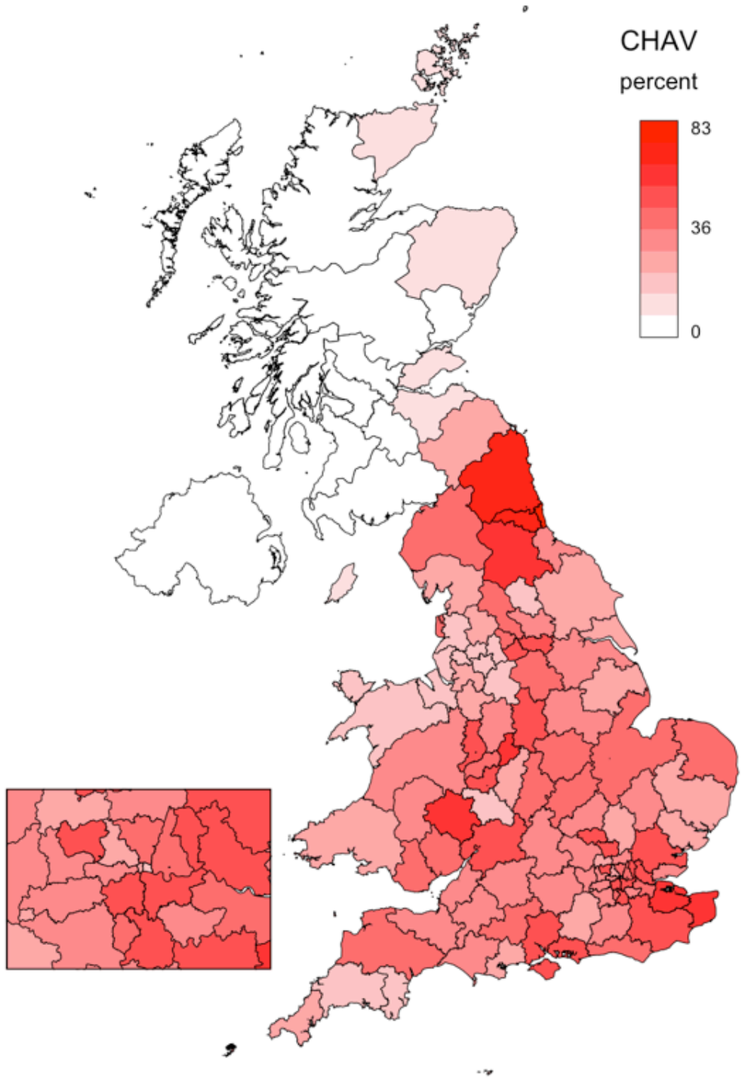
We then compared and correlated (Spearman) these maps to the BBC voices maps to assess similarity.

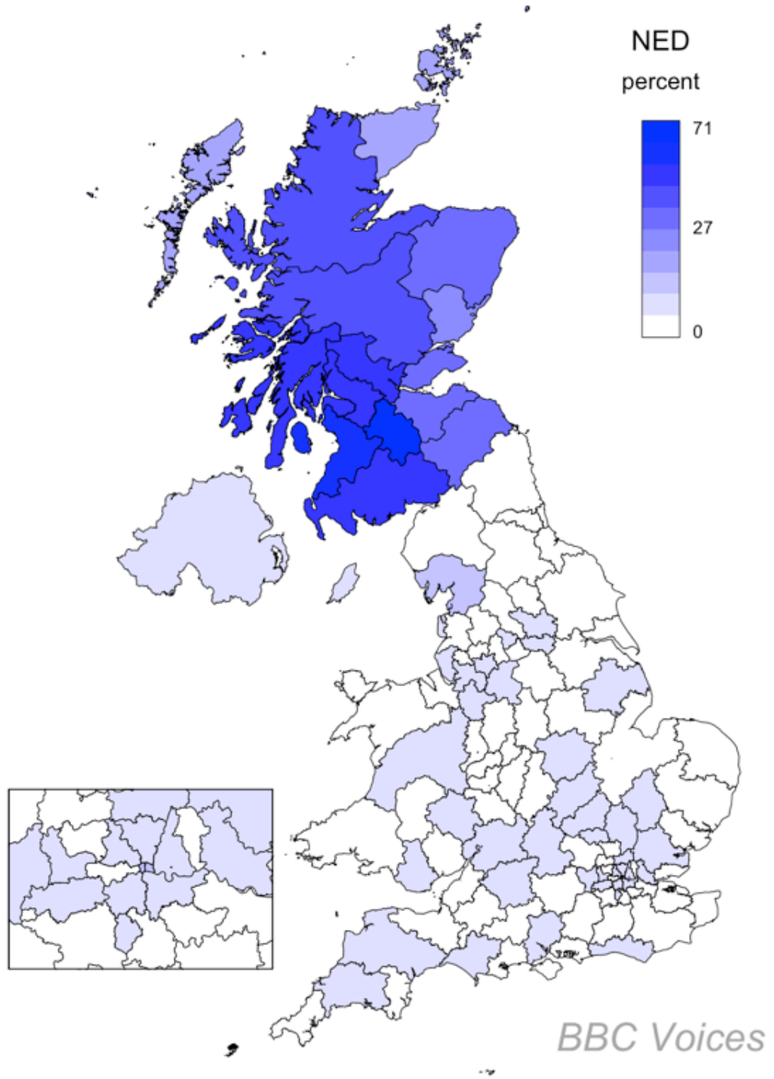
boiling	roasting	hot	baked	sweating
freezing	chilly	nippy	cold	
shattered	knackered			
sick	poorly	ill		
chuffed	happy	made.up		
pissed.off	angry			
play	lake			
skive	bunk	wag		
chuck	lob			
whack	smack	thump	wallop	belt
kip	sleep	snooze	nap	doze
pissed	wasted			
pregnant	expecting			
skint	broke	poor		
loaded	minted	well.off		
mad	nuts	crazy	mental	bonkers
fit	gorgeous	pretty	hot	
ugly	minger			
mardy	grumpy	stroppy	moody	
baby	bairn	wean	kid	little.one
mum	mam	mummy	ma	
nanny	granny	grandma		
grandad	grandpa	grampa		
mate	pal	friend	buddy	
chav	ned			
clothes	gear	clobber	kit	
trousers	pants	jeans		
pumps	daps	trainers		
living.room	lounge	sitting.room	front.room	
sofa	settee	couch		
loo	bog	toilet		
alley	path	pavement		
drizzle	spit	shower		
pour	chuck	bucket		
stream	brook	burn	beck	

Twitter: *Chav*/*Ned*

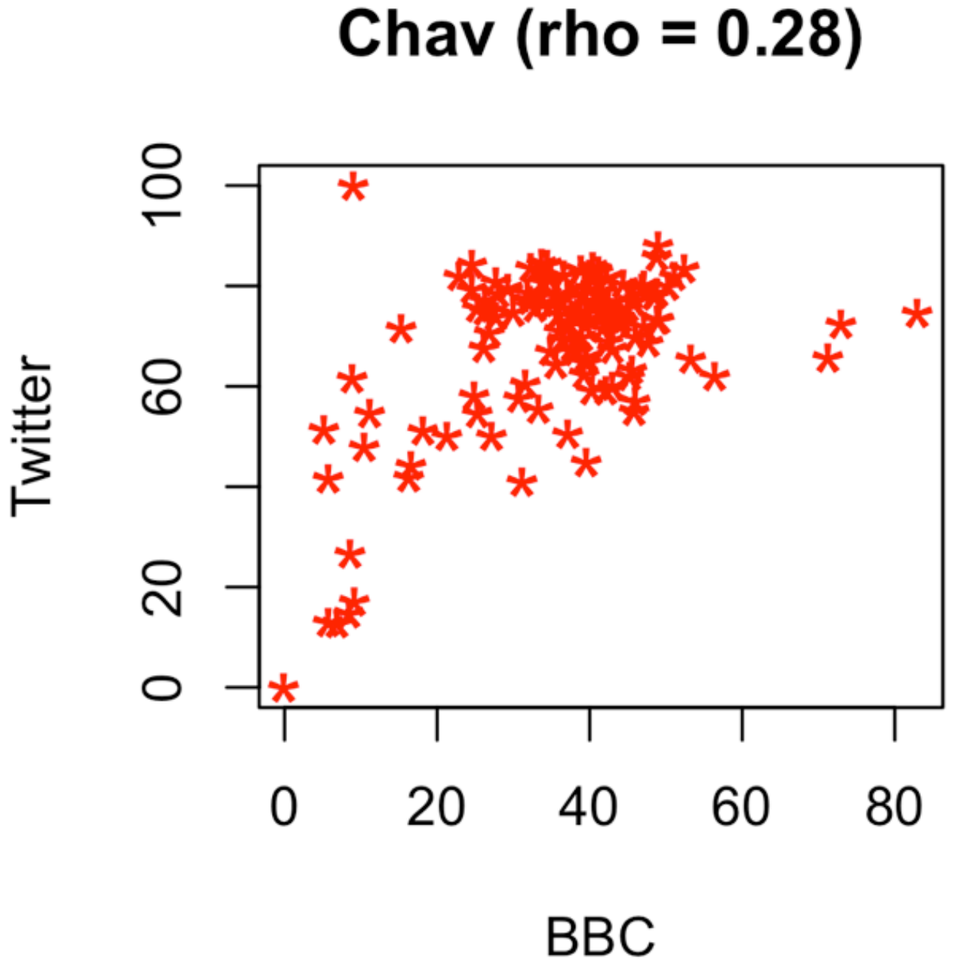


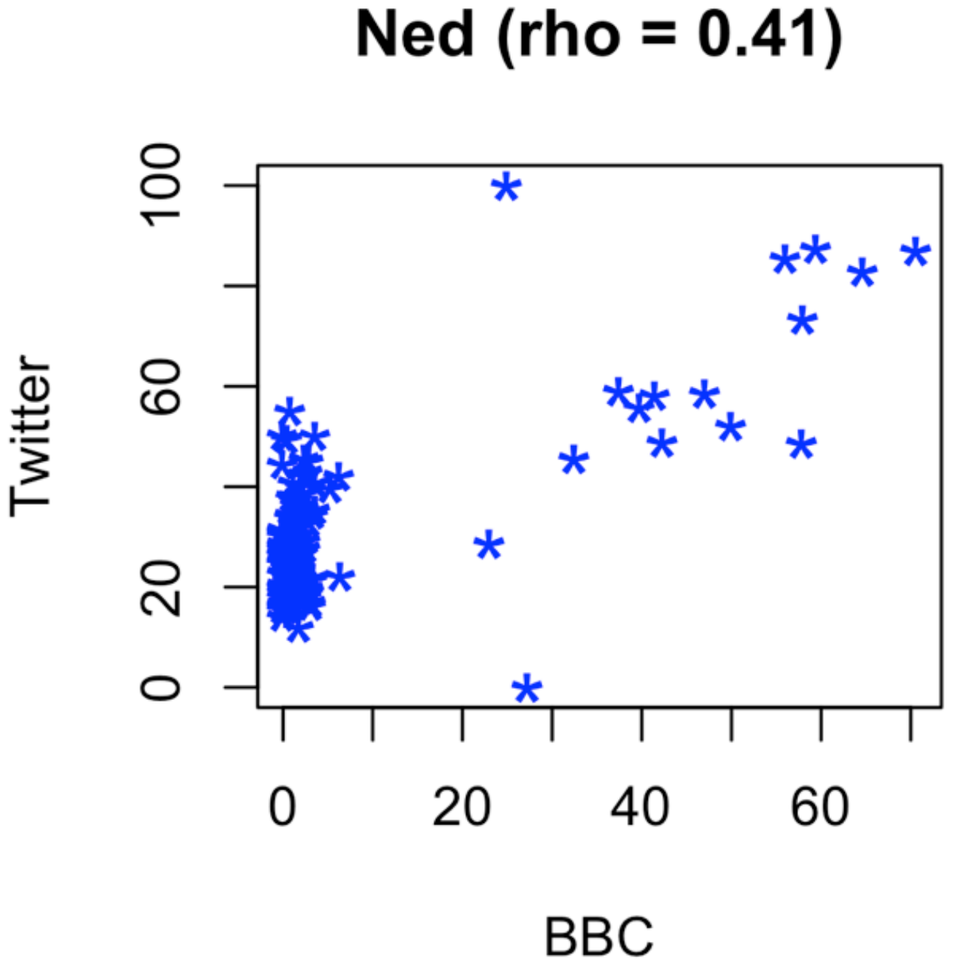
BBC Voices: *Chav*/*Ned*



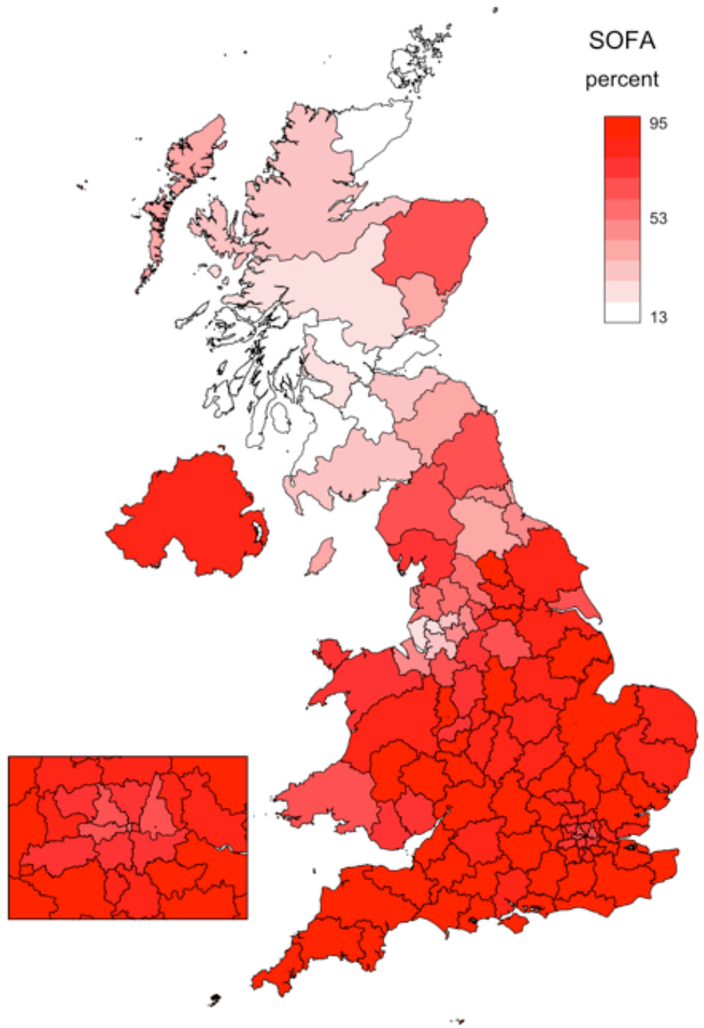


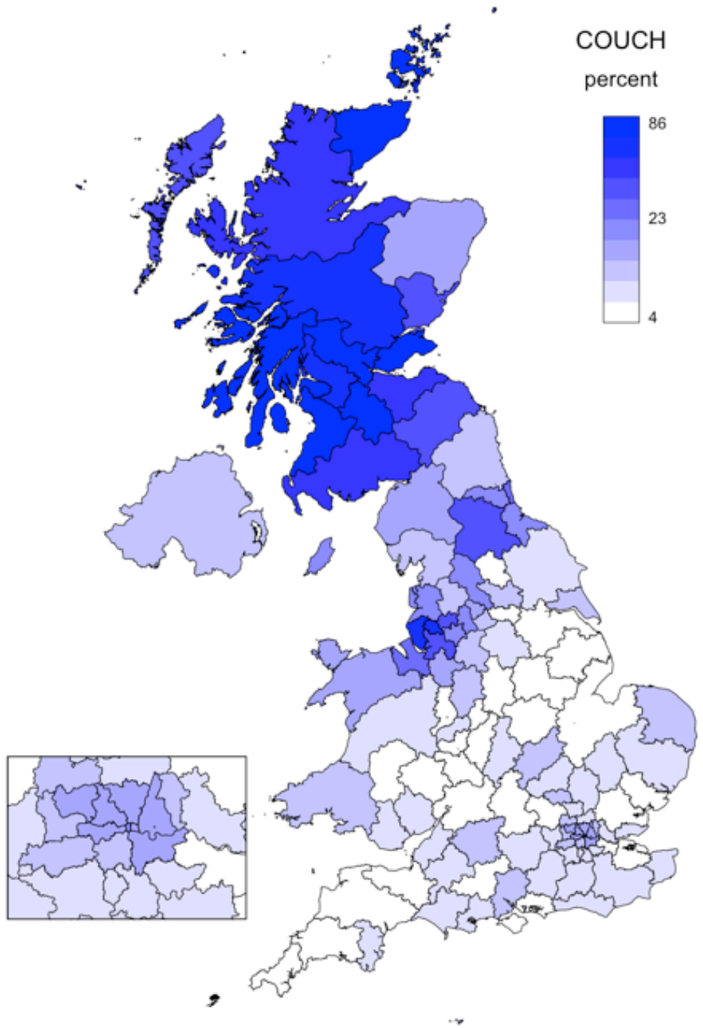
Comparison: *Chav*/*Ned*

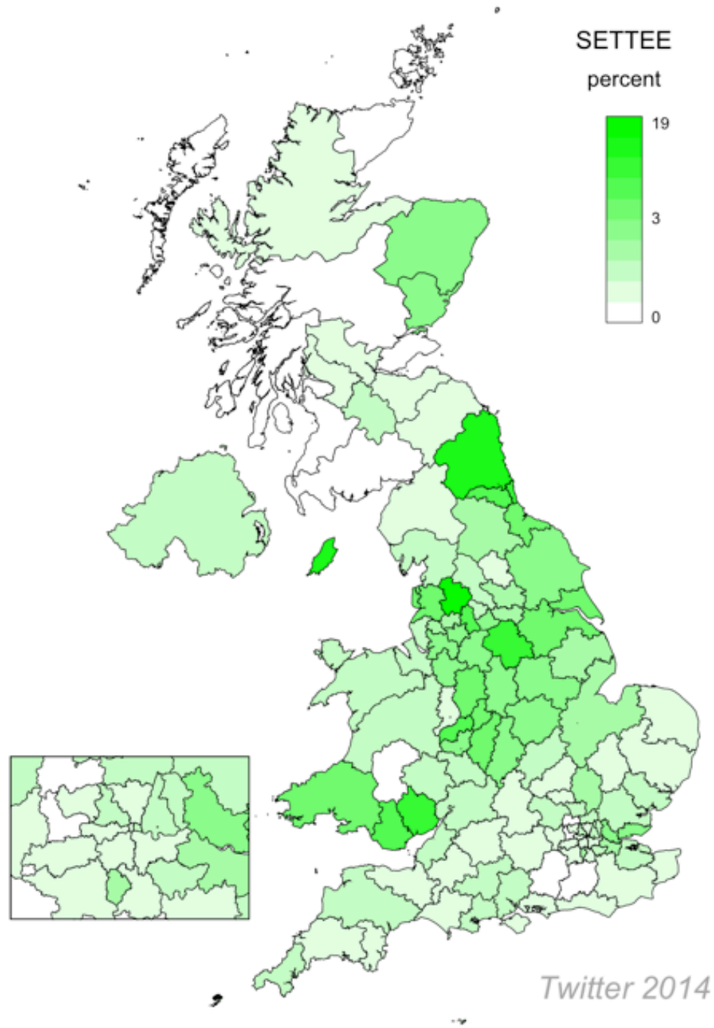




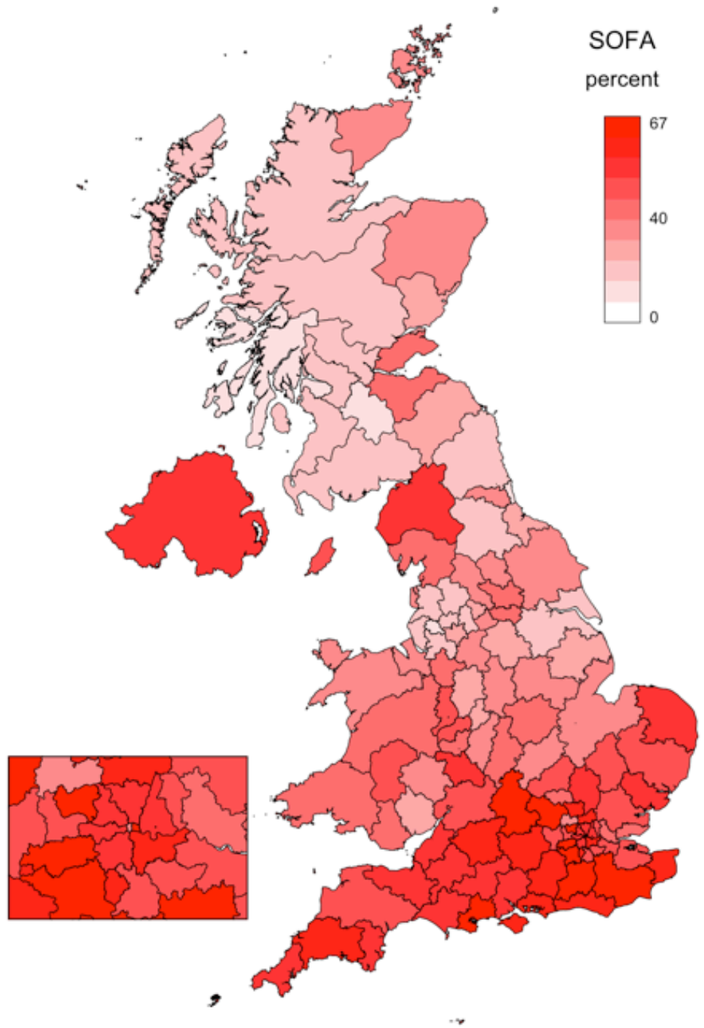
Twitter: *Sofa*/*Couch*/*Settee*

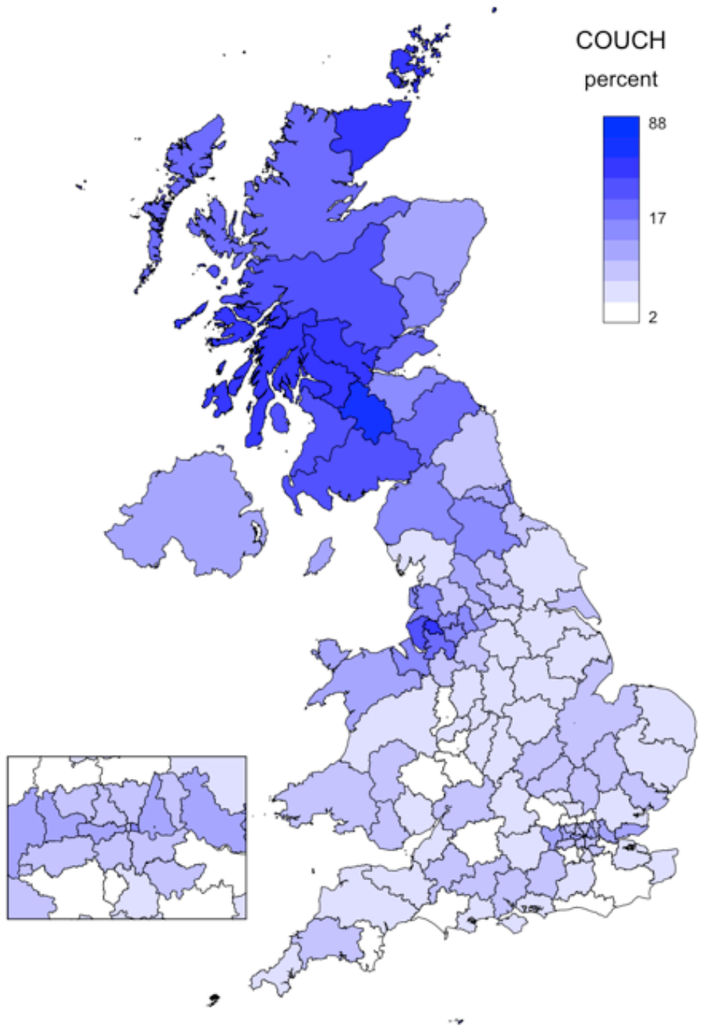


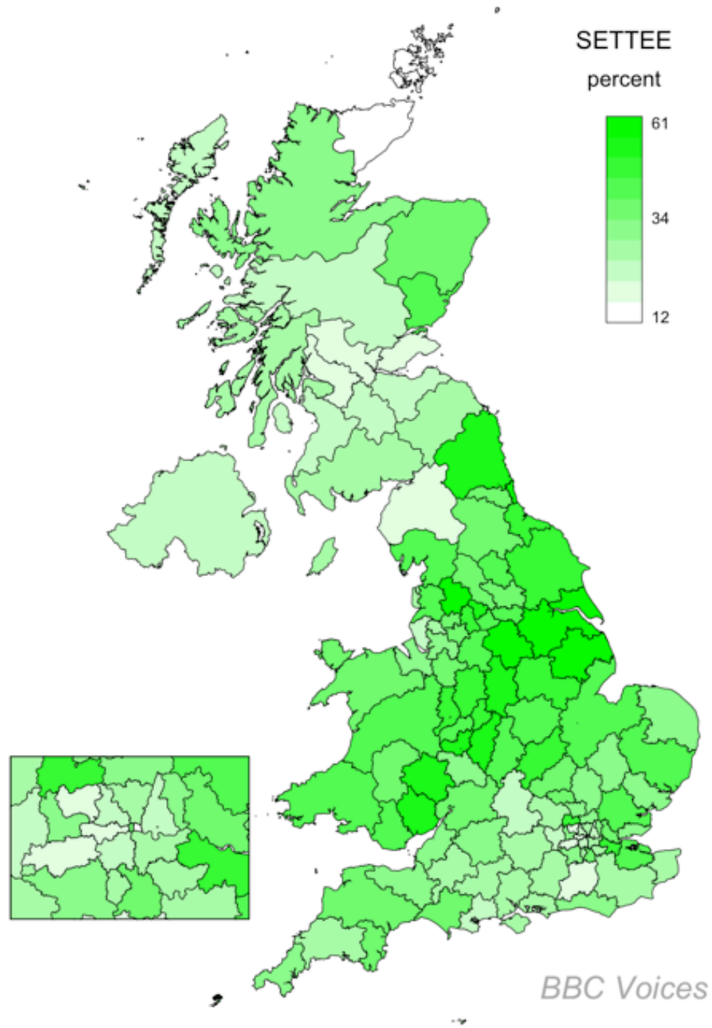




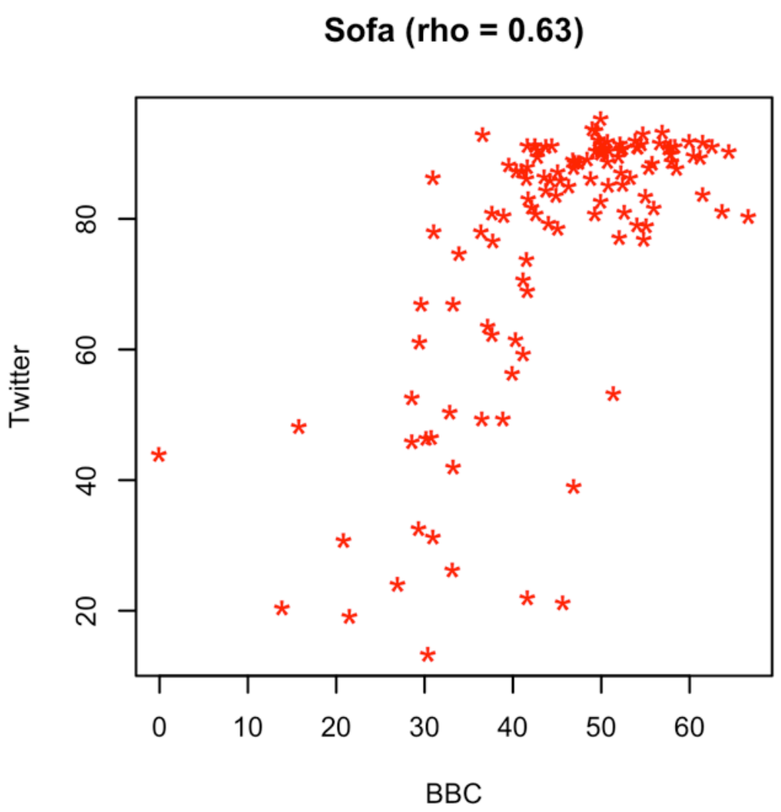
BBC Voices: *Sofa*/*Couch*/*Settee*

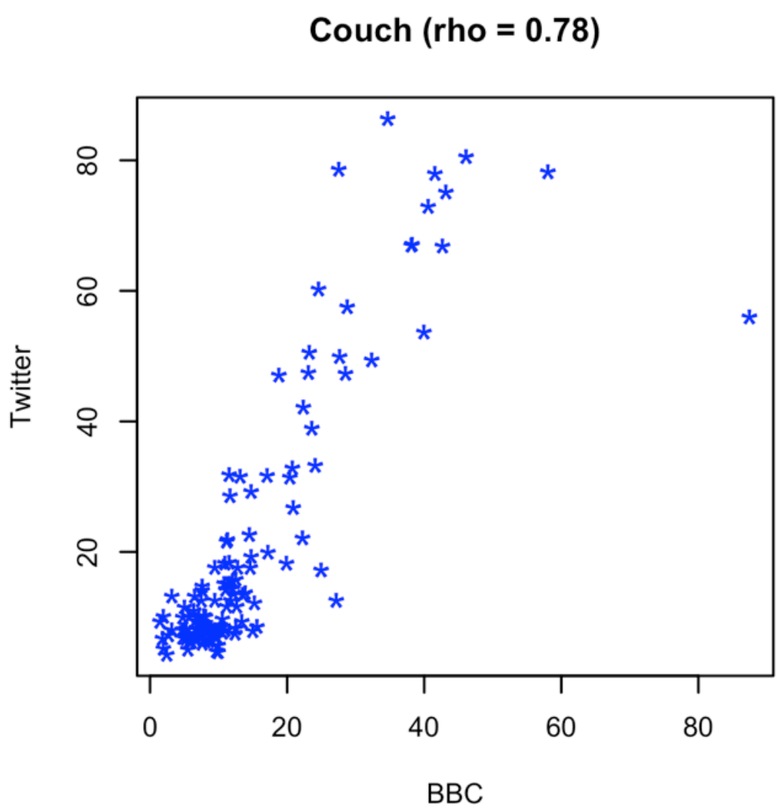


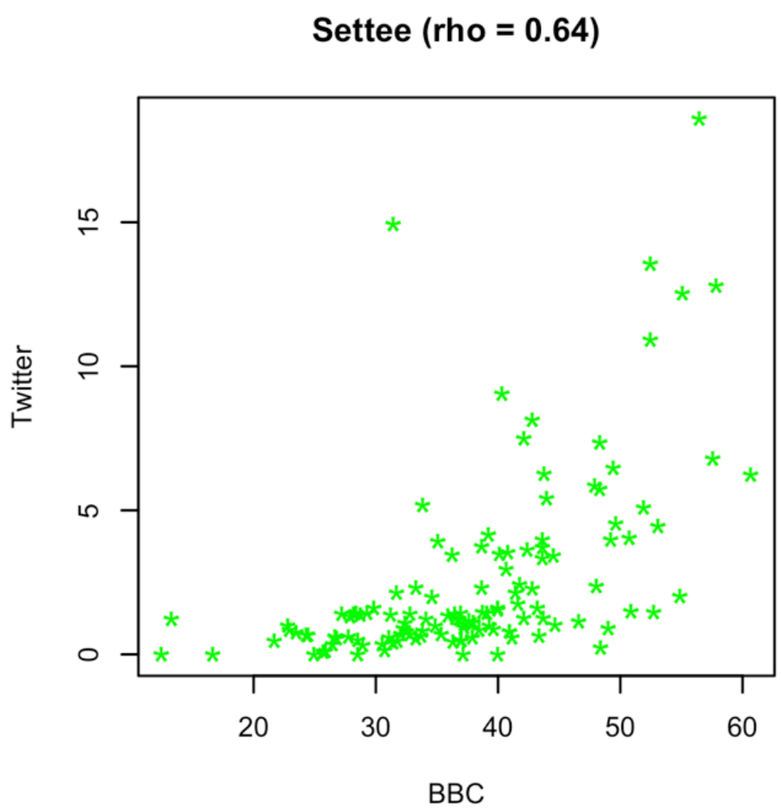




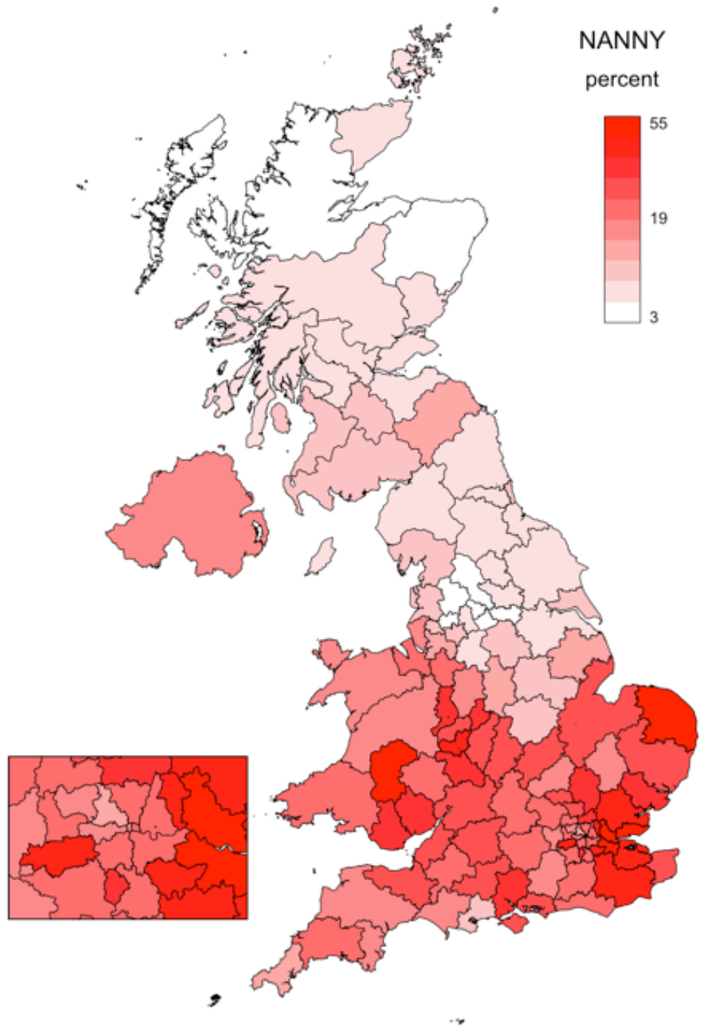
Comparison: *Sofa*/*Couch*/*Settee*

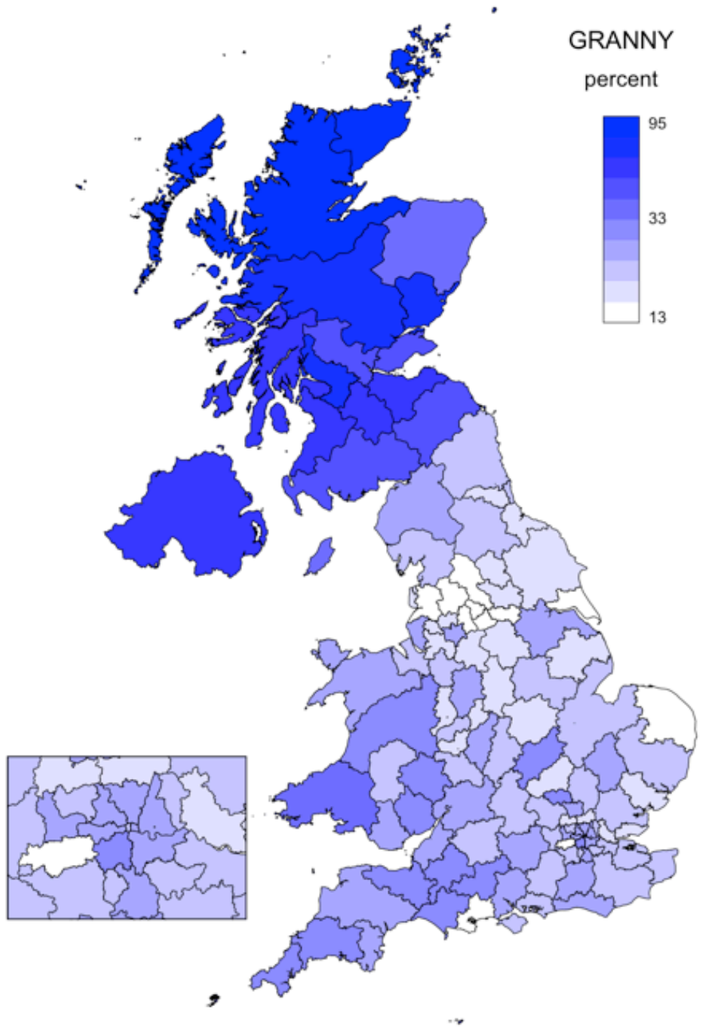


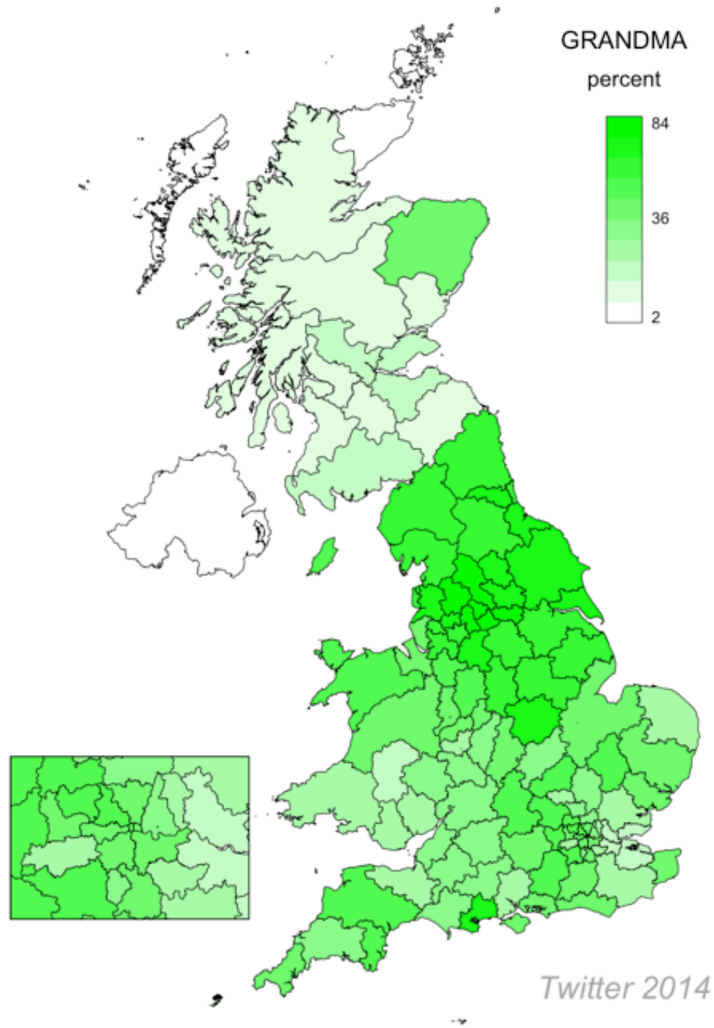




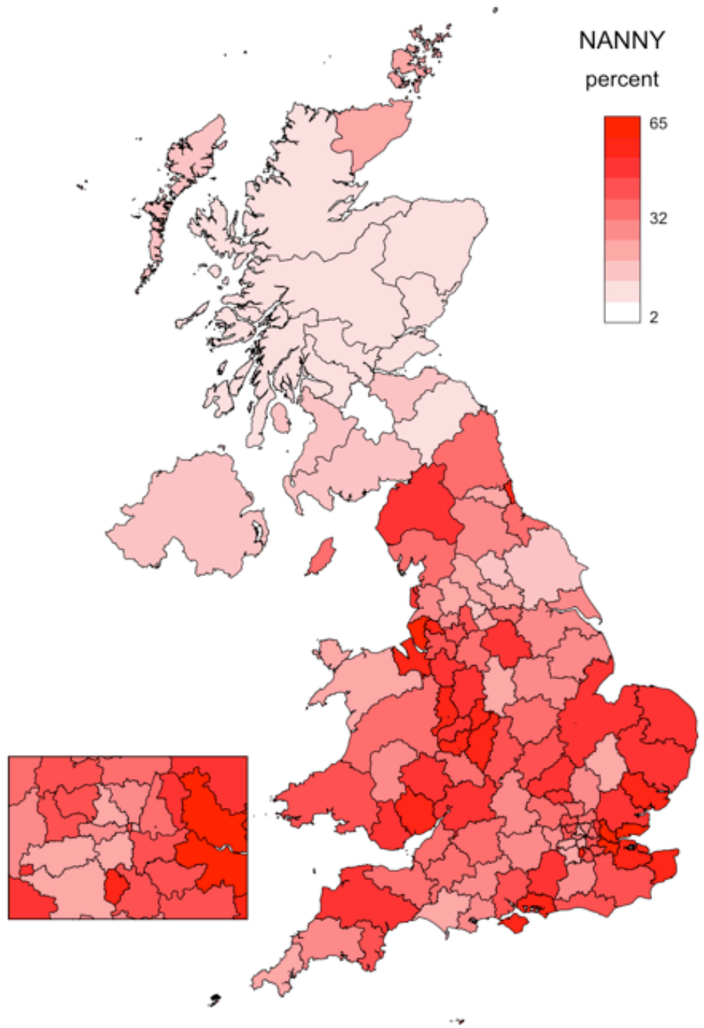
Twitter: *Nanny*/*Granny*/*Grandma*

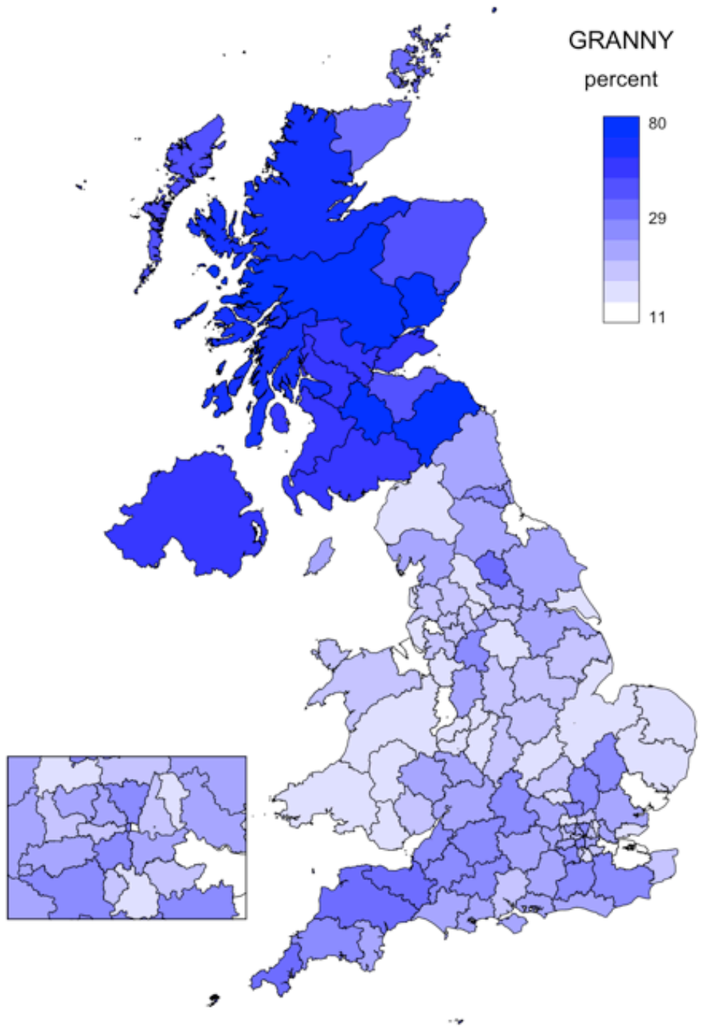


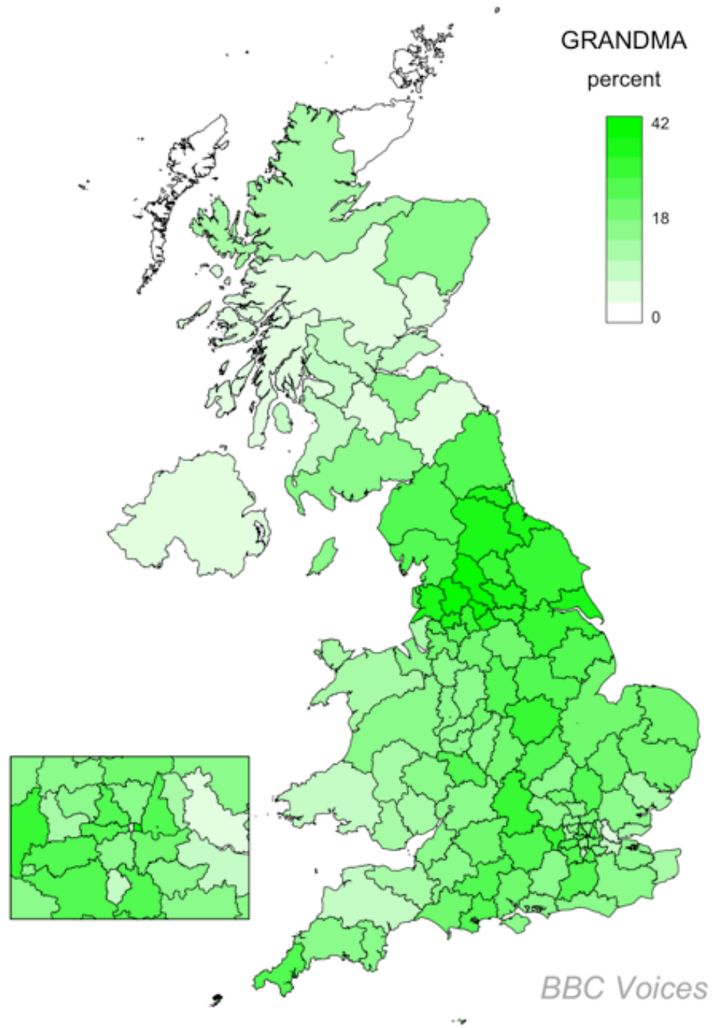




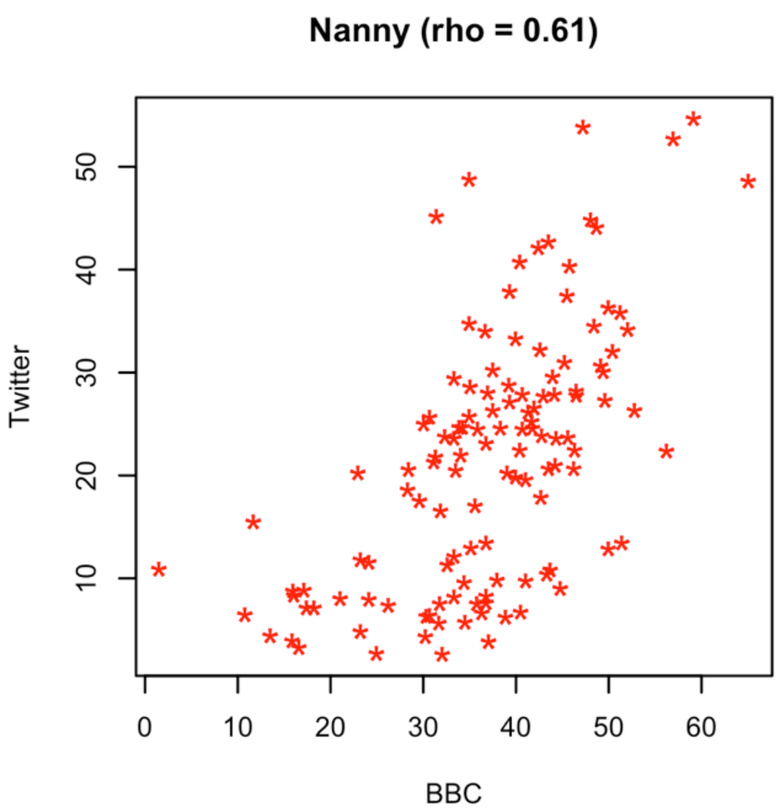
BBC Voices: *Nanny*/*Granny*/*Grandma*

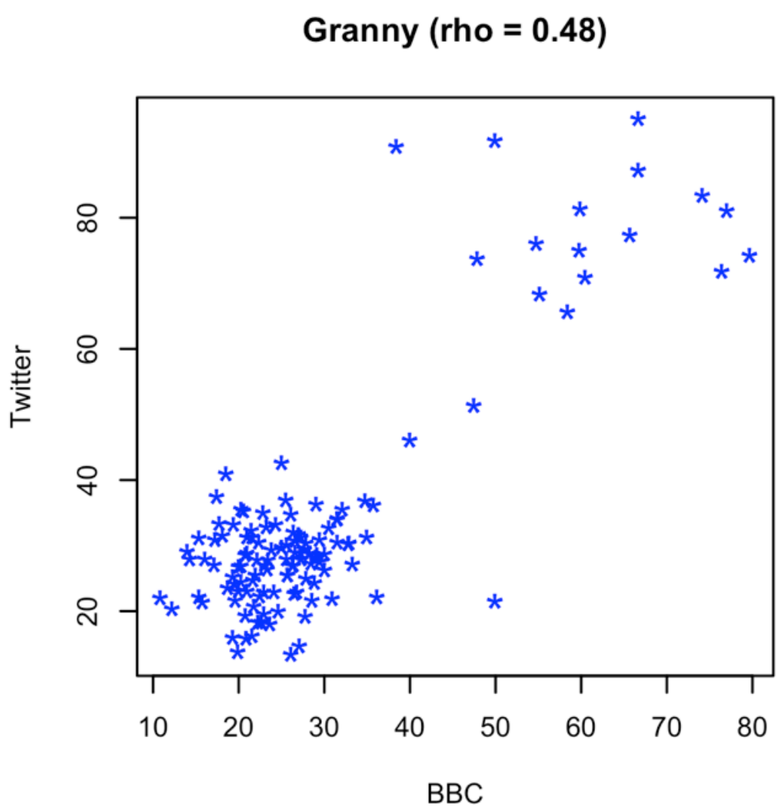


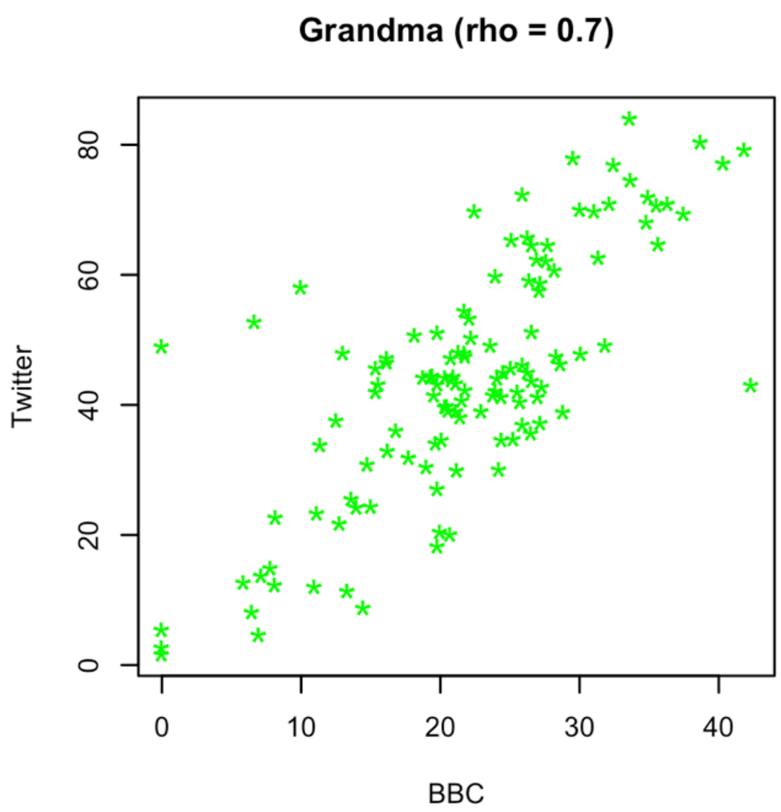




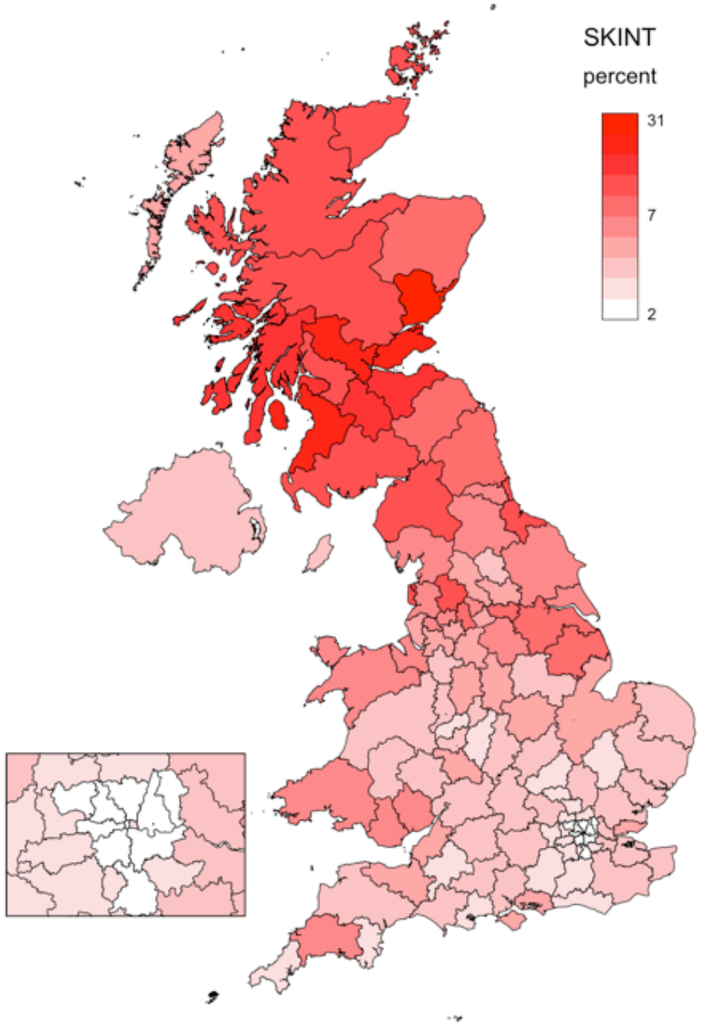
Comparison: *Nanny*/*Granny*/*Grandma*

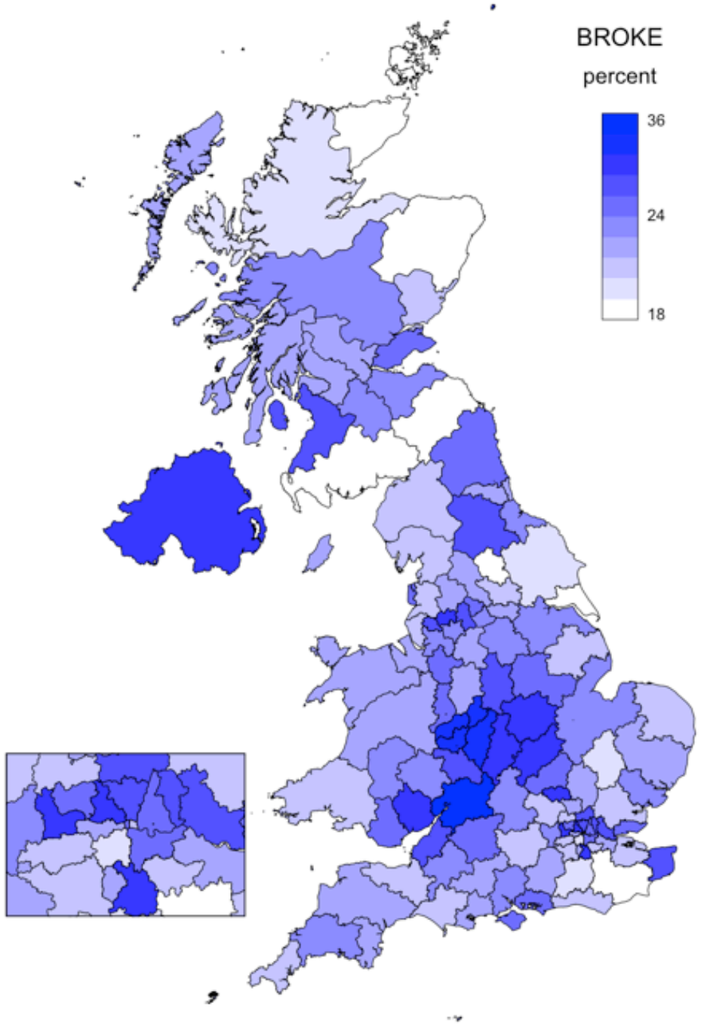


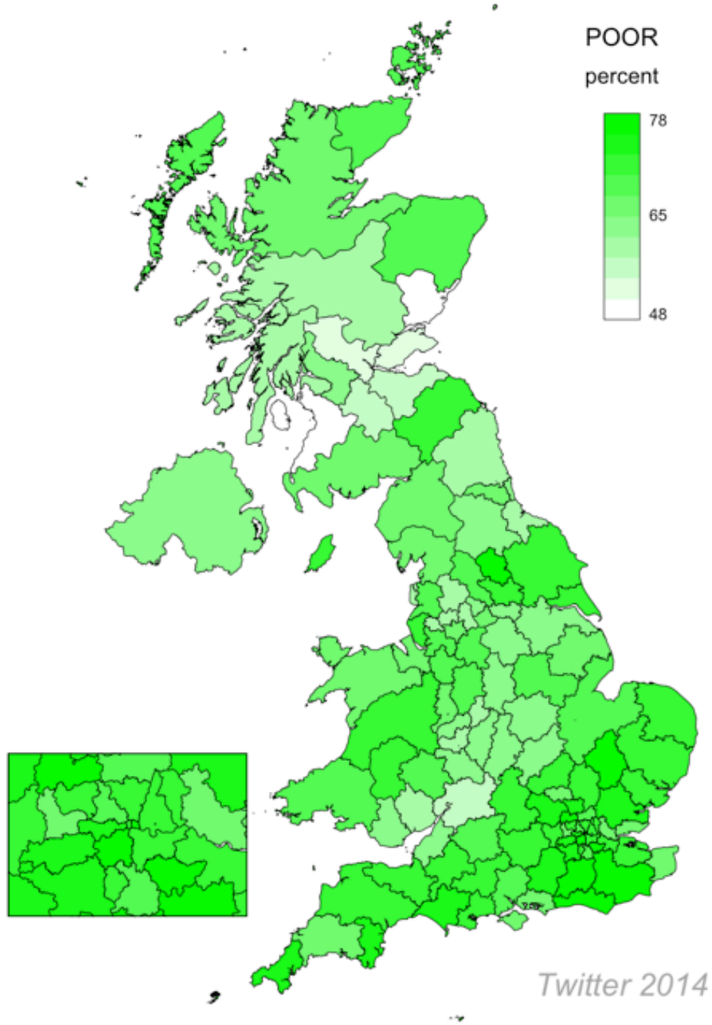




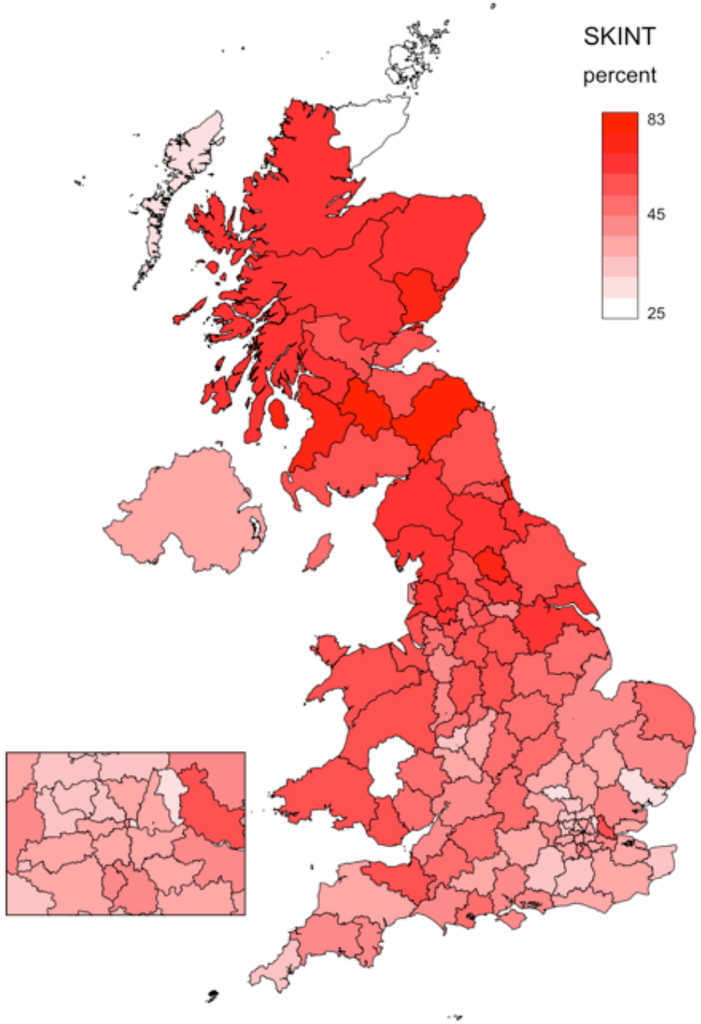
Twitter: *Skint*/*Broke*/*Poor*

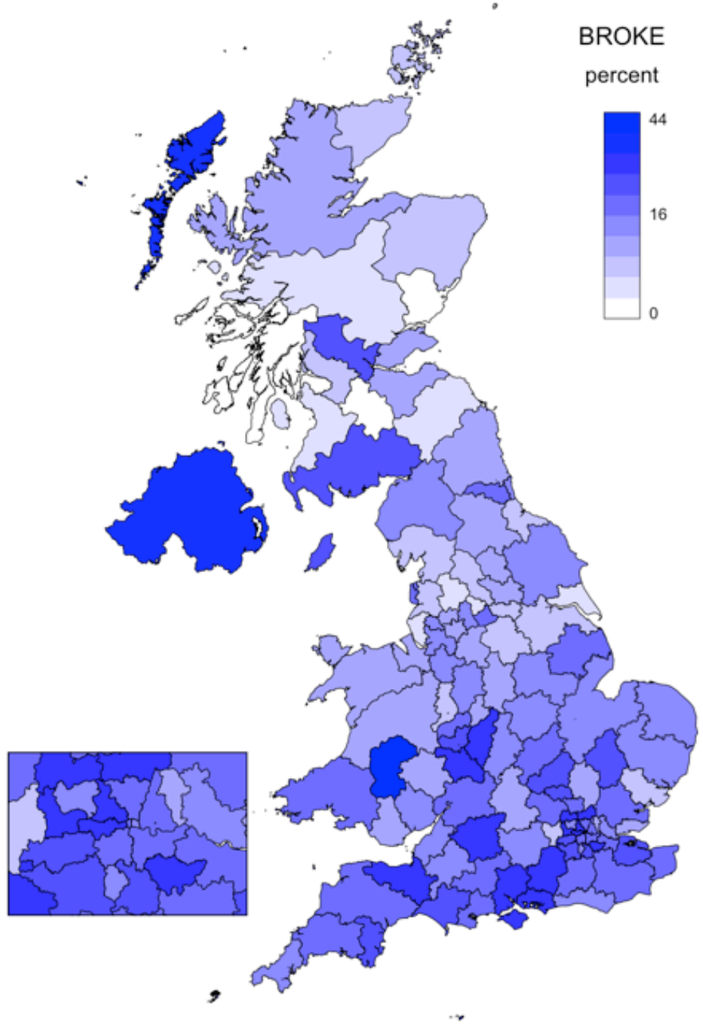


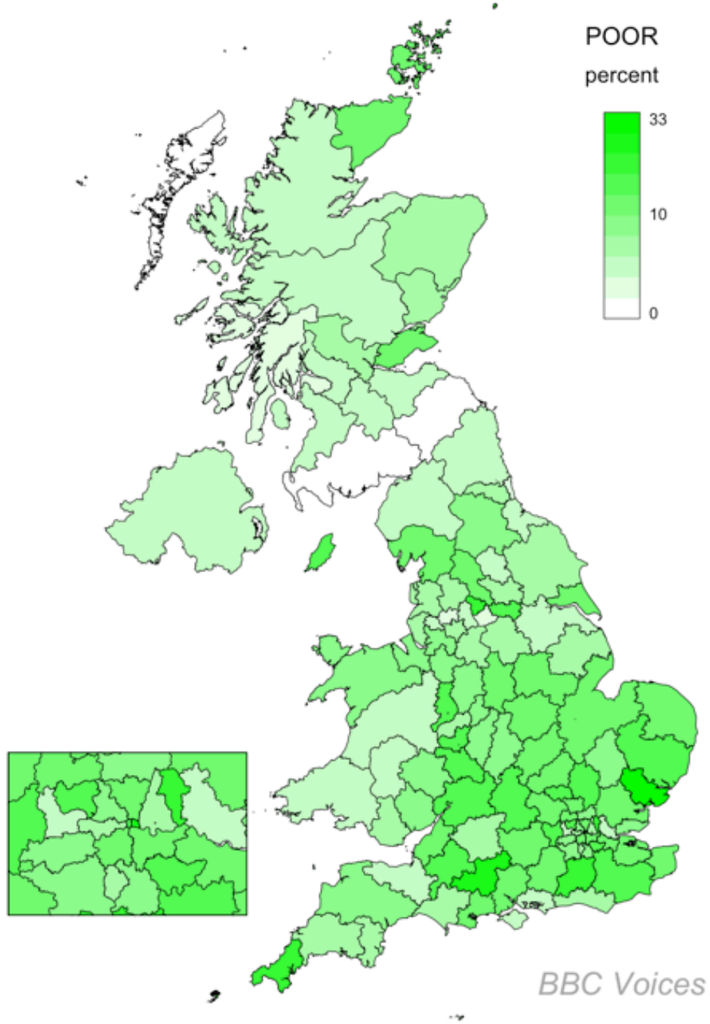




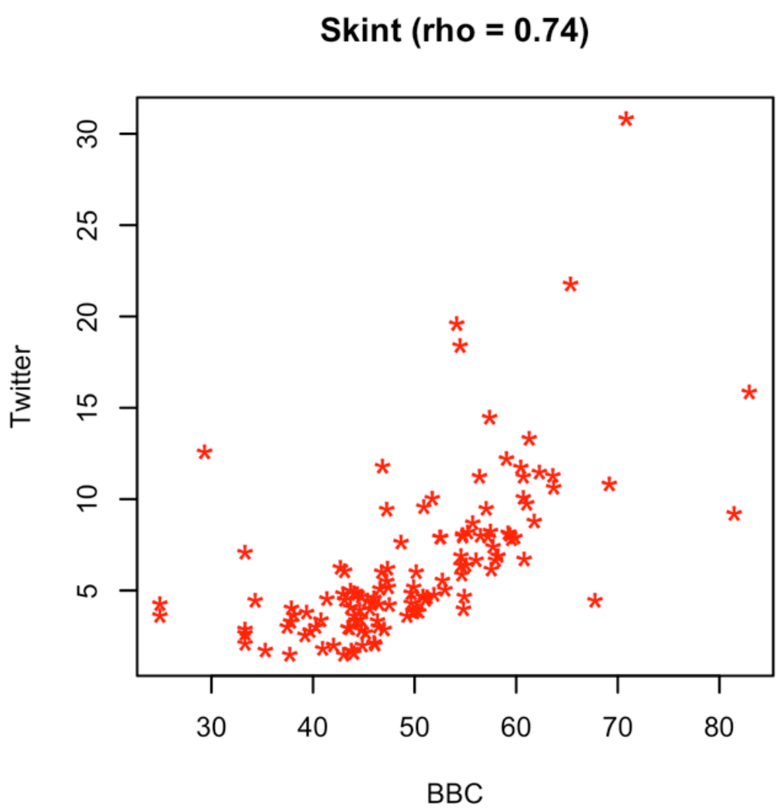
BBC Voices: *Skint*/*Broke*/*Poor*

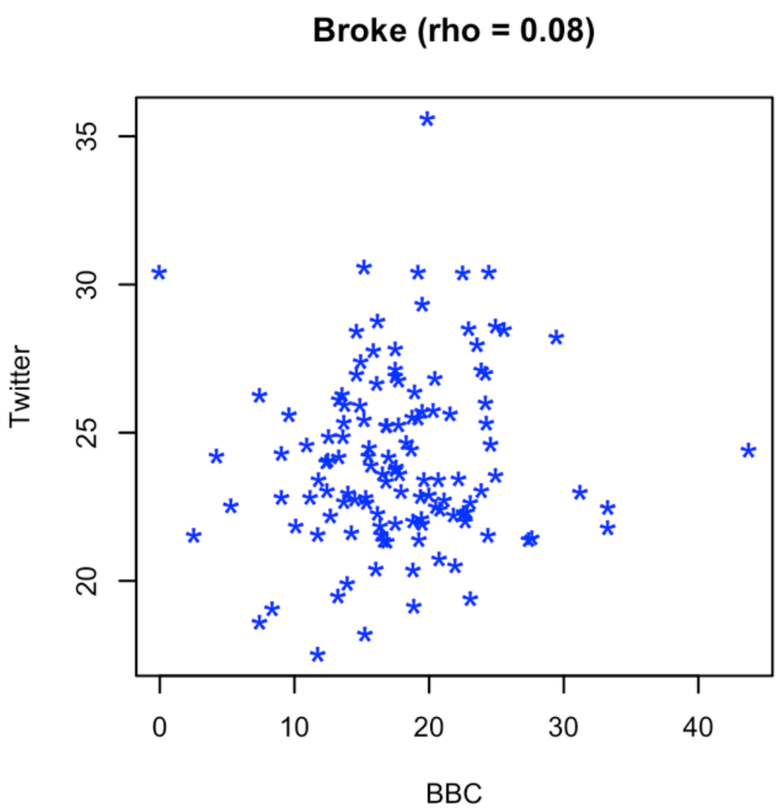


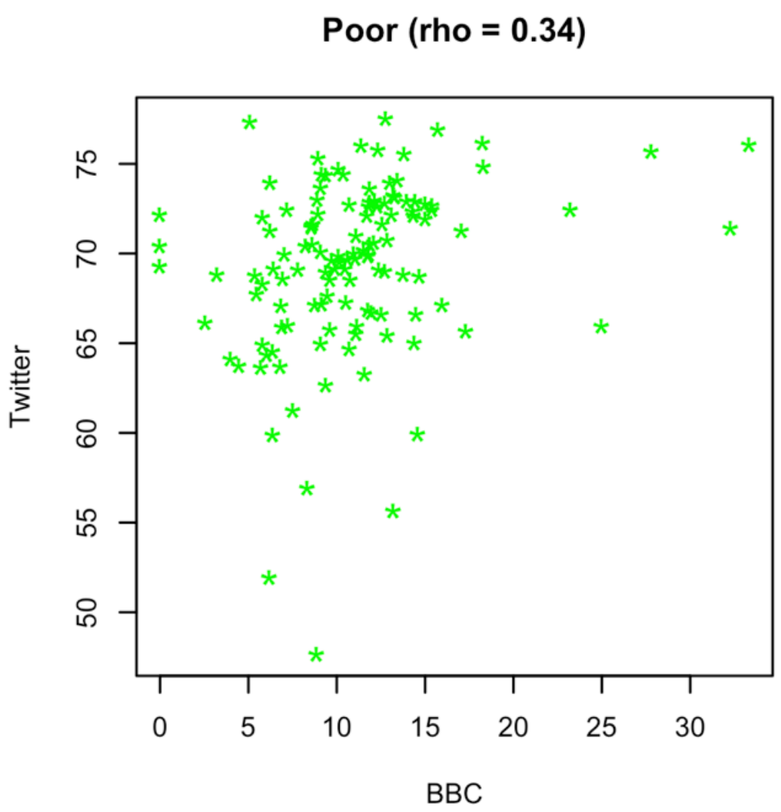




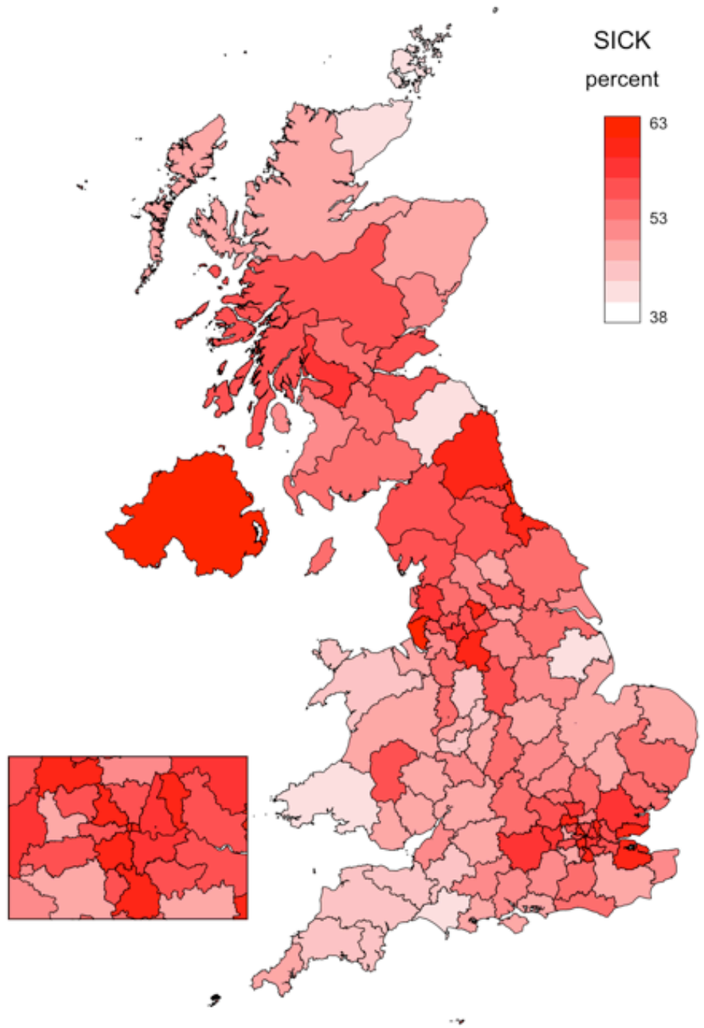
Comparison: *Skint*/*Broke*/*Poor*

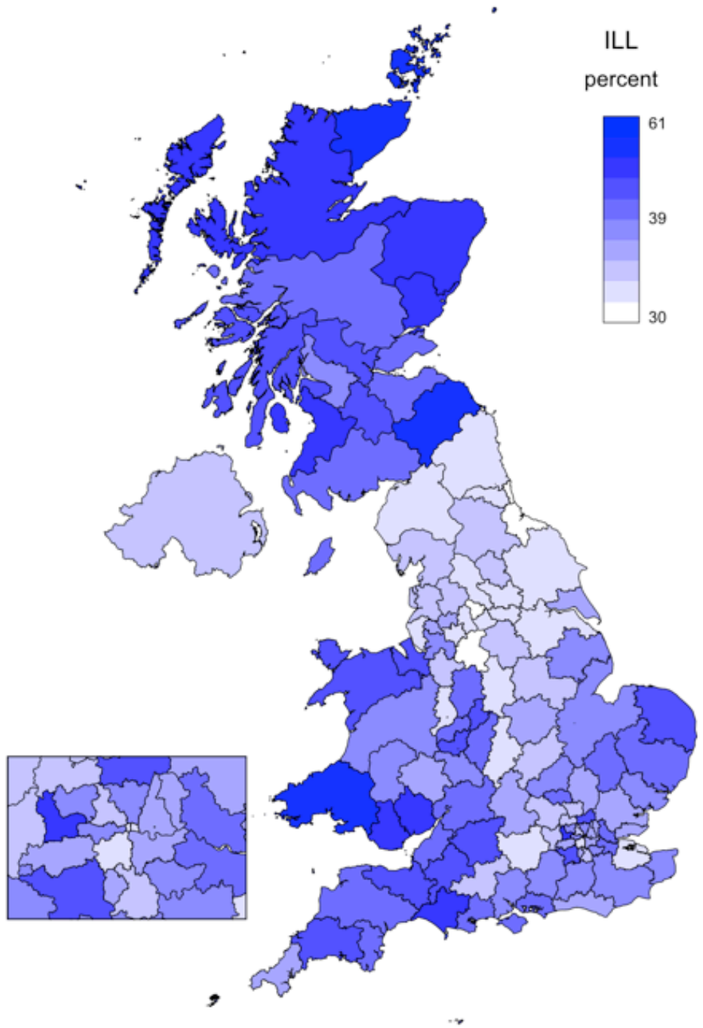


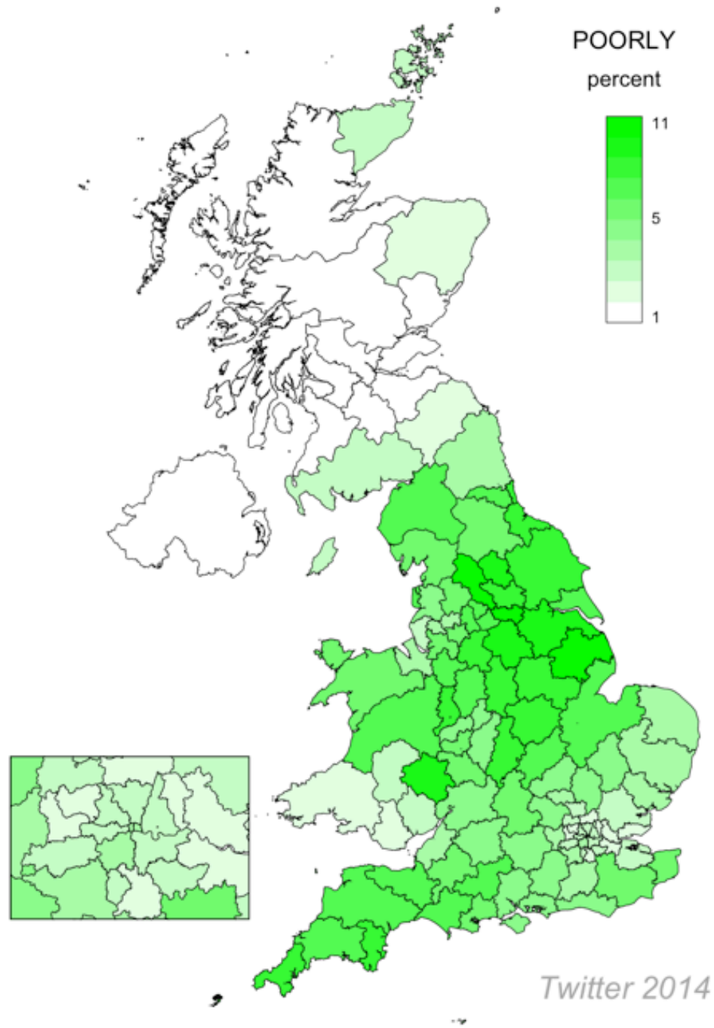




Twitter: *Sick*/*Ill*/*Poorly*

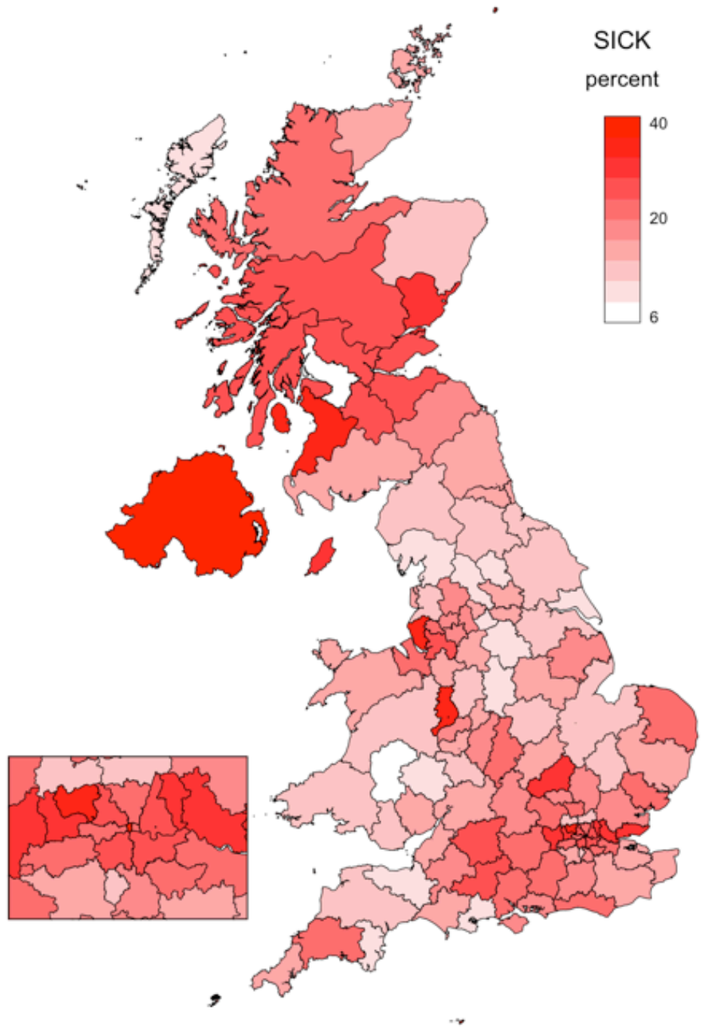


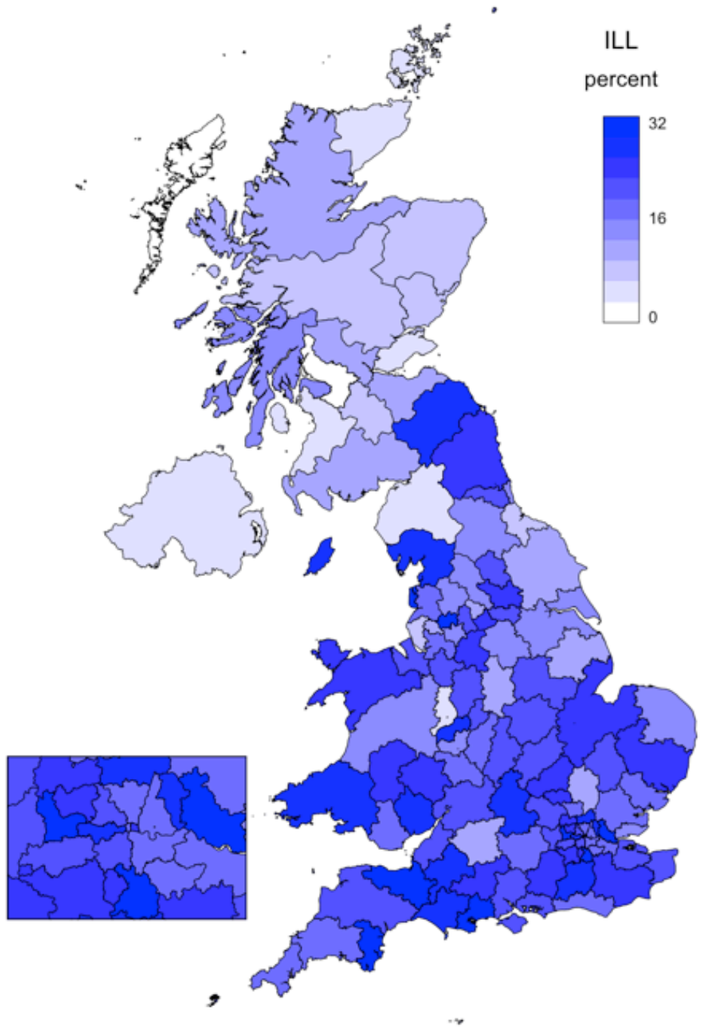


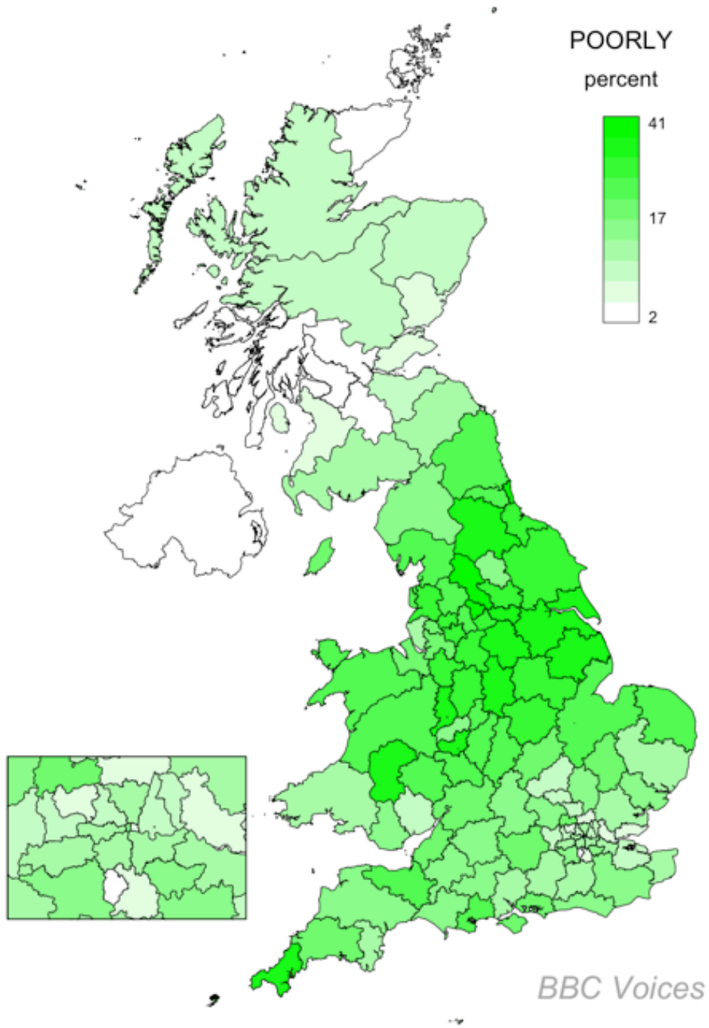


Twitter 2014

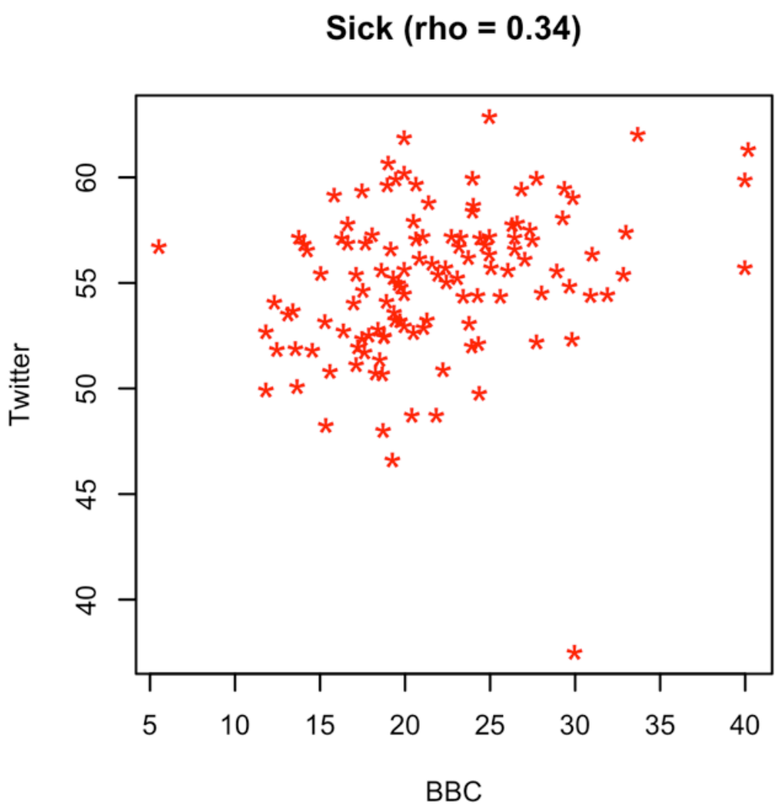
BBC Voices: *Sick*/*Ill*/*Poorly*

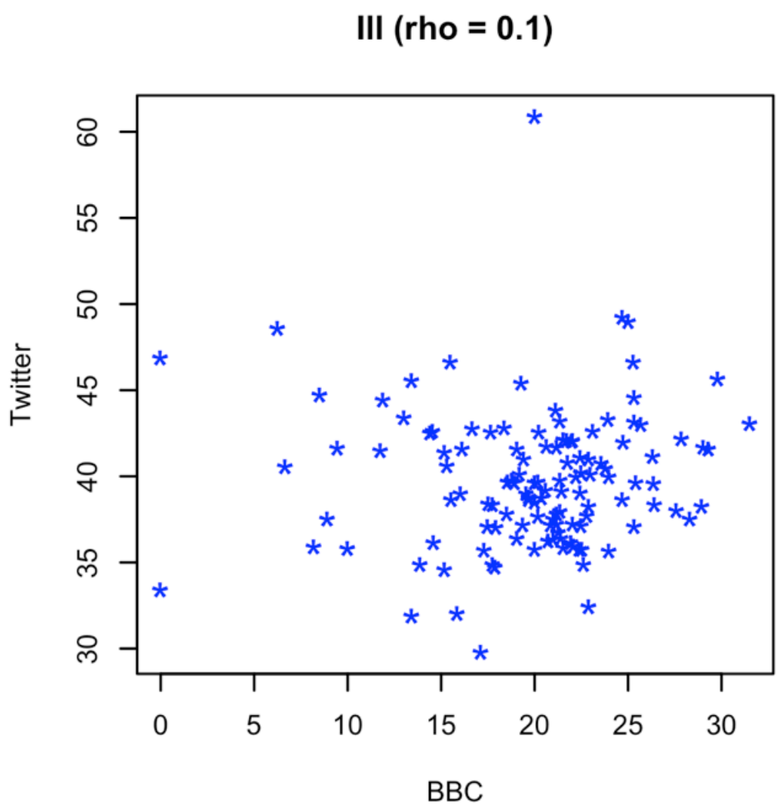


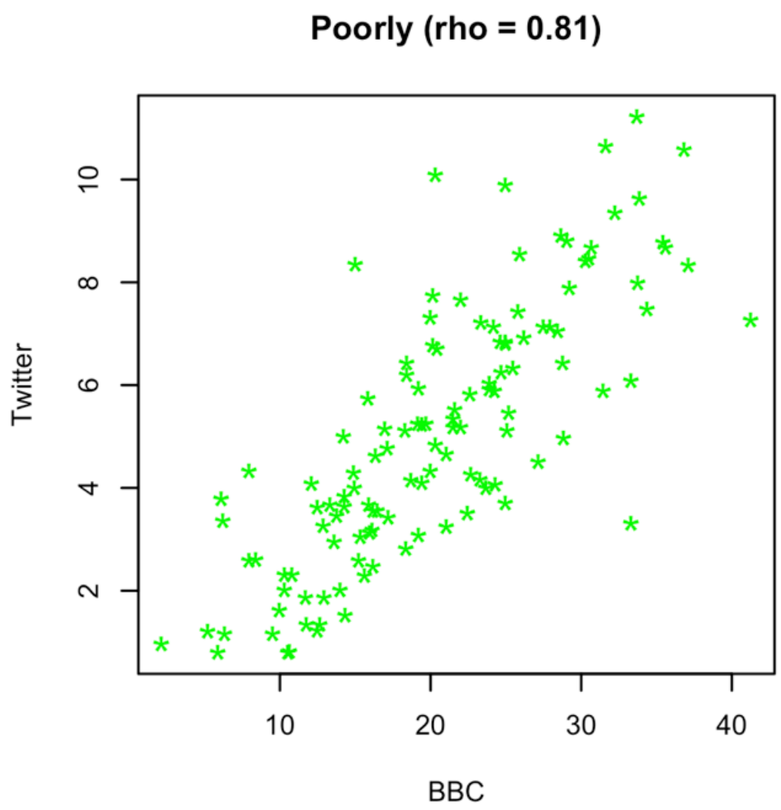




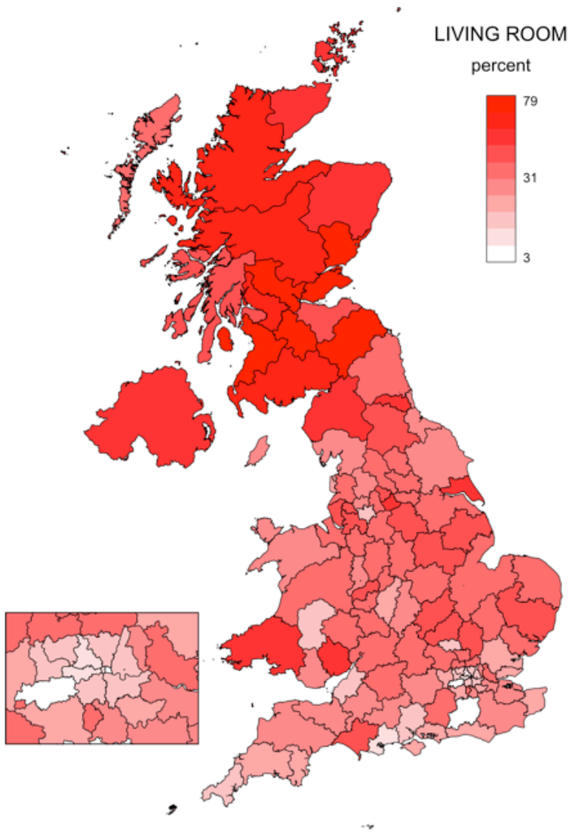
Comparison: *Sick*/*Ill*/*Poorly*

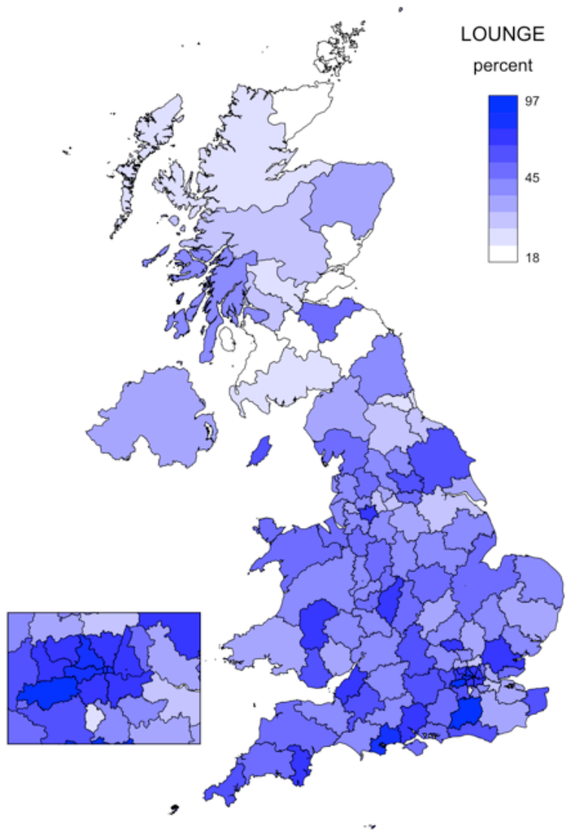


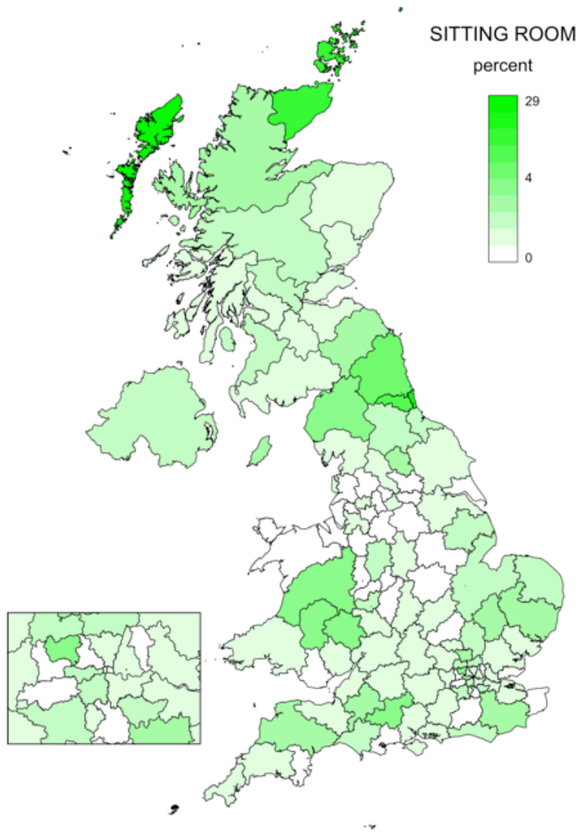


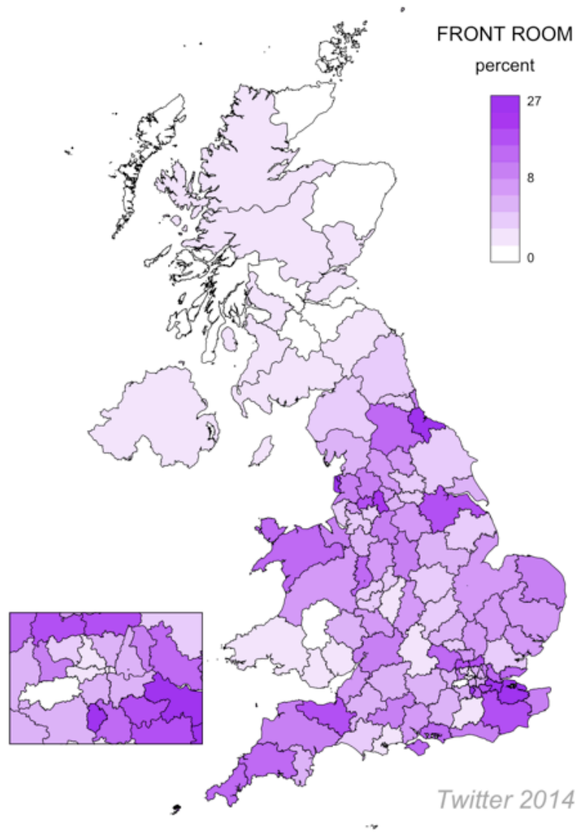


Twitter: *Living Room*/*Lounge*/*Sitting Room*/*Front Room*

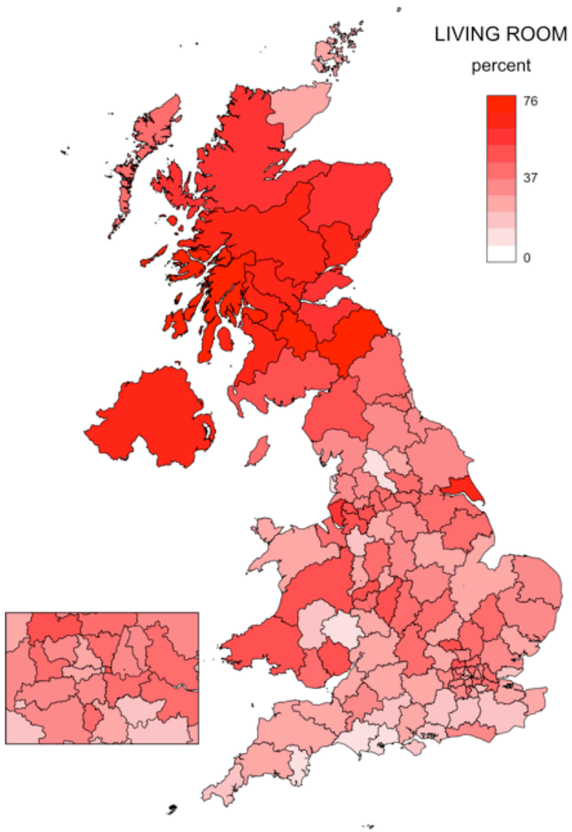


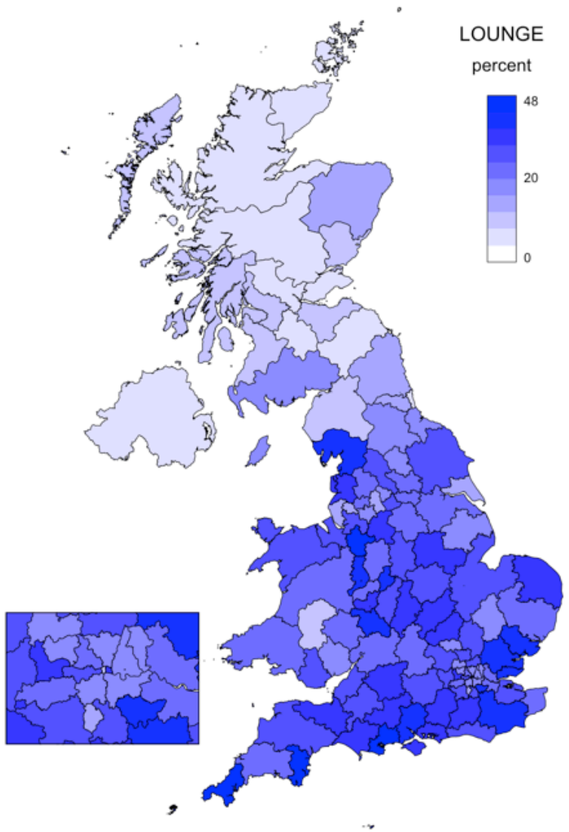


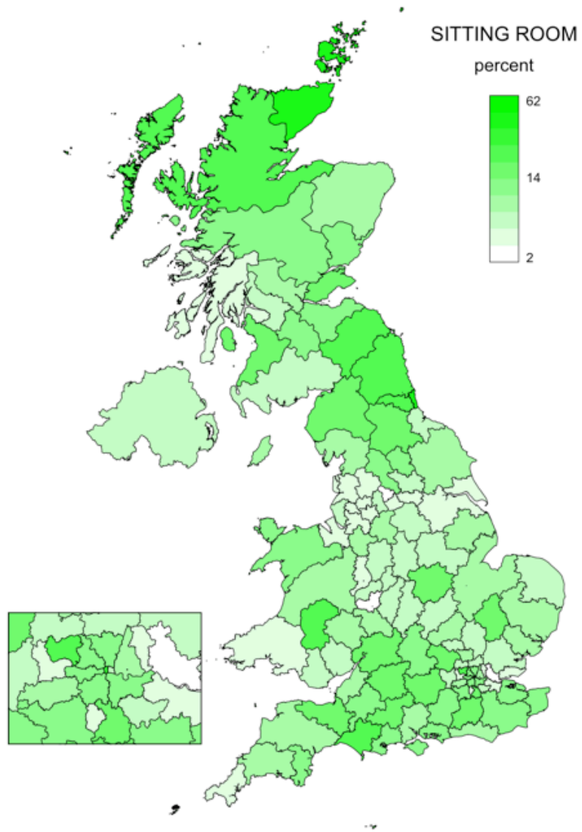


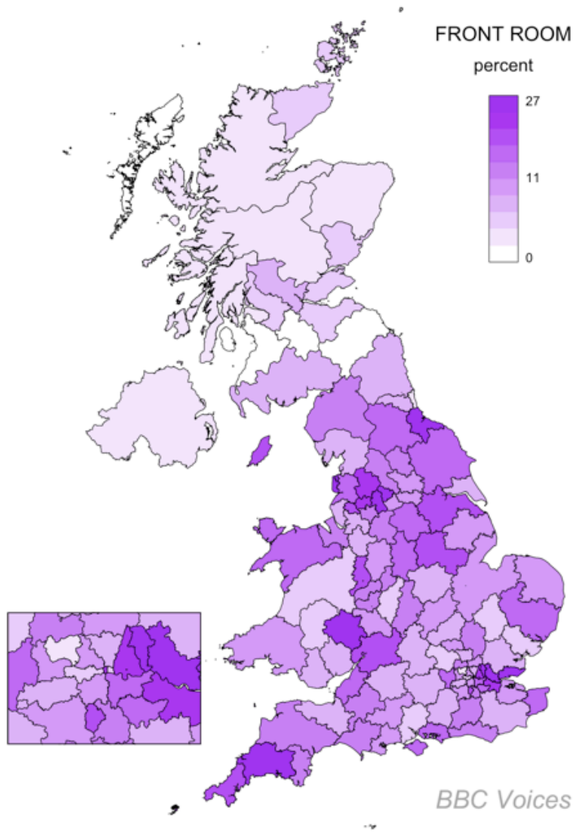


BBC Voices: *Living Room*/*Lounge*/*Sitting Room*/*Front Room*

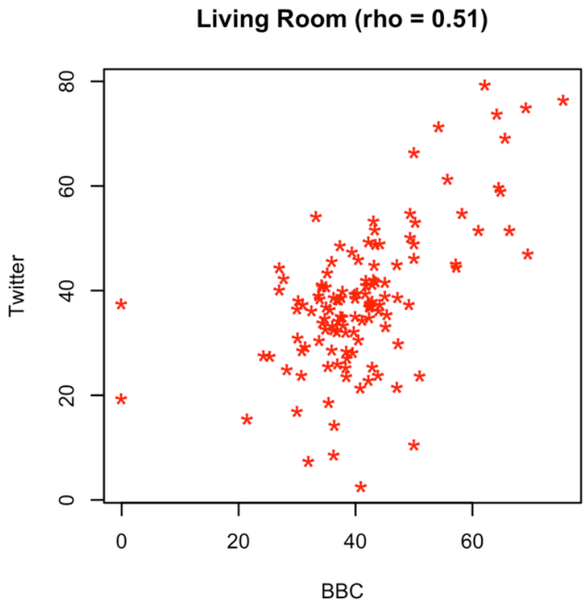


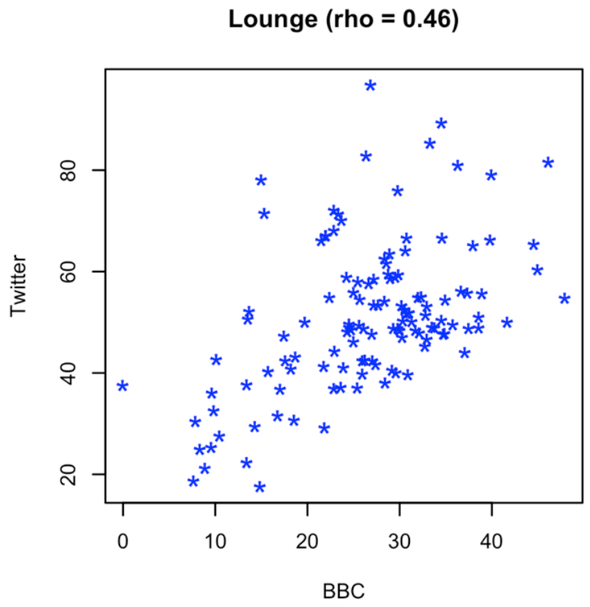


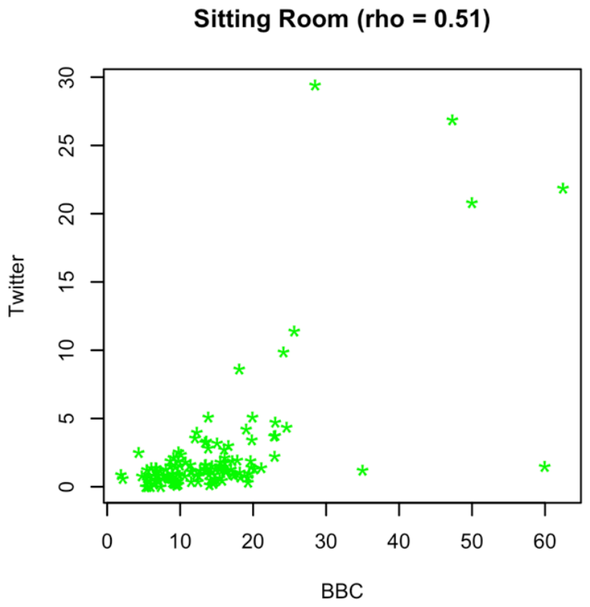


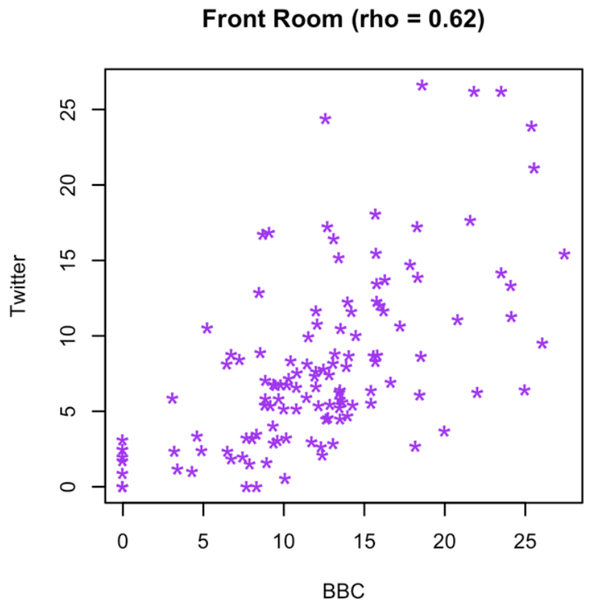


Comparison: *Living Room*/*Lounge*/*Sitting Room*/*Front Room*

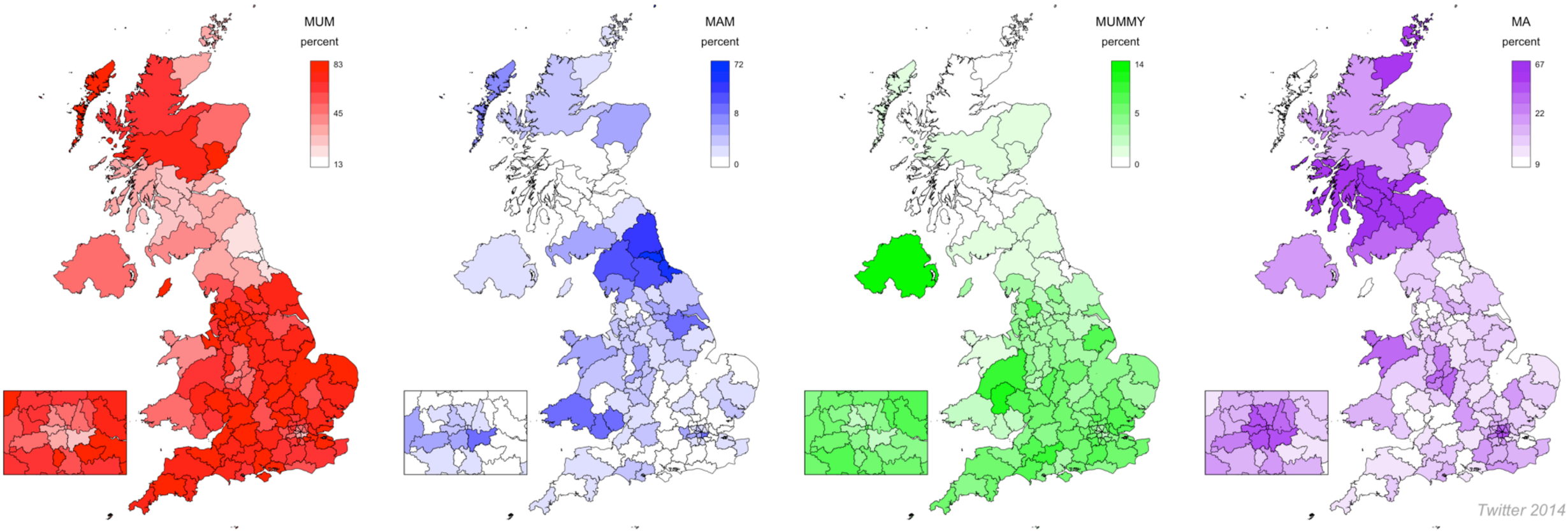




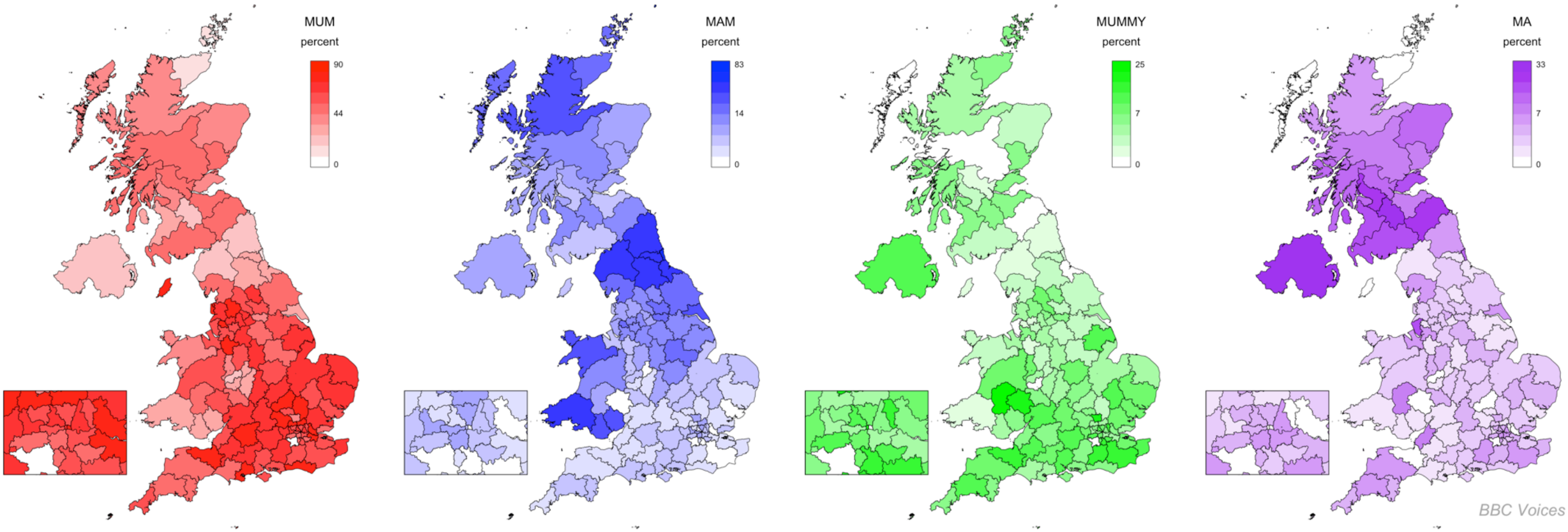




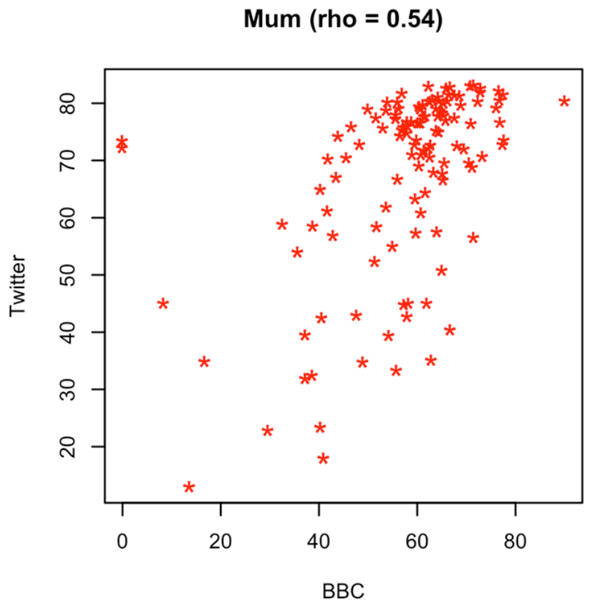
Twitter: *Mum*/*Mam*/*Mummy*/*Ma*

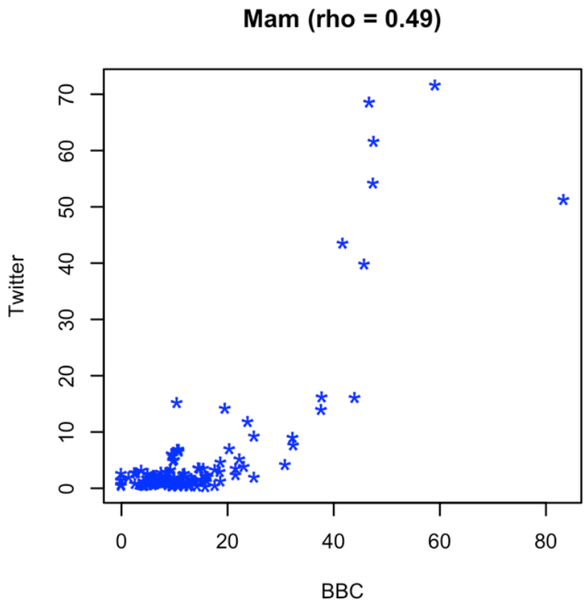


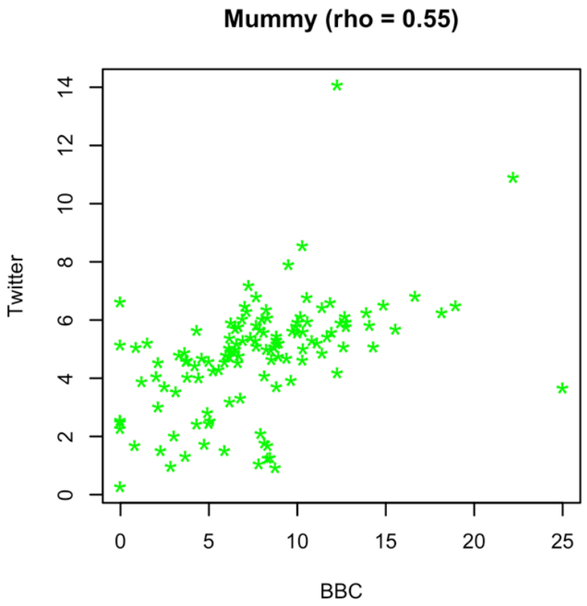
BBC Voices: *Mum*/*Mam*/*Mummy*/*Ma*

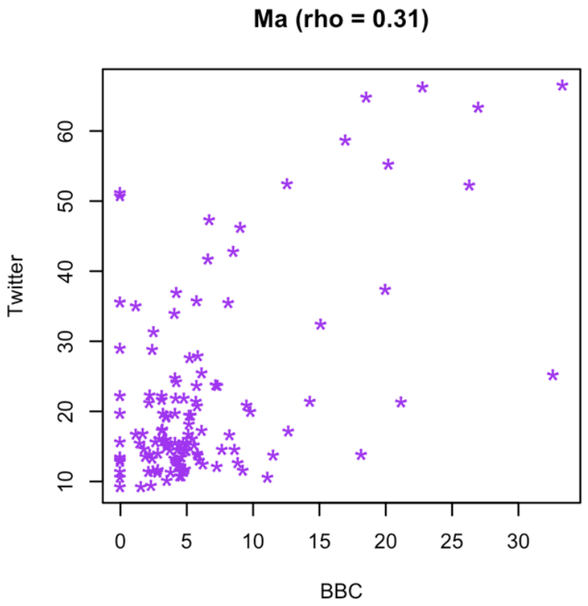


Comparison: *Mum*/*Mam*/*Mummy*/*Ma*









boiling	roasting	hot	baked	sweating
freezing	chilly	nippy	cold	
shattered	knackered			
sick	poorly	ill		
chuffed	happy	made.up		
pissed.off	angry			
play	lake			
skive	bunk	wag		
chuck	lob			
whack	smack	thump	wallop	belt
kip	sleep	snooze	nap	doze
pissed	wasted			
pregnant	expecting			
skint	broke	poor		
loaded	minted	well.off		
mad	nuts	crazy	mental	bonkers
fit	gorgeous	pretty	hot	
ugly	minger			
mardy	grumpy	stroppy	moody	
baby	bairn	wean	kid	little.one
mum	mam	mummy	ma	
nanny	granny	grandma		
grandad	grandpa	grampa		
mate	pal	friend	buddy	
chav	ned			
clothes	gear	clobber	kit	
trousers	pants	jeans		
pumps	daps	trainers		
living.room	lounge	sitting.room	front.room	
sofa	settee	couch		
loo	bog	toilet		
alley	path	pavement		
drizzle	spit	shower		
pour	chuck	bucket		
stream	brook	burn	beck	

	Rho	
	>.50	16%
	>.30	43%
	>.20	53%
	>0	86%
	<0	14%

Median Rho = .24

Worst Matches

Cold (for cold weather)
Nippy (for cold weather)
Baked (for hot weather)
Belt (for hit)
Bucket (for rain)
Pretty
Wasted (for drunk)
Pregnant
Pissed off (for angry)
Angry
Toilet

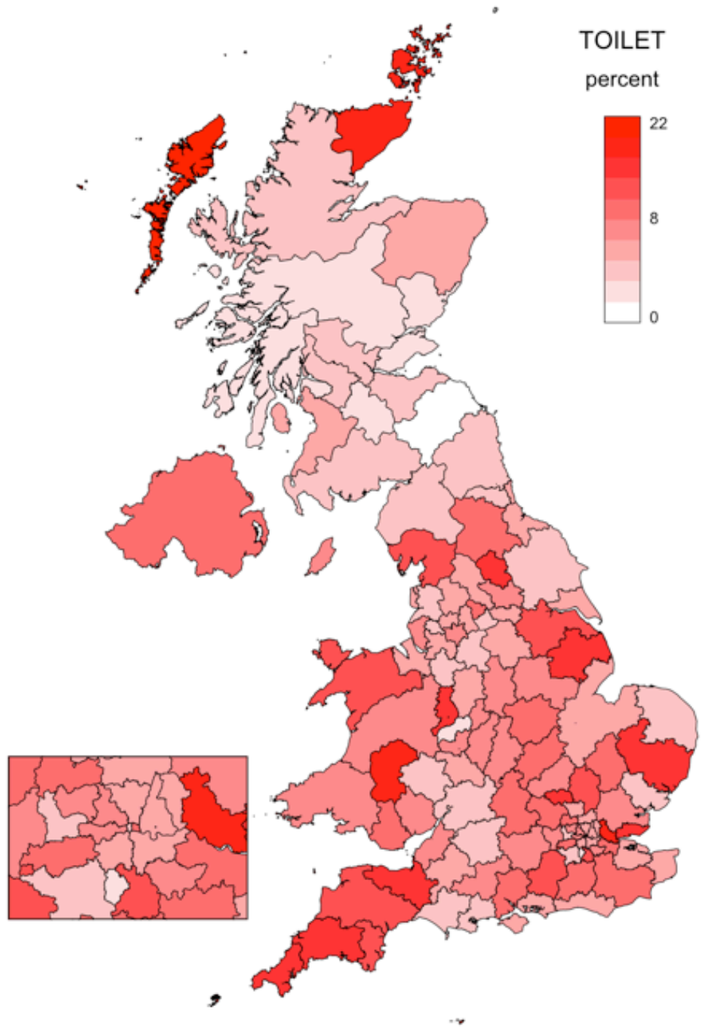
The worst matches are of three basic types:

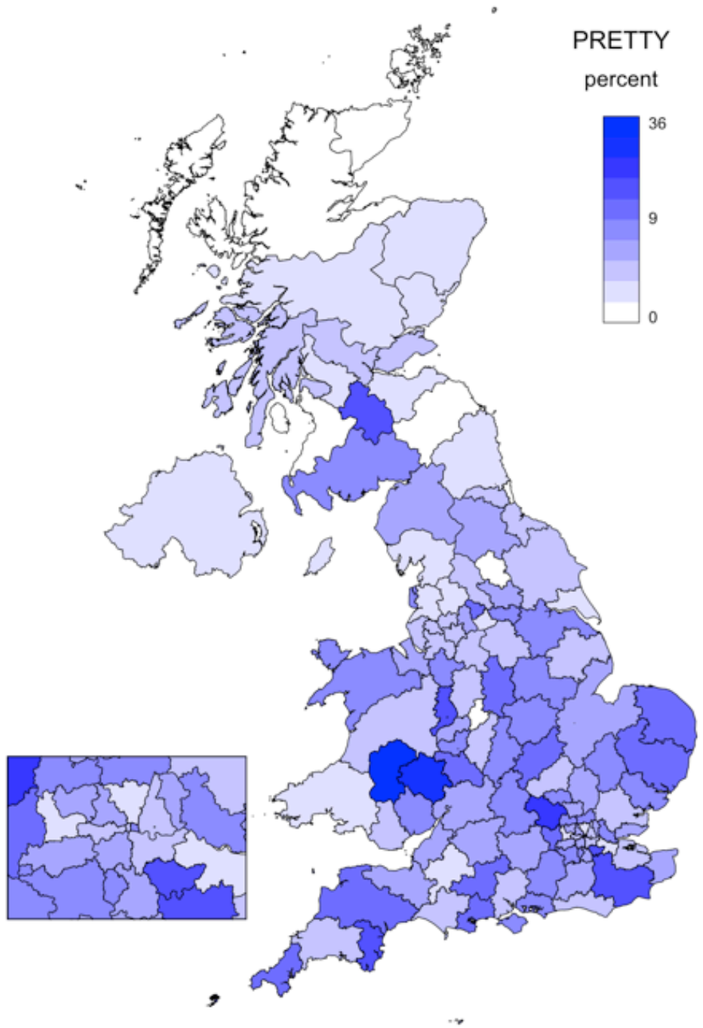
Polysemous words, where the target meaning is uncommon.

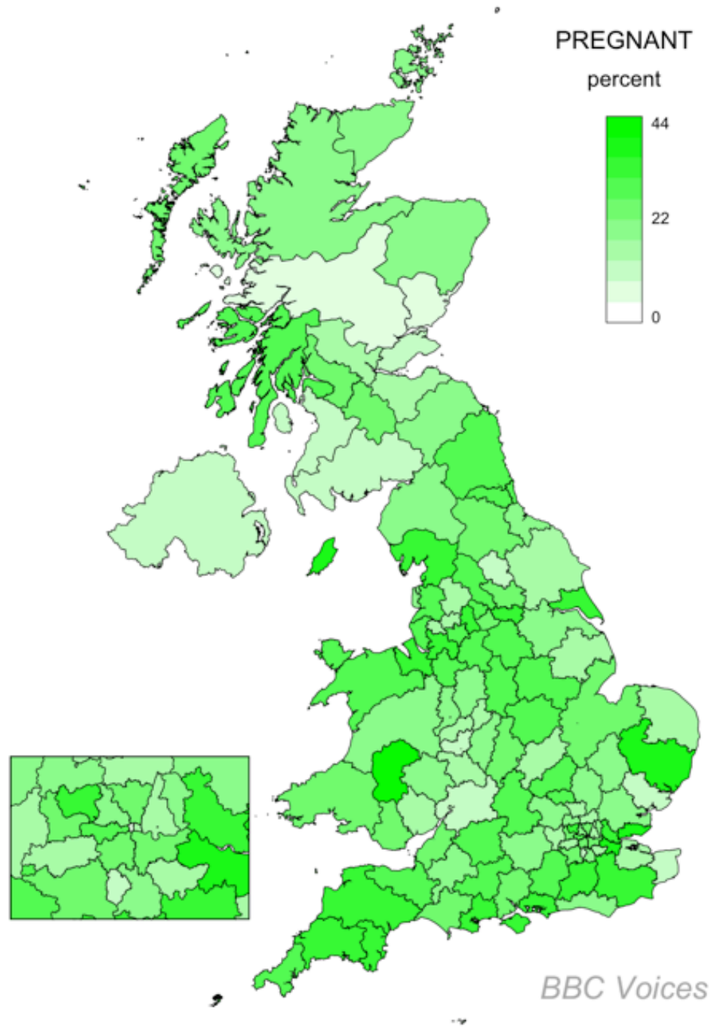
Frequent words, which are the general form in the UK.

Variants in alternation with words of these other two types.

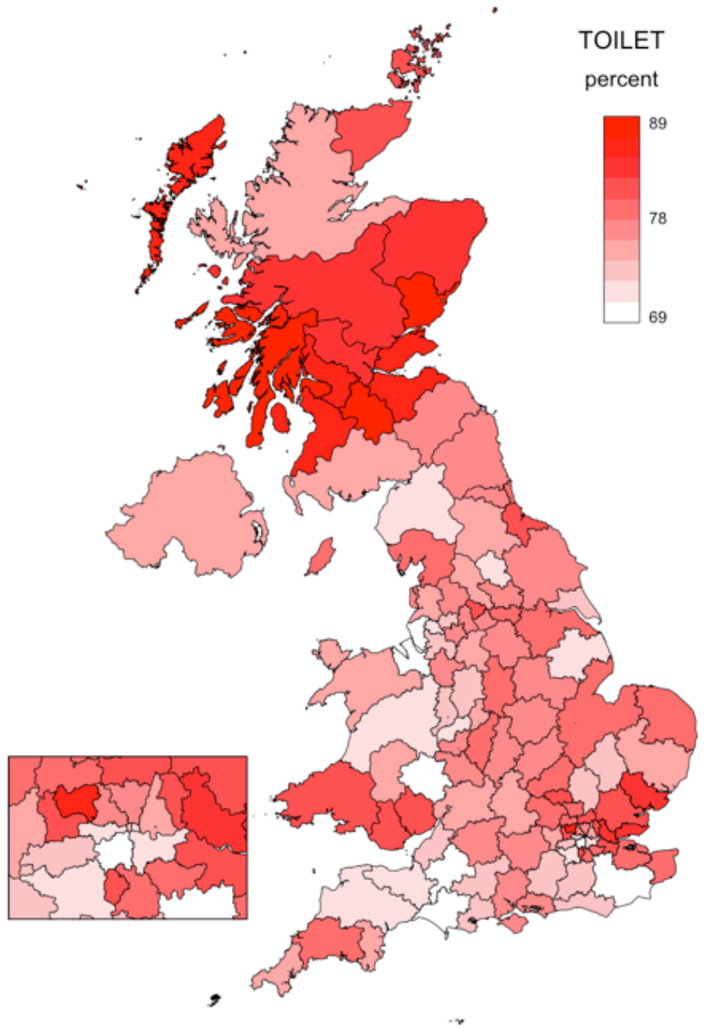
BBC Voices: *Toilet*/*Pretty*/*Pregnant*

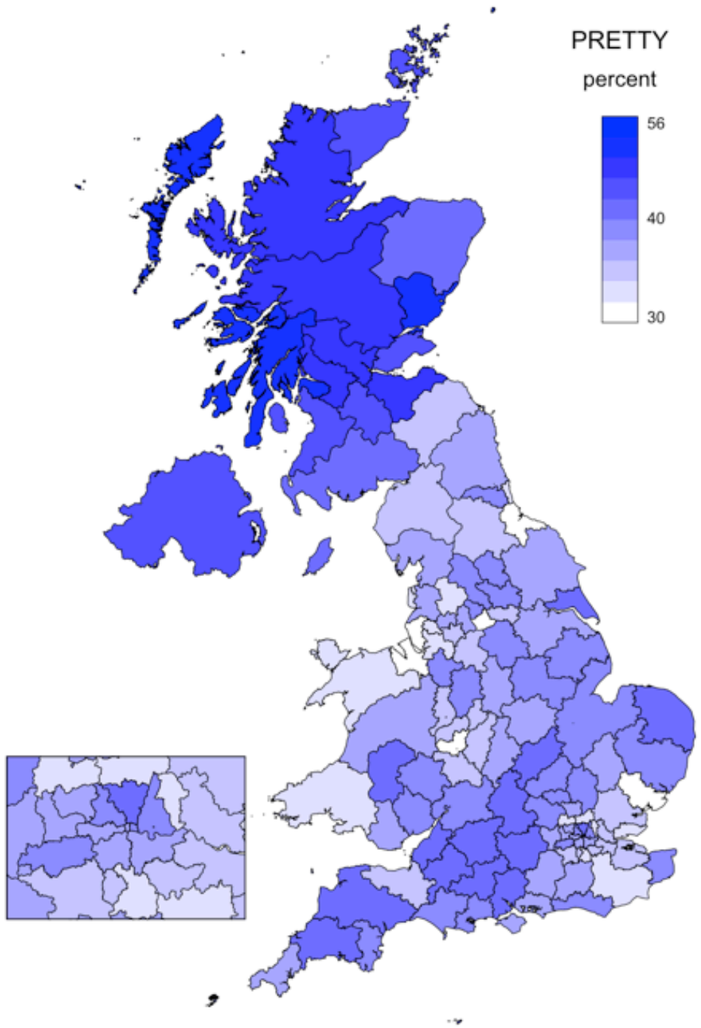


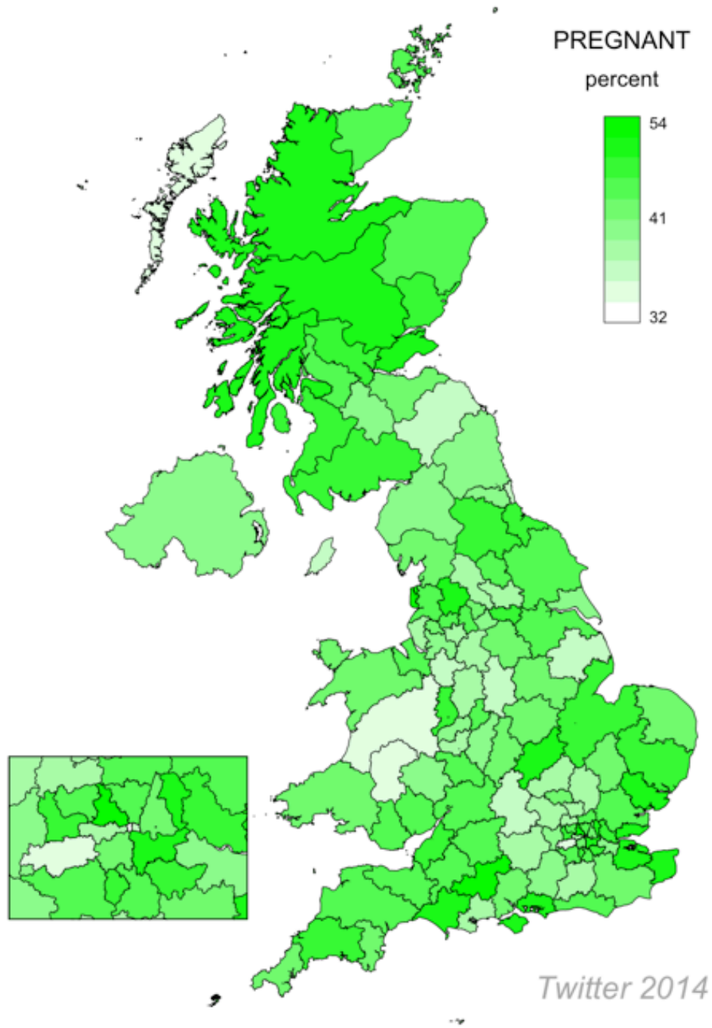




Twitter: *Toilet*/*Pretty*/*Pregnant*







Twitter 2014

Methodological Conclusion

Overall many of the Twitter maps align with the BBC Voices Maps.

The matches tend to be weakest when the survey did not identify regional patterns or when the word is highly polysemous and the target meaning is rare.

Polysemy appears to be the main issue with Twitter maps.

This can be dealt with by looking at words in context or using other methods for word sense disambiguation.

These problems are solvable, especially with more data.

Theoretical Conclusions

Given that maps from Twitter and surveys do largely align, it would appear that regional lexical patterns are largely independent of register.

Presumably this is because the same general extra-linguistic forces produce regional patterns in lexis across communicative situation, i.e. regional patterns in culture and topography.

Assessing the Use of Social Media for Mapping Lexical Variation in British English

Jack Grieve, Aston University (@JWGrieve)

Chris Montgomery, University of Sheffield (@montgomerychris)

Andrea Nini, University of Manchester (@and_nini)

Diansheng Guo, University of South Carolina (@Diansheng_Guo)

6 June 2017

ICLAVE 9

Malaga, Spain