THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Mapping Great Britain's semantic footprints through a large language model analysis of Reddit comments

Citation for published version:
Berragan, C, Singleton, A, Calafiore, A & Morley, J 2024, 'Mapping Great Britain's semantic footprints through a large language model analysis of Reddit comments', *Computers, Environment and Urban Systems*, vol. 110, 102121, pp. 1-12. https://doi.org/10.1016/j.compenvurbsys.2024.102121

Digital Object Identifier (DOI):
10.1016/j.compenvurbsys.2024.102121

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Computers, Environment and Urban Systems

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

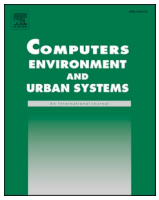
Download date: 11. May. 2024

Computers, Environment and Urban Systems 110 (2024) 102121

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus



Mapping Great Britain's semantic footprints through a large language model analysis of Reddit comments

Cillian Berragan [a,*], Alex Singleton [a], Alessia Calafiore [b], Jeremy Morley [c]

[a] University of Liverpool, Geographic Data Science Lab, UK
[b] University of Edinburgh, Edinburgh College of Art, UK
[c] Ordnance Survey, UK

ARTICLE INFO

Keywords:
Vernacular geography
Semantics
Social media
Natural language processing

ABSTRACT

Observed regional variation in geotagged social media text is often attributed to dialects, where features in language are assumed to exhibit region-specific properties. While dialects are seen as a key component in defining the identity of regions, there are a multitude of other geographic properties that may be captured within natural language text. In our work, we consider locational mentions that are directly embedded within comments on the social media website Reddit, providing a range of associated semantic information, and enabling deeper representations between locations to be captured. Using a large corpus of geoparsed Reddit comments from UK-related local discussion subreddits, we first extract embedded semantic information using a large language model, aggregated into local authority districts, representing the semantic footprint of these regions. These footprints broadly exhibit spatial autocorrelation, with clusters that conform with the national borders of Wales and Scotland. London, Wales, and Scotland also demonstrate notably different semantic footprints compared with the rest of Great Britain.

1. Introduction

The prevalence of social media data for use in geographic research has generated a renewed interest in the concept of 'place' (Purves, Winter, & Kuhn, 2019; Wagner, Zipf, & Westerholt, 2020; Westerholt, Mocnik, & Zipf, 2018), as contributions to social media are theorised to capture informal knowledge that represents a place-based understanding of geography (Goodchild & Li, 2011; Sui & Goodchild, 2011). In the context of language, this place-based knowledge is generated through 'vernacular geography', which describes the natural language used when informally describing geographic locations (Gao et al., 2017; Goodchild & Li, 2011; Hollenstein, 2008; Waters & Evans, 2003). This informal knowledge incorporates biases regarding locations, better representing human perceptions of geography, compared with formal administrative definitions. In this sense, associations of geography drawn from social media capture place through a 'bottom-up' approach, building knowledge through experience rather than administrative formalisations (Agnew, 2005; Sui & Goodchild, 2011). While many works have considered the formalisation of place through geotagged social media data, few have considered how the semantic properties of text may reveal geographic heterogeneity between regions, generated directly through vernacular geography. The components of vernacular geography are closely coupled with the identity of regions, where culture, topics, and general perceptions are captured through the language associated with locational mentions in text (Buttimer, 2015; Paasi, 2003).

A multitude of works have considered the geographic variation in geotagged social media text (Arthur & Williams, 2019; Doyle, 2014; Eisenstein, O'Connor, Smith, & Xing, 2014; Gonçalves & Sánchez, 2014; Huang, Guo, Kasakoff, & Grieve, 2016; Pérez, Aleman, Kalinowski, & Gravano, 2019; Russ, 2012), focussing primarily on how dialect variation is captured through differences in the vocabulary (lexicons) of contributors over geographic space. For example, Tweet lexicons originating in the North East of England are noticeably different compared with the South (Arthur & Williams, 2019). While dialects do demonstrate geographic heterogeneity, they only present one component of language that may exhibit geographic variation and do not directly contribute properties associated with vernacular geography. This limitation stems primarily from the reliance of these works on geotagged social media, where the textual content rarely relates to the geotagged

* Corresponding author.
E-mail addresses: c.berragan@liverpool.ac.uk (C. Berragan), ucfnale@liverpool.ac.uk (A. Singleton), acalafio@ed.ac.uk (A. Calafiore), jeremy.morley@os.uk (J. Morley).

https://doi.org/10.1016/j.compenvurbsys.2024.102121
Received 20 November 2023; Received in revised form 16 April 2024; Accepted 19 April 2024
Available online 26 April 2024
0198-9715/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

location (Kropczynski et al., 2018), meaning dialects are the only explainable trait that results in geographic heterogeneity.

In our work, we instead consider the ability to compare the geographic variation in semantic information relating to locational mentions embedded directly within social media text. This approach means that instead of solely focussing on dialects, our work captures language directly associated with locations, contributed by the vernacular geography of users. While lexical approaches explore the vocabulary of a language, we instead generate sentence embeddings using new developments in natural language processing, which generate contextual semantic representations of text, using a large language model (Devlin, Chang, Lee, & Toutanova, 2019; Hu et al., 2020). These embeddings are therefore able to distinguish between nuanced differences in how locations are discussed, building representations of words that incorporate their surrounding context, and utilising human knowledge learned by the large language model. Notably, unlike lexicons, embeddings associated with unique words generated by a large language model have different representations, depending on their surrounding context. This is particularly important in our use-case, where general topics like 'restaurants' are frequently discussed in location forums, but differences in the way they are discussed is influenced by the distinctive culture of each location.

We name these representations the 'semantic footprints' of locations; capturing semantic traces relating to locations, contributed by individuals through a subset of their digital footprints (Walden-Schreiner, Leung, & Tateosian, 2018). We then analyse these semantic footprints, to determine whether they form geographically cohesive clusters, through an analysis of their spatial autocorrelation. We then investigate whether observed clusters of semantic footprints correspond with associated national identities. To achieve this, we utilise the emergent properties of large language models (LLMs), where a task known as zero-shot classification enables models to assign labels to text, without any annotated training data. We query an LLM to attribute a specific sub-nationality within the United Kingdom to each of our comments and explore whether the varying strength of these nationalities correlate with differences in our semantic footprints.

Section 2 first gives an overview of work exploring semantic variation in social media text, regional identities, and how our approach differs to related work. Section 3 describes our data, then outlines the processing used to generate semantic footprints and describes our geographic analysis of these footprints. Section 4 presents our results and Section 5 concludes with suggestions for future work.

2. Geographic variation in social media text

While formal geographic regions within Great Britain are typically designed for administrative and political purposes, they are non-restrictive in how populations can move between them. The level of geographic cohesion between regions across Great Britain is often studied from the context of mobility, where data sources like Census or transport records describe the physical movement of populations and individuals across geographic space (Rae, 2009; Titheridge, Achuthan, Mackett, & Solomon, 2009), or through non-physical networks using phone records (Lambiotte et al., 2008; Reades, Calabrese, & Ratti, 2009; Sobolevsky et al., 2013; Y. Zheng, 2015), and social media (Arthur & Williams, 2019; Lengyel, Varga, Ságvári, Jakobi, & Kertész, 2015; Sui & Goodchild, 2011). When these networks are examined, cohesive clusters develop, which broadly appear to correlate with administrative boundaries (Arthur & Williams, 2019; Ratti et al., 2010).

Alternatively, many works have taken advantage of the abundance of geotagged social media text, to examine regional differences in dialects (Arthur & Williams, 2019; Doyle, 2014; Eisenstein et al., 2014; Gonçalves & Sánchez, 2014; Han, Cook, & Baldwin, 2012; Huang et al., 2016; Russ, 2012; Zheng, Han, & Sun, 2018). Many of these works have noted that, like online or physical networks, geographically cohesive properties emerge, which appear to correlate with administrative boundaries (Arthur & Williams, 2019; Eisenstein et al., 2014; Gonçalves & Sánchez, 2014; Huang et al., 2016). These results conform with the idea that dialects are an important component in the identity of regions (Haesly, 2005; Llamas, 2009; Llamas & Watt, 2014). Despite this, dialects only present a single component of language that contributes to a sense of geographic identity between regions (Haesly, 2005; Middleton & Freestone, 2008), ignoring the wealth of vernacular geography that may also be captured in text (Berragan, Singleton, Calafiore, & Morley, 2023;Evans & Waters, 2007; Sui & Goodchild, 2011).

Studies that consider dialect variation in social media text only consider geotags to be a geographically relatable feature of this data source. Given social media communication comprises a broad range of topics that do not necessarily relate to locational discussion, these geotags and associated text are unlikely to be directly related. Any observed regional variation is therefore only attributable to the dialect of the contributing author, with the assumption that the author is a resident in the geotagged location. In contrast to this approach, locational mentions embedded directly within text present an alternative method to explore how the language regarding locations varies geographically. Place names embedded within text directly can also be related with the surrounding context of their use, capturing the vernacular geography of contributing users (Evans & Waters, 2007; Sui & Goodchild, 2011). Lexicons associated with locations identified in this manner therefore incorporate a broad range of topics, associations, and cultural information, rather than solely dialects, more broadly capturing the components of language that contribute to the identity of locations (Haesly, 2005). In our work, we therefore extract place names from a collection of UK specific comments taken from the social media website Reddit, where coordinate information was attributed to comments through a process called geoparsing (Purves et al., 2019), allowing for us to explore the geographic heterogeneity of text associated with identified locations.

While past works have primarily considered the statistical comparison between location-based lexicons, where word counts are associated with aggregate regions generated through geotagged Tweets, this approach is limited when considering the more nuanced semantic variations in vernacular geography. Recent progress in natural language processing have led to the development of large language models (LLMs) which are able to capture deep contextual semantic information from text, through sentence and word embeddings (Devlin et al., 2019). Unlike a lexical approach, where word order and semantic information is not captured, these embeddings act as numerical representations of text which incorporate contextual semantic information in depth. Embeddings that are more semantically similar are closer together in their embedding space, meaning, like lexicons, these embeddings may be statistically compared. We therefore generate sentence embeddings for each geoparsed comment in our corpus, which are then aggregated by location, forming what we call a semantic footprint. These footprints represent the collective geographic knowledge of each individual user in our corpus, built through their vernacular geography, capturing informal, place-based information through their perception of discussed locations (Goodchild & Li, 2011; Sui & Goodchild, 2011).

In this work, we generate a new comparative measure between regions in the UK through an examination of text associated with locations, extracted from comments on the social media website Reddit. While past work has examined variation between regions from the perspective of social media networks, or by examining lexicons associated with geotagged social media messages, we examine regional variations derived from geoparsed embeddings generated from a large language model. Unlike using geotags, which ascribe linguistic features such as dialect to specific locations, our method instead captures any comment that mentions a location alongside its semantic context. Quantified information therefore does not reflect dialects associated with locations, but common semantic associations, embedding cultural information, or location-specific topics and opinions. Given users mentioning locations are not necessarily residents, these semantic

associations represent a collective informal geographic knowledge generated through the vernacular geography of people across the UK, embedding their general semantic footprint.

3. Methodology

The following section first introduces our main data source; the social media website Reddit, from which we access a collection of geoparsed user-submitted comments. Following this, we detail our methodology for generating semantic footprints from each of these comments, and how we analyse the geographic properties of these footprints.

3.1. Data

Reddit is a public discussion, news aggregation social network, and among the top 20 most visited websites in the United Kingdom. In 2020, Reddit had around 430 million active monthly users, comparable to the number of Twitter[1] users (Murphy, 2019; Statista, 2022). Reddit is divided into separate independent *subreddits* each with specific topics of discussion, where *users* may submit *posts* which each have dedicated nested conversation threads that users can add *comments* to. Subreddits cover a wide range of topics, and in the interest of geography, they also act as forums for the discussion of local places. The United Kingdom subreddit acts as a general hub for related topics, notably including a list of smaller and more specific related subreddits. This list provides a 'Places' section, a collection of local British subreddits, ranging in scale from country (`/r/England`), region (`/r/thenorth`, `/r/Teeside`), to cities (`/r/Manchester`) and small towns (`/r/Alnwick`). In total there are 213 subreddits that relate to 'places' within the United Kingdom.[2] We use the corpus generated by Berragan et al. (2023), which consists of a collection of all Reddit comments taken from each UK related subreddit (Baumgartner, Zannettou, Keegan, Squire, & Blackburn, 2020), with place names identified by a custom transformer-based named entity recognition model.[3] In total 8,282,331 comments were extracted, submitted by 490,535 unique users, between 2011 and 01-01 and 2022-04-17. Table gives an example entry from this geoparsed Reddit corpus.

There are a total of 40,429 unique locations in this corpus, with a highly skewed distribution in mentions. Many locations were only mentioned a single time (37%), while 'London' was mentioned in 283,521 comments. To reduce this skew, we sampled any location mentioned >5000 times, retaining only up to 5000 randomly sampled comments per location. The goal with this processing was to ensure that our generated embeddings did not simply become biased towards the word embedding for a single location, and instead capture a broader sense of an aggregate region. In our data subset, we find that 1% of users (1734) mention 29% of our place names. This subset leaves a total of 852,461 comments containing place names. Comments range from 1 to 3555 words in length, with a mean length of 79. Table gives an overview of the number of comments, word count and number of places that were identified within each administrative region of the UK.

3.2. Generating and analysing geographic footprints

Statistical comparisons between two or more distinct texts first relies on an appropriate method for processing the text into a numerical format. Typically, a Term Frequency-Inverse Document Frequency (TF-IDF) approach is used to generate document embeddings (Daniel & James, 2007), which assigns word importance based on the frequency of mentions within a corpus. TF-IDF however does not have the capability to capture broader semantic information, given that there is no knowledge of the meaning behind words. Large Language Models (LLMs) instead are pre-trained on a very large corpus of natural language text, which, alongside their architecture, enables them to more appropriately consider semantic information (Devlin et al., 2019). As with TF-IDF, text is input into these models and output as a numerical representation, which embeds words as high dimensional vectors, capturing contextual semantic information.

This approach differs from past work that only considered a lexical analysis, where semantic information and context is not preserved, instead building vectors that act as semantic representations of locations identified in our corpus, which we name 'semantic footprints'. Given semantic information is preserved, locational embeddings are able to reflect the deeper associations between geographic locations, built from a multitude of contexts and perspectives, forming an aggregate representation. Any geographically cohesive relationships between footprints therefore demonstrate a direct association between geography and language, which hasn't been captured previously.

Once we generate these footprints we first explore how they produce emerging spatial structures from the bottom-up, generating clusters of small-scale geographic units to capture larger scale aggregations based on semantic information. In this analysis we find that our generated spatial structures broadly conform with larger scale administrative aggregations. We therefore then consider a top-down approach, using these larger administrative regions to generate a comparative analysis of aggregate footprints. To derive explainable characteristics of observed differences between these regions, we observe how national identities can be captured through text, and how these identities vary geographically.

3.3. Creating embeddings

We first create semantic embeddings for each comment in which a location was mentioned, using the `sentence-transformers` Python library (Reimers & Gurevych, 2019), with the `all-mpnet-base-v2` model.[4] With our selected embedding model, we then performed the following steps to generate embeddings for each Local Authority District (LAD) in Great Britain.

1. Masked any place name with a generic token: 'PLACE' (using place name text spans included in the corpus).
2. Generate sentence embeddings for each comment.
3. Group embeddings by LAD using identified locations, and mean-pooling.

To visualise the outputs from this processing we consider an example comment $s_1 = $ "I live in London.", shown on Eq. (1).

$$\mathbf{1.}\begin{array}{l} s_i = \text{'I live in London'} \\ \downarrow \\ s_i = \text{'I live in PLACE'} \end{array} \quad \mathbf{2.}s_i \rightarrow \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\mathbf{3.}LAD_j = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,t} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,t} \end{bmatrix} \rightarrow \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \\ \vdots \\ \overline{x_n} \end{bmatrix} \tag{1}$$

In Eq. (1), n is the `sentence-transformers` embedding dimension (768), and t is the total number of unique comments that relate to locations within a single LAD region (LAD_j). Values (x_i) in step **2.** are model weights that represent the embedding for the comment s_i, capturing semantic information. Fig. 1 demonstrates this process visually.

Given each *LAD* has a variable number of comments associated with

[1] Now known as X
[2] https://www.reddit.com/r/unitedkingdom/wiki/british_subreddits
[3] Berragan et al. (2023)
[4] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

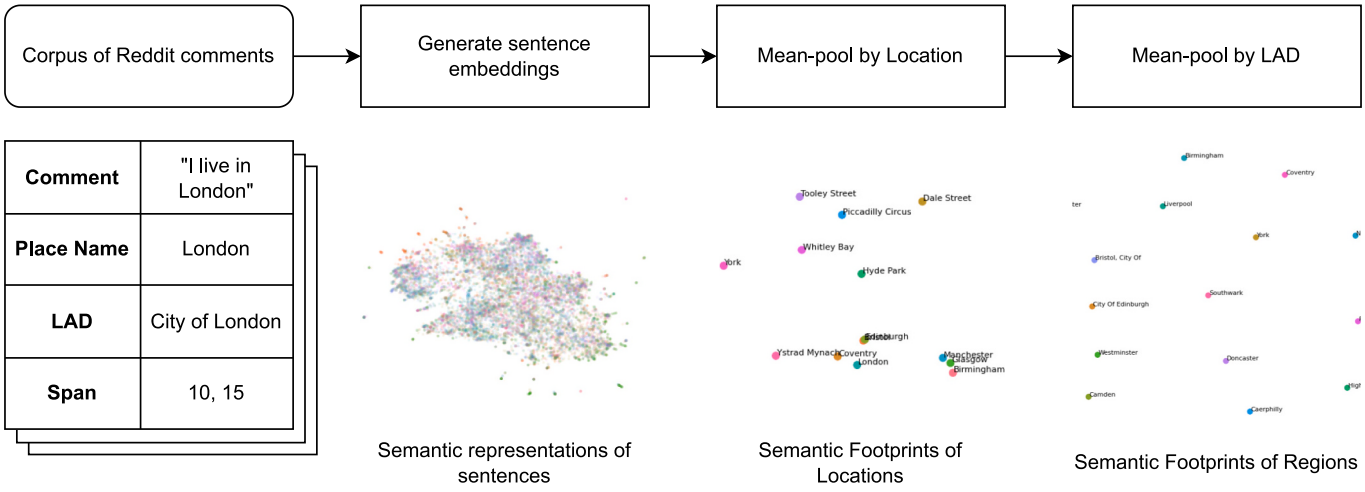


Fig. 1. Workflow diagram showing Reddit Corpus processed into sentence embeddings, then aggregated into location and LAD semantic footprints.

them, we process associated embeddings into a 'semantic footprint' representation of a fixed size, so that they may be directly compared. To achieve this, all embeddings associated with comments relating to locations within a LAD_j are processed into a one-dimension vector of size 1×768. The most common approach for this dimensionality reduction uses 'mean-pooling'; taking the mean across all embeddings, which is common in tasks like topic analysis (Reimers & Gurevych, 2019).

Place name spans provided by our corpus include all names identified as place names by the corpus, regardless of whether they are geographically grounded, meaning points of interest like restaurants, or shop names are also excluded from our embeddings. By masking place names, we ensure that no comment embeddings accidentally incorporate geographically grounded information. For example, comments in South Eastern local authorities are likely to frequently mention London, given they are geographically proximal. Embeddings for these locations would therefore capture an association through the mention of London, rather than general semantic information. For our work, we want to exclude any geographic information, ensuring that embeddings solely capture semantic associations.

Given that transformers are a relatively new architecture in natural language processing, and the creation of these models require significant computational resources and training time, their use to date has been limited in related research. Our choice to use the transformer architecture stems from the emphasis we place on the extraction of nuanced and contextual semantic information, which is lost with lexical count-based methods like TF-IDF. It should be noted however that while TF-IDF methods are less complex, they are typically more interpretable; for instance, words that contribute importance to an embedding may be extracted from a TF-IDF model. The numerical representations of any text generated by transformers are not directly interpretable in this manner. The following section therefore analyses our semantic footprints with respect to their numerical representations, rather than through their lexicons.

3.4. Spatial clustering and autocorrelation

It is reasonable to assume that there are LADs within our corpora that generate embeddings that capture similar semantic properties. A typical method to group unlabelled multi-variate data based on shared properties uses unsupervised clustering (Likas, Vlassis, Verbeek, & J., 2003; Sinaga & Yang, 2020). Therefore, to explore whether geographically cohesive clusters appear within our semantic embeddings, we generate hierarchical clusters, which are non-geographically bounded, using agglomerative clustering. This clustering method automatically determined the optimal number of clusters to be three, using distance threshold of zero. These clusters were visualised geographically, to examine whether geographically cohesive groupings occurred. The proportion of clusters present within each administrative region (RGN)[5] in Great Britain was also plotted to determine whether clusters appeared to correlate with administrative boundaries.

To quantify the level of spatial autocorrelation that our embeddings exhibit, we consider the Moran's I metric, which identifies the spatial relationship between each observation and its geographic neighbours (Anselin, 1995; Rey, Arribas-Bel, & Wolf, 2023). Moran's I values are generated based on the strength of correlation between values and the aggregate values of their geographic neighbours, known as their spatial lag. Higher Moran's I values therefore denote a stronger spatial autocorrelation. Given that Moran's I analysis requires univariate data, we explore global spatial autocorrelation of our semantic footprints decomposed into two dimensions using 'Uniform Manifold Approximation and Projection' (UMAP) (McInnes, Healy, & Melville, 2020), and plot both dimensions against their spatial lag, giving two distinct global Moran's I values. UMAP is selected over alternative algorithms like t-distributed Stochastic Neighbor Embedding (t-SNE) as it has been found to outperform t-SNE for downstream tasks, and is capable of preserving the global structure of the data (Allaoui, Kherfi, & Cheriet, 2020; McInnes et al., 2020).

We then consider how localised levels of high spatial autocorrelation may be identified through a Local Indicators of Spatial Autocorrelation (LISA) analysis. Instead of single global values, LISA analysis determines whether each unique LAD polygon exhibits a statistically significant level of spatial autocorrelation, and assigns a local Moran's I value for each.

It is important to note that the magnitude of our embeddings do not convey any definable information, values therefore only highlight differences in semantic information between regions, rather than importance. For example, an embedding value of 0 is not less important than a value of 1 or -1.

3.5. Semantic similarity

Following our analysis of LAD semantic footprints, we explore our semantic footprints from a top-down perspective, aggregating LADs into established large-scale RGNs across Great Britain, taking the mean of the collective semantic footprints. Each RGN is therefore represented by a single 768 dimension semantic footprint embedding. We then calculate

[5] The highest tier of sub-national division in England. For Scotland and Wales we use the full national extents.

the cosine similarity between each RGN embedding, demonstrating the level of inter-region semantic cohesion across Great Britain. Eq. (2) shows how the cosine similarity is calculated; the angle between two non-zero vectors determined through their dot product, divided by the product of their lengths.

$$Cosine(x, y) = \frac{x \cdot y}{|x||y|} \tag{2}$$

Cosine similarity is a common metric for comparing embeddings, as it is invariant to the magnitude of the vectors, and only considers the direction. This is required as the magnitude of embeddings is not meaningful, and only the direction of the vector conveys information. For example, the embedding for the 'South East' cannot be twice as important as the embedding for the 'North West'.

3.6. Capturing national identities through text

To generate explainable characteristics of any geographically distinct semantic footprints generated in our analysis, we consider how a language model associates national identities with the semantic properties of text. In our approach we mirror qualitative data collection methodologies in political science research, where individuals are typically queried their chosen national identity (Griffiths, 2022; Haesly, 2005); instead generating the categorisations of comments by querying a large language model (LLM).

LLMs are pre-trained on a large corpus of natural language text, building representations of this text that emulate a human understanding of language. The underlying theory is that these representations capture the collective knowledge of humans that contributed the natural language text used to build them. Therefore, in addition to factual information, when posed with non-deterministic questioning, these models are able to contribute the biased information that is incorporated into their model weights.

Recent research has noted on the ability to perform zero-shot classification using LLMs, where class predictions may be made without the model ever having previously seen the labels (Wei et al., 2022; Wei et al., 2022). While research has considered the use of questionnaires to query the strength of national identities within the UK (Griffiths, 2022; Haesly, 2005), an LLM may instead be used. For example, an LLM may be questioned whether it personally feels a sequence of text appears to be 'British', 'English', 'Scottish', or 'Welsh'. Through this zero-shot classification, we are able to determine the strength of national identity associated with each region in our work, to examine whether this appears to correlate with any cohesion between the semantic footprints that we generate. Importantly, we are also able to generate confidence values from the chosen LLM, allowing for the strength of these national identities to be captured.

Semantic information within our comments is expected to capture both explicit information contributed by users; for example stating 'London is a British city', in addition to implicit semantic information that exists within language. For example the phrase 'bonnie Scotland' may suggest a strong identity due to the inclusion of Scottish slang.[6] Unlike our semantic footprints, we do not mask place name mentions in these embeddings, enabling the model to make its own decisions regarding place name mentions.

To identify regional identities through semantic information, we build on the emergent properties of large language models, which enable a task known as 'Zero-Shot Classification'. This allows models to predict a class that was not seen during training, by generating a prompt that contains the labels required. For this task we select the `typeform/distilbert-base-uncased-mnli` model,[7] which is tailored towards zero-shot classification, therefore generating slightly different embeddings compared with those used for our semantic footprints. For our task the following gives an example prompt with a portion of a comment taken from our corpus, where the Scottish colloquial slang 'gonnae' is used:

```
Classify the following input text into one of the following four categories:
[British, English, Scottish, Welsh]
Input Text: My favourite was in Livingston: 'Rab, I'm gonnae find you.'
```

The output would then be given as a sequence of confidence values for each label:

```
'labels': ['Scottish', 'British', 'Welsh', 'English']
'scores': [0.761, 0.144, 0.052, 0.043].
```

4. Results

Fig. 2(a) shows clusters of each 363 LAD transformer embeddings, UMAP decomposed into two dimensions, indicating embeddings that share similar semantic properties. These clusters appear to broadly correlate with three distinct regions within Great Britain, where cluster 0 most closely identifies with England, 1 with London and surrounding areas, and 2 with Scotland and Wales (Fig. 2(b-c)). The few areas that appear as cluster 0 in Wales and Scotland are major urban centres like Cardiff, Glasgow, and Edinburgh. Overall these clusters appear to be geographically restricted, and even broadly conform with administrative regions like the Welsh and Scottish borders.

These findings appear to share similarities with past work that has observed strong 'boundary effects', where lexical similarity between geotagged Tweets often correlates with administrative boundaries (Arthur & Williams, 2019; Bailey, Cao, Kuchler, Stroebel, & Wong, 2018; Li et al., 2021; Yin, Kann, Yu, & Schütze, 2017). Our embeddings also exhibit the general geographically coherent patterns that have been observed in geographical lexical variations in social media (Arthur & Williams, 2019; Doyle, 2014; Eisenstein et al., 2014; Gonçalves & Sánchez, 2014; Huang et al., 2016; Pérez et al., 2019; Russ, 2012). Notably, unlike dialects, where a geographic component is expected, the geographic association of our general semantic embeddings has not been demonstrated in past work. Results therefore demonstrate that despite no pre-existing geographic information like geotags or place names, general text associated with locations appears to embed a geographic component. The distinct change in clusters at the borders of Scotland and Wales conforms with our hypothesis that the vernacular geography that exists within social media text embeds components that contribute to the strength of national identities (Haesly, 2005).

As noted however, major cities in Wales and Scotland Glasgow, Edinburgh and Cardiff share a cluster with English LADs rather than their respective country, suggesting that these locations are more semantically connected with the rest of Great Britain. This observation mirrors the results of work that considered co-occurring locational mentions between cities, where shared city mentions in text often appear irrespective of distance, and across administrative borders [anonymised]. This deviation from the relative semantic isolation of Scotland and Wales from England appears to be reflective of the nature of major cities, given they tend to share stronger physical geographic connections across a larger geographic scope, and more influential cultural connections compared with rural areas, captured in our work through shared semantic traits with the cluster associated with England.

Cluster 1 presents in areas surrounding London and suggests distinctiveness of this region relative to the rest of Great Britain. This is interesting given London's extensive connectivity relative to the rest of the country, and the general sense of strong association with other cities, given it is the capital city [anonymised]. Our results therefore suggest that despite London's importance nationally, semantic information is able to capture a deeper context that dissociates it from other regions. This effect may be due to factors unique to London, for example its prominence globally, influencing both tourism and business external to

[6] See 'Scottish English' or 'Scots'; (Stuart-Smith, 2008)

[7] https://huggingface.co/typeform/distilbert-base-uncased-mnli

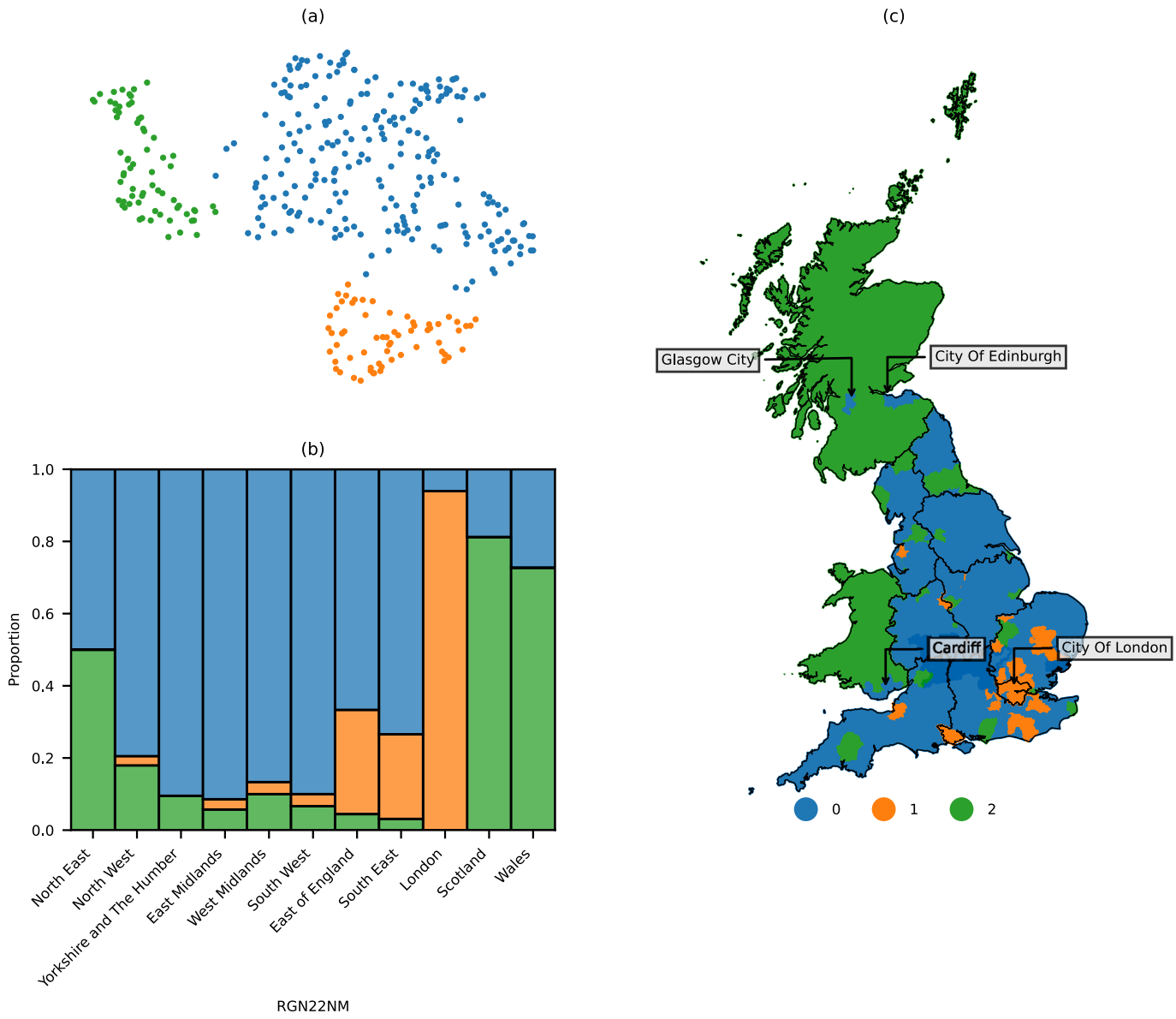


Fig. 2. Semantic footprints associated with 363 LAD corpora, coloured by hierarchical agglomerative clusters where K = 3. (a) LAD footprints UMAP decomposed into two dimensions. (b) Proportion of LADs within clusters by RGN. (c) Geographic location of LAD clusters.

the United Kingdom, which alter the cultural landscape of the city. The isolated characteristics of London are particularly observable through its economic differences, where high costs of living have generated the need for a 'London weighting'[8] of salaries (Hirsch, 2016).

The following section formalises the level of geographic coherence that the embeddings exhibit, and highlights the key locations that drive the relationship between text and geography.

4.1. Moran's I analysis

To quantify whether our embeddings demonstrate spatial autocorrelation, we consider the Moran's I metric, which identifies the spatial relationship between each observation and its geographic neighbours (Anselin, 1995). Given that this analysis requires univariate data, we explore global spatial autocorrelation of our UMAP decomposed embeddings, computing the spatial lag for both dimensions. On Fig. 3, we plot both values for each LAD semantic footprint in Great Britain, against the spatial lag of these values. A higher correlation between the semantic footprints values and their spatial lag indicates a stronger level of global spatial autocorrelation, resulting in a higher Moran's I value. Fig. 3 shows a positive correlation between the PCA decomposed embedding values and their spatial lag, resulting in Moran's I values of 0.31 and 0.39. This indicates a reasonably strong spatial autocorrelation with both embedding dimensions, confirming that semantic footprints are typically more similar between nearby locations. While the Moran's I values for both dimensions are similar, their cosine similarity is negative (−0.11), meaning these two decomposed dimensions capture distinctly different semantic traits.

While spatially coherent results have been demonstrated from the perspective of dialects on social media (Arthur & Williams, 2019; Doyle, 2014; Eisenstein et al., 2014; Gonçalves & Sánchez, 2014; Huang et al., 2016; Pérez et al., 2019; Russ, 2012), we have demonstrated that this phenomenon can also be captured from general semantic information. Notably, while dialects have always been considered to have strong geographical grounding (Trudgill, 2004), it is more surprising that general semantic information regarding locations similarly exhibits this relationship.

[8] https://en.wikipedia.org/wiki/London_weighting

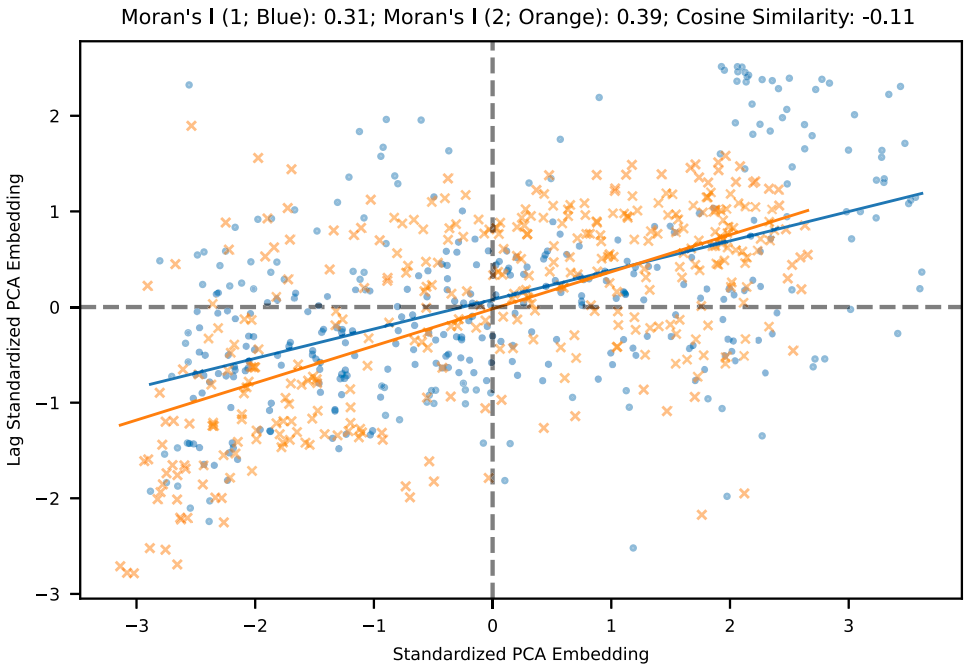


Fig. 3. Moran's I Plot: LAD embeddings decomposed into 2 dimensions and standardised against their spatial lag.

To explore local indicators of spatial autocorrelation (LISA) we plot each decomposed embedding on Fig. 4(a/d), each local Moran's I value on (b/e) and all significant ($p < 0.05$) HH and LL LISA quadrants on (c/f). Note that only selecting significant p values on Fig. 4(c/f) ensures that no regions are included that have values that could demonstrate autocorrelation even if randomly distributed geographically. From Fig. 4(c/f), we can see that notable large areas with significant levels of spatial correlation include;

- Scotland
- Wales
- London and surrounding LADs
- the South West; towards Cornwall

As demonstrated by the low cosine similarity between our UMAP embeddings, they appear to capture distinctly different semantic information. London for example only appears in dimension 0, while dimension 1 captures broader spatial autocorrelation across Scotland and Wales. In Scotland we can see that from both LISAs, Glasgow and Edinburgh represent areas of HL/LH, where semantic information in these cities is not the same as surrounding LADs, an effect that is also captured in some LADs surrounding London. England overall appears to be a less semantically cohesive country based on this analysis, where most LADs do not contribute significant levels of spatial autocorrelation.

These results again demonstrate geographic cohesion between semantic footprints, which notably appear to correspond with the national boundaries of Wales and Scotland. This mirrors the observations of past work where dialect differences appeared to correlate with administrative boundaries (Arthur & Williams, 2019; Bailey et al., 2018; Li et al., 2021; Yin et al., 2017). In addition to Wales and Scotland, we have also identified a notable grouping in the South West, which potentially reflects the Cornish identity (Deacon, 2007), as well as a grouping associated with London.

4.2. Semantic similarity and identity

Given the regions highlighted as having strong spatial autocorrelation in their semantic footprints appear to broadly conform with the administrative regions of Wales, Scotland, and London, we examine these footprints from a top-down analysis using pre-defined larger scale aggregations.

Fig. 5 compares the cosine similarity between each RGN embedding, allowing for inter-regional cohesion to be explored. The North West has the overall highest level of cosine similarity, displaying comparatively high similarity with most regions across England, excluding London. London has the lowest overall similarity, only sharing positive cosine similarity values with the South and South East of England. As expected, Scotland and Wales have low overall cosine similarity values, with Wales sharing even lower similarity with respect to London and the South East compared with Scotland. Mean values show clearly that the least cohesive regions appear to be London, Wales, and Scotland, three regions that are also those with the strongest levels of spatial autocorrelation.

Excluding London, the North East is the region in England with the lowest overall cosine similarity with the rest of Great Britain. This is perhaps reflective of distinct differences with this region, for example the distinctly lower gross value added (GVA) compared with other regions (Fenton, 2018), or the general sense of strong identity that is often noted by residents (Middleton & Freestone, 2008). Alternatively, the North West is home to nationally influential urban conurbations, especially between Manchester and Liverpool (Oguz & Walton, 2022), likely generating the highest overall semantic similarity of this region compared with the rest of the UK. Comparatively, the East of England, South East and London are neighbouring regions that share high similarities with each other, but exhibit low similarity with the rest of Great Britain, suggesting there are semantic components that distinguish this region of the country from the rest. There is a slightly higher mean similarity with respect to Scotland compared with Wales, due to higher similarities with regions in England, like the North West and South East. Major urban centres in Scotland are relatively well connected to Great Britain through rail routes, and Edinburgh and Glasgow are historically important UK cities, captured by their distinct difference in embedding values during the spatial autocorrelation analysis. This factor likely increases the cosine similarity of Scotland with regions in England, while Wales in this sense is less directly associated with the rest of the UK.

To determine whether regional identities generated by a large language model align with these semantically isolated regions in our analysis, we plot the distribution of regional identities identified

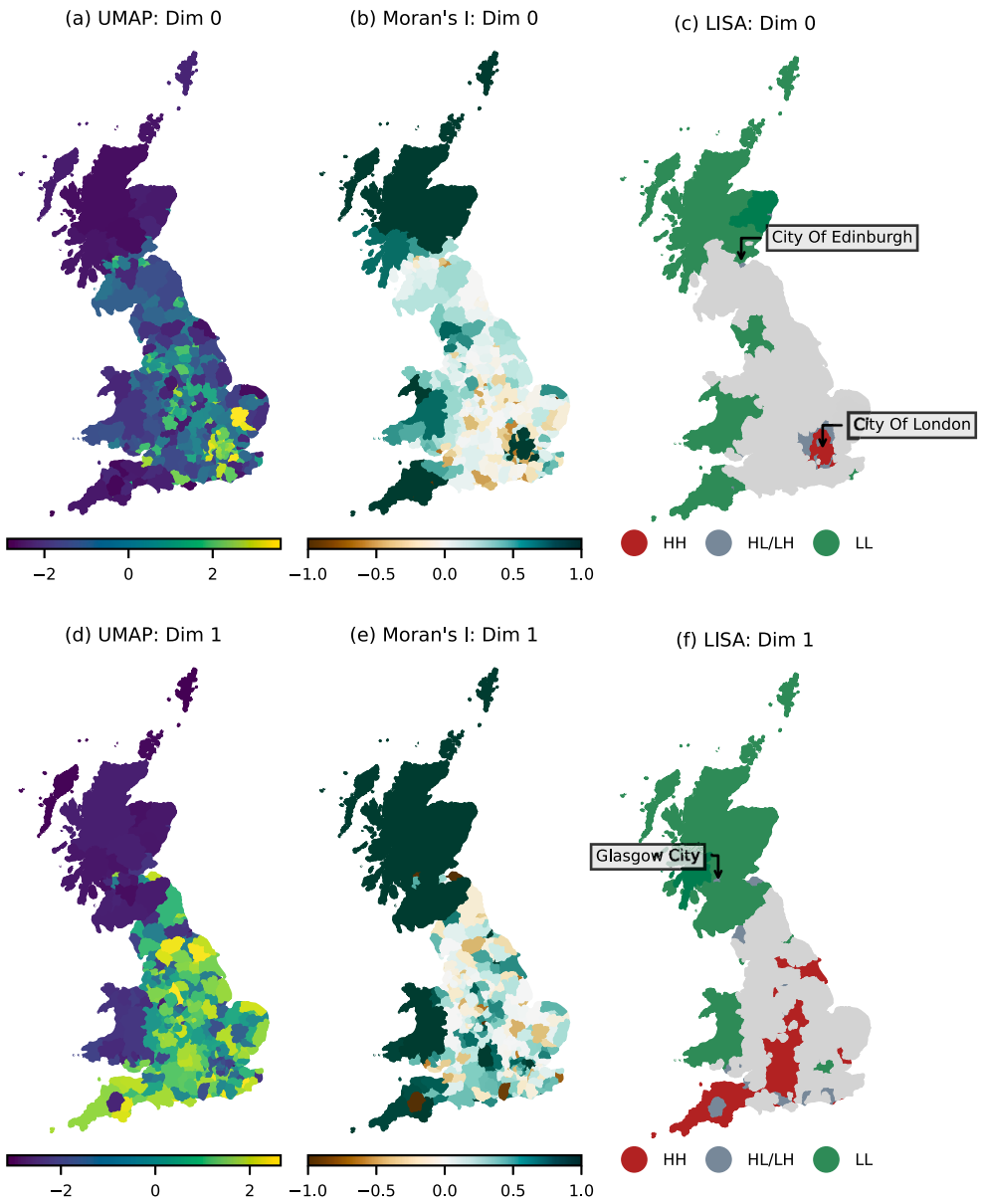


Fig. 4. Local Indicators of Spatial Auto-correlation (LISA). (a/d) 1 dimensional embedding values. (b/e) Local Moran's I values (Is). (c/f) LISA HH and LL significant values ($p < 0.05$), both are included as the value of embeddings do not convey information.

through our zero-shot classification on Fig. 6.

Across each region, the 'English' identity is always lower than 'British', suggesting that regions within England are typically more strongly associated with the United Kingdom[9] than solely England. Unlike English regions however, comments relating to both Scottish and Welsh locations are more strongly associated with their respective nationalities. However, comments relating to Welsh locations appear on average to have stronger confidence values with respect to the British classification, compared with Scottish locations. Similar observations have been captured from qualitative interviewing, where Welsh residents similarly appear to more strongly associate themselves with the British identity, compared with Scottish residents (Carman, Johns, & Mitchell, 2014; Haesly, 2005; Llamas, 2009; Llamas & Watt, 2014). Of the English regions, London has a distinctly higher average confidence value of both British and English identities compared with all other regions. Notably given the semantic footprints for Scotland, Wales, and London also have the lowest overall cosine similarity values, these differences in generated identity compared with other regions are a likely component in their semantic differences.

4.3. General observations

Unlike typical representations of the North-South divide within England (Jewell, 1994), semantic differences appear to be influenced primarily by proximity to London. Unlike typical representations of this divide, the South West of England therefore appears to be distinct from the South East, with a stronger association with the North. South Eastern regions however do share lower similarity to the Midlands and North of England, which conforms with a typical view of the English North-South divide.

In a similar sense, Scotland and Wales demonstrate distinctly more cohesive semantic properties compared with England, exhibiting high spatial autocorrelation, like London. In traditional linguistic research,

[9] Note that despite etymologically relating to 'Great Britain', the term 'British' refers to 'belonging to or relating to the United Kingdom of Great Britain and Northern Ireland'

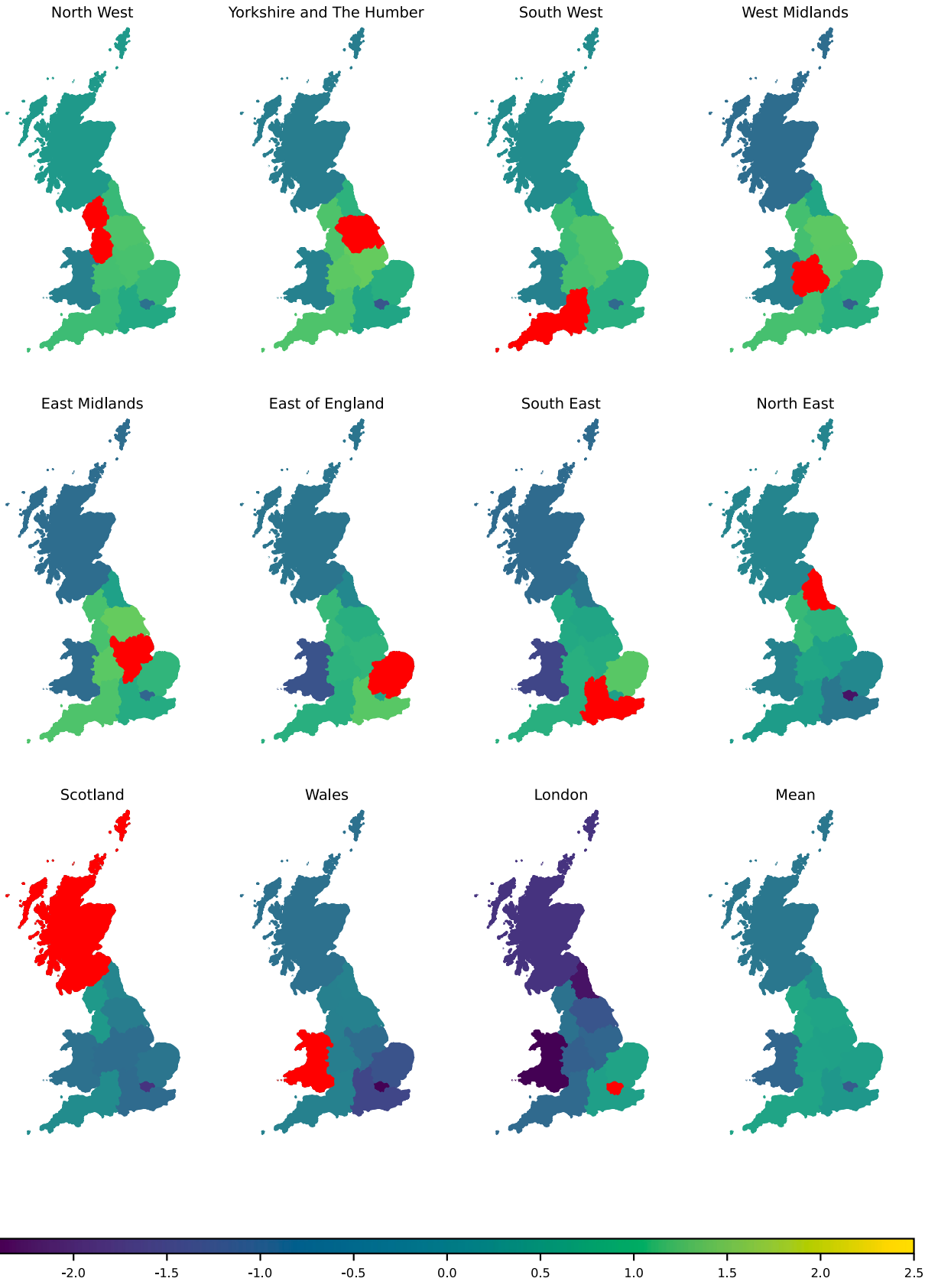


Fig. 5. Scaled cosine similarity of embeddings for administrative regions across the UK. Higher values indicate greater cosine similarity. Regions shown in descending order by mean cosine similarity value.

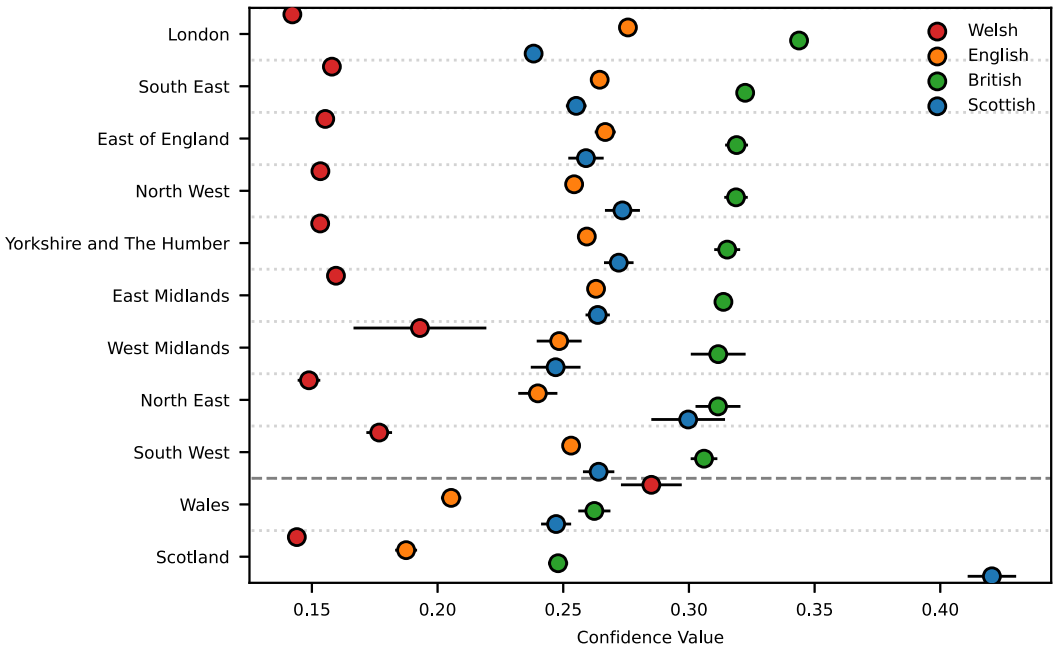


Fig. 6. Zero Shot classification of each corpus into regional identities; British (Green), English (Orange), Scottish (Blue), Welsh (Red). Values show mean confidence value across each comment, lines indicate standard error. Descending order by British confidence. The dashed line separates English regions from Scotland and Wales. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the spoken dialect across England is known to vary considerably (Chambers & Trudgill, 1998; Deacon, 2007; Knowles, 1973; MacKenzie, Bailey, & Turton, 2022), which captures the distinct localised identities that exist across geographic space. In our analysis this is mirrored through the variation in semantic footprints for LADs across the UK, where spatial autocorrelation is generally low, and highly localised to regions like London. The high spatial autocorrelation within Wales and Scotland appears to capture the stronger sense of national identity that these constituent countries exhibit in our analysis, and is a common qualitative observation in political science research (Carman et al., 2014; Haesly, 2005).

As demonstrated in past work that has examined both physical and non-physical networks, our observed semantic information similarly appears to correlate with pre-defined administrative boundaries, particularly the national boundaries of Scotland and Wales (Arthur & Williams, 2019; Bailey et al., 2018; Li et al., 2021; Yin et al., 2017). The distinct difference in footprints between each constituent country in the UK conforms with the idea that vernacular geography captures a sense of identity, given our zero-shot classification demonstrates distinct nationalities between Scotland and Wales, unlike English regions where the generated national identity is typically considered British rather than English. Notably however, the slightly stronger British identity within Wales has been observed previously through qualitative interviewing (Carman et al., 2014; Haesly, 2005), suggesting that even the nuanced properties of text appear to correlate with the true perceptions of individuals. It is also worth noting that, given that place names themselves are masked within our embeddings, these distinct differences are not simply the result of differences in place names. Welsh names are often derived from the Welsh language, and as such are often distinctly different compared with English place names, which may have influenced the results of past lexical work.

Despite most locations across Scotland and Wales appearing disconnected with the rest of the UK, major cities like Glasgow and Edinburgh are more semantically similar, a distinction that was also observed when the distance decay of locational co-occurrences in text was examined [anonymised]. This suggests that these cities do appear to be typically more semantically connected with the UK, regardless of geographic distance and borders, while other locations typically share semantic properties within the same nation, captured through stronger spatial autocorrelation.

Internal migration patterns within the UK are primarily influenced by family ties, rather than economic factors, employment, or education (Thomas, 2019). The observations made in our work demonstrate that this sense of belonging to regions influences the geographically cohesive nature of our semantic footprints. While populations have the ability to distribute evenly across geographic space, they are often reluctant to move far. Local inhabitants within regions develop an identity associated with their home region, traditionally captured in language through dialect variation, and demonstrated in our work through broader semantic associations, which embed contextual meaning, incorporating the cultural variation of regions.

5. Conclusions and future work

Our paper demonstrates a new method to compare aggregate semantic information for local authorities and regions within the UK, from Reddit comments that mention geoparsed locations, which we name semantic footprints. When examining the semantic footprints of each LAD in the UK, we find that geographically cohesive clusters appear, with significant levels of spatial autocorrelation. Clusters broadly conform with the national borders of Scotland and Wales, while London also appears to be semantically distinct from the rest of England. Our approach shows the extent to which vernacular geographies map to established national and regional boundaries of the UK. The bottom-up identities that emerge from the text appear to correspond with these politically defined boundaries in regions like Scotland, Wales and London, providing a nuanced view of the way UK geographies may be represented; built from the vernacular geographies of social media users.

Geoparsing methods contribute an additional geographic dimension to non-geotagged social media data, allowing for a much larger repository of informal natural language geographic text to be used for research. Future work may consider the use of Reddit comment data to derive notable urban areas of interest (Chen, 2019). This area of research in particular would benefit from methodologies focussing on the extraction of fine-grained locations from text, which at present is a challenging task (J. Han et al., 2018).

Data statement

The data used to produce the analysis in this paper is available at the FigShare DOI https://doi.org/10.6084/m9.figshare.25304575.v1

Funding

This work was supported by the Economic and Social Research Council (ES/P000401/1).

CRediT authorship contribution statement

Cillian Berragan: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alex Singleton:** Writing – review & editing, Supervision, Conceptualization. **Alessia Calafiore:** Writing – review & editing, Supervision, Conceptualization. **Jeremy Morley:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

None.

Data availability

A FigShare link has been added to the Author Statement

References

Agnew, J. (2005). *Space: place in cloke P and Johnston R eds spaces of geographical thought*.

Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020). Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study. In A. El Moataz, D. Mammass, A. Mansouri, & F. Nouboud (Eds.), *Image and signal processing* (pp. 317–325). Springer International Publishing. https://doi.org/10.1007/978-3-030-51935-3_34.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis, 27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Arthur, R., & Williams, H. T. P. (2019). The human geography of twitter: Quantifying regional identity and inter-region communication in England and Wales. *PLoS One, 14*(4), Article e0214466. https://doi.org/10.1371/journal.pone.0214466

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives, 32*(3), 259–280. https://doi.org/10.1257/jep.32.3.259

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. *The Pushshift Reddit dataset (arXiv:2001.08435). arXiv*. (2020). https://arxiv.org/abs/2001.08435.

Berragan, C., Singleton, A., Calafiore, A., & Morley, J. (2023). Transformer based named entity recognition for place name extraction from unstructured text. *International Journal of Geographical Information Science, 37*(4), 747–766. https://doi.org/10.1080/13658816.2022.2133125

Buttimer, A. (2015). Home, reach, and the sense of place. In *The human experience of space and place* (pp. 166–187). Routledge.

Carman, C., Johns, R., & Mitchell, J. (2014). *More Scottish than British: The 2011 Scottish parliament election*. Springer.

Social differentiation and language. In Chambers, J. K., & Trudgill, P. (Eds.), *Dialectology* (2nd ed.,, (pp. 57–69). (1998) pp. 57–69). Cambridge University Press. https://doi.org/10.1017/CBO9780511805103.007.

Chen, M. (2019). *Understanding the dynamics of urban areas of interest through volunteered geographic information. 21*. https://doi.org/10.1007/s10109-018-0284-3

Daniel, J., & James, H. M. (2007). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. prentice hall*.

Deacon, B. (2007). County, nation, ethnic group? The shaping of the Cornish identity. *The International Journal of Regional and Local Studies, 3*(1), 5–29. https://doi.org/10.1179/jrl.2007.3.1.5

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. *BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [Cs]*. (2019). https://arxiv.org/abs/1810.04805.

Doyle, G. (2014). Mapping dialectal variation by querying social media. Proceedings of the. In *14th conference of the European chapter of the association for computational linguistics* (pp. 98–106). https://doi.org/10.3115/v1/E14-1011

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PLoS One, 9*(11), Article e113114. https://doi.org/10.1371/journal.pone.0113114

Evans, A. J., & Waters, T. (2007). Mapping vernacular geography: Web-based GIS tools for capturing 'fuzzy' or 'vague' entities. *International Journal of Technology, Policy and Management, 7*(2), 134–150.

Fenton, T. (2018). Regional economic activity by gross value added (balanced), UK - Office for National Statistics. https://www.ons.gov.uk/economy/grossvalueaddedgva/bulletins/regionalgrossvalueaddedbalanceduk/1998to2017.

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J.-A., McKenzie, G., … Yan, B. (2017). A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science, 31*(6), 1245–1271. https://doi.org/10.1080/13658816.2016.1273357

Gonçalves, B., & Sánchez, D. (2014). Crowdsourcing dialect characterization through twitter. *PLoS One, 9*(11), Article e112074. https://doi.org/10.1371/journal.pone.0112074

Goodchild, M. F., & Li, L. (2011). *Formalizing space and place*.

Griffiths, J. D. (2022). Scrutinizing relative territorial identity measures. *Publius: The Journal of Federalism, 53*(1), 133–151. https://doi.org/10.1093/publius/pjac011

Haesly, R. (2005). Identifying Scotland and Wales: Types of Scottish and welsh national identities. *Nations and Nationalism, 11*(2), 243–263. https://doi.org/10.1111/j.1354-5078.2005.00202.x

Han, B., Cook, P., & Baldwin, T. (2012). *Geolocation prediction in social media data by finding location indicative words. 18*.

Han, J., Sun, A., Cong, G., Zhao, W. X., Ji, Z., & Phan, M. C. (2018). Linking fine-grained locations in user comments. *IEEE Transactions on Knowledge and Data Engineering, 30*(1), 59–72. https://doi.org/10.1109/TKDE.2017.2758780

Hirsch, D. (2016). *London weighting and London costs-a fresh approach?*.

Hollenstein, L. (2008). *Capturing vernacular geography from georeferenced tags*.

Hu, S., He, Z., Wu, L., Yin, L., Xu, Y., & Cui, H. (2020). A framework for extracting urban functional regions based on multiprototype word embeddings using points-of-interest data. *Computers, Environment and Urban Systems, 80*, Article 101442. https://doi.org/10.1016/j.compenvurbsys.2019.101442

Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2016). Understanding U.S. regional linguistic variation with twitter data analysis. *Computers, Environment and Urban Systems, 59*, 244–255. https://doi.org/10.1016/j.compenvurbsys.2015.12.003

Jewell, H. M. (1994). *The north-south divide: The origins of northern consciousness in England*. Manchester University Press.

Knowles, G. O. (1973). *Scouse: The urban dialect of Liverpool*.

Kropczynski, J., Coche, J., Obeysekare, E., Bénaben, F., Grace, R., Halse, S., Montarnal, A., & Tapia, A. (2018). *Identifying Actionable Information on Social Media for Emergency Dispatch* (p. 11).

Lambiotte, R., Blondel, V. D., de Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., & Van Dooren, P. (2008). Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications, 387*(21), 5317–5325. https://doi.org/10.1016/j.physa.2008.05.014

Lengyel, B., Varga, A., Ságvári, B., Jakobi, Á., & Kertész, J. (2015). Geographies of an online social network. *PLoS One, 10*(9), Article e0137248. https://doi.org/10.1371/journal.pone.0137248

Li, Z., Huang, X., Ye, X., Jiang, Y., Martin, Y., Ning, H., … Li, X. (2021). Measuring global multi-scale place connectivity using geotagged social media data. *Scientific Reports, 11*(1), 14694. https://doi.org/10.1038/s41598-021-94300-7

Likas, A., Vlassis, N., Verbeek, J., & J.. (2003). The global k-means clustering algorithm. *Pattern Recognition, 36*(2), 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2

Llamas, C. (2009). *Language and identities*. Edinburgh University Press.

Llamas, C., & Watt, D. (2014). Scottish, English, British?: Innovations in attitude measurement. *Lang & Ling Compass, 8*(11), 610–617. https://doi.org/10.1111/lnc3.12109

MacKenzie, L., Bailey, G., & Turton, D. (2022). Towards an updated dialect atlas of British English. *Journal of Linguistic Geography, 10*(1), 46–66. https://doi.org/10.1017/jlg.2022.2

McInnes, L., Healy, J., & Melville, J. *UMAP: Uniform manifold approximation and projection for dimension reduction (arXiv:1802.03426). arXiv*. (2020). https://arxiv.org/abs/1802.03426.

Middleton, C., & Freestone, P. (2008). *The impact of culture-led regeneration on regional identity in north East England*.

Murphy, N. (2019). Reddit's 2019 year in review - upvoted. https://www.redditinc.com/blog/reddits-2019-year-in-review/#content.

Oguz, S., & Walton, A. (2022). Productivity in towns and travel to work areas, UK - Office for National Statistics. https://www.ons.gov.uk/economy/.

Paasi, A. (2003). Region and place: Regional identity in question. *Progress in Human Geography, 27*(4), 475–485. https://doi.org/10.1191/0309132503ph439pr

Pérez, J. M., Aleman, D. E., Kalinowski, S. N., & Gravano, A. *Exploiting User-Frequency Information for Mining Regionalisms from Social Media texts (arXiv:1907.04492). arXiv*. (2019). https://arxiv.org/abs/1907.04492.

Purves, R. S., Winter, S., & Kuhn, W. (2019). Places in information science. *Journal of the Association for Information Science and Technology, 70*(11), 1173–1182. https://doi.org/10.1002/asi.24194

Rae, A. (2009). From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census. *Computers, Environment and Urban Systems, 33*(3), 161–178. https://doi.org/10.1016/j.compenvurbsys.2009.01.007

Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., … Strogatz, S. H. (2010). Redrawing the map of Great Britain from a network of human interactions. *PLoS One, 5*(12), Article e14248. https://doi.org/10.1371/journal.pone.0014248

Reades, J., Calabrese, F., & Ratti, C. (2009). Eigenplaces: Analysing cities using the space – Time structure of the mobile phone network. *Environment and Planning. B, Planning & Design, 36*(5), 824–836. https://doi.org/10.1068/b34133t

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural*

language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 3980–3990). https://doi.org/10.18653/v1/D19-1410

Rey, S., Arribas-Bel, D., & Wolf, L. J. (2023). *Geographic data science with python*. CRC Press.

Russ, B. (2012). *Examining large-scale regional variation through online geotagged corpora*.

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access, 8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Sobolevsky, S., Szell, M., Campari, R., Couronné, T., Smoreda, Z., & Ratti, C. (2013). Delineating geographical regions with networks of human interactions in an extensive set of countries. *PLoS One, 8*(12), Article e81707. https://doi.org/10.1371/journal.pone.0081707

Statista. (2022). Most Popular Social Networks Worldwide as of January 2022, Ranked by Number of Monthly Active Users. In Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

Stuart-Smith, J. (2008). *Scottish English: Phonology. Varieties of English. 1* pp. 48–70).

Sui, D., & Goodchild, M. F. (2011). The convergence of GIS and social media: Challenges for GIScience. *International Journal of Geographical Information Science, 25*(11), 1737–1748. https://doi.org/10.1080/13658816.2011.604636

Thomas, M. J. (2019). Employment, education, and family: Revealing the motives behind internal migration in Great Britain. *Population, Space and Place, 25*(4), Article e2233. https://doi.org/10.1002/psp.2233

Titheridge, H., Achuthan, K., Mackett, R. L., & Solomon, J. (2009). Assessing the extent of transport social exclusion among the elderly. *Journal of Transport and Land Use, 2*(2). https://doi.org/10.5198/jtlu.v2i2.44

Trudgill, P. (2004). *Dialects* (2nd ed.). Routledge. https://doi.org/10.4324/9780203314609

Wagner, D., Zipf, A., & Westerholt, R. (2020). *Place in the GIScience community – An indicative and preliminary systematic literature review*. https://doi.org/10.5281/zenodo.3628855

Walden-Schreiner, C., Leung, Y.-F., & Tateosian, L. (2018). Digital footprints: Incorporating crowdsourced geographic information for protected area management. *Applied Geography, 90*, 44–54. https://doi.org/10.1016/j.apgeog.2017.11.004

Waters, T., & Evans, A. J. (2003). *Tools for web-based gis mapping of a "Fuzzy" vernacular geography. 10*.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., … Le, Q. V. (2022). *Finetuned language models are zero-shot learners (arXiv:2109.01652). arXiv*. https://doi.org/10.48550/arXiv.2109.01652

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., … Fedus, W. (2022). *Emergent abilities of large language models (arXiv:2206.07682). arXiv*. https://doi.org/10.48550/arXiv.2206.07682

Westerholt, R., Mocnik, F.-B., & Zipf, A. (2018). *Introduction to the PLATIAL'18 workshop on platial analysis*. https://doi.org/10.5281/zenodo.1475267

Yin, W., Kann, K., Yu, M., & Schütze, H. *Comparative study of CNN and RNN for natural language processing. arXiv:1702.01923 [Cs]*. (2017). https://arxiv.org/abs/1702.01923.

Zheng, X., Han, J., & Sun, A. (2018). A survey of location prediction on twitter. *IEEE Transactions on Knowledge and Data Engineering, 30*(9), 1652–1671. https://doi.org/10.1109/tkde.2018.2807840

Zheng, Y. (2015). Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology, 6*(3), 1–41. https://doi.org/10.1145/2743025