Mapping News Geography: A Computational Framework for Classifying Local Media Through Geographic Coverage Patterns

Simona Bisiani[1], Agnes Gulyas[2], and Bahareh Heravi[1]

[1] Institute for People-Centred AI, University of Surrey, Guildford, United Kingdom
[2] School of Creative Arts and Industries, Canterbury Christ Church University, Canterbury, United Kingdom

Abstract

This study introduces a novel computational framework for defining local news outlets through their geographic coverage patterns. This approach addresses the growing disconnect between legacy spatial markers (e.g., newsroom location or circulation cut-offs) and the geographic dimension of news coverage amidst media ownership consolidation and digital transformation. We develop a four-step pipeline consisting of data sampling, geoparsing, feature engineering, and clustering analysis. Our approach employs large language models for toponym disambiguation and develops eight spatial metrics across four dimensions: spatial extent, administrative reach, spatial heterogeneity, and distance decay. We test this pipeline on a sample of more than 465,000 articles from 360 UK digital local news outlets. Clustering analysis of more than 1.3 million locations reveals six distinct outlet types, ranging from hyperlocal and metropolitan to national outlets. This classification reflects different scales and structures of news provision in the UK's evolving media landscape. The method offers a scalable, open-source approach for mapping local news coverage and understanding the scope of local news providers as a function of their coverage, with implications for media geography and policy, ownership studies, and the computational humanities.

Keywords: local media, computational journalism, geographic coverage, geoparsing, media geography, outlet classification, content analysis, spatial analysis, clustering algorithms, news localisation

1 Introduction

Understanding the spatial focus of local media is crucial for studying the rise of "news deserts" [2; 7; 39; 43; 46; 54; 61], the influence of local journalism on political participation [21], community engagement [20], and the deterrence of corruption [13]. Prevailing methods to assign spatial references to news outlets include outlet self-declarations of their coverage area [4; 55], newsroom addresses [43], or audience dispersion [18]. While useful, these indicators provide only indirect evidence of which communities are actually represented in news content.

The salience of this shortcoming has heightened as a result of increasing media ownership consolidation and digital transformation, which have centralised news production [53] and increased both physical and editorial distance between journalists and local audiences [30; 31; 41]. As a result, established markers of *localness* often fail to represent the true geography of news provision [11]. While recent studies advocate for content-driven approaches to mapping local news in light

Simona Bisiani, Agnes Gulyas, and Bahareh Heravi. "Mapping News Geography: A Computational Framework for Classifying Local Media Through Geographic Coverage Patterns." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 845–866. https://doi.org/10.63744/PmuIcNvSnDuo.

of these conditions [18; 49], there remains no operational framework for empirically deriving a news outlet's spatial focus from its own reporting.

This work addresses this gap by introducing a computational approach to infer the editorial geography of news outlets using the locations mentioned in their published content. We ask:

RQ1: *How can a news outlet's spatial focus be computationally inferred from its content?*

RQ2: *What coverage patterns and spatial typologies emerge from such an analysis?*

Our contributions are summarised here:

1. We develop and validate a scalable, modular, and reproducible pipeline that infers the geographic focus of news outlets based on the locations mentioned in their reporting, moving beyond traditional proxies like self-declared coverage areas or newsroom addresses. This pipeline shifts from data collection, to geoparsing, then feature engineering, and finally clustering analysis.

2. We test this pipeline on a large stratified sample of articles from UK local news outlets and derive a six-cluster typology of news outlets. Our analysis uncovers six distinct spatial typologies (Major Regional Daily, County and Regional, Major City and Regional, Market Town and Rural, Local, Hyperlocal) that span five orders of magnitude in coverage area and are derived by a combination of spatial properties.

3. Our work makes original use of large language models (LLMs) to aid in several data annotation tasks. We favour open-source, locally-run LLMs as a means to make this pipeline reproducible and accessible. In doing so, we contribute to emerging research of responsible LLMs usage in media research.

2 Background

2.1 Spatiality of Local News and Media Provision

Each news organisation operates serving a geographically bounded audience [48]. This is particularly true for local media, which targets spatially organised communities [15]. Historically, researchers have located local media using markers such as circulation thresholds [52], self-declared coverage areas [4; 55], newsroom locations [43], or audience distributions [18]. However, structural transformations in this market have fundamentally altered the relationship between these markers and news coverage, meaning that the distance between the where an outlet appears to be, and the where it actually talks about in the news, has increased [11; 30]. The concept of "ghost newspapers" (outlets that exist nominally but rely largely on syndicated or wire content rather than original, community-centred reporting [1]) exemplifies this phenomenon.

Media ownership consolidation, propelled by declining print circulation, advertising revenue loss, and digital monetisation challenges [47], is often viewed as a driver of decrease in community-centric reporting [22]. Large conglomerates, such as Reach PLC—which controls roughly 13% of UK local media [57]—centralise newsroom functions into digital hubs servicing multiple regions [44], often resulting in newsroom closures and diminished local reporter presence [53]. This centralisation reduces the "visibility" and "sensibility" of local journalism—its attunement to the communities it ostensibly serves [41]. The geographic principle of "distance decay" predicts that increased physical distance weakens relationships, unless offset by targeted efforts [32]. Empirical studies confirm this: for example, Swedish municipalities without editorial offices experienced declines in original and community news coverage after newsroom centralisation [30].

Digital transformation complicates this landscape further. While digital platforms enable theoretically boundary-less audience reach, structural factors—such as newsroom location and brand

identity—continue to shape the spatial distribution of journalistic production [45]. Yet digital news consumption often privileges national or sensationalistic content aimed at mass audiences over nuanced, place-based reporting [8]. Particularly, intensified usage of content syndication across ownership networks [14] and following a news-value system driven by consumption metrics [34], are two practices that have further diluted geographical focus in local news.

These conditions have led to the emergence of novel forms of journalistic production. *Hyperlocals* are borne out of dissatisfaction with mainstream media and the desire to fill gaps in local news coverage [42]. These outlets tend to have an intentionally narrow geographic focus (e.g., a small city or even a neighbourhood), and are constituting an increasingly important part of the local news ecosystem, often times being the sole source of local news for a specific community [19].

These spatial dimensions of journalism are critical: media coverage influences democratic participation and community identity by shaping a "sense of place" through representation of events, people, and voices [23]. Communities with weaker media infrastructures suffer from increased corruption, lower political engagement, and poorer public resource management [13; 20; 21; 51]. Patterns of news coverage often reveal inequalities along urban-rural and socio-demographic lines [59].

Geographic coverage reflects the aggregate outcome of countless editorial decisions about newsworthiness, resource allocation, and community relevance—revealing the implicit geographic logic that guides news production [64]. Unlike audience metrics, which may be influenced by algorithmic distribution or marketing strategies, or editorial statements, which may not reflect actual practice, location mentions in news content provide a direct trace of where outlets choose to focus their journalistic attention. We therefore conceptualise "local" news as journalism that demonstrates sustained geographical specificity in its content—that is, news production characterised by consistent attention to particular places, communities, and administrative jurisdictions. This definition encompasses outlets ranging from hyperlocal community blogs to regional newspapers, unified by their commitment to place-based reporting rather than their organisational structure or ownership.

2.2 Computational Approaches to News Geography

Recent advances in computational methods have significantly expanded our ability to study local news geography [33; 39; 50; 62; 63]. Geoparsing techniques, including emerging methods leveraging large language models (LLMs) [3], enable precise extraction of geographic locations from news text. Despite this, many studies reduce spatial analysis to simple frequency counts or binary presence measures aggregated at administrative levels [39; 62]. This approach overlooks a fundamental geographic principle that "near things are more related than distant things" [58]: that space is relational rather than purely categorical.

Some research has begun applying spatial statistics, such as Moran's I and geographically weighted regression, to explore socio-demographic factors influencing newspaper survival and news coverage patterns [50; 63]. These analyses reveal spatial heterogeneity and autocorrelation, suggesting that journalistic geography is neither random nor uniform. Studies leveraging social media data have further delineated local, regional, and national outlets based on audience spatial clustering around newsroom locations [18], operationalising the concept of distance decay within audience geography.

2.3 Limitations and Research Gap

While these contributions mark important progress, existing research tends to focus on either audience geography or simplistic spatial proxies, often neglecting the geography of news content.

There is a lack of systematic, multidimensional frameworks combining content-based spatial features such as geographic extent, concentration, proximity, and alignment with political boundaries to characterise and classify subnational media outlets.

Furthermore, common computational studies typically overlook relational spatial patterns, reducing geography to counts within predefined administrative units, and fail to incorporate spatial clustering or distance decay effects in their analyses of content. As [49, p.21] notes, understanding "the news that each outlet produces—and whether that news covers the entire area the outlet claims to serve" is a crucial direction for future research.

To address this gap, our work develops a transferable and scalable computational method that integrates geoparsing with spatial clustering techniques to infer the spatial focus of news outlets directly from their published content. This approach moves beyond traditional proxies to provide an empirically grounded classification of local news provision based on actual reporting geography. By doing so, we aim to enable researchers to reassess local news geographies, examine the impact of media ownership on editorial scope, and revisit the relevance of geographic proximity as a core journalistic value in the digital age.

3 Methodology

The research design comprises four analytical stages: (1) data collection, (2) geoparsing, (3) feature engineering, and (4) coverage typologies identification. Following [16], we conceptualise "local" media as outlets targeting subnational, spatially identifiable audiences, thus we incorporate both large and small outlets in the study. We adopt as a case study the commercial and independent digital news sector in the United Kingdom, whose well-documented media landscape [4; 7; 57]—with its mix of regional newspapers, city news sites, and hyperlocal digital outlets—provides an ideal setting for examining spatial dimensions of local news distribution.

3.1 Data

Our analysis draws upon the UKTwitNewsCor dataset [6], the largest openly available article collection for UK local news, containing over 2.5 million articles published between 2020-2022 across 360 UK local outlets. This Twitter-derived corpus captures news actively disseminated through social media by commercial and independent digital news outlets. The dataset covers 47% of eligible UK local news domains active in 2022, distributed across 94% of Local Authority Districts (LADs). Although this dataset does not cover the entire population of outlets, it offers meaningful coverage across a variety of publishers and locations (see [6] for more detail). The dataset's spatial coverage proves particularly valuable for our study. This extensive geographic representation, combined with the three-year temporal span, minimises seasonal or event-driven distortions in coverage patterns while capturing enduring spatial reporting tendencies.

Due to computational constraints, we implemented a two-stage stratified cluster sampling. First, we stratified across geography and publishers to ensures balanced geographic and organisational representation [37]. Then, we sampled temporally from each of these subgroups to minimise seasonal patterns in news coverage. We randomly selected eight weekdays per stratum-year (two per quarter) using the constructed week method [38]. The resulting sample (465,105 articles, 18.35% of corpus) maintains complete spatial and publisher coverage while preserving within-day editorial relationships.

3.2 Geoparsing

To extract geographic locations from the articles, we applied a geoparsing pipeline developed in a companion study [5]. First, we extracted locations using SpaCy's transformer-based named entity

recognition model [24]. Location entities were then linked to real-world coordinate-candidates using two popular gazetteers for the UK: Ordnance Survey Open Names and OpenStreetMap. The identification of the correct set of coordinates for each location, a task known as toponym disambiguation [26], was found by presenting a large language model (LLM), gemma2-9b [56], with the location, its candidate options, and additional context to guide the selection. Context included the article and the website of the news outlet. Given LLMs' limitations in directly processing geographic coordinates in relation to text [40], we reformulated the disambiguation task around administrative boundaries. We mapped each coordinate set to its Local Authority District and asked the LLMs to classify which administrative district corresponds to a given toponym. This approach leverages the fact that LLMs can use their stronger contextual interpretation capabilities rather than performing precise spatial computations. This pipeline has been validated on a gold standard dataset consisting of 182 articles and 1,313 locations extracted from the same corpus, achieving high classification accuracy (F1 = 0.88, Accuracy = 0.82) and falling just behind state-of-the-art, which leverages fine-tuning[25]. Appendix A provides additional information regarding this procedure.

The geoparsing pipeline processed 465,105 articles, identifying location mentions in 365,714 articles (78.6%). From these articles, 1,657,425 location mentions were extracted, with 1,388,720 (83.8%) successfully geocoded to specific coordinates. The final analytical sample comprises 347,291 articles containing geocoded locations, representing 158,288 unique geographic locations across the UK. Pipeline performance varied by outlet, with the majority (75%) achieving geocoding success rates above 75% (Figure 1b). The relationship between total mentions and unique locations (Figure 1c) reveals that outlets with higher mention volumes tend to have proportionally fewer unique locations, reflecting both the finite nature of local geography and the tendency for outlets to repeatedly reference key locations within their coverage areas.

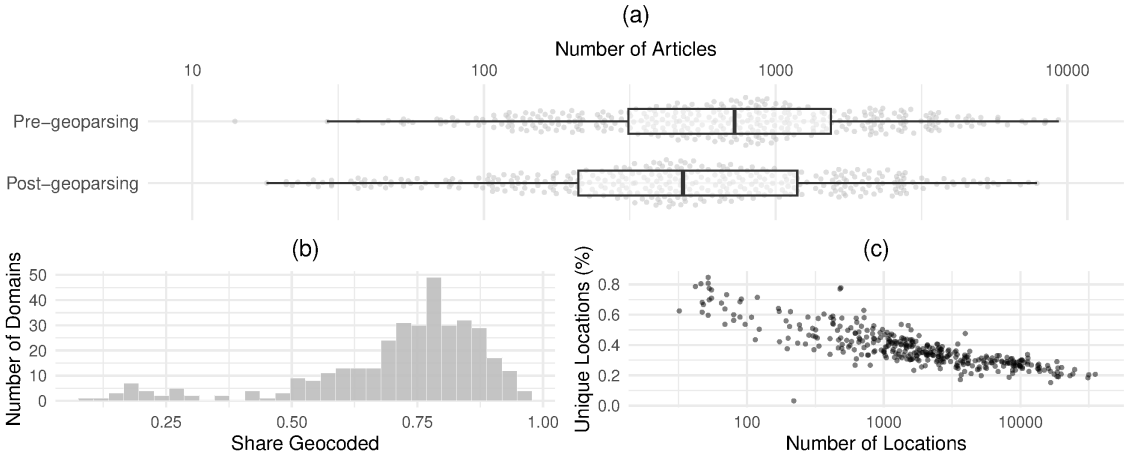

Figure 1: (a) Box plots showing the number of articles for each outlet before and after geoparsing. (b) Histogram of the share of successfully geocoded articles across domains. (c) Relationship between total number of locations and percentage of unique locations.

3.3 Feature Engineering

To systematically characterise local news coverage patterns, we develop a four-dimensional framework drawing on concepts from previous studies in news geography: distance decay [18; 30], spatial extent and presence across administrative boundaries [36; 62], and spatial heterogeneity [50; 63]. These dimensions operationalise different aspects of how news outlets construct their geographic coverage:

1. **Spatial Extent**: This dimension captures the geographic footprint and shape of an outlet's coverage area, and distinguish between outlets with compact versus dispersed coverage patterns. We operationalise it using two metrics calculated on the top 75% most frequently mentioned locations (i.e., after filtering out the bottom quartile by frequency), capturing geographic scope of the most relevant locations, accounting for occasional mentions of distant locations (e.g., Westminster or Downing Street): (1) the *convex hull area* enclosing these locations (measured in km^2, where larger values indicate broader territorial coverage), and (2) the *radial extent* from the most frequently mentioned location to the furthest location (measured in km, indicating the maximum reach of regular coverage).

2. **Administrative Reach**: This dimension reflects how news coverage aligns with and crosses administrative boundaries. We measure this through (3) the total number of Lower Super Output Areas (LSOAs) Districts in which locations by news outlets are situated (higher counts indicate wider coverage), and (4) the *Gini coefficient* of attention across different socio-demographic area types based on the LSOAC 2021/22 classifications.[1] This metric ranges from 0 (perfectly equal coverage across demographic areas) to 1 (highly concentrated coverage), revealing whether outlets provide balanced coverage across diverse communities or focus on particular demographic contexts. Appendix B provides additional details about LSOAs the LSOAC classification system.

3. **Spatial Heterogeneity**: This dimension assesses the diversity and clustering patterns of locations mentioned in news coverage. We use (5) *Shannon's entropy* to quantify the evenness of coverage across administrative districts — higher values indicate more balanced attention across districts, whilst lower values suggest concentration on particular districts. Complementing this, we calculate (6) *Moran's I* with a 50 km spatial weights matrix to detect spatial autocorrelation. Positive values indicate spatially clustered coverage (nearby locations mentioned together), negative values suggest dispersed patterns (nearby locations mentioned separately), and values near zero indicate random spatial distribution.

4. **Distance Decay**: This dimension examines how coverage intensity diminishes with distance from an outlet's primary location, operationalising the principle that proximity often increases relevance in local journalism. We identify each outlet's focal point through an automated information extraction pipeline using OpenAI's o4-mini model with web search capabilities (see Appendix C for methodology). Manual validation of a 20% random sample achieved 98% agreement with human annotations. Distances from this focal point to all mentioned locations are then computed using the Haversine formula, with metrics including (7) the *coefficient of variation* of distances (higher values indicate more scattered coverage patterns), and (8) the *percentage of locations within ten kilometres* of the primary location (higher percentages indicate more localised coverage).

Technical details for these metrics are provided in Appendix D.

3.4 Clustering Analysis for Spatial Typology Identification

To identify empirical typologies of news outlet spatial coverage, we implemented an unsupervised clustering approach designed to let structural patterns emerge from the data. Because the geometry of spatial strategies in news coverage is unknown a priori, we systematically compared five different clustering algorithms: k-means (spherical/equally-sized clusters) [28], hierarchical clustering (Ward and complete linkages), Divisive Analysis (DIANA) for capturing hierarchical or nested

[1] Retrieved from: `https://data.geods.ac.uk/dataset/lsoac`.

structures, and HDBSCAN for density-based clusters and explicit outlier detection [10]. This ensemble allows us to test robustness across a range of plausible spatial organisational models (see Appendix E for technical details).

3.4.1 Dimensionality and Feature Selection

High correlations between spatial features and large number of variables introduce well-known clustering challenges [60]. To address this, we test different configurations of our feature dataset: (1) all engineered features as a baseline; (2) a reduced set of minimally correlated features based on pairwise analysis ($r > 0.8$) [35]; and (3) Principal Component Analysis (PCA) to create orthogonal components explaining 80%, 90%, and 95% of total variance [29]. PCA revealed that coverage patterns reflect multiple independent dimensions, from local intensity to geographic scope and administrative complexity, rather than a simple local-to-national continuum (see Figure 2).

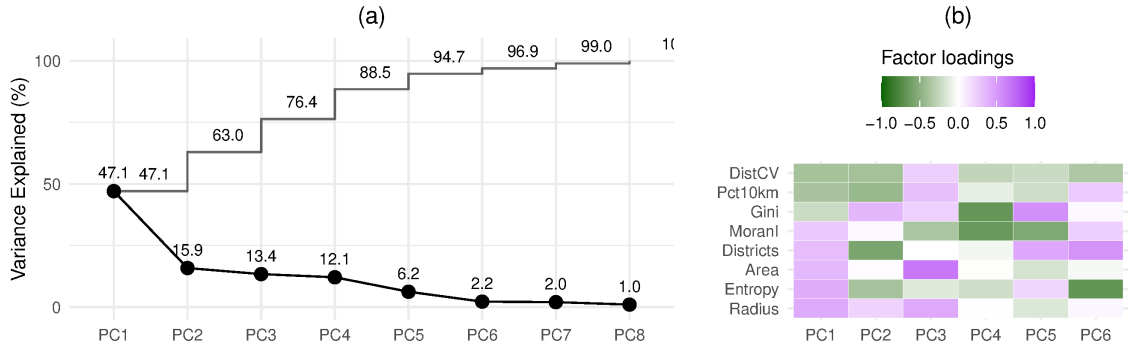

Figure 2: Principal Component Analysis of news coverage geography. (a) Scree plot showing variance explained by each component (dots) and cumulatively (steps); (b) Factor loadings for principal components.

3.4.2 Experimental Design and Validation

To ensure that typology discovery was robust to both algorithm and feature selection, we systematically evaluated 25 combinations (5 algorithms $\times$ 5 feature sets). All features were z-score standardised prior to analysis. For methods requiring a set number of clusters, we tested values from $k = 3$ to 6, prioritising solutions with both interpretability and good silhouette widths; silhouette was our primary internal quality metric, while the elbow method and Adjusted Rand Index (ARI) assessed solution consistency [27].

Cluster characterisation was performed by examining the mean values of the original spatial features within each cluster, preserving interpretability in terms of spatial scale, concentration, proximity, and heterogeneity. Further diagnostic details and stability analysis are documented in Appendix E.

4 Results

4.1 Spatial Properties of News Coverage

News outlets in our sample display substantial diversity in the spatial characteristics of their coverage (Table 1).

Some outlets exhibit tightly localised coverage areas (as little as 4.5 km^2), while others reach over 123,000 km^2; there is similar variation in the number and distribution of administrative areas referenced. Coverage is rarely uniform: most outlets concentrate attention in a limited set of location types, with some demographic communities (e.g., Legacy Communities, Low-Skilled

Variable (Units)	Mean	Median	SD	Min	Max
Spatial Extent					
Convex hull area (km²)	4,556.68	1,988.77	10,731.07	4.46	123,047.90
Radius of 75% mentions (km)	37.54	31.81	29.65	1.49	222.62
Administrative Reach					
Number of districts covered (count)	346.28	235	331.07	6	1,881
Gini coefficient (0–1)	0.53	0.54	0.11	0.06	0.84
Spatial Heterogeneity					
Shannon entropy	3.75	3.81	0.73	0.96	5.94
Moran's I (−1 to +1)	0.04	0.03	0.04	−0.20	0.17
Distance Decay					
Median distance to mentions (km)	17.52	14.77	14.63	0.64	103.52
Distance coefficient of variation (unitless)	2.00	1.89	0.72	0.62	5.79
Share of mentions within 10 km (%)	41.0	40.0	23.0	0.0	96.0

Notes: Statistics computed across 358 media outlets.

Table 1: Descriptive Statistics of Features

Migrant and Student Communities) disproportionately under-represented (Figure 4a). Moderate to high spatial inequality (Gini ≈ 0.5) is common, and spatial clustering (Moran's I ≈ 0) is typically low, indicating that mention patterns generally lack strong clustering or dispersion. Shannon entropy values further show that coverage is usually moderate in spatial diversity. Distance decay patterns also vary widely: some outlets achieve 75% coverage within 25 km of their focal point, while others require more than 100 km, reflecting both hyperlocal and regional scales of operation (Figure 4b). The proportion of highly local coverage (within 10 km) ranges from none to nearly all, with the distance coefficient of variation spanning 0.62 to nearly 6.

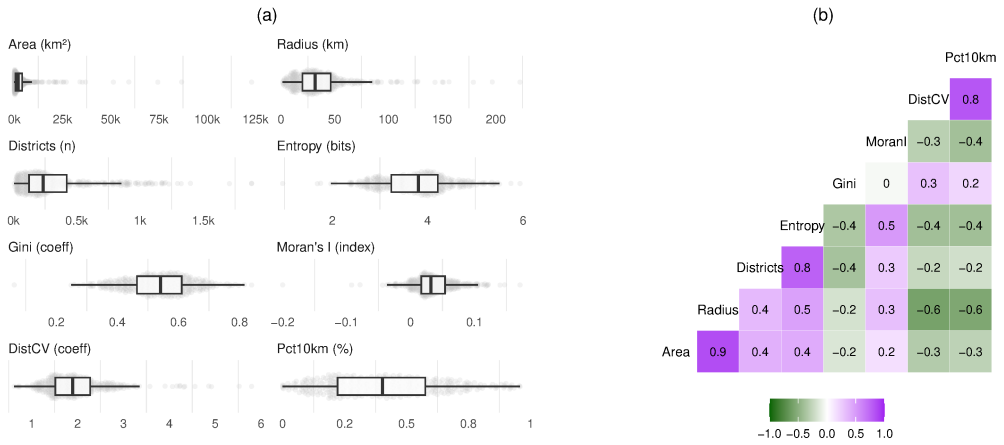

Figure 3: Distribution and correlation analysis of landscape metrics. (a) Distribution of eight landscape metrics across study areas, showing individual data points (grey dots) overlaid with boxplots indicating median, quartiles, and range. Values are displayed with abbreviated units for readability (k = thousands). (b) Pearson correlation matrix between landscape metrics, with correlation coefficients displayed and color-coded from dark green (strong negative correlation, -1.0) to purple (strong positive correlation, +1.0).

Correlation analysis reveals several key structural relationships (Figure 3b). The strongest positive correlation exists between area and radius (r = 0.9), confirming that larger coverage areas correspond to greater radial reach from the primary location. Strong positive correlations (r = 0.8) between Shannon's entropy and the number of districts shows that outlets mentioning more

districts also tend to distribute their coverage more evenly across them. Conversely, strong negative correlations are observed between radius and the percentage of mentions within 10 km (r = -0.6), suggesting that outlets with greater radial extent focus less on their immediate vicinity. Finally, we observe a moderate positive correlation between Entropy and Moran's I (r = 0.5), that outlets with more evenly distributed coverage across districts tend to exhibit regionally clustered patterns of attention, reflecting geographically balanced but locally coherent reporting.

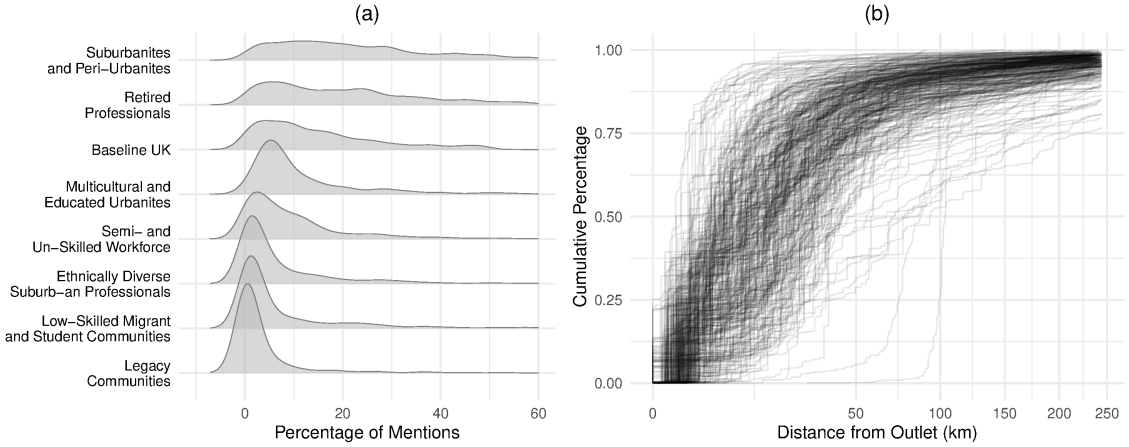

Figure 4: (a) Demographic representation across LSOAs. (b) Cumulative percentage distributions of distances of location mentions from primary location of news outlet.

4.2 Clustering

We systematically evaluated clustering solutions, finding that k-means with a highly reduced, decorrelated set of features yields the most interpretable, balanced spatial typologies.[2] The optimal solution partitions outlets into six clusters, hierarchically arranged by area, administrative reach, and spatial diversity (see Table 2). This empirically supports the existence of multi-scalar news systems ranging from hyperlocal (typ. < 40 km^2) through major city and county outlets, up to major regional dailies ($> 70,000$ km^2).

Cluster Type	N	Area (km²)	Districts	Gini	Entropy	Moran's I	DistCV	Pct10km (%)
Hyperlocal	15	33	75	0.70	2.64	-0.001	4.34	89.5
Local	152	1,216	254	0.54	3.55	0.025	2.31	53.9
Market Town and Rural	38	2,715	88	0.53	3.13	-0.015	1.47	29.1
County and Regional	90	5,769	234	0.56	3.87	0.070	1.56	24.2
Major City and Regional	58	7,389	936	0.45	4.68	0.051	1.66	30.3
Major Regional Daily	5	79,014	1,119	0.41	5.38	0.051	1.17	23.0

Notes: Cluster means shown. DistCV = Coefficient of Variation of distances; Pct10km = Percentage of location mentions within 10 km of outlet.

Table 2: Media Outlet Spatial Reach Typologies

1. *Hyperlocal* outlets represent the smallest organisational scale with minimal spatial footprints (mean coverage 33 km²) and the highest concentration of mentions near their focal location (89.5% within 10 km). Examples include *brixtonblog.com* (South London), *ec1echo.co.uk*

[2] For transparency, we have built an interactive dashboard that allows users to explore the results from this set of experiments: `https://simonabisiani-clustering-analysis-dashboard.share.connect.posit.cloud/`.

(Central London postcode), and *greatergovanhill.com* (Glasgow neighbourhood). These outlets exhibit extreme distance variability (DistCV = 4.34), though this metric is sensitive to the small mean distances characteristic of hyperlocal coverage. Spatial autocorrelation is negligible (Moran's I = -0.001), indicating mentions neither cluster nor disperse systematically beyond the compressed geographic scale.

2. *Local* outlets constitute the modal category (N = 152), characterised by intermediate coverage area (mean 1,216 km²) and moderate administrative penetration (254 districts). Examples include *cambridge-news.co.uk*, *nottinghampost.com*, and *oxfordmail.co.uk*. These outlets demonstrate notable distance variability (DistCV = 2.31) with substantial local concentration (53.9% of mentions within 10 km), indicating coverage focused on a primary urban centre with extensions to surrounding areas. Weak positive spatial autocorrelation (Moran's I = 0.025) suggests modest spatial clustering of coverage attention.

3. *Market Town and Rural* outlets serve smaller settlements and dispersed rural hinterlands (N = 38), with mean coverage area of 2,715 km² but the lowest administrative penetration (88 districts). Examples include *brecon-radnor.co.uk* (rural Welsh counties), *cambrian-news.co.uk*, and *newry.ie* (border town). These outlets show moderate entropy (3.13) and 29.1% of mentions within 10 km of their focal location. Slight negative spatial autocorrelation (Moran's I = -0.015) indicates coverage dispersed across geographic space rather than concentrated in contiguous zones, consistent with outlets serving scattered rural communities or isolated market towns.

4. *County and Regional* outlets (N = 90) occupy an intermediate position with mean coverage of 5,769 km² across 234 districts. Examples include *southwalesargus.com*, *ardrossanherald.com*, *thewestmorlandgazette.co.uk*, and *bordercountiesadvertizer.co.uk*. These outlets exhibit notably the highest spatial autocorrelation in the typology (Moran's I = 0.070), indicating location mentions cluster spatially where neighbouring locations receive similar coverage attention. Combined with higher entropy (3.87) and moderate Gini coefficient (0.56), this suggests coverage patterns extending across multiple administrative units with spatially coherent zones of attention rather than the dispersed patterns observed in Market Town outlets.

5. *Major City and Regional* outlets combine substantial coverage extent (mean 7,389 km²) with exceptionally high administrative complexity (936 districts), creating the highest districts-to-area ratio in the typology. Examples include *belfasttelegraph.co.uk*, *birminghammail.co.uk*, and *manchestereveningnews.co.uk*. These outlets show the lowest Gini coefficient (0.45) alongside high entropy (4.68), indicating relatively even coverage distribution across diverse administrative units. Positive spatial autocorrelation (Moran's I = 0.051) suggests coverage attention clusters in spatially contiguous patterns despite the administrative complexity.

6. *Major Regional Daily* outlets represent the apex of geographic reach, with mean coverage of 79,014 km² spanning 1,119 districts. Examples include *dailyrecord.co.uk*, *scotsman.com*, and *blackpoolgazette.co.uk*. This small group (N = 5) exhibits the highest spatial diversity (entropy = 5.38) and lowest inequality (Gini = 0.41), combined with the lowest distance variability (DistCV = 1.17). Positive spatial autocorrelation (0.051) persists despite vast geographic extent, indicating coverage attention maintains spatial structure across extensive territories.

To validate our results, we projected our clustered data on the PCA components generated on the full feature dataset. The PCA projection confirms our clustering validity, with clusters occupying distinct regions of the reduced dimensional space (Figure 5). The biplot reveals that spatial

scale (Area, Districts, Entropy) forms the primary dimension of variation, whilst circulation concentration metrics (DistCV, Pct10km) and spatial clustering (Moran's I) contribute orthogonal dimensions. This dimensional structure validates our typology's theoretical foundation, demonstrating that media outlets' spatial reach patterns reflect fundamental trade-offs between geographic scale, administrative complexity, and audience concentration.

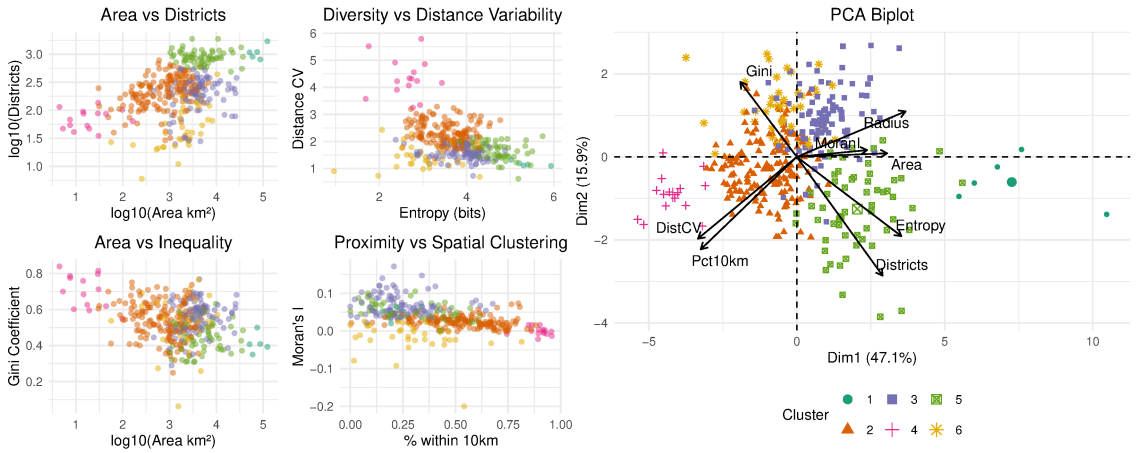

Figure 5: Cluster analysis validation across key variable relationships. Left panels show scatter plots of highly correlated and discriminative variable pairs, with clusters colour-coded to demonstrate separation. Right panel shows PCA biplot validation with variable loadings, confirming that clusters separate cleanly in the reduced dimensional space.

5 Discussion and Conclusion

This study shows that computational analysis of location mentions in news texts can provide a valid approach to classify subnational media into meaningful news outlet typologies, thus offering a novel approach to define local news [17]. We find six empirically distinct scales and types of UK news outlets, confirming that subnational media organisation is inherently multi-scalar and reflective of differences in geography and socio-demographic characteristics across rural-urban spectrums. Computational approaches to measure spatial heterogeneity also reveal organisation forms (e.g. urban clustering) that would be invisible using simple spatial thresholds. Our results do not validate whether, or how, the reported coverage overlaps with stated areas of coverage. Yet distinctive geographic scopes across outlets show that in order to understand how different communities are served by their local news providers, it is important we ask how that coverage is spatially organised.

We identify three primary contributions of this work. First, it advances the field of media geography by operationalising local news provision through extraction and analysis of spatial references in news coverage. Second, it introduces scalable natural language processing techniques for geographic inference in journalism studies, enabling systematic empirical evaluation of editorial strategies. Third, it provides practical tools for identifying and responding to the effects of media consolidation, enabling regulators to assess whether ownership changes genuinely threaten local coverage diversity. In doing so, this work contributes to ongoing efforts to render the structure and functioning of local media systems more visible and intelligible. As [41, p.12] states: "Governments have a powerful role to play in securing the sustainability of local news and, as such, have a responsibility to hold to account those news outlets that purport to be *local*".

This work contributes new evidence of spatial organisational structures in the UK's commercial and independent subnational media market. In doing so, we offer a new angle from which to

investigate whether outlets or publishers have diluted local news coverage in response to digital transformation or ownership consolidation. This approach is particularly valuable for identifying ”ghost newspapers” that maintain local branding while providing minimal community-specific coverage [1], and for detecting the spatial reorganisation of news provision under ownership consolidation.

This research opens several pathways: integrating with audience data to understand the relationship between geographic focus and audience demand, enabling content-based analyses of local news provision based on a systematic, multi-dimensional framework. Future research could build upon this framework by incorporating editorial weighting to distinguish between central coverage and passing mentions, developing methods to identify coverage gaps within outlets' stated territories, and systematically mapping spillover patterns where outlets cover unexpected geographic areas—advancing towards a more nuanced understanding of local news provision.

It is perhaps useful to also indicate what this study does not do. Firstly, it does not incorporate temporal aspects into the analysis (despite the longitudinal span covering three years), yet this would constitute a valuable future work extension to understand shifts in coverage attention distributions. Secondly, this study does not contribute a transferable set of parameters or heuristics for deriving outlet typologies: different datasets, reflective of their markets, geographic settings, national contexts will likely produce uniquely relevant clustering profiles. Here, we offer a modular (features can be tested and included/excluded as one sees fit), scalable (this analysis can be done at national scale as demonstrated here, but the focus could be on smaller as well as larger datasets and contexts), and accessible (all software used, beyond GPT4o-mini in the outlet primary location identification step, is free and open-source). Beyond media, our framework can be applied to other domains, from cultural and heritage institutions to public services, to analyse how organisations construct and manage geographic scope in their communication.

5.1 Limitations

Our study is limited by its UK focus, which may affect generalisability. Future research should test the framework in other national, regional, or linguistic contexts. We acknowledge that our geoparsing pipeline, while achieving comparably satisfactory results [25], is not error-free and may under-represent certain place types or ambiguous references: further technical refinement is warranted. The UKTwitNewsCor dataset, while extensive, is sourced from Twitter, which may introduce biases in outlet representation. Specifically, commercial outlets with active social media strategies (e.g., Reach PLC) may dominate, while smaller hyperlocals or print-focused publishers could be underrepresented [6]; articles shared on Twitter may prioritise ”clickworthy” topics (e.g., crime, events) over routine local governance coverage, potentially inflating mentions of certain locations (e.g., city centers) and underrepresenting others (e.g., rural districts). Future work could mitigate this by attempting a different approach to data collection, incorporating direct RSS feeds or news coverage from outlets' archives, where available.

These limitations underscore the need to intersect spatial analysis with content quality metrics (e.g., [9]'s proximity discourse analysis) and audience data to evaluate whether geographic coverage aligns with democratic needs. For instance, an outlet with broad spatial reach but superficial reporting may offer less value than a hyperlocal with deep, critical coverage. In doing so, researchers should be careful in associating proximity with positive representation. Extreme geographic proximity has been shown to paradoxically reduce journalistic quality, as editors prioritise place-making narratives over critical coverage to maintain community relations [12]. This creates situations where the most geographically proximate outlets may provide inferior democratic accountability compared to more distant ”local” providers. As such, while this work has contributed a deeper understanding of the *where* of local news, at the outlet-level, this dimension will require intersecting with *what*, *how*, *why* and *when* to generate meaningful understanding of inequalities

in news coverage.

Data Availability Statement

The analyses in this study can be reproduced through the following, permanently-stored, open-access code repository: https://doi.org/10.7910/DVN/T7SE5F. We have also developed a platform for users to interact with the data and explore how different outlets' coverage spread across a map: https://simonabisiani-local-news-map-explorer.share.connect.posit.cloud/.

Acknowledgements

The classification of census tracts data used in this research have been provided by the Geographic Data Service, a Smart Data Research UK Investment, under project ID GeoDS 3139, ES/Z504464/1.

References

[1] Abernathy, Penelope Muse. "News deserts and ghost newspapers: Will local news survive?" UNC Hussman School of Journalism and Media. 2020. URL: https://www.usnewsdeserts.com/reports/news-deserts-and-ghost-newspapers-will-local-news-survive/.

[2] Abernathy, Penelope Muse. "The rise of a new media baron and the emerging threat of news deserts". Center for Innovation and Sustainability in Local Media, University of North Carolina Chapel Hill. Chapel Hill, NC, 2016.

[3] Aubin Le Quéré, Marianne, Wang, Siyan, Fatima, Tazbia, and Krisch, Michael. "Towards Identifying Local Content Deserts with Open-Source Large Language Models". In: *Computation + Journalism Symposium 2024. Northeastern University.* 2024.

[4] Bisiani, S. and Heravi, B. "Uncovering the State of Local News Databases in the UK: Limitations and Impacts on Research". In: *Journalism and Media 4*, no. 4 (2023), pp. 1211–1231. DOI: 10.3390/journalmedia4040077.

[5] Bisiani, Simona, Gulyas, Agnes, and Heravi, Bahareh. "Towards Efficient and Accessible Geoparsing of UK Local Media: A Benchmark Dataset and LLM-based Approach". In: *Computational Humanities Research 1* (2025). DOI: 10.1017/chr.2025.10012.

[6] Bisiani, Simona, Gulyas, Agnes, Wihbey, John, and Heravi, Bahareh. "UKTwitNewsCor: A Dataset of Online Local News Articles for the Study of Local News Provision". In: *Proceedings of the International AAAI Conference on Web and Social Media 19*, no. 1 (2025), pp. 2371–2384. DOI: 10.1609/icwsm.v19i1.35940.

[7] Bisiani, Simona and Mitchell, Joe. "UK Local News Report April 2024". Public Interest News Foundation. London, 2024. URL: https://www.publicinterestnews.org.uk/_files/ugd/cde0e9_31f2ee78fff64c3e8616e7eafdf28f99.pdf (visited on 07/09/2024).

[8] Boczkowski, Pablo J. and Mitchelstein, Eugenia. *The News Gap: When the Information Preferences of the Media and the Public Diverge*. MIT Press, 2013. 317 pp. ISBN: 978-0-262-01983-5.

[9] Bros, Victor and Gatica-Perez, Daniel. "Decoding community proximity discourse: a mixed-methods comparative analysis of online local and national newspapers in Romandy, Switzerland". OSF. 2025. DOI: 10.31219/osf.io/fah5g_v3.

[10] Dash, Manoranjan and Liu, Huan. “Feature selection for clustering”. In: *Knowledge Discovery and Data Mining*, ed. by Takao Terano, Huan Liu, and Arbee L.P. Chen. Springer Verlag, 2000, pp. 110–121. DOI: 10.1007/3-540-45571-x_13.

[11] Franklin, Bob and Cushion, Stephen. “Downgrading the ‘local’in local newspapers’ reporting of the 2005 UK general election”. In: *Local Journalism and Local Media*. Routledge, 2006, pp. 256–269.

[12] Freeman, Julie. “Differentiating distance in local and hyperlocal news”. In: *Journalism 21*, no. 4 (2020). SAGE Publications, pp. 524–540. DOI: 10.1177/1464884919886440.

[13] Gao, Pengjie, Lee, Chang, and Murphy, Dermot. “Financing dies in darkness? The impact of newspaper closures on public finance”. In: *Journal of Financial Economics 135*, no. 2 (2020), pp. 445–467. DOI: 10.1016/j.jfineco.2019.06.003.

[14] Garz, Marcel and Ots, Mart. “Media consolidation and news content quality”. In: *Journal of Communication* (2025), jqae053. DOI: 10.1093/joc/jqae053.

[15] Gasher, Mike and Klein, Reisa. “Mapping the Geography of Online News”. In: *Canadian Journal of Communication 33*, no. 2 (2008), pp. 193–212. DOI: 10.22230/cjc.2008v33n2a1974.

[16] Gulyas, A. “Local news deserts”. In: *Reappraising Local and Community News in the UK: Media, Practice, and Policy*. Routledge, 2021, pp. 16–28. DOI: 10.4324/9781003173144-2.

[17] Gulyas, Agnes and Baines, David. “Introduction: Demarcating the field of local media and journalism”. In: *The Routledge Companion to Local Media and Journalism*. Routledge, 2020. ISBN: 978-1-351-23994-3.

[18] Hagar, Nick, Bandy, Jack, Trielli, Daniel, Wang, Yixue, and Diakopoulos, Nicholas. “Defining Local News: A Computational Approach”. Computation + Journalism Symposium. 2020.

[19] Harcup, Tony. “Asking the Readers: Audience research into alternative journalism”. In: *Journalism Practice 10*, no. 6 (2016). Routledge, pp. 680–696. DOI: 10.1080/17512786.2015.1054416.

[20] Hayes, D. and Lawless, J.L. “As local news goes, so goes citizen engagement: Media, knowledge, and participation in us house elections”. In: *Journal of Politics 77*, no. 2 (2015), pp. 447–462. DOI: 10.1086/679749.

[21] Hayes, Danny and Lawless, Jennifer L. “The Decline of Local News and Its Effects: New Evidence from Longitudinal Data”. In: *The Journal of Politics 80*, no. 1 (2018), pp. 332–336. DOI: 10.1086/694105.

[22] Hendrickx, Jonathan and Ranaivoson, Heritiana. “Why and how higher media concentration equals lower news diversity – The Mediahuis case”. In: *Journalism 22*, no. 11 (2021), pp. 2800–2815. DOI: 10.1177/1464884919894138.

[23] Hess, Kristy and Waller, Lisa. “River Flows and Profit Flows: The powerful logic driving local news”. In: *Journalism Studies 17*, no. 3 (2016), pp. 263–276. DOI: 10.1080/1461670X.2014.981099.

[24] Honnibal, Matthew, Montani, Ines, Van Landeghem, Sofie, and Boyd, Adriane. “spaCy: Industrial-strength Natural Language Processing in Python”. In: (2020). DOI: 10.5281/zenodo.1212303.

[25] Hu, Xuke, Kersten, Jens, Klan, Friederike, and Farzana, Sheikh Mastura. “Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge”. In: *International Journal of Geographical Information Science* (2024), pp. 1–28. DOI: 10.1080/13658816.2024.2405182.

[26] Hu, Xuke, Sun, Yeran, Kersten, Jens, Zhou, Zhiyong, Klan, Friederike, and Fan, Hongchao. “How can voting mechanisms improve the robustness and generalizability of toponym disambiguation?” In: *International Journal of Applied Earth Observation and Geoinformation 117* (2023), p. 103191. DOI: 10.1016/j.jag.2023.103191.

[27] Hubert, Lawrence and Arabie, Phipps. “Comparing partitions”. In: *Journal of Classification 2*, no. 1 (1985), pp. 193–218. DOI: 10.1007/BF01908075.

[28] Jain, Anil K. “Data Clustering: 50 Years Beyond K-means”. In: *Machine Learning and Knowledge Discovery in Databases*, ed. by Walter Daelemans, Bart Goethals, and Katharina Morik. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 3–4. ISBN: 978-3-540-87479-9.

[29] Jolliffe, Ian T and Cadima, Jorge. “Principal component analysis: a review and recent developments”. In: *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences 374*, no. 2065 (2016), p. 20150202.

[30] Karlsson, Michael and Rowe, Erika Hellekant. “Local Journalism when the Journalists Leave Town”. In: *Nordicom Review 40* (s2 2019), pp. 15–29. DOI: 10.2478/nor-2019-0025.

[31] Kekezi, Orsa and Mellander, Charlotta. “Geography and consumption of local media”. In: *Journal of Media Economics 31*, no. 3 (2018). Routledge, pp. 96–116. DOI: 10.1080/08997764.2020.1871250.

[32] Kent, Josh, Leitner, Michael, and Curtis, Andrew. “Evaluating the usefulness of functional distance measures when calibrating journey-to-crime distance decay functions”. In: *Computers, Environment and Urban Systems 30*, no. 2 (2006), pp. 181–200. DOI: 10.1016/j.compenvurbsys.2004.10.002.

[33] Khanom, Asma, Kiesow, Damon, Zdun, Matt, and Shyu, Chi-Ren. “The News Crawler: A Big Data Approach to Local Information Ecosystems”. In: *Media and Communication 11*, no. 3 (2023). DOI: 10.17645/mac.v11i3.6789.

[34] Kormelink, Tim Groot and Meijer, Irene Costera. “What clicks actually mean: Exploring digital news user practices”. In: *Journalism 19*, no. 5 (2018), pp. 668–683. DOI: 10.1177/1464884916688290.

[35] Kyriazos, Theodoros and Poga, Mary. “Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions”. In: *Open Journal of Statistics 13*, no. 3 (2023), pp. 404–424. DOI: 10.4236/ojs.2023.133020.

[36] Lindgren, April. “News, Geography and Disadvantage: Mapping Newspaper Coverage of High-needs Neighbourhoods in Toronto, Canada”. In: *Canadian Journal of Urban Research 18*, no. 1 (2009). Institute of Urban Studies, University of Winnipeg, pp. 74–97. URL: https://www.jstor.org/stable/26193230.

[37] Long, Marilee, Slater, Michael D., Boiarsky, Greg, Stapel, Linda, and Keefe, Thomas. “Obtaining Nationally Representative Samples of Local News Media Outlets”. In: *Mass Communication and Society 8*, no. 4 (2005), pp. 299–322. DOI: 10.1207/s15327825mcs0804_2.

[38] Luke, Douglas A., Caburnay, Charlene A., and Cohen, Elisia L. “How Much Is Enough? New Recommendations for Using Constructed Week Sampling in Newspaper Content Analysis of Health Stories”. In: *Communication Methods and Measures 5*, no. 1 (2011), pp. 76–91. DOI: 10.1080/19312458.2010.547823.

[39] Madrid-Morales, Dani, Rodríguez-Amat, Joan Ramon, and Lindner, Peggy. “A Computational Mapping of Online News Deserts on African News Websites”. In: *Media and Communication 11*, no. 3 (2023). DOI: 10.17645/mac.v11i3.6857.

[40] Mai, Gengchen et al. “On the Opportunities and Challenges of Foundation Models for GeoAI (Vision Paper)”. In: *ACM Transactions on Spatial Algorithms and Systems 10*, no. 2 (2024), pp. 1–46. DOI: 10.1145/3653070.

[41] McAdam, Alison and Hess, Kristy. “Re-asserting the Value of Local “News Presence” for Small-town News Outlets in a Digital era”. In: *Journalism Practice 0*, no. 0 (2024), pp. 1–16. DOI: 10.1080/17512786.2024.2433659.

[42] Metzgar, Emily T., Kurpius, David D., and Rowley, Karen M. “Defining hyperlocal media: Proposing a framework for discussion”. In: *New Media & Society 13*, no. 5 (2011), pp. 772–787. DOI: 10.1177/1461444810385095.

[43] Metzger, Zach. “The State of Local News - 2024 Report”. Local News Initiative, Medill School of Media, Journalism, and Integrated Marketing Communications, Northwestern University. Evanston, IL, 2024. URL: https://localnewsinitiative.northwestern.edu/projects/state-of-local-news/2024/report/ (visited on 10/28/2024).

[44] Moore, Martin and Ramsay, Gordon Neil. “Local News in National Elections: An “Audit” Approach to Assessing Local News Performance During a National Election Campaign”. In: *Digital Journalism 0*, no. 0 (2024), pp. 1–20. DOI: 10.1080/21670811.2024.2333827.

[45] Mosco, Vincent. *The Political Economy of Communication*. Sage, 2009, pp. 1–280. URL: https://www.torrossa.com/en/resources/an/4913650 (visited on 07/07/2025).

[46] Negreira-Rey, María-Cruz, Vázquez-Herrero, Jorge, and López-García, Xosé. “No People, No News: News Deserts and Areas at Risk in Spain”. In: *Media and Communication 11*, no. 3 (2023), pp. 293–303. DOI: 10.17645/mac.v11i3.6727.

[47] Noam, Eli M., Collaboration, The International Media Concentration, Noam, Eli M., and Collaboration, The International Media Concentration. *Who Owns the World’s Media?: Media Concentration and Ownership around the World*. Oxford, New York: Oxford University Press, 2016. 1440 pp. ISBN: 978-0-19-998723-8.

[48] Picard, R.G. *Media Economics: Concepts and Issues*. Commtext Series. SAGE Publications, 1989. ISBN: 978-0-8039-3502-0.

[49] PINF. “Deserts, Oases and Drylands”. 2023. URL: https://www.publicinterestnews.org.uk/map (visited on 10/06/2023).

[50] Qin, Abby Youran. “Where Is Local News Dying Off?: Mechanisms Behind the Formation of Local News Deserts in the United States”. In: *Journalism & Mass Communication Quarterly* (2024), p. 10776990241277885. DOI: 10.1177/10776990241277885.

[51] Ramos, G., Torre, L., and Jerónimo, P. “No Media, No Voters? The Relationship between News Deserts and Voting Abstention”. In: *Social Sciences 12*, no. 6 (2023). DOI: 10.3390/socsci12060345.

[52] Ramsay, Gordon and Moore, Martin. “Monopolising local news: Is there an emerging local democratic deficit in the UK due to the decline of local newspapers?” King’s College London. 2016. DOI: 10.18742/pub01-026.

[53] Sharman, David. “Reach plc to close all bar 15 of its newspaper offices - Journalism News from HoldtheFrontPage”. HoldtheFrontPage. 2021. URL: https://www.holdthefrontpage.co.uk/2021/news/publisher-to-close-all-bar-15-offices-leaving-dailies-without-base-on-patch/ (visited on 12/02/2024).

[54] Silva, C.E.L. da and Pimenta, A. “Local news deserts in Brazil: Historical and contemporary perspectives”. In: *The Routledge Companion to Local Media and Journalism*. Routledge, 2020, pp. 44–53.

[55] Stonbely, Sarah. “What Makes for Robust Local News Provision? Structural Correlates of Local News Coverage for an Entire U.S. State, and Mapping Local News Using a New Method”. In: *Journalism and Media 4*, no. 2 (2023), pp. 485–505. DOI: 10.3390/journalmedia4020031.

[56] Team, Gemma et al. “Gemma 2: Improving Open Language Models at a Practical Size”. arxiv. 2024. DOI: 10.48550/arXiv.2408.00118.

[57] The Media Reform Coalition. “Who Owns the UK Media?” 2025. URL: https://www.mediareform.org.uk/media-ownership/media-ownership-2025 (visited on 05/28/2025).

[58] Tobler, W. R. “A Computer Movie Simulating Urban Growth in the Detroit Region”. In: *Economic Geography 46* (sup1 1970). Routledge, pp. 234–240. DOI: 10.2307/143141.

[59] Usher, N. “Putting “Place” in the Center of Journalism Research: A Way Forward to Understand Challenges to Trust and Knowledge in News”. In: *Journalism and Communication Monographs 21*, no. 2 (2019), pp. 84–146. DOI: 10.1177/1522637919848362.

[60] Verleysen, Michel and François, Damien. “The Curse of Dimensionality in Data Mining and Time Series Prediction”. In: *Computational Intelligence and Bioinspired Systems*, ed. by Joan Cabestany, Alberto Prieto, and Francisco Sandoval. Berlin, Heidelberg: Springer, 2005, pp. 758–770. ISBN: 978-3-540-32106-4. DOI: 10.1007/11494669_93.

[61] Verza, Sofia, Blagojev, Tijana, Da Costa Leite Borges, Danielle, Kermer, Jan Erik, Trevisan, Matteo, and Reviglio della Venaria, Urbano. “Uncovering news deserts in Europe: risks and opportunities for local and community media in the EU”. Technical Report. European University Institute. 2024. DOI: 10.2870/741398.

[62] Vogler, Daniel, Weston, Morley, and Udris, Linards. “Investigating News Deserts on the Content Level: Geographical Diversity in Swiss News Media”. In: *Media and Communication 11*, no. 3 (2023). DOI: 10.17645/mac.v11i3.6794.

[63] Wang, Ryan Yang. “The Geography of Newspaper Circulations: A Spatial Taxonomy of “News(Paper) Deserts” in the United States”. In: *Media and Communication 11*, no. 3 (2023), pp. 304–317. DOI: 10.17645/mac.v11i3.6856.

[64] Weiss, Amy Schmitz. “Journalism Conundrum: Perceiving Location and Geographic Space Norms and Values”. In: *Westminster Papers in Communication and Culture 13*, no. 2 (2018). DOI: 10.16997/wpcc.285.

A Geoparsing Methodology

This appendix provides additional details on the geoparsing pipeline employed in this study, which was developed and validated in a companion paper that established accessible prompt-based methods for geoparsing UK local media [5].

A.1 Pipeline Development and Validation

The geoparsing methodology was developed using the Local Media UK Geoparsing (LMUK-Geo) dataset, a newly created gold standard corpus of 182 UK local news articles containing 1,313 toponyms. The companion study addressed the lack of benchmarking resources for UK local media geoparsing, which presents unique challenges due to fine-grained geographies and colloquial place names underrepresented in existing international datasets. The pipeline development involved three main stages: **Toponym Recognition:** The companion study employed a two-stage annotation process. The first 100 articles were manually annotated using Prodigy, while the second 100 articles used SpaCy's transformer-based NER model (en_core_web_trf). When validated against manual annotations, SpaCy achieved F1-score = 0.94, precision = 0.97, and recall = 0.91, demonstrating sufficient accuracy for automated extraction. **Candidate Generation:** For each identified toponym, coordinate candidates were generated by querying two gazetteers: Ordnance Survey Open Names (covering Great Britain) and OpenStreetMap Nominatim (providing additional coverage, particularly Northern Ireland). This produced 7,374 candidate locations across the dataset. For 52 toponyms that returned no matches, coordinates were manually retrieved using Google Maps with article context. **Toponym Disambiguation:** The companion study tested two approaches: (1) contextual disambiguation, where LLMs selected from the same candidate options presented to human annotators, and (2) few-shot classification, where LLMs identified Local Authority Districts without candidate lists. Given LLMs' limitations with coordinate calculations [40], all approaches reformulated disambiguation as administrative boundary classification. The companion study systematically evaluated four open-source LLMs: Gemma2 (9B), Llama3.1 (8B), Qwen2 (7B), and Mistral (7B), testing various prompt configurations, metadata combinations, and temperature settings. The optimal configuration used a detailed prompt with one-shot examples, including only the publishing outlet's domain as metadata context. Individual model performance varied significantly, with Mistral performing notably worse across all configurations. The study implemented majority voting across the three best-performing models, but Gemma2-9b emerged as the strongest individual performer, achieving: Accuracy: 88.2%; Precision: 88.0%; Recall: 88.1%; F1-score: 82.2%. When compared against a state-of-the-art fine-tuned approach by [25], the prompt-based method achieved competitive performance while offering greater accessibility by eliminating resource-intensive fine-tuning requirements.

A.2 Adaptation for Large-Scale Application

For the present study's analysis of the sample extracted from UKTwitNewsCor, we adapted the validated methodology with the following considerations: **Model Selection:** Given computational resource constraints and Gemma2-9b's strong individual performance in the validation study, we deployed this single model rather than the ensemble approach. This decision was justified by Gemma2's robust standalone results and the impractical computational overhead of running multiple models across the full corpus. We implemented the optimal configuration identified in the validation study: Gemma2-9b with the detailed prompt format and domain metadata only. The administrative boundary classification approach was maintained, mapping all locations to Local Authority Districts before coordinate assignment. This methodology represents the first large-scale application of prompt-based LLM geoparsing to UK local media, demonstrating the practical scalability of the approach developed in the companion study (withheld for review).

B Output Area Classification Overview

The UK Output Area Classification (OAC) 2021/22 categorises neighbourhoods into hierarchical groups—Supergroups, Groups, and Subgroups—based on 2021 Census data (England and Wales), modelled 2022 data (Scotland), and apportioned 2021 data (Northern Ireland). These were derived

using clustering on 58 sociodemographic variables, including ethnicity, housing, and occupation, with assignments determined by proximity to cluster centroids. The classification contains 8 Supergroups, 21 Groups, and 52 Subgroups, aggregated to LSOA, DZ, and SDZ levels for analysis. This paper utilises the Supergroups to infer broader demographic trends, summarised as follows:

1. **Retired Professionals**: Affluent, ageing populations in rural/low-density areas; characterised by high homeownership, detached housing, car ownership; predominantly White, UK-born retirees with managerial/professional backgrounds.

2. **Suburbanites and Peri-Urbanites**: Middle-aged, home-owning families in suburban and rural-urban fringe areas; skilled/professional occupations, high education, low ethnic diversity, strong Christian affiliation.

3. **Multicultural and Educated Urbanites**: Young-to-middle-aged, ethnically diverse urban dwellers; high private renting, student populations, degree-level education; concentrated in inner cities, especially London.

4. **Low-Skilled Migrant and Student Communities**: Young, transient renters in high-density terraces/flats; overrepresented in low-skilled jobs, unemployment, ethnic minorities; common in industrial towns and outer London.

5. **Diverse Suburban Professionals**: Multi-ethnic professionals in outer suburbs; high homeownership, managerial roles, degree qualifications; families with children and moderate religious diversity.

6. **Baseline UK**: Modal UK characteristics with mixed housing tenure and intermediate occupations; overrepresented in south London; high unemployment and ethnic diversity.

7. **Semi- and Unskilled Workforce**: Deprived, UK-born industrial communities; social renting, low education, elementary occupations; high disability rates; prevalent in former industrial regions.

8. **Legacy Communities**: Ageing, isolated populations in flats and social housing; low skills, high unemployment and disability; concentrated in coastal and remote areas.

C Prompt

```
"Your task is to provide JSON-formatted spatial information regarding UK local
    news websites.

Note: The news outlet may no longer be active, but provide an analysis based on
    any available information. Acceptable sources include: the domain itself,
    archived content (Wayback Machine), Wikipedia entries, local council websites
    , media directories, or third-party references to the publication. Return
    your response in valid JSON format with exactly these fields:

{
"domain": "the domain",
"coverage_area_description": "Brief description of geographical coverage",
"primary_location": "Single location name representing the main/central coverage
    area",
"status": "active/inactive/unknown",
"confidence": "high/medium/low",
"scope": "small/medium/large"
```

```
}

Requirements:
- coverage_area_description: Maximum 3 sentences.
- primary_location: If multiple locations are equally central, choose the main
    town/city or most central location that best represents the coverage area.
- scope: Use the following logic: Small: Coverage focuses on a single town or
    city. Medium: Coverage spans multiple towns, cities, or villages. Large:
    Coverage includes entire regions, counties, or large geographic areas.
- ensure primary_location can be geocoded (e.g., 'Manchester', 'Scottish Borders
    ', 'West Sussex') by solely providing location name and no additional text.
    If ambiguous, use the format "City, County".
- If no information is found, use "unknown" for relevant fields.

Example Input: example-gazette.co.uk
Example Output:
{
"domain": "example-gazette.co.uk",
"coverage_area_description": "Serves the market town of Exampleville and
    surrounding villages in North Yorkshire. Coverage includes local council news
    , community events, and business updates. Primary focus on the Exampleville
    district and nearby rural areas.",
"primary_location": "Exampleville, North Yorkshire",
"status": "active",
"confidence": "high",
"scope": "medium"
}

Now provide an output for this domain: {}"
```

D Technical Notes

All analyses were conducted in R (version 4.4.1) using spatial, statistical, and data wrangling libraries, including `tidyverse`, `sf`, `spdep`, `concaveman`, `ineq`, `geosphere`, and `DescTools`. To ensure stability in estimates, two outlets mentioning fewer than 10 locations were removed from the analysis. Spatial data were processed using the British National Grid (EPSG:27700) and WGS84 (EPSG:4326) projections as appropriate. Administrative boundary data for England, Scotland, Wales, and Northern Ireland were sourced from respective national statistical offices.

D.1 Metric Definitions

D.1.1 Spatial Extent Metrics

Radius measures the geographic reach from an outlet's primary location (the location with the most mentions). We calculated the smallest radius containing 75% of all location mentions by frequency. The 75% threshold captures core coverage areas while excluding occasional outlier mentions.

Area measures the size of the convex polygon enclosing all locations within the 75% radius, calculated in km^2. This provides a two-dimensional measure of geographic footprint.

D.1.2 Administrative Reach Metrics

Districts tallies the number of unique administrative districts covered by an outlet's location mentions, indicating administrative breadth.

Gini measures inequality in how mentions distribute across demographic area types (ONS Supergroup categories). Values range from 0 (perfectly equal distribution across area types) to 1

(mentions concentrated in one area type). Location mentions were assigned to districts via nearest-neighbour spatial joins, then classified by demographic supergroup.

D.1.3 Spatial Heterogeneity Metrics

Entropy measures how evenly location mentions spread across districts. Higher values indicate more uniform geographic attention across administrative units, while lower values suggest concentration in fewer districts. Values approach $\log(D)$ for perfectly even distribution and 0 for complete concentration.

Moran's I measures spatial autocorrelation—whether nearby locations receive similar coverage attention. Positive values indicate clustering (neighbouring locations get similar mention frequencies), negative values indicate dispersion, and zero indicates random spatial distribution. We used a 50 km distance threshold, selected based on sensitivity analysis showing stable outlet rankings compared to 75 km (Spearman's $\rho = 0.95$) but greater divergence at 25 km ($\rho = 0.75$).

D.1.4 Distance Decay Metrics

Coefficient of variation measures how variable distances are relative to mean distance from the outlet's primary location. Higher values indicate more scattered coverage patterns.

Proportion within 10 km calculates the fraction of all location mentions falling within 10 km of the outlet's primary location, indicating local coverage intensity.

All distances calculated using the Haversine formula for geographic coordinates.

E Clustering Experiments

To identify optimal clustering solutions for media outlet spatial typologies, we conducted a comprehensive experimental evaluation across multiple algorithmic approaches and feature representations. This appendix documents the systematic comparison that informed our methodological choices and validates the robustness of our findings.

Table 3 reveals systematic trade-offs between clustering quality and solution interpretability. DIANA clustering consistently achieved the highest silhouette scores (0.533–0.541) across all PCA configurations, yet produced severely imbalanced solutions with most outlets concentrated in single dominant clusters. Whilst these solutions demonstrate strong statistical separation, they offer limited analytical insight into outlet diversity.

Density-based approaches (HDBSCAN) identified substantial outlier populations (47–170 outlets) across all feature configurations, reflecting the presence of outlets with unique spatial characteristics that resist categorisation. However, the resulting cluster structures often comprised small, highly specialised groups that, whilst statistically coherent, provided insufficient coverage for comprehensive typological analysis.

The minimal feature set consistently outperformed dimensionally reduced alternatives in terms of identifying nuances subgroups in the data. K-means clustering on the minimal dataset achieved average statistical quality (silhouette = 0.335) but nonetheless produced six well-differentiated clusters with meaningful size distributions.

Cross-validation analysis revealed some degree of consistency in cluster identification across different algorithmic approaches when applied to the minimal feature set. K-means, Ward hierarchical clustering, and complete linkage methods produced similar cluster numbers (6, 6, and 6 respectively) with comparable balance scores (0.58, 0.41, and 0.54), demonstrating robust identification of underlying outlet typologies independent of specific algorithmic assumptions.

Dataset	Method	Silhouette	Clusters	Outliers	Cluster Sizes
Top performing experiments					
PCA (80%)	DIANA	0.541	3	0	347, 9, 2
PCA (90%)	DIANA	0.537	3	0	347, 9, 2
Full	DIANA	0.534	3	0	348, 8, 2
PCA (95%)	DIANA	0.533	3	0	347, 9, 2
Minimal	HDBSCAN	0.490	2	47	8, 303
Small dataset					
Minimal	**K-means**	**0.335**	**6**	**0**	**5, 152, 90, 15, 58, 38**
Minimal	Ward	0.331	6	0	102, 140, 20, 76, 8, 12
Minimal	DIANA	0.343	3	0	260, 91, 7
Minimal	Complete	0.305	6	0	180, 97, 20, 53, 5, 3
PCA dataset					
PCA (80%)	Complete	0.456	3	0	340, 15, 3
PCA (80%)	Ward	0.300	6	0	23, 125, 123, 66, 8, 13
PCA (80%)	K-means	0.297	6	0	49, 8, 92, 134, 52, 23
Full feature dataset					
Full	Complete	0.434	3	0	340, 15, 3
Full	K-means	0.256	3	0	8, 167, 183
Full	Ward	0.223	3	0	222, 83, 53
Full	HDBSCAN	0.195	4	170	7, 6, 6, 169

Notes: Selected results from the combinations tested. Bold indicates chosen solution. DIANA = Divisive Analysis; Complete = Hierarchical clustering with complete linkage; Ward = Hierarchical clustering with Ward linkage; HDBSCAN = Hierarchical Density-Based Spatial Clustering. PCA percentages indicate variance retention.

Table 3: Clustering Method Performance Comparison