Buildings & Cities

Urban data: harnessing subjective sociocultural data from local newspapers

SPECIAL COLLECTION: DATA POLITICS IN THE BUILT ENVIRONMENT

RESEARCH

ubiquity press

FILIPE MELLO ROSE

JUIWEN CHANG

*Author affiliations can be found in the back matter of this article

ABSTRACT

As data-based governance becomes mainstream, social and cultural interactions that characterise urban life are at risk of being ignored in decision-making practices if only supposedly objective, quantifiable data are used. In this context, this article conceptualises subjective sociocultural data as a data form that considers a city's intangible and unquantifiable social and cultural aspects. A methodology is proposed for collecting and using subjective sociocultural data by highlighting local press articles as a potential data source. A pilot application conducted in Hamburg, Germany, demonstrates a potential integration of subjective sociocultural data into urban planning processes by analysing over 2500 local newspaper articles. The findings reveal that local journalism can be a data source for understanding diverse social and cultural interactions between citizens and urban places. This street-level information from local newspaper articles can (1) provide urban planners with an overview of newspaper mentions of any specific urban areas, (2) support the identification of local debates, and (3) aid in the observation of emerging places of sociocultural interactions. This approach can support the diverse government and non-government stakeholders engaged in data-based governance to better account for intangible sociocultural aspects of urban life.

PRACTICE RELEVANCE

This research supports governance actors in dealing with the epistemological limitations of purposely gathered and/or objective data by conceptualising a new—currently untapped—data type: subjective sociocultural data sourced from local journalism. By using geographical text analysis on local newspaper articles, urban planners and decision-makers gain access to a wealth of street-level information, local debates and temporal dynamics of urban issues. This approach provides a comprehensive understanding of intangible and unquantifiable aspects of urban life, allowing for more informed and context-sensitive decision-making. The practical benefits include identifying diverse uses of urban spaces, capturing local public debates, and tracking the emergence and disappearance of places in the public sphere, possibly leading to more effective and inclusive urban planning practices.

CORRESPONDING AUTHOR:

Filipe Mello Rose

Digital City Science, HafenCity University Hamburg, Hamburg, DE; Laboratory of Knowledge Architecture, Technical University Dresden, Dresden, DE

f.mellorose@mail.de

KEYWORDS:

cities; data-based governance; data politics; datafication; local journalism; planning; sociocultural data; subjective data; urban data; urban governance

TO CITE THIS ARTICLE:

Mello Rose, F., & Chang, J. (2023). Urban data: harnessing subjective sociocultural data from local newspapers. *Buildings and Cities*, 4(1), pp. 369–385. DOI: https://doi.org/10.5334/bc.300

1. INTRODUCTION: BETTER DATA FOR URBAN GOVERNANCE

Many new trends in urban planning, such as algorithmic governance, smart cities and digital twins, rest on data-based urban governance. In this type of urban governance, 'a complex system of actors, relationships, processes, and technologies' is designed to decide on urban issues based on diverse sets of available and analysable data (Maffei *et al.* 2020: 124). Data-based planning is a crucial instrument for addressing the upcoming environmental and societal challenges (Batty & Yang 2022: 146) as it allows comprehensive and interdisciplinary analyses of complex problems by making them 'more "legible" for governing' (Mejias & Couldry 2019: 4). At the same time, the datafication of urban governance raises various issues concerning data validity, as well as the reliability, representativeness and interoperability of available data (Bunders & Varró 2019).

Crucially, data-based urban governance is limited by all urban problems not being 'equally knowable and solvable' by data-based governance practices (Bunders & Varró 2019). Many key intangible and unquantifiable aspects of urban life (*e.g.* the role of family-owned corner shops in a local neighbourhood) play a vital role in the day-to-day urban experiences of citizens (Jacobs 1961/1992). Likewise, urban places such as parks or pavements 'mean nothing divorced from their practical, tangible uses' (Jacobs 1961/1992: 111). This information is hardly quantifiable and is thus ignored in data-based governance. These limitations of data-based governance can at least be mitigated by including broader and more variegated datasets (Batty 2019). In this sense, the present article conceptualises *subjective sociocultural data* as an important yet frequently ignored or underused input for data-based governance. With this conceptualisation and an exemplar pilot methodology, the aim of this paper is to support the diverse stakeholders engaged in data-based governance (including government and non-government actors) to better account for intangible and unquantifiable aspects of urban life in their decision-making.

Subjective data are a particular data form that explicitly involves human judgment in its collection. Sociocultural data depict aspects 'related to the different groups of people in society and their habits, traditions, and beliefs' (Cambridge University Press n.d.). Jointly, subjective sociocultural urban data depict self-reported and evaluated accounts of a city's social and cultural practices (*i.e.* the interactions forming a city's social and cultural fabric). In pursuing the research goal of broadening the data sources for urban governance, this article conceptualises the utility of subjective sociocultural urban data and proposes a methodology for collecting and using this data type. This conceptualisation and methodology rest on a critical review of the literature on the different types of data used in urban governance and a synthesis of conceptual discussions on subjective data. Moreover, local press articles were identified as a potential and largely untapped source of subjective sociocultural urban data, which, once geo-referenced and synthesised, can inform local governance decisions.

Empirically, this study reflects on a pilot study that aims to integrate subjective sociocultural data into urban planning by drawing on a research project carried out with the state-owned company for real estate management and land assets of the German city of Hamburg: the *Landesbetrieb Immobilienmanagement und Grundvermögen Hamburg* (LIG). The pilot study developed a geo-parsing module with subjective sociocultural data that aims to complement 'objective' data used in decision-making regarding public real estate policies. This study draws on data from 2539 local newspaper articles from Hamburg published online between 2018 and 2022. The articles were automatically analysed for spatial references in Hamburg (*i.e.* geo-parsed and geocoded) and clustered around 10 algorithmically generated topics. The pilot use cases show that local journalism can (1) provide urban planners with an overview of newspaper mentions of any specific urban areas, (2) support the identification of local debates, and (3) aid in the observation of emerging places of sociocultural interactions.

In conceptualising local press articles as a source of subjective sociocultural urban data for urban governance, this study draws on and contributes to research on datafication processes in governance. However, as 'datafication is cross-disciplinary in nature' (Lomborg *et al.* 2020), this study also draws on multiple academic fields. First, it draws on media studies to argue for the general utility of (local) journalism as a data source for the data-based governance (Battistini *et *al.* 2013; Sarın & Uluğtekin 2019). Second, it draws on digital natural language processing (NLP) methods for geographical text analysis and topic modelling (Cai 2021; Middleton *et al.* 2018; Porter *et al.* 2015).

This article is structured as follows. Following this introduction, the paper conceptualises and situates subjective sociocultural data among other urban data forms used for urban planning. The methodology is then described and the results of an empirical pilot application are reported. This empirical section sets out the NLP methods and three potential use cases of the outlined techniques. The last section reviews and discusses the use of subjective sociocultural data for urban governance.

2. EXPLICITLY SUBJECTIVE DATA FOR DATA-BASED GOVERNANCE

2.1 FROM GOVERNMENT DATA TO DATA-BASED GOVERNANCE

In contrast to the traditional governments' use of self-collected data, data-based governance draws on a variety of data sources. These range from outcomes of algorithms and the Internet of Things (Coletta & Kitchin 2017), as well as large datasets, dashboards or surveillance systems to create a more efficient administration and governance of places (Bunders & Varró 2019; Kitchin & McArdle 2017; Valdez *et al.* 2018; Zuboff 2019). Moreover, with the increasing prevalence of sources in data-based governance, many stakeholders are able and needed to provide or analyse data. Most notably, international platform corporations now also inform decision-making and policy development with their datasets and analytical tools (*e.g.* Rettberg 2020). Along with numerous other factors, this datafication has thus further advanced the shift from *traditional government* to *governance* by beyond-the-state 'horizontal associational networks of private (market), civil society […] and state actors' (Swyngedouw 2005: 1992). Data-based governance, thus, coordinates this 'beyond-the-state governing' arrangement by providing and analysing significant and diverse datasets.

This datafication improves the capacity for anticipatory planning in a governance system that includes an ever greater variety of actors (beyond traditional government administrations) by using data 'to revolutionize the process of policy analysis' (Maffei *et al.* 2020: 124). While data-based governance can thus represent new business opportunities for companies collecting data as a by-product (*e.g.* Rettberg 2020), it can also allow researchers or civil society organisations to put governments and corporations under greater scrutiny (*e.g.* Dalton 2019). Crucially:

> datafication enables us to deal with lines of inquiry that were difficult if not impossible to pursue before.
>
> (Lomborg *et al.* 2020: 207)

In this sense, data-based governance can also foster a more collaborative and inclusive approach to governance, where a network of actors work together to identify and address complex societal challenges. By providing and analysing data that are of interest to urban governance, these non-government actors thus, directly or indirectly, engage in expanding networks of data-based governance.

2.2 ONTOLOGY OF DATA SOURCES OF DATA-BASED GOVERNANCE

Before digitalising public administration, economic and social activities, data-based governance relied primarily on *purposely collected data*, such as structured surveys and observation data. This type of data source includes data deliberately produced by government institutions (*e.g.* data from census and statistics offices) or non-government organisations supporting data-based governance with their data. For instance, data-based governance often draws on specialised survey institutes or universities to collect data with diverse formal methods (online surveys, in-person interviews, online tracking, *etc.*). A trait of these purposely collected data or survey-type data is, however, that its controlled collection processes reduce the data's 'scope, temporality, and size, and are quite inflexible in their administration and generation' (Kitchin & McArdle 2016: 2).

Digitalising social, economic and governmental activities has enabled the collection of usable data as a by-product. As public administrations become more digitised, more data are processed digitally in government bureaucracies and become a potential new data source (Kitchin & McArdle 2016). This type of data, which are gathered as a by-product, is either observed (*i.e.* as a result of people using technology) or inferred (*i.e.* consolidated information from existing data sources) (Taylor & Richter 2015). Data that originate as a by-product of government activities also include records, such as budgets and spending data, and data on assets as well as from registries (*e.g.* registration of residences, financial activities or vehicles). Crucially, this data source type is not limited to any *ex-ante* specification of purpose in data-based governance, as the possible knowledge gains are defined after the data are collected during data preparation (Kitchin & McArdle 2016) (Table 1).

Table 1: Conceptualisation of different data sources and collection forms.

	GOVERNMENT DATA	NEW DATA-PRODUCERS IN DATA-BASED GOVERNANCE
Purposely collected data (Structured)	Periodically collected data on the polity, *e.g.* census data, surveys of statistics administrations	Data from non-state data collection companies and research institutes, *i.e.* market research, university research
Data as a collected as a by-product (Unstructured)	Data collected as a by-product of government administrations, *e.g.* permit data, data from public companies (*i.e.* health insurance, unemployment insurance, public transport), budget data	Data collected as a by-product of ordinary economic and social activities, *e.g.* social media companies, targeted advertisements, customer reviews

All data (sources) have natural epistemological limits. Purposely collected data are limited to addressing societal problems known before the data collection based on scientific theory (Kitchin 2014). While mostly free of sampling and statistical errors, this data type has a low potential to fulfil unanticipated usages (Kitchin & McArdle 2016). Unstructured data collected as a by-product, in contrast, are prone to sampling and statistical errors and biases originating from unequal uses of a service or product but allow exploratory research. Traditional data sources in urban governance generally relied on the accessibility of purposely collected data, such as surveys and censuses. As purposely collected data require an *ex-ante* definition of research problems (*i.e.* before data collection), data-based governance must also consider data sources allowing greater exploratory research to identify new research problems.

2.3 SUBJECTIVE AND SOCIOCULTURAL DATA

Sociocultural data depict 'society and their habits, traditions, and beliefs' (Cambridge University Press n.d.). They feature structured and statistical cultural, ethnic, religious or demographic information, as well as (more subjective) data on cultural practices, day-to-day activities and routines, and beliefs. In this article, sociocultural data are essentially data that allow one to understand better a given area's social and cultural fabric.

Data-based governance primarily draws on supposedly 'objective' data as impersonal and neutral 'raw information' that are best collected with the least 'subjective' human involvement possible (Rieder & Simon 2016). This applies to sociocultural data that are often reduced to variables for ethnic, religious or class identification which can be assessed somewhat objectively. Subjective data, in contrast, significantly and explicitly involve human judgment in its production. Despite a greater risk for biases, this data type is used as purposely collected data to measure social phenomena such as happiness, affection or wellbeing (*e.g.* Kahneman & Krueger 2006; Macků *et al.* 2020). Subjective measures are relevant for two main reasons. First, in the absence of (more) objective measures, subjective ones can (at least partially) depict social phenomena on which data are scarce. Second, subjective measurements allow one to:

> capture changes in both the explicit and the implicit components of the variable being measured and, therefore, [...] can be better suited for the study of broadly defined concepts.
>
> (Jahedi & Méndez 2014: 3)

Data on the sociocultural aspects of urban life meet both reasons for using subjective measures. For one, sociocultural factors of urban life offer few alternative 'objective' measures. Most urban interactions and practical, tangible uses of places are difficult to quantify objectively without using (sometimes far-fetched) proxy measures. For another, the sociocultural aspects of urban life are a broadly defined concept that refers to the intangible aspects, such as the diverse social and cultural interactions enriching urban life. Allowing different actors to determine the relevant social and cultural interactions will likely better depict the local urban social and cultural fabric. Moreover, while statistical and systematically quantified sociocultural data allow large-scale analyses, small-scale subjective data can help planners to better understand the social conditions shaping an area's urban (sociocultural) fabric. Urban sociocultural data depict day-to-day activities, routines, values and public spaces' practical, tangible uses. In this sense, this article understands subjective sociocultural data as collections of descriptions of social interactions and cultural life in an area. This includes accounts of cultural and civic events, the uses of public spaces, or popular meanings of places (*e.g.* the importance of traditional corner shops or bars to a community) relevant to understanding the local urban fabric.

Combining different types of data sources allows for triangulating information and complementing analyses with different perspectives. In this sense, adding data gathered as a by-product to data-based governance eases exploratory, unexpected lines of enquiry, while subjective sociocultural data expand data-based governance to better account for a given area's social and cultural fabric.

2.4 LOCAL JOURNALISM AS A DATA SOURCE FOR URBAN GOVERNANCE

This study conceptualises and empirically tests the use of local press articles as a source of subjective sociocultural urban data. In doing so, it draws on multiple studies in which press articles serve as an explicitly subjective source of geographical and sociocultural data. For instance, Voukelatou *et al*. (2021: 300) find that 'news data are a new promising data source for the further exploration of subjective well-being'. Battistini *et al*. (2013: 157) use online news feeds as data sources to efficiently map geohazards (*i.e.* earthquakes, landslides and floods) under the assumption that whenever geohazards have 'relevant consequences, news is reported on the Internet'.

Gregory & Paterson (2020) elaborate on this practice under the title of geographical text analysis, for which the scholars geo-parse historical press articles to map historical poverty in the UK. The scholars' study on historical poverty demonstrates how analysing textual data can provide insights into geographical patterns (Gregory & Paterson 2020). Ozgun & Broekel (2021) use regional newspapers to enquire about the regional difference in attitudes to innovation by quantifying the mentions and sentiments of different regional newspaper articles. However, while most studies focus on the national or at least regional level, Paterson (2020: 66) points out that:

> references to place, especially at more local levels, can make the impact(s) of abstract concepts […] appear more concrete by situating them within definable geographical boundaries.

Other studies have used similar methodologies to mobilise social media data for data analysis in environmental (Ghermandi & Sinclair 2019) and tourism research (Chen *et al.* 2021). However, in contrast to social media data, press articles have at least a minimal degree of quality control. Moreover, while interactions on social media are highly diverse and frequently use colloquial or group-specific language, newspaper articles are typically written for broader audiences and require less context-specific language or knowledge about the authors to be intelligible. In this spirit, applying geographical text analysis to local press articles appears as an auspicious way of introducing new sources of subjective sociocultural data to data-based urban governance.

The use of newspaper articles as a source of sociocultural data is grounded on the assumption that the content produced by newspapers at least generally depicts cultural practices, day-to-day activities and routines, and beliefs of a given place. The way media outlets 'frame an issue shapes how people understand and remember it' (Ozgun & Broekel 2021: 3). Moreover, media outlets have nested interests that influence their framing and categorisations of what is ‘newsworthy’. However, while the press might advance specific agendas, ‘they do not do so independent of their audience’ (3). As most media organisations are commercially driven, the choice of newsworthiness and the tone of reporting news is strongly linked with attempts to gain and retain the largest possible audiences (Scheufele 1999; Gentzkow & Shapiro 2010; Agirdas 2015). As people seem to ‘selectively expose themselves to attitude-consistent information’ (Earle & Hodson 2022: 9), aligning to the political orientation and reporting priorities of a target readership is a commercial necessity and a response to public demand (Gentzkow & Shapiro 2010). While national news outlets tend to become increasingly partisan to win audience loyalty in a politically divided market (Gentzkow & Shapiro 2010), this is not the case for local newspapers due to their limited geographic competition (Agirdas 2015).

3. EMPIRICAL PILOT APPLICATION: SOURCING LOCAL JOURNALISM FOR URBAN GOVERNANCE

In the following, this study describes and discusses an empirical pilot application of the possible uses and sources of sociocultural data for data-based governance conceptualised above. While this pilot application has been developed in partnership with a state-owned company for a specific case (*i.e.* to improve data-based governance of public real-estate assets), the methodology (including the used algorithms) is available as open-source code on GitHub. This way, the practical implementation of sourcing local journalism for urban governance can be tested and improved in other contexts.

3.1 CONTEXT OF THE EMPIRICAL PILOT APPLICATION

This study is situated within a broader research project that investigates new modes of data-based planning conducted with Hamburg’s state-owned company for real estate management and land assets, the LIG. The German city of Hamburg is no exception to the trend of datafication of urban governance, and the LIG aims to improve its governance activities with new digital tools for spatial analysis. To develop a method of digital data-based real estate and land management in the city-state, a collaboration between the LIG and HafenCity University has engaged in the creation of a digital platform for the evaluation of land parcels. This platform will consolidate various data sources and inform decisions regarding the acquisition and sale of land by public administrations. More precisely, the digital tool allows the comparison of land parcels based on their surroundings (*i.e.* the proximity to various urban amenities) and their connectivity (*i.e.* isochrones in the city) in the ‘LIG-Finder’ module. In addition, the ‘geo-parsing’ module provides a spatial overview of geocoded texts. This geo-parsing module displays geocoded documents from the state parliament and multiple regional and local newspapers. This includes the *Elbe Wochenblatt*, a hyperlocal free newspaper that this study uses as a pilot application.

Elbe Wochenblatt is a hyperlocal advertisement-based newspaper from Hamburg. Its articles are disseminated online and in seven weekly neighbourhood-level print editions, distributed free of charge to all households (unless they opt out) within a specific delivery zone (Figure 1). As an advertisement-based newspaper, it provides a platform for and depends on local businesses to advertise their products and services. *Elbe Wochenblatt* features short articles using colloquial language on local politics, sports, civic life, cultural events and businesses. Next to *Hamburger Wochenblatt*, *Elbe Wochenblatt* is the second largest free newspaper in Hamburg with approximately 300,000 weekly printed copies.[1] However, the media landscape in Hamburg is also—and probably more substantially—shaped by regional and Hamburg-based national newspapers, such as *MOPO/BILD*, *Hamburger Abendblatt* and *Zeit*.[2]

Elbe Wochenblatt’s accessibility and high volume of publications from 2018 to 2021 make it a valuable test case for the pilot application of the LIG-Finder’s geo-parsing module. During that period, the newspaper was majoritively owned by Madsack Mediengruppe, with a minority stake held by Funke Mediengruppe. In 2021, the newspaper was fully acquired by Funke Mediengruppe, which integrated the previously independent editorial office with other local newspapers from

Hamburg (notably with *Hamburger Wochenblatt*) in January 2023. While Madsack Mediengruppe is regionally focused on northern Germany, Funke Mediengruppe is the 10th largest media group in Germany (Institut für Medien- und Kommunikationspolitik 2022).

This study uses all 3511 newspaper articles published online (and in at least one of the seven neighbourhood print editions) from 2018 to 2021. The time frame was selected due to the newspaper's above-average publishing activity. The focus is on articles authored by the newspaper's editorial staff to safeguard a consistent format and focus on the newspaper's target areas. This excludes submissions and commentaries from readers that vary vastly in format, making an automated corpus analysis difficult at this piloting stage. *Elbe Wochenblatt* was selected for this pilot application because of (1) its hyperlocal focus and descriptions of neighbourhood life in their reporting and (2) the availability of a large corpus of articles. At the time of data collection, most other neighbourhood newspapers in Hamburg covered smaller areas. In this sense, using local journalism as sociocultural data is exemplified using *Elbe Wochenblatt* as a paradigmatic case (Flyvbjerg 2006: 232). This study does not aim to discuss the media practices of the newspaper beyond a necessary contextualisation.

3.2 NATURAL LANGUAGE PROCESSING (NLP) OF LOCAL JOURNALISM

As unstructured data and a by-product of economic and social activities, local newspaper articles require significant and meticulous preparation to provide tangible benefits to urban decision-makers as sociocultural data. The preparation and curation of the unstructured data draw on the practices of the emerging field of NLP, advancing its use in data-based urban governance (*e.g.* Cai 2021). NLP is an algorithmically enabled analytical practice that combines linguistics, computer science and artificial intelligence to 'structure large volumes of unstructured data' (Cai 2021: 1). This study uses two primary methods of NLP to transform unstructured newspaper data into a source of subjective sociocultural urban data: geo-parsing and topic modelling.

3.2.1 Geo-parsing

The geo-parsing process allows the extraction of location names and coordinates from the newspaper articles' unstructured textual data. The 'parsing' process separated texts into a computer-readable format to detect terms that can be linked to geographical identifiers. In other words, geo-parsing involves two subprocesses that first extract location identifiers (*i.e.* place names) and then geocodes them (Wang *et al.* 2020). First, toponym recognition and geotagging identify names of geographical locations in the text based on name entity recognition. This task locates and classifies 'named entities' in unstructured text into predefined categories (*e.g.* person names, organisations, locations, time expressions, quantities, monetary values and percentages). After comparing multiple name entity recognition classification libraries, this study used spaCy, an open-source software library available for over 70 languages.[3] Second, this study uses Nominatim[4]—an open-source geocoder to convert place names into geocoordinates—to geocode the geographical entities identified in the previous step. The algorithm builds a statistical model estimating the probability of a named entity referring specific geolocated place (Middleton *et al.* 2018). Entries with a score below a probability of 0.95 were checked manually. The geocoding of the 3511 newspaper articles identified 10,320 mentions of 1718 places in Hamburg.

After the geo-parsing process, the database of newspaper articles required further preparation and data cleaning. A significant portion of place mentions refers to entities linked to areas across multiple neighbourhoods (*i.e.* districts, regions or Hamburg as a whole). As these place names are somewhat arbitrarily geocoded to a point within the area they represent, these place mentions clutter neighbourhood-level analysis with information related to areas outside the neighbourhoods.[5] For this reason, the study excludes all mentions of place names referring to geographical areas above the neighbourhood level. This step excludes 1665 mentions of 'Hamburg', 356 mentions of the general parts of Hamburg (*e.g.* West Hamburg, South Hamburg) and 845 mentions of one of the seven districts. This filtering reduces the database to 7903 mentions of 1689 places.

Elbe Wochenblatt is distributed in an area that covers the entire west and south of Hamburg, including 34 of 104 official neighbourhoods (Figure 1). This area accounts for 41% of Hamburg's total area and 35.3% of Hamburg's population. The neighbourhoods serviced by the newspaper are highly diverse across various metrics ranging from population density to employment. For instance, the area of delivery includes the three wealthiest and the three poorest neighbourhoods of the city.

The geocoding confirms the hyperlocal focus of the newspaper's reporting, as over 89.5% of all place mentions in the database are in the delivery zone. Moreover, the vast majority (2539 articles, or 72.3%) of all 3511 analysed articles mention at least one place that can be located within a neighbourhood where the newspaper's print edition is delivered. Therefore, the following study focuses on the areas in which the newspaper's print version is distributed.

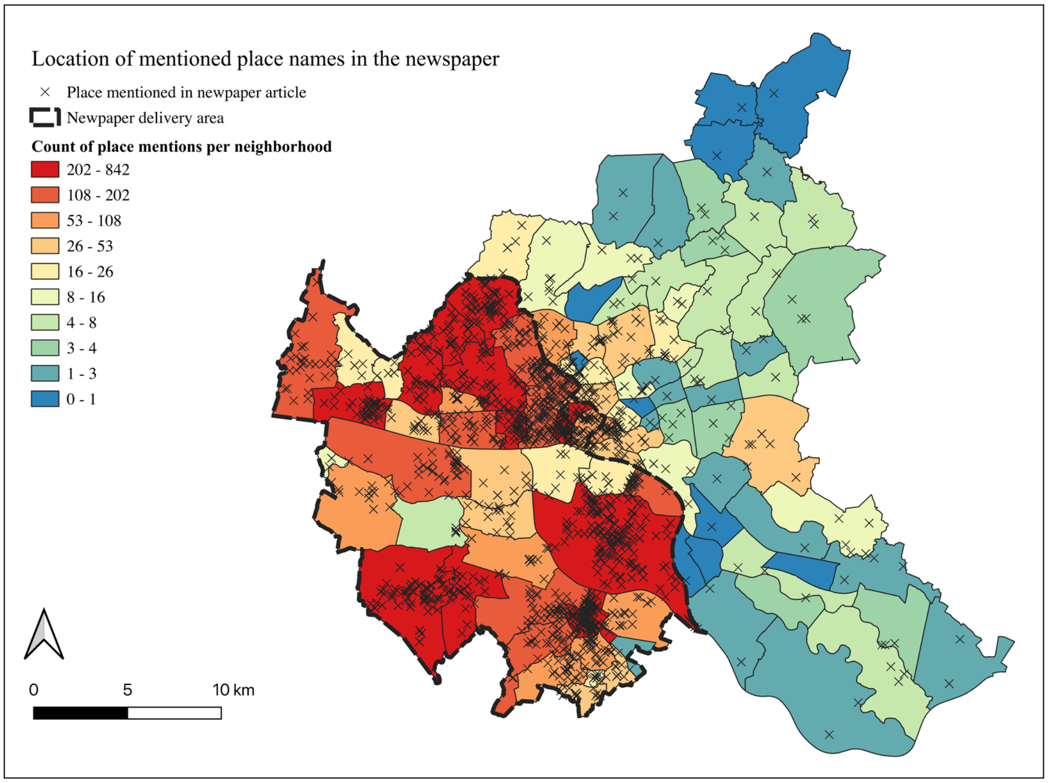
Location of mentioned place names in the newspaper
× Place mentioned in newpaper article
Newpaper delivery area
Count of place mentions per neighborhood
202 - 842
108 - 202
53 - 108
26 - 53
16 - 26
8 - 16
4 - 8
3 - 4
1 - 3
0 - 1
0
5
10 km

Figure 1: Spatial distribution of the place mentions in the local newspaper.

The 'cleaned' corpus of newspaper articles thus includes 2539 articles mentioning 1424 different urban places in the delivery area a total of 6665 times. Tests of the text corpus and the parsed and geocoded places indicate that the number of times places in each neighbourhood are mentioned significantly correlates to that neighbourhood's population (Pearson correlation = 0.667; $p < 0.01$) and area (Pearson correlation = 0.546; $p < 0.01$). No significant statistical relationship exists between the number of place mentions in an area and the area's median income, housing prices or other tested socioeconomic factors (*i.e.* share of migrants, share of higher education degrees).

3.2.2 Topic modelling

Topic modelling is an important step in NLP that allows the identification of different topics and topic distributions in large corpora of text. While there are several topic modelling methods, the most popular are latent dirichlet allocation (LDA) (Blei *et al.* 2003) and structured topic modelling (STM) (Roberts *et al.* 2019). The process follows the general procedure of clustering co-occurring lists of keywords into a predefined number of topics in the first step without needing any prelabelled data. To improve the certainty of identifying the correct topic, this study used both methods to identify 10 topics (*i.e.* co-occurring lists of keywords) in the text corpus of local newspaper articles. While both major topic modelling methods produce sensible results, a close analysis of the resulting 10 keyword lists forming each topic suggested the use of STM, as it allows the inclusion of additional information. In the case of this study, sentiment scores regarding the article's style and tone were added to the analysis. These sentiment scores rest on 'TextBlobDE'[6] and the 'Valence Aware Dictionary and Sentiment Reasoner'.[7]

Table 2: Results of the topic modelling.

Note: CDU = Christian Democratic Union (*Christlich Demokratische Union*); SPD = Social Democratic Party (*Sozialdemokratische Partei*).

ID	TOPIC LABEL	ARTICLES	PLACE MENTIONS	KEYWORDS IN ORDER OF LIKELIHOOD (ROOT WORDS)
1	Pandemic	325	416	corona; vaccine; virus; test; pool; infect; centre; offer; hygiene; ship; trip; number; mask; pandemic; distance
2	Health and social care	275	1,318	health; hospital; help; also; care; work; homeless; employee; office; need; support; service; centre; contact; will
3	Mobility	233	1,515	traffic; bridge; park; station; railway; construct; car; road; plan; street; tree; residence; work; area; can
4	Sports	231	960	team; club; sport; game; league; will; football; championship; player; play; time; coach; goal; point; start
5	Politics	223	638	district; green; SPD; office; CDU; citizen; member; elect; initial; board; federal; politic; association; parliament; left
6	Schools and education	183	531	school; student; children; work; young; elementary; project; class; lesson; parent; teacher; educate; district; young; learn
7	Housing	151	391	build; new; plan; area; district; house; tenant; develop; rent; apartment; centre; euro; property; construct; office
8	Concerts and events	134	385	music; ticket; artist; theatre; perform; stage; artist; play; concert; show; band; musician; film; choir
9	Portraits of neighbourhood figures	93	277	like; father; photo; time; wife; friend; remember; even; still; just; now; always; live; want; know
10	Civic life	79	234	festival; museum; o'clock; children; offer; café; organ; church; event; open; market; culture; place; visitor; take
0	*No clear topic*	*1,588*	*2,671*	

In the second step, the topic modelling algorithm calculates the likelihood of an individual article being linked to each of the 10 topics (*i.e.* 'topic score'). This analysis was based on the entirety of each article. As the sum of the 10 topic scores is equal to 1, the articles having a topic score > 0.5 (more than all other nine topics combined) in any given topic are considered to be significantly related to that topic. Articles without a topic score > 0.5 are understood not to be significantly linked to any of the 10 topics (*e.g.* combining multiple topics). Table 2 shows the topics identified in the topic modelling, the number of articles (with a respective topic score of at least 0.5) and the number of places mentioned in articles concerned with that topic. The number of place mentions varies significantly across different topics. For example, while topic 1, 'pandemic', was the most frequent topic in the corpus, articles clearly linked to topic 4, 'mobility', have the most place mentions. Due to the high threshold to be considered related to a particular topic, about 45% of articles are not linked to any topic.

3.3 EXEMPLAR USE CASES OF SOURCING LOCAL JOURNALISM FOR URBAN GOVERNANCE

The following three use cases illustrate how the application of NLP to local newspaper articles can generate data to better comprehend a given area's social and cultural fabric. While the first use case focuses on geo-parsing, the second demonstrates the application of topic modelling in combination with geocoding. The third use case illustrates a combination of both methods to consider changes in sociocultural patterns over time.

3.3.1 Use case 1: Street-level reporting on urban places

The geocoding of local newspaper articles provides street-level information on how places are discussed in local media (Figure 2). More precisely, the geocoded visualisation of place mentions provides an overview of how given urban places and areas are portrayed in the local news. This visualisation thus allows urban planners to access large amounts of urban sociocultural data on a given area without engaging in extensive keyword searches for all the area's streets and place names. With the geocoded visualisation, it is possible to quickly identify social, civic or cultural uses of specific urban places. For instance, certain streets occasionally serve as venues for neighbourhood festivals, while other places are frequently used as meeting places for civil society organisations. In addition, other articles also include information on the following sociocultural issues:

- reports about traditional shops and establishments risking going out of business
- reports about citizens' frustrations with public transport deficiencies or high traffic volumes
- reports about nuisances from social uses of public space (*e.g.* noise pollution)
- information about the (informal) history of places.

While most of this information is—in some form—available to public authorities, these data are rarely aggregated or made accessible via such an interactive city map. The aggregated and geocoded newspaper articles allow urban planners to include more sociocultural information in their data-based decision-making. However, while the map provides a first summarising overview of existing sociocultural data, actors using this tool must use it as a starting point for analyses that go beyond the excerpts of the newspaper articles. Nevertheless, urban planners can use these geocoded and visualised local sociocultural data as an overview of the sociocultural relevance that urban spaces might have to a local population. This way, planners can identify possible blockades in development projects early on. For instance, planners at LIG can use such visualisations to consider sociocultural information in their decisions on parcel sales and acquisitions. This means that sales of land that threaten the destruction socioculturally important places might be recognised earlier and avoided.

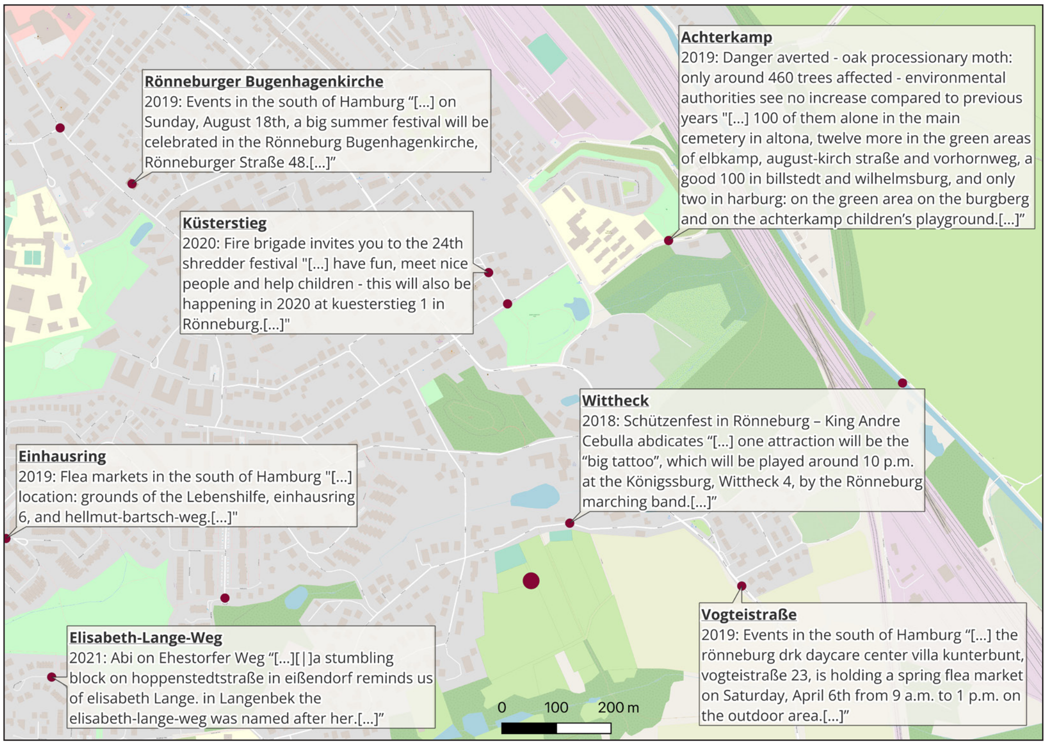
Rönneburger Bugenhagenkirche
2019: Events in the south of Hamburg "[...] on Sunday, August 18th, a big summer festival will be celebrated in the Rönneburg Bugenhagenkirche, Rönneburger Straße 48.[...]"
Küsterstieg
2020: Fire brigade invites you to the 24th shredder festival "[...] have fun, meet nice people and help children - this will also be happening in 2020 at kuestersteig 1 in Rönneburg.[...]"
Achterkamp
2019: Danger averted - oak processionary moth: only around 460 trees affected - environmental authorities see no increase compared to previous years "[...] 100 of them alone in the main cemetery in altona, twelve more in the green areas of elbkamp, august-kirch straße and vorhornweg, a good 100 in billstedt and wilhelmsburg, and only two in harburg: on the green area on the burgberg and on the achterkamp children's playground.[...]"
Witttheck
2018: Schützenfest in Rönneburg – King Andre Cebulla abdicates "[...] one attraction will be the "big tattoo", which will be played around 10 p.m. at the Königssburg, Wittheck 4, by the Rönneburg marching band.[...]"
Einhausring
2019: Flea markets in the south of Hamburg "[...] location: grounds of the Lebenshilfe, einhausring 6, and hellmut-bartsch-weg.[...]"
Elisabeth-Lange-Weg
2021: Abi on Ehestorfer Weg "[...][|]a stumbling block on hoppenstedtstraße in eißendorf reminds us of elisabeth Lange. in Langenbek the elisabeth-lange-weg was named after her.[...]"
Vogteistraße
2019: Events in the south of Hamburg "[...] the rönneburg drk daycare center villa kunterbunt, vogteistraße 23, is holding a spring flea market on Saturday, April 6th from 9 a.m. to 1 p.m. on the outdoor area.[...]"
0
100
200 m

Figure 2: Illustrative example of the use of local journalism to generate sociocultural data for urban governance translated into English.

3.3.2 Use case 2: Topic modelling to identify local debates

In combination, topic modelling and geo-parsing allow the identification of divergences in what is discussed in different neighbourhoods. Aggregating topic modelling and geo-parsing results highlights differences in how places and areas are discussed in the local press (Figure 3). While place names in some neighbourhoods are mostly mentioned in the context of reports on cultural events and local civic activities, most place names in other areas are mostly linked to debates about mobility issues.

Articles revolving around traffic, construction works and public transport, on average, mention 3.6 geocoded place names. Mobility is thus the most recurring theme in 20 of 38 neighbourhoods in the newspaper's delivery zone. In contrast, local civic activities are the dominant topic in most neighbourhoods in the southern suburbs of Hamburg. Here the newspaper reports on local (flea) markets, street festivals, and meetings in cafés and churches.

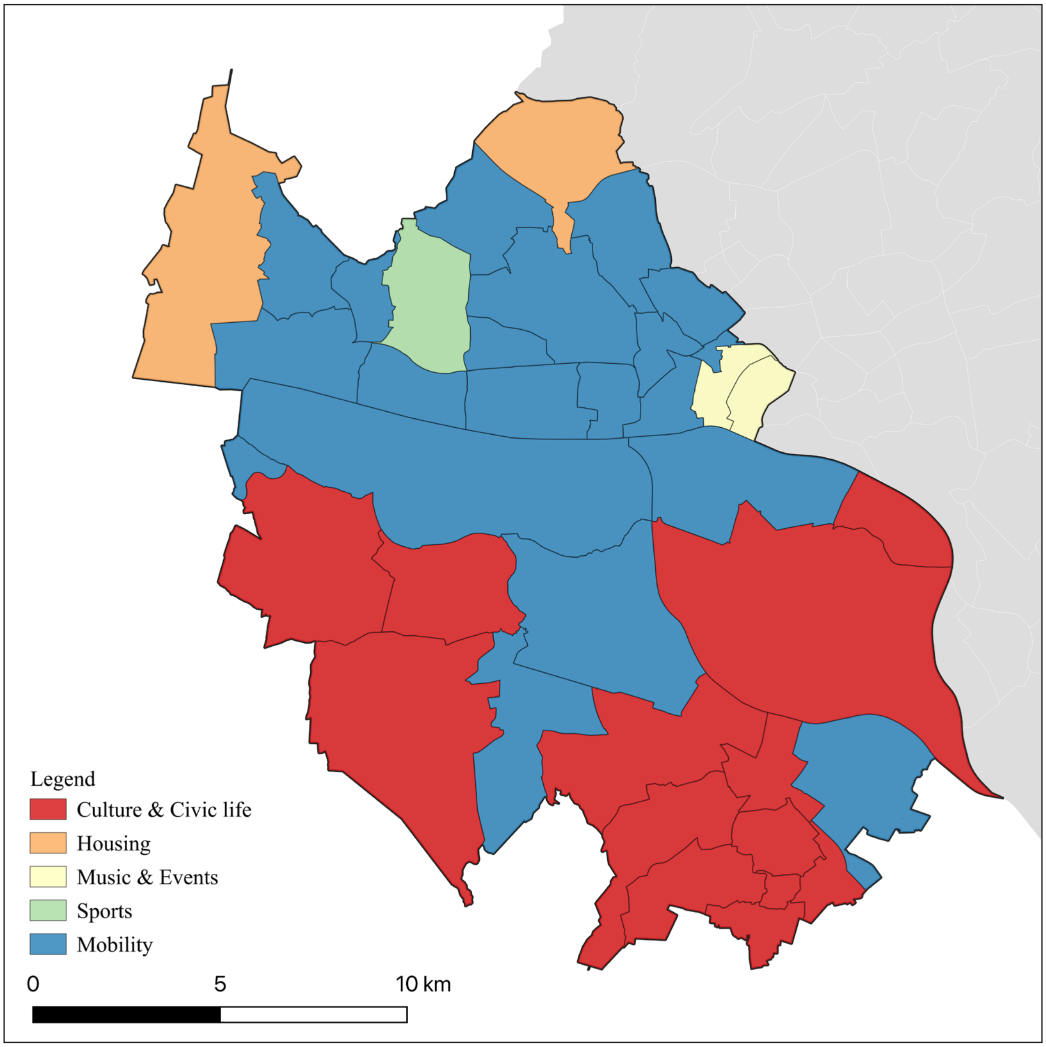

Figure 3: Most frequent topics of articles mentioning places in a given neighbourhood.

In the city centre, the newspaper reports on more formalised events in theatres and concert halls taking place in the area. The city centre includes the entertainment districts in Neustadt and St. Pauli, known for their vibrant nightlife. Sports is the central topic in the relatively peripheral neighbourhood of Osdorf. The newspaper mentions places in this area mainly in relation to its many amateur sport clubs that use the neighbourhood's numerous sporting grounds. Moreover, place names in two neighbourhoods are mentioned mostly in relation to housing, construction and real estate issues. Multiple large construction and urban development projects are occurring in these two areas, leading to disputes between civil society organisations, government officials and real estate investors.

Some topics likely carry greater policy relevance to governance stakeholders than others. In this sense, mapping articles reporting on urban planning and housing issues might provide relevant information for any administration concerned with planning and housing. However, a better understanding of dominant local topics can help urban planners to situate an area's perceived strengths and weaknesses more generally. Urban planners can complement their analysis of an area's needs with information on recurring problems (*e.g.* failures in public transportation and traffic jams) and potentials (*e.g.* vibrant civic life). In combination, geocoding and topic modelling of newspaper articles thus provide planners with a database of articles related to a given area and topic. This allows planners to understand better the critical issues related to each area of analysis.

3.3.3 Use case 3: Identification of emerging topics and places

Changes in topic relevance over time are closely tied to the sociocultural relevance gained or lost by certain places. Leveraging article publication dates and topic modelling data thus allows planners to examine how shifts in relevant topics influence the mentions of specific places in newspaper articles.

For instance, the COVID-19 pandemic represents an opportune moment to study these changes in topics and their impact on sociocultural depictions of urban places. The pandemic significantly impacted the places that had sociocultural information attached to them. As the pandemic itself became a topic (topic 1), more articles were concerned with health and social issues (topic 2). In contrast, local sports (topic 4), concerts and events (topic 8) and civic life (topic 10) became less relevant topics. These topic changes had significant implications for the places mentioned in the newspaper. Of 1129 places mentioned in 2018 and 2019, 652 (58%) were not mentioned in any article in 2020 and 2021. At the same time, 295 places not mentioned previously were brought up for the first time (at least since 2017) during the pandemic. Comparing the places that were mentioned for the first time or considerably more[8] with those that were no longer or considerably less[9] mentioned finds that the shift in topics is directly related to the places being mentioned. A chi-square test ($p < 0.001$) confirms the impression by showing a statistically significant relationship between the relevance of a place in the local newspaper during the pandemic and the topic to which it is related (Table 3).

TOPICS IN WHICH …	NEW OR CONSIDERABLY MORE RELEVANT PLACES	MORE OR LESS EQUALLY RELEVANT PLACES	SIGNIFICANTLY LESS OR NO LONGER RELEVANT PLACES
Mentioned in relation to topics 1 and 2	116	264	98
Mentioned in relation to topics 4, 8 and 10	51	542	920
Mentioned in relation to other topics or no clear article topic	507	2,311	1,856

Table 3: Linkages between place mentions and topics shown by an analysis of the COVID-19 pandemic on local journalism.

Note: Topic numbers correspond to those shown in Table 2.

When mapped, an overview of places gaining new relevance can indicate important sociocultural dynamics that might otherwise be overlooked. With this information, urban planners can account for new initiatives (*e.g.* by civil society actors) that vitalise public spaces (gaining positive attention) or identify places undergoing (reported) deterioration (gaining negative attention).

4. DISCUSSION AND CONCLUSIONS

This article conceptualised and provided a pilot methodology for subjective sociocultural data as novel input for data-based governance. This conceptualisation and pilot methodology aims to support the diverse stakeholders engaged in data-based governance to better account for intangible sociocultural aspects of urban life in their decision-making.

Data-based urban governance requires transforming diverse social and cultural processes into urban data (Mejias & Couldry 2019). Broadening the range of interactions that serve as data sources is crucial in improving the visibility of social and cultural aspects in data-based governance. Natural language processing (NLP) of local press articles allows the creation of sociocultural datasets that draw attention to the sociocultural meaning of urban places for habits, traditions and beliefs. This way, the datafication of governance can be better equipped to pursue previously impossible lines of enquiry (Lomborg *et al.* 2020: 207). More globally, this study highlights the potential uses of (sociocultural) data that are collected as a by-product of economic and social activities for data-based governance (*e.g.* also including corpora from social media companies, targeted advertisements and customer reviews). Moreover, this research also highlights how new urban (digital) geographies can be studied with geographical text analysis, even if these tools and analytical practices have only found limited attention from researchers so far (*e.g.* Gregory & Paterson 2020).

In practical terms, the methodology outlines how using NLP and topic modelling on a corpus of local newspaper articles can (at least) (1) provide urban planners with an overview of newspaper mentions of given urban areas, as well as (2) support the identification of local debates and (3) of emerging places of sociocultural interactions. The use of local journalism as an additional data source for Hamburg's state-owned company for real estate management and land assets should encourage urban planners to also consider sociocultural data in their decision-making. This way, decisions on sales and acquisitions of public land can also be based on subjective sociocultural information such as those outlined in the use cases. However, it is essential to recognise that, in particular, due to the evident ambiguities and situatedness (*i.e.* the context-specificness) of the provided information, subjective sociocultural data should never be used for automated decision-making. Instead, data-based governance still requires significant human judgment to evaluate these sociocultural data jointly with other 'traditional' forms of data (*e.g.* indices of real estate prices, closeness to amenities and transport, and quantifiable demand for different urban development types).

While much of the localised data displayed in use case 1 might be available in different government departments (*e.g.* knowledge on flea markets/street festivals, public construction works, or local establishments going out of business), the geolocated aggregation of these data and distribution to other governance stakeholders are often lacking. Using different methods of NLP on unstructured local journalism data might allow different governance stakeholders to anticipate emerging issues in an exploratory (Bunders & Varró 2019) and in a collaborative way.

To put this into practice, all stakeholders in data-based governance must be capable of understanding and adapting to the workings and limitations of (sociocultural) data sources. Due to openly accessible technological tools, diverse actor types can carry out exploratory research with textual information (*i.e.* on local newspaper articles). Civil society actors could therefore be empowered by introducing sociocultural data (and the respective analyses) into data-based governance. For instance, civil society actors and marginalised groups could use NLP and geographical text analysis of local journalism to identify and criticise the geographies of (specific) media outlets and how these relate to geographically unequal public attention to social issues (*e.g.* by mapping spaces that are 'forgotten' by local media). More generally, civil society actors could use similar sociocultural data to contest supposedly objective metrics of urban life that rely only on easily quantifiable information (*i.e.* use sociocultural metrics for social and civic life to contrast classical economic measures based on monetary transactions). Such alternative and subaltern uses of the conceptualisation and methodology outlined in this article deserve their own enquiries and pilot applications in the future.

The use of local journalism, in particular, and NLP, in general, as a source of sociocultural data has multiple limitations. Particularly if misunderstood or abused, this conceptualisation and use of sociocultural data can have negative implications and raise ethical dilemmas. First, it still needs to be clarified to what extent hyperlocal press outlets depict day-to-day social and cultural activities in a representative manner. The underlying assumption of this conceptualisation and methodology is that the content produced by media outlets at least somewhat depicts public opinions and sociocultural uses of urban places (Ozgun & Broekel 2021) This assumption needs to be confirmed with greater research. Second, using journalistic material to infer sociocultural data can further empower key opinion-shaping newspapers and publishers that use their journalistic activities for broader policy goals. Even if the sources of newspaper data are diversified (which might not be possible everywhere) and biases seem to be mitigated through data cleaning and preparation, different media companies might have similar nested interests that influence their understanding of what is considered 'newsworthy'. As commercial enterprises, publishers strongly rely on funding from advertisers and possibly benefit from exaggerating problems to attract a greater readership and attention on social media platforms (*e.g.* through sensationalistic 'click baiting'). If journalism becomes the only source of sociocultural data, omissions of 'not-newsworthy' events might lead to new data gaps worsened by a facade of completeness in available data. In this sense, expanding the corpus with more sources and critically examining the geography of news coverage is crucial to avoid gatekeeping of news outlets and publishers. Third, including more sources, however, such as blogs from civil society organisations and content from social media platforms, might encounter limits linked to geographical text analysis and NLP. In including a greater diversity of text corpora, the (geographical and social) contexts for which the texts are written also become more diverse. As NLP cannot grasp irony and multiple meanings across various contexts properly, errors in geo-parsing and topic modelling will likely increase. Geographical identifiers, for instance, might refer to various scales, and place names can mean different locations depending on context. This pilot study accounted for these problems by excluding or re-defining problematic place names manually and by focusing on a neighbourhood-level newspaper with content on the Hamburg area only. However, such issues might be exacerbated in larger, less prespecified corpora (*i.e.* notably in social media texts where context is often implicit). Fourth, using NLP combined with spatial analyses generally enables new possibilities for government surveillance. As with other data collection and analysis methods, using newspaper articles to infer sociocultural data raises ethical debates regarding the right to privacy and the general desirability of data-based urban governance.

Making governance decisions solely based on subjective sociocultural data would be unreasonable. Yet, sourcing local journalism as an additional source of data may prove helpful in considering intangible sociocultural aspects of urban places in decision-making. In this sense, sourcing sociocultural data from local newspapers might also lead to general reflections regarding the limited diversity of data sources in data-based governance. In addition, explicitly subjective data should support an understanding of data-based governance that aims to provide optimal information for human judgment as the final case-to-case decision-making instance rather than taking automated decisions. In a careful and transparent application, adding data layers depicting previously excluded subjective sociocultural aspects of urban life would thus improve data-based governance.

NOTES

[1] See https://www.hamburg.de/kostenlos/8284674/kostenlose-stadtmagazine/.

[2] These newspapers are also being tested as possible sources of sociocultural data. However, as these newspapers focus less on Hamburg and its neighbourhood life, applying NLP to them is technically challenging and less likely to provide a significant amount of data.

[3] See https://spacy.io/.

[4] See https://nominatim.org/.

[5] For example, the place name 'Hamburg' is mentioned in 1665 articles, leading to 1665 data points being geocoded at the city centre, regardless of whether the articles actually refer to the city centre.

[6] See https://pypi.org/project/textblob-de/.

[7] See https://pypi.org/project/vaderSentiment/.

[8] Defined as being mentioned at least twice more during the period 2020–21 than during 2018–19.

[9] Defined as being mentioned at half as much during the period 2020–21 than during 2018–19.

ACKNOWLEDGEMENTS

This study was part of the 'LIG-Finder II' research project, which aims to develop new tools and methods for data-based urban land governance. Special thanks to Jörg Rainer Noennig, who made this project possible. The authors also thank the participants at the special issue's paper workshop, and the two anonymous reviewers for their extensive and helpful comments on earlier versions of this paper. However, the authors are solely responsible for the content of this article.

AUTHOR AFFILIATIONS

Filipe Mello Rose orcid.org/0000-0002-8348-5028
Digital City Science, HafenCity University Hamburg, Germany; Laboratory of Knowledge Architecture, Technical University Dresden, Germany

Juiwen Chang orcid.org/0000-0002-7896-5609
Digital City Science, HafenCity University Hamburg, Germany

COMPETING INTERESTS

The authors have no competing interests to declare. The *Landesbetrieb Immobilienmanagement und Grundvermögen Hamburg (LIG)* had no influence on the content of this article.

FUNDING

This research was made possible by the *Landesbetrieb Immobilienmanagement und Grundvermögen Hamburg (LIG)*.

REFERENCES

Agirdas, C. (2015). What drives media bias? New evidence from recent newspaper closures. *Journal of Media Economics*, 28(3), 123–141. DOI: https://doi.org/10.1080/08997764.2015.1063499

Battistini, A., Segoni, S., Manzo, G., Catani, F., & **Casagli, N.** (2013). Web data mining for automatic inventory of geohazards at national scale. *Applied Geography*, 43, 147–158. DOI: https://doi.org/10.1016/j.apgeog.2013.06.012

Batty, M. (2019). A map is not the territory, or is it? *Environment and Planning B: Urban Analytics and City Science*, 46(4), 599–602. DOI: https://doi.org/10.1177/2399808319850652

Batty, M., & **Yang, W.** (2022). *A digital future for planning spatial planning reimagined*. Digital Task Force for Planning.

Blei, D. M., Ng, A. Y., & **Jordan, M. I.** (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(January), 993–1022. DOI: https://doi.org/10.5555/944919.944937

Bunders, D. J., & **Varró, K.** (2019). Problematizing data-driven urban practices: Insights from five Dutch 'smart cities'. *Cities*, 93, 145–152. DOI: https://doi.org/10.1016/j.cities.2019.05.004

Cai, M. (2021). Natural language processing for urban research: A systematic review. *Heliyon*, 7(3), e06322. DOI: https://doi.org/10.1016/j.heliyon.2021.e06322

Cambridge University Press. (n.d.). Sociocultural. In *Cambridge advanced learner's dictionary & thesaurus*. https://dictionary.cambridge.org/dictionary/english/sociocultural

Chen, J., Becken, S., & **Stantic, B.** (2021). Harnessing social media to understand tourist mobility: The role of information technology and big data. *Tourism Review*, 77(4), 1219–1233. DOI: https://doi.org/10.1108/TR-02-2021-0090

Coletta, C., & **Kitchin, R.** (2017). Algorhythmic governance: Regulating the 'heartbeat' of a city using the Internet of Things. *Big Data & Society*, 4(2), 205395171774241. DOI: https://doi.org/10.1177/2053951717742418

Dalton, C. M. (2019). Rhizomatic data assemblages: Mapping new possibilities for urban housing data. *Urban Geography*, 41(8), 1090–1108. DOI: https://doi.org/10.1080/02723638.2019.1645553

Earle, M., & **Hodson, G.** (2022). News media impact on sociopolitical attitudes. *PLoS ONE*, 17(3), e0264031. DOI: https://doi.org/10.1371/journal.pone.0264031

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, 12(2), 219–245. DOI: https://doi.org/10.1177/1077800405284363

Gentzkow, M., & **Shapiro, J. M.** (2010). What drives media slant? Evidence from U.S. daily newspapers. *Econometrica*, 78(1), 35–71. DOI: https://doi.org/10.3982/ECTA7195

Ghermandi, A., & **Sinclair, M.** (2019). Passive crowdsourcing of social media in environmental research: A systematic map. *Global Environmental Change*, 55, 36–47. DOI: https://doi.org/10.1016/j.gloenvcha.2019.02.003

Gregory, I. N., & **Paterson, L. L.** (2020). English language and history: Geographical representations of poverty in historical newspapers. In *The Routledge handbook of English language and digital humanities* (pp. 418–439). Routledge. DOI: https://doi.org/10.4324/9781003031758-22

Institut für Medien- und Kommunikationspolitik. (2022). Ranking—Die zehn größten deutschen Medienkonzerne 2021 (June). *Statista*. https://de.statista.com/statistik/daten/studie/194686/umfrage/die-10-groessten-medienkonzerne-in-deutschland/

Jacobs, J. (1992). *The death and life of great American cities* [1961]. Vintage/Penguin Random House.

Jahedi, S., & **Méndez, F.** (2014). On the advantages and disadvantages of subjective measures. *Journal of Economic Behavior & Organization*, 98, 97–114. DOI: https://doi.org/10.1016/j.jebo.2013.12.016

Kahneman, D., & **Krueger, A. B.** (2006). Developments in the measurement of subjective well-being. *Journal of Economic Perspectives*, 20(1), 3–24. DOI: https://doi.org/10.1257/089533006776526030

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 2053951714528481. DOI: https://doi.org/10.1177/2053951714528481

Kitchin, R., & **McArdle, G.** (2016). What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, 3(1), 2053951716631130. DOI: https://doi.org/10.1177/2053951716631130

Kitchin, R., & **McArdle, G.** (2017). Urban data and city dashboards: Six key issues. In *Data and the city* (Vol. 1). Routledge. DOI: https://doi.org/10.4324/9781315407388-9

Lomborg, S., Dencik, L., & **Moe, H.** (2020). Methods for datafication, datafication of methods: Introduction to the Special Issue. *European Journal of Communication*, 35(3), 203–212. DOI: https://doi.org/10.1177/0267323120922045

Macků, K., Caha, J., Pászto, V., & **Tuček, P.** (2020). Subjective or objective? How objective measures relate to subjective life satisfaction in Europe. *ISPRS International Journal of Geo-Information*, 9(5), art. 5. DOI: https://doi.org/10.3390/ijgi9050320

Maffei, S., Leoni, F., & **Villari, B.** (2020). Data-driven anticipatory governance. Emerging scenarios in data for policy practices. *Policy Design and Practice*, 3(2), 123–134. DOI: https://doi.org/10.1080/25741292.2020.1763896

Mejias, U. A., & **Couldry, N.** (2019). Datafication. *Internet Policy Review*, 8(4). https://policyreview.info/concepts/datafication. DOI: https://doi.org/10.14763/2019.4.1428

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., & **Kompatsiaris, Y.** (2018). Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems*, 36(4), 40:1–40:27. DOI: https://doi.org/10.1145/3202662

Ozgun, B., & **Broekel, T.** (2021). The geography of innovation and technology news—An empirical study of the German news media. *Technological Forecasting and Social Change*, 167, 120692. DOI: https://doi.org/10.1016/j.techfore.2021.120692

Paterson, L. L. (2020). Mapping austerity: Geographical text analysis of UK place-names in *The Guardian* and *The Daily Telegraph*. In *Multimodal approaches to media discourses*. Routledge. DOI: https://doi.org/10.4324/9780367332907-4

Porter, C., Atkinson, P., & **Gregory, I.** (2015). Geographical text analysis: A new approach to understanding nineteenth-century mortality. *Health & Place*, 36, 25–34. DOI: https://doi.org/10.1016/j.healthplace.2015.08.010

Rettberg, J. W. (2020). Situated data analysis: A new method for analysing encoded power relationships in social media platforms and apps. *Humanities and Social Sciences Communications*, 7(1), art. 1. DOI: https://doi.org/10.1057/s41599-020-0495-3

Rieder, G., & **Simon, J.** (2016). Datatrust: Or, the political quest for numerical evidence and the epistemologies of big data. *Big Data & Society*, 3(1), 2053951716649398. DOI: https://doi.org/10.1177/2053951716649398

Roberts, M. E., Stewart, B. M., & **Tingley, D.** (2019). stm: An R Package for structural topic models. *Journal of Statistical Software*, 91, 1–40. DOI: https://doi.org/10.18637/jss.v091.i02

Sarın, P., & **Uluğtekin, N.** (2019). Analyzing newspaper maps for earthquake news through cartographic approach. *ISPRS International Journal of Geo-Information*, 8(5), art. 5. DOI: https://doi.org/10.3390/ijgi8050235

Scheufele, D. (1999). Framing as a theory of media effects. *Journal of Communication*, 49(1), 103–122. DOI: https://doi.org/10.1111/j.1460-2466.1999.tb02784.x

Swyngedouw, E. (2005). Governance innovation and the citizen: The Janus face of governance-beyond-the-state. *Urban Studies*, 42(11), 1991–2006. DOI: https://doi.org/10.1080/00420980500279869

Taylor, L., & **Richter, C.** (2015). Big data and urban governance. In J. Gupta, K. Pfeffer, H. Verrest, & M. Ros-Tonen (Eds.), *Geographies of urban governance* (pp. 175–191). Springer. DOI: https://doi.org/10.1007/978-3-319-21272-2

Valdez, A.-M., Cook, M., & **Potter, S.** (2018). Roadmaps to utopia: Tales of the smart city. *Urban Studies*, 55(15), 3385–3403. DOI: https://doi.org/10.1177/0042098017747857

Voukelatou, V., Gabrielli, L., Miliou, I., Cresci, S., Sharma, R., Tesconi, M., & **Pappalardo, L.** (2021). Measuring objective and subjective well-being: Dimensions and data sources. *International Journal of Data Science and Analytics*, 11(4), 279–309. DOI: https://doi.org/10.1007/s41060-020-00224-2

Wang, J., Hu, Y., & **Joseph, K.** (2020). NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages. *Transactions in GIS*, 24(3), 719–735. DOI: https://doi.org/10.1111/tgis.12627

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

TO CITE THIS ARTICLE:
Mello Rose, F., & Chang, J. (2023). Urban data: harnessing subjective sociocultural data from local newspapers. *Buildings and Cities*, 4(1), pp. 369–385. DOI: https://doi.org/10.5334/bc.300

Submitted: 10 February 2023
Accepted: 05 June 2023
Published: 30 June 2023

Buildings and Cities is a peer-reviewed open access journal published by Ubiquity Press.