Estimating socioeconomic indicators using online data

by

Sirasit Lochanachit

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Warwick Business School

June 2020

Contents

List of Tables

List of Figures

List of Symbols and Abbreviations

Symbols

Linear regression

Y	Output, dependent, response or outcome variable
X	Input, independent, predictor, feature, or explanatory variable
p	Number of variables
β	Coefficient or parameter
ϵ	Error term
N	Normal distribution
μ	Mean of the distribution
σ^2	Variance of the distribution
$\hat{\beta}$	Estimated coefficient from the model
$\hat{y}$	Estimated value of the response
e	Error/residual - The difference between the actual outcome and the predicted outcome

Linear mixed effects model

α_i	Intercept/mean of the random-effect variable
s	Fixed effect variable

Model accuracy

r	Pearson correlation
Cor	Correlation between variables
n	Number of observations
$\bar{x}$	Mean of variable x
$\bar{y}$	Mean of variable y
x_i	The ith observation of variable x
y_i	The ith observation of variable y
$\sum$	Summation
R^2	R-squared
f_i	Forecast of ith observation of variable y
p_i	Percentage error of ith observation

Ridge, LASSO, and elastic net

λ	Tuning parameter
j	The jth parameter
α	Elastic net penalty

Logistic regression

p	Probability
e	Euler's number
L	Likelihood
$\prod$	Product

Model selection

L	Maximised value of the likelihood function
d	Total number of parameters
σ^2	Estimate of the variance
n	Number of data points

Kendall's rank correlation

τ	tau correlation coefficient
N	Number of data points
p	p-value

Autoregressive model

ΔY_t	Change in unemployment rate from the previous month
ΔX_t	Change in *Google Trends*'s search volumes from the previous year
Φ	Regression coefficients of changes in search volumes of *Google Trends*
β	Regression coefficients of changes in unemployment rates
c	Intercept/mean

Abbreviations

AIC	Akaike Information Criterion
API	Application Programming Interface
AR	Autoregressive model
ARIMA	Autoregressive Integrated Moving Averages model
BCP	Best Current Practices
BIC	Bayesian Information Criterion

CLD	Chromium Language Detector
DM	Diebold-Mariano model accuracy test
FDR	False Discovery Rate
GLM	Generalised Linear Model
GPS	Global Positioning System
IP	Internet Protocol
JSA	Jobseeker's Allowance
LASSO	Least Absolute Shrinkage and Selection Operator
LFS	Labour Force Survey
LOOCV	Leave-one-out Cross-validation
LSOA	Lower Layer Super Output Area
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MSOA	Middle Layer Super Output Area
OLS	Ordinary Least Squares
ONS	Office for National Statistics
PCR	Principal Component Regression
RMSE	Root Mean Squared Error
RSS	Residual Sum of Squares
TSS	Total Sum of Squares
URL	Uniform Resource Locator
VPN	Virtual Private Network

Acknowledgments

I would like to express my greatest appreciation to Professor Suzy Moat and Professor Tobias Preis for their continuous support throughout my PhD career. Their guidance, patience, and knowledge have been valuable assets that contribute greatly to the completion of this thesis.

I would also like to thank the Thai government for their sponsorship of my PhD studies. Without their funding, my PhD journey would not have been possible.

Special thanks to my girlfriend for giving me the strength, motivation, and endless moral support through challenging times.

Also, I would like to thank God for everything in my life.

Finally, I wish to thank my parents for their love, encouragement, and for always believing in me.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

Abstract

Policymakers and businesses need a good understanding of the current state of society to make fully informed decisions. In contrast to traditional approaches to measuring human behaviour, which can be expensive, time-consuming and subject to delay, data on collective online behaviour, such as what people are searching for on *Google*, is available publicly, rapidly and at low cost. Studies into online behaviour may therefore be able to provide useful insights into collective human behaviour in the real world.

Here, we investigate whether online data from social media platforms, such as *Instagram* and *Twitter*, and search engine data, specifically data from *Google*, can help estimate key characteristics of society. In particular, we seek to infer the number of people speaking various languages across different urban areas based on publicly exchanged messages on the photo-sharing platform *Instagram*. We find that such data can help estimate the spatial distribution of language usage in Greater London. In a parallel analysis, we investigate whether *Twitter* data is similarly useful. However, our results suggest that data from *Instagram* is more valuable, as a higher number of posts to the service contain location data.

We also investigate whether online data can be used to help estimate economic activity. Specifically, we focus on unemployment rates in the United Kingdom and draw on data retrieved from *Google Trends*. Our findings reveal that *Google* search data can help generate quicker estimates of the current level of unemployment before official data is released. We also find that, according to some performance metrics, a variable selection technique based on an elastic net can improve model performance.

This thesis highlights the potential for inferences generated from online data to complement official statistics, for example by providing quicker estimates before official figures are released. We suggest that rapid, low-cost measurements of collective human behaviour from publicly available data may provide valuable new insights for policymakers and businesses alike.

Chapter 1

Introduction

In this digital age, people are actively engaging in online activities such as searching for information, communicating, and sharing stories. These activities generate online data that are being collected by online service providers, resulting in huge datasets that contain information on human behaviour at a collective level. Furthermore, some of these datasets are publicly available and can be accessed at low cost. For researchers, this becomes a new source of information opening up new opportunities for studying human behaviour and society at a large scale.

Gaining a better understanding of current human activity patterns at the collective level is crucial for shaping the future. Governments and businesses need to make decisions and require up-to-date information that captures a detailed snapshot of our society. Many key measurements of society are currently obtained from surveys. For example, in England and Wales, the Census is carried out every decade to collect valuable measurements of the status of society [1]. However, the time and effort required for such surveys mean that results are only published after a long period of time [2–6]. Official statistics bodies around the world have been looking for alternative approaches to estimate statistics that is cheaper and quicker to publish, complementing more traditional approaches [7–10]. Therefore, it is increasingly important to obtain the best possible estimates in a timely fashion to support decision making on important matters.

Text, images, audio and video files are all forms of data that people have been generating through online activities. To be able to gain a better understanding of these large amounts of data, appropriate methods for data collection, analysis, and presentation are necessary. For example, spatial data could be used to investigate the question "How many people are employed in each boroughs of London?". In contrast, time series data could be used to ask "How many people are employed in

London during each month or year?".

Appropriate methods are required to work with spatial and temporal data. Online data from social media platforms and search engines can provide such data in order to produce statistics across space and statistics across time to understand human behaviour at a collective level. Based on the types of data described, this thesis will investigate statistics across space and statistics across time using online data and official statistics. First, it will investigate one example that involves spatial data. In relation to this, *Instagram* is a large social media platform that allows users to upload photos and videos which can include information on where the photos and videos were taken. *Twitter* is another platform in which users can post and interact with short messages or 'tweets'. Similar to *Instagram*, users can attach a real-world location to their tweets at the time of posting. For both platforms, information on the real-world location can be either specified directly by users or determined automatically by GPS enabled device such as a smartphone. Photos or tweets that contain geographic location details are also known as geotagged photos or tweets. By using location information from these two data sources, a spatial data analysis can be conducted to produce statistics across space. This thesis will also investigate another example that involves estimating official statistics across time. *Google* is an online search engine that allows users to submit queries to search for information online. Aggregated search volume data on queries from *Google* are available in a time series format, which can be used to produce statistics across time.

Computational social scientists have recently started to exploiting the advantage of the large amount of online datasets to investigate human behaviour in the real world. Chapter 2 explores a wide range of examples pf previous research that study the potential of online data to infer human activity patterns at a collective level, including obtaining quicker estimates of important measurements, such as incidences of influenza. We cover a number of studies demonstrating that data from social networking websites such as *Facebook*, social media for microblogging (e.g. *Twitter*), photo sharing platforms (e.g. *Instagram* and *Flickr*) and search engines and online encyclopedias (e.g.*Google* and *Wikipedia*) can be used to gain a better understanding of collective human behaviour. These studies also highlight the usefulness of online data which can complement current approaches and help obtain quicker estimates of national statistics. Chapter 3 reviews the UK's official statistics data sources for both language usage and unemployment rates. We then describe statistical approaches used in the thesis. This chapter also provides information on various error measures and model selection techniques that will be used to quantify the usefulness of online data and compare its estimating performance in terms of accuracy.

To begin with, this thesis will investigate the online photo sharing platform, *Instagram*, to identify the relationship between online data and official statistics. There are currently more than one thousand spoken languages across the world. The ability to estimate how many people speak a particular language, such as Japanese, across regions would help policymakers in planning local public services, including language translation needs. Language usage statistics could also provide insights into the cultural population of local areas, which can be of value for businesses when making inferences about potential customers. In politics, early estimates of statistics on language usage could help potential politicians to start a campaign aiming to support cultural communities and increase the number of cultural facilities, which could subsequently bring social and economic benefits to the area. In Chapter 4, we start with estimating national statistics across space, specifically language usage statistics across areas in Greater London and Greater Manchester, using *Instagram* data. We focus on languages that are commonly used on *Instagram*. Our findings reveal that there are some languages in our *Instagram* data that can be used to help generate estimates, complementing the official data from the Census, which is published every ten years.

Apart from *Instagram*, we consider another online data source, *Twitter*, to evaluate whether it can be used to help obtain quicker estimates, including exploring broader spatial units of measurement. In Chapter 5, we build on and extend the analysis in Chapter 4 by using geotagged *Twitter* data from Greater London Our results show that, overall, *Twitter* data provides improvement in estimates more than baseline models although it has a weaker effect compared to the *Instagram* analysis. This may be due to a lower availability of posts with geotagged coordinates on *Twitter* compared to *Instagram*. In addition, we conduct the analysis on place-tagged tweets published to estimate the number of people who speak a particular language across different boroughs in Greater London. Our results reveal there is only one language for which there is evidence that *Twitter* data can provide better estimates than the baseline model for borough-level analysis.

Overall, this thesis provides evidence that *Instagram* and *Twitter* data reflect Census data for some languages. Social media data may therefore be useful for estimating the size of smaller cultural communities in urban settings by measuring language usage. In contrast to the spatial online data used to infer how languages are used in different parts of cities, we also investigate the potential of using online data to estimate national statistics across time. The unemployment rate is a key economic indicator, as it can be used to track economic performance. However, in the UK, it is usually released with two months delay. For example, in mid-May 2020, the most recent data relates to February 2020. In particular at times of crisis, it is clear that data that is several months old does not provide sufficient insight into the current economic wellbeing of the country. Quicker measurements allow policymakers to rapidly plan labour market policies. It also allows businesses and jobseekers to monitor up-to-date labour market performance, including periods of recession and recovery.

In Chapter 6, we investigate the usefulness of *Google* search data from January 2004 to February 2017 in generating current estimates of unemployment rates in the United Kingdom, before official figures are released. Obtaining early estimates before the official release of such figures is known as nowcasting. Search data obtained from *Google Trends* are used to complement the official data in the model to generate the estimates. We find that the nowcasting models incorporating search data generate more accurate estimates than a model with official data only. Our results suggest that *Google* search data can help nowcast the UK's unemployment rate. Previous research has focused on a single search term or small groups of *Google* search terms. To obtain a broader set of search terms provided by *Google Trends*, we consider a novel machine learning and regularisation technique. This technique selects a number of relevant or important *Google* keywords to be used in nowcasting models. Our findings show that according to some performance metrics, variable selection techniques, especially an elastic net, can help select the optimal model to estimate current unemployment rates. This underlines the usefulness of machine learning technique on selecting relevant keywords objectively, providing quicker estimates of key economic indicators at low cost.

Finally, in Chapter 7, the main conclusions of this thesis are presented.

Chapter 2

Background

As introduced in Chapter 1, the current approach to retrieve key measurements of the society (e.g. unemployment rates or language statistics) requires human effort to collect and process data, which can be time-consuming and costly. The availability of online data opens up new opportunities to study human behaviour and society at a large scale. Therefore, this thesis will investigate whether online data can be used to estimate national statistics across space and statistics across time. Statistics across space requires spatial data, such as the Census, to analyse while time series data, such as historical monthly figures of unemployment rates, is necessary for statistics across time.

Traditionally, the official figures representing the current state of the economy are normally released with a delay. The figures are often in a time series format, revealing trends and magnitudes of key economic quantities over time. Alternative data sources, such as business surveys, are published more quickly than the official figure and they can be used to monitor current activity. Economists, therefore, tackle this challenge by looking for a signal of change in direction using other data sources which are accessible and more recent than the official figures. This approach, known as nowcasting [4, 11, 12], allows economists to predict the present to obtain early estimates before the official figure is released. The fundamental of nowcasting is building a regression model to obtain the estimates given the training data - either official figure or other data sources or both. Even though the official figure has a time lag, the more recent data from other data sources can be used to estimate the missing gap. In relation to research, nowcasting opens a new opportunity for researchers to use other data sources, which can help close publication gaps in official figures, to investigate and gain new insights into the current state of society. Recent studies have investigated whether online data from search engines and social media might provide a good indicator before official figures are available at low cost [4, 5, 13–21]. It has been revealed that online data can complement official data in obtaining the best possible current estimates. Thus, nowcasting with online data can potentially help close reporting gaps in official statistics.

This thesis considers whether online data could be used to provide quicker estimates by analysing data obtained from social media platforms such as *Instagram* and *Twitter* and search engines (e.g. *Google*). To gain a better understanding of the potential of online data, the following section explores a wide range of relevant previous research statistics. This chapter will begin with exploring two broad categories of online data that are publicly available: Social media data and data on online information gathering. In each category, it will start with describing previous studies that have used online data to gain new insights, it will be followed by studies that have specifically considered nowcasting to obtain quicker estimates.

1 Social networking data

The advancement of communication devices and Internet technology, such as email, mobile phones, and instant messaging, has allowed people to interact instantly without the need to physically meet. This also includes social networking where users can engage in social interactions in a public manner. These activities generate data and these datasets are being collected by online service providers. Social scientists are interested in gaining new insights into human behaviour at a large scale from these new data sources, which are publicly available at low cost. Social networking services such as *Instagram* and *Twitter* provide a platform on which people can express their opinions and share their stories through short text messages, photos, audio and video messages in any language. This section will describe previous studies that have used online data from various social media platforms such as *Instagram*, *Twitter*, *Facebook* and *Flickr* to obtain new insights into collective human behaviour. Based on each online data source, it will also explore previous studies that have focused on obtaining best possible estimates, including nowcasting.

1.1 Instagram photos

Instagram is another large social media platform with almost a billion users around the world. *Instagram* was founded in 2010 as an online platform which allows users to upload photos and videos with the option to include geographic location. *Instagram* data is made publicly available through the *Instagram* API but with restricted access, users need to request to access first.

Previous research has argued that *Instagram*'s social network properties are different to *Twitter* and *Flickr* as users typically post once a week, and users prefer to use the share location feature with other users to indicate the location the photo was taken [22]. Previous studies on data from *Instagram* have investigated its potential in estimating mobility patterns deriving from geolocation that are tagged with the uploaded picture [23]. Mobility patterns involves monitoring people move from place to place. These patterns can be estimated based on the real-world geographic location that are indicated either directly by the users on the platform or from the mobile devices that have GPS (Global Positioning System) technology enabled. By introducing techniques to analyse the social network and visualising the online data, the studies suggested that *Instagram* datasets might be suitable for finding popular regions of cities, be able to capture cultural signature behaviour, and are less susceptible to changes over time. The authors also suggest that *Instagram* data can complement other online data sources to provide more information about human movement patterns in the city and urban social behaviour. Similarly, another study estimated the number of attendees during a football match based on the number of *Instagram* users who shared photos from within the stadium during the time period of the event [24].

Another research area involving *Instagram* datasets are socio-cultural characteristics [25–29]. For example, colour usage and hue intensities patterns in images were compared between those from New York City and Tokyo [26]. Visual data of *Instagram* photos can imply specific cultural characteristics based on colour and hue identification which are different between New York City and Tokyo. Another study visualised *Instagram* data across 13 different cities to analyse data at multiple spatial and temporal scales. Analysing the photos visually, the researchers explored how image composition in *Instagram* photos can have impact on the relationship between the viewers and the photographers [28]. It also increases the viewer's desire in social connection and to share their own experience through photos. Interestingly, a recent study has shown that photos with faces are more likely to gain attention from other users regardless of the number of faces appearing in the photo; age; and gender [27]. Previous research has analysed *Instagram* photos and corresponding texts (e.g. captions and comments) containing healthy lifestyle hashtags. [29]. The results suggested that such images are associated with positive feelings and promote a healthy lifestyle through body appearance of the subject.

Geographic data from *Instagram* images shows a specific spatial distribution of images and main hotspots, which implies possibilities to generate information on citizens' preferred locations such as green space or natural park [30]. This information can support city planners when making decisions about policies and transportation as people's preferences are captured in online data.

Through analysing the textual information of *Instagram* photos and labelling photos subjectively by a human, another study reveals that *Instagram* posts with a negativity percentage from 60% to 70% are less likely to encounter online bullying [31]. Moreover, through linguistic and psychological analysis, *Instagram* posts with certain linguistic contents such as death, appearance, religion, and sexuality as well as image contents relating to drugs are highly associated with cyberbullying. Lastly, a machine learning model can identify cyberbullying based on textual information only while image data can be used to detect occurrences of cyberbullying.

Having examined previous research using *Instagram* data, the amount of research relating to nowcasting is limited. Moreover, modelling spatial data using vast quantities of geotagged data, including texts, from *Instagram* photos has not yet been looked at. Therefore, there is an opportunity to use *Instagram* photos data in this way.

1.2 Twitter data

Twitter is an online platform where users can share opinions publicly via short messages, called "tweets". A tweet was previously limited to 140 characters in order to encourage active engagement with easily digestible information rather than long pieces of text. This was changed in 2017, and the limit is now 280 characters. Users can choose to follow other users by using a subscribing feature called "following". One of the widely used features of *Twitter* is retweeting, allowing tweets to be shared by other users with a single click. The nature of retweets enables rapid transfer of information to the public. *Twitter* data are available through the *Twitter* public API (Application Programming Interface). This has allowed many researchers to analyse *Twitter* messages.

The online flow of information based on where the message is sent to can disclose fundamental social, political, and economic behaviour - trends and patterns. However, inferring both language and location from short messages imposes an issue since the best method to do this remains unclear.

Determining exact location of the user at the time of submitting the tweet based on a tweet is a significant challenge. One way to obtain this information is to retrieve profile information that is specified by a user (e.g. 'London, United Kingdom') when setting up a *Twitter* account. However, a user can type in any text, which in turn can be in any language. Thus, it is difficult to obtain an accurate geolocation from profile location alone. Alternatively, some researchers narrow down tweets by restricting theirs analysis to tweets that come with geolocation - the real-world location of the user at the time of the tweet as determined by the GPS enabled device. Geolocation information depends on the user's privacy settings for their *Twitter* account. A single tweet contains either a pair of latitude and longitude coordinates representing an exact real-world location or a rectangular bounding box depicting approximate location or both. Tweets that contain latitude/longitude coordinates are known as "geotagged" tweets [32]. A bounding box is a set of coordinates that covers the geographic area as a rectangular box rather than a single coordinate point. The exact location information is provided by the device that was used to upload the tweet message. Specifically, this information is inherited from either the Global Positioning System (GPS) implemented by the user's device or by finding the user's location via the Internet Protocol (IP) address [33, 34]. However, the user's IP address can be masked by using a Virtual Private Network (VPN) or the Tor browser which preserves users' anonymity online [35].

Researchers have found that geotagged tweets with precise location constitute a small proportion only [33]. Specifically, out of 19 million tweets, less than one percent have geolocation information. This poses a challenge of forming a representative sample of the broader population in terms of content in messages as the number of data points is cut down. Furthermore, the sample could be biased by factors such as age and level of income. Having collected 144 million geotagged tweets in the US, containing 2.6 million unique users, a previous study investigated the relationship between counts of unique *Twitter* users and 2010 Census population counts [36]. By using the smallest granularity level available from the Census, the results reveal that the geotagged tweets are non-randomly distributed over the US population. As a result, it is suggested that geotagged tweets in the US have a population bias. In addition, they are influenced by a higher level of income, the preference of younger people, and whether area is urban or not. Therefore, the low proportion of geotagged tweets in comparison to overall tweets poses a challenge for researchers, as they have to find a method to infer locations from non-geotagged tweets, such as tweets with bounding box coordinates that are tagged by *Twitter*.

To determine language of short messages, methods employed on *Twitter* data in previous studies vary. *Twitter* has its own language classifier in accordance to Best Current Practices (BCP 47) for the Internet community. The key challenge of identifying the language of short texts or tweets is that language classifier algorithms are often trained on long sentences or whole documents [37]. Moreover, the use of URLs and hashtags in tweets introduces complexity. URLs and hashtags were found to be strongly related with retweet rates amongst 74 million tweets [38]. In addition, more than half of all retweeted tweets contain URLs, while only 19 percent of all tweets have URLs.

Some researchers aggregate tweets based on user account to create longer messages so that the classifier can produce a more accurate answer [37]. Alternatively, there is an automated tool called Compact Language Detection (CLD) which was developed by *Google*. After providing a text in any language to CLD, it reports one or more detected languages. It also indicates the corresponding percentage share of the detected language in the original text. The CLD has been used in detecting languages in blog parts [39]. However, the CLD's accuracy for short messages remains unclear.

Due to the vast quantity of tweets being made available, numerous studies using *Twitter* messages have investigated a variety of research areas applying different techniques for collecting location and language data on *Twitter*. Human behaviour can be influenced by opinions as people's beliefs and perceptions of reality, and the decisions we make, are largely conditioned on how others perceive and evaluate the world. People usually seek others' opinion before making an informed decision and this also applies to organisational decision making. Through collecting 1.6 million tweets for two months in 2009, researchers have investigated social influence on user's interactions, such as the spreading of information, based on user characteristics - number of followers, friends, and tweets [40]. Based on a linear regression tree and resampling techniques, their results reveal that having a large number of followers is one of the main factors that could lead to a higher number of retweets. A logistic regression model built on 10,000 randomly sampled tweets estimated the probability of retweeting. The number of followers was found to correspond with the amount of retweets, implying social influence [38].

Two of the active fields are sentiment analysis and opinion mining which analyse people's attitudes and emotions towards various entities such as individuals, objects, and services [41]. *Twitter* data have been used to determine positive, negative and neutral sentiments using natural language processing techniques [42–44]. In contrast, the collective mood (positive/negative, calm, alert, sure, vital, kind, and happy) classified in tweets through mood detecting tools has been used to estimate the stock market, specifically the Dow Jones Industrial Average [45]. Mood states were linked to signal changes in stock market values. Another study found that the sentiment expressed in tweets can be used to predict information flow relating to terrorist events [46]. Positive and supporting tweets tend to be shared more often than negative tweets following a terrorist event.

One of the key characteristics of social media is that users have the capability to become active content producers. Therefore, *Twitter* is also widely used as a news sharing platform since users can participate in news production and diffusion. People who are interested in certain topics such as sport and politics are likely to express opinions and emotions through tweets. In the US, researchers have found that users tend to be supportive of those who have similar political view through retweeting [47]. Also, politicians who have extreme ideological positions have a large number of followers [48]. Using text analysis software on 100,000 tweets relevant to politics, researchers found evidence that the number of messages mentioning a political party corresponds with the election result in Germany, suggesting the user-generated content as an indicator of political sentiment [49].

A growing number of studies in recent years have studied language statistics across space using *Twitter* data. Previous studies in this field have focused on determining the language and geographic location of tweets [33], and mapping languages across different geographic scales ranging from country to city level [50]. Another study compared the most commonly used languages within the top ten countries that tweet the most in a particular year [51]. Different communication patterns of eight popular Twitter languages were investigated [52]. The effect of language usage on online social ties [53] and the average length of tweets across different regions [54] were also studied. These studies suggest that spatial online data, especially *Twitter*, which is a popular data source for scholars, has the potential to estimate language statistics. Nevertheless, the question whether the distribution of languages retrieved from online data reflects official statistics data such as the Census remains unclear.

This section introduced several approaches to detect language and location in short messages. However, the objective of this thesis is not to compare the best approach in identifying them, but to highlight the importance of using such methods to gain an understanding of the spatial distribution of languages on *Twitter*. Moreover, having discussed previous studies focusing on *Twitter* data, it remains unclear whether language usage in tweets can improve official estimates. Also, modelling spatial data using vast quantities of geotagged tweets and bounding box coordinates has not yet been looked at. Therefore, this presents an opportunity to investigate the relationship of language patterns in official statistics and online data, specifically *Twitter* data.

Nowcasting with Twitter data

Online data opens up a new opportunity for researchers to investigate whether online data can be used to generate quicker estimates of key quantities, which might otherwise only be known at a later date, which is commonly referred to as "nowcast".

This would allow academics to gain early estimates of collective human behaviour in the real-world. One example for such a behaviour is how people move from one location to another. The capability of online data to infer such mobility patterns led to a range of studies using *Twitter* data [55–57]. Previous research has shown that *Twitter* data can also be used to infer levels of rainfall in a given location and time [58].

One of the active fields aiming to nowcast with tweets is disease detection and monitoring. Amongst various diseases, Dengue is a viral disease transmitted by mosquitoes, which can lead to fatality. It is difficult to predict since it is costly to build real-time monitoring systems in many cities and regions where this is a problem. In Brazil, reports of dengue cases are frequently delayed by 3 to 4 weeks and often longer [59, 60]. It has been suggested that *Twitter* data can provide quicker estimates, complementing traditional disease-surveillance systems, at low cost [61, 62]. Previous research has demonstrated that *Twitter* can be used as a real-time source for information on dengue activity at population level as tweets relevant to the disease have a positive correlation with the reported cases [60, 63].

By comparing *Twitter*, *Google Trends* and *Wikipedia* as potential online data sources for dengue activity, researchers have also discovered that there is a correlation between the number of disease mentions in social networks and physician visits [64]. Out of the mentioned online data sources above, the authors also found that *Twitter* can play a more crucial role to nowcast dengue activity at city and country level. A combination of these online data sources might offer more timely information on dengue activity which would benefit public health officials.

Twitter datasets are publicly available and therefore attracted many social researchers in various research subfields. In the context of nowcasting, many studies on human mobility patterns, natural events, and disease monitoring have been conducted using *Twitter* data. They underline the usefulness of online data in nowcasting real-world events before official figures are released. However, the number of studies on estimating language statistics across space are very limited. This presents an opportunity for this thesis.

1.3 Facebook data

Facebook is an online social networking website that allows users to share messages, comments, photographs and videos. These contents as well as user information can be shared with other users publicly or can be limited to a certain group of friends. *Facebook* had more than two billion users in June 2017 and it has attracted social scientists to study information from this very large data source [65]. However, *Facebook* heavily limits the access to their data due to privacy concerns. In order to overcome this issue, researchers started to collect user data through *Facebook* apps, which request user consent for accessing and analysing the data.

Facebook provides user content posted on the social network through the "News Feed" feature, which is a proprietary algorithm. It provides personalised content according to a user's information consumption preferences. A number of studies involving *Facebook* data have investigated the existence of a social media echo chamber, in which online users tend to consume information from sources that support their beliefs or other people who share similar viewpoints [66–69]. Having investigated video contents and user comments on *Facebook*, one study reveals that the two distinct and conflicting types of narratives (i.e. conspiracy-like and scientific news) encourage echo chambers [66]. In addition, through a statistical learning model using commenting patterns under posted videos, it can be determined which one of the two conflicting narratives a user prefers with good precision. Using quantitative analysis on *Facebook* data, researchers have found that online users tend to have a limited number of preferred news sources or *Facebook* pages, creating a community structure [67]. Another study suggests that exposure to news and opinions are more influenced by users' choices rather than the ranking of content by the *Facebook* News Feed algorithm [70]. Therefore, the effect of algorithms in encouraging the creation of echo chambers remains controversial as the algorithm is being developed dynamically.

The concept of echo chambers has been applied to other subjects such as the spread of information including false information known as "fake news". Recent studies on fake news detection analyse *Facebook* posts considering users' preferences to scientific or conspiracy-like pages. As a result, a post can be classified as "fake" or "non-fake" on the basis of users, who like the post [71]. By comparing scientific and conspiracy-like news that are published on *Facebook*, it is suggested that users who are continuously exposed to unverified rumours are more likely to spread the false information [68, 72–75]. These rumours are continuously circulating within communities that support such a viewpoint, and, as a consequence, become widespread and keep emerging on social platforms. Previous research provides evidence that confirmation bias plays a major role in the spreading of misinformation online [76, 77]. This can subsequently generate echo chambers and polarised communities that have similar patterns of information consumption [72, 74, 75, 77] and disregard contrary information that attempts to prove that their beliefs are invalid [78]. Additionally, engaging in a discussion with corrections tends to create negative sentiment in the polarised group [70, 73]. Using keyword extraction techniques, a previous study has provided evidence that echo chambers encourage the spread of false information in various categories including environment, diet, health and politics [79].

Facebook and its political effect on general election results has attracted attention from the public, specifically during the 2016 US presidential election. As a result, numerous studies investigated whether social media has played a crucial role in influencing political campaigns and election results. Social media was first used for political activities during the 2008 US presidential election. A study relevant to the mentioned election used linear regressions on undergraduate student data and revealed that political activity on *Facebook* has a positive relationship with political participation in the real-world [80], suggesting that political activity on *Facebook* helps increase in political participation. Through quantitative content analysis on a sample of *Facebook* groups that are associated with the 2008 US presidential candidates, a study has found evidence that engagement between users and positive/negative perspective towards presidential candidates also plays a role in the amount of support on the platform [81]. Since then, social media has become a critical tool for political campaigns worldwide and researchers have attempted to discover the effects of social media on users during elections, such as perception towards the candidates and participation in political activities, including election results and political campaign strategies [82–86]. *Facebook* allows politicians to quickly express ideas and share their private daily life to gain new followers and possible voters at low cost [82]. Reaching voters and targeting messages across different audiences in terms of political orientations (e.g. conservative and liberal) and interests are possible by paying *Facebook* for advertisements. Researchers have discovered that personalised political campaigns according to users' gender, geographic location, and political ideology can attract voters who were initially undecided [87]. A study on the UK 2015 general election revealed that the Conservative party used *Facebook* to target messages to specific voters, possibly exceeding spending limits [88]. This would undermine the principles of fair and open elections in the UK because political parties that have lower funds would reach fewer voters through *Facebook.* Through content analysis from US candidates' pages focusing on three categories - political advertising, emotional appeals, and social endorsement - a previous study has identified different strategies used by candidates such as attacking the opponent by highlighting the opponent's weakness or acclaiming (expressing candidate's strength and why people should vote for) during the 2008 and 2012 general elections [89].

Since 2016, *Facebook* has allowed users to express their emotional reactions when viewing posted content on the network. This allows social scientists to investigate the expressed emotion and sentiment using both reactions data and textual information. The text used in messages is usually classified as positive, negative or neutral, implying user's perception towards the content. Using deep learning techniques processing posted content and reactions from public pages of supermarket chains, a study has shown that textual content by users can be used to predict users' reaction to a new post (e.g. angry, love, and etc.). [90]. In addition to *Facebook* reactions, researchers have found evidence suggesting that emoticon usage, such as ":)", can be used to detect user sentiment as there is a positive relationship between emoticon usage and reactions [91]. Therefore, it is suggested that *Facebook* reaction data can help provide more information for emotion detection and classification complementing the analysis of textual messages [92].

Due to limited access on *Facebook* data, it is challenging to obtain *Facebook* data at population level including time series and geolocation data. Therefore, the majority of research on *Facebook* focuses on content consumption patterns, community networks, the spread of misinformation, and controversial issues, such as political campaign and social media influence. Moreover, studies that focus on nowcasting to obtain quicker estimates using *Facebook* are very limited. For these reasons, this thesis will evaluate other data sources that can be used to complement official statistics.

1.4 Flickr data

An online photo sharing platform called *Flickr* allows users to upload photos and share them amongst friends or other users. For each photo, *Flickr* records meta information of the photos including the location, time and camera settings with which it was taken, as well as uploaded title, description and tags. Users have the option to include location data when uploading from a smartphone or a camera in which GPS is enabled. Although users can manually add a geographic location after photos have been successfully uploaded to *Flickr*, it is likely that GPS enabled devices contribute the majority of geotagged photographs [19]. *Flickr* datasets are made publicly available. They can be accessed through an API.

Researchers have investigated *Flickr* as a potential online data source for various studies. A recent study examined whether *Flickr* photos can be used to quantify the presence of art in an urban area such as London [93]. The study demonstrates that the number of *Flickr* photos with the word "art" attached is positively correlated with relative increases in property prices within London. At the same time, mapping unpleasant and pleasant sound in urban areas is possible using geotagged *Flickr* photos with sound-related tags [94]. Researchers combined geotagged photo tags from *Flickr* and *Instagram* and geotagged tweets from *Twitter* to show that online data can be mapped to smell-related words, classified into ten categories, in urban landscapes [95]. These studies suggest that online data could help city planners and policy makers to create smart urban cities by gaining more understanding of environmental issues.

Nowcasting with Flickr data

A study on *Flickr*, identifying protests through photo descriptions across different countries, reveals a positive relationship between the number of photos that are tagged with protest and the number of protest reports in a newspaper [96]. This implies that online social media data can be used to detect real-world events. *Flickr* data has also been used to monitor collective interest in large-scale disasters, such as Hurricane Sandy in 2012. Researchers found that there is a relationship between the number of photos uploaded to *Flickr* with a relevant tag specific to Hurricane Sandy and the level of disaster severity during the incident [97].

Human mobility patterns can be investigated using the location information where a photo was taken and its timestamp which is attached to *Flickr* photos. The sequence of geotagged photos can infer users' travel patterns, including destinations, allowing researchers to study and gain more understanding of tourism [98]. In the area of tourism, a previous study suggests that *Flickr* data can help estimate visitor numbers to recreational locations worldwide such as natural parks and museums using *Flickr* photos [99]. Researchers have also discovered that geotagged photos from *Flickr* can be used to estimate tourism demand at the city level [100]. Furthermore, by inferring travel patterns based on geolocation information of *Flickr* photos, researchers have found evidence of a correlation between the number of visitors to the UK based on *Flickr* estimates and the official foreign visitors' estimates in the UK [19]. Similarly, another relevant study on movement patterns of individuals between major cities in the UK have also identified a link between the number of journeys between cities based on *Flickr* estimates and the official data [20]. A previous study modelled international travel flows to quantify travelling interactions (e.g. travel distance and number of countries visited) between countries [101].

Using the information on when and where the photo was uploaded, researchers were able to utilise this information to investigate online behaviour and real-world events. These studies have underlined the usefulness of *Flickr* photo data in detecting and monitoring collective human behaviour.

1.5 Summary of social media data

Previous studies have demonstrated that the analysis of online social media data can provide insight into current or recent past behaviour in the real world. These online data sources have highlighted the usefulness and the opportunity in obtaining accurate and quick estimates.

Since social media services and their functions vary, investigating different types of social media can provide additional evidence. In addition, each social media site might have a specific population due to various preferences or interests of users. Previous studies have focused on social network (i.e. *Facebook*) and social media sites that provide short messaging services (e.g. *Twitter*). Photo sharing sites have a large number of users too. The usefulness of *Flickr* photos in obtaining estimates has been extensively investigated by extracting relevant information, such as messages or locations from photos. However, some data sources such as *Instagram* have been less popular within the scientific community. While both *Flickr* and *Instagram*'s main function is photo sharing, *Instagram* has more active users compared to *Flickr* [102, 103]. Also, *Flickr*'s strength is to offer a photo library for bloggers and professional photographers. On the other hand, *Instagram* attracts users through the use of social features such as stories. In addition, the majority of *Instagram* users tends to be the younger generation (under age 35) [104, 105]. This implies that both *Flickr* and *Instagram* provide photo sharing services to different population and age groups. A more extensive study on *Instagram* would help close the gap in the literature and provide a better understanding in using online social media data.

A study on modelling a spatial data source that provides a vast amount of geotagged photos, including texts, and its relationship with language usage statistics across space, is missing. Therefore, this presents an opportunity for this thesis to investigate *Instagram* data. Furthermore, *Twitter* is also a spatial data source that provides geotagged tweets. The amount of research on estimating language statistics across space using *Twitter* is also limited. Therefore, investigating *Twitter* could help to obtain a better understanding of social media data and better estimates of language usage statistics. A study focusing on both data sources will provide a response to one of the main objectives of this thesis, which is to investigate whether online data can be used to estimate statistics spatially. In contrast, statistics across time often requires data in a time series format to detect a trend or pattern. It is challenging to obtain social media data across time due to limited access. It requires effort to collect these datasets over a long period. The next section will explore another type of online data sources that publicly offer time series data.

2 Data on online information gathering

This section will introduce two of the most commonly used online data sources for online information gathering: *Google* and *Wikipedia*. For each online data source, we will explore previous studies in this broad research area and then focus on nowcasting in particular.

2.1 Google

To search for information online, people often submit a query to a search engine. One of the most popular search engines is *Google*. Data on search queries tends to reveal people's interests. For instance, it is likely that people who are without a job will search for "job" using a search engine. *Google* is a widely used website that dominates the world's search engine market and handles 90 percent of all queries made in 2017 [106, 107]. Therefore, search volume data from *Google* is likely to cover a wide range of Internet users in the world. This highlights the potential of Internet search data as a data source for the analysis of collective human behaviour.

Search query data that users send to *Google* each day is aggregated and made publicly accessible on the *Google Trends* website. On *Google Trends*, search volume data is available from 2004 onwards and can be restricted to a country or region. The website reports search volume for a given query as an index relative to the highest search volume during the specified time period within the particular geographical region. There is no information on the absolute number of searches and queries with very low search volume are not reported due to privacy reasons. The maximum search volume the period is normalised to 100. The rest is scaled proportionally. In other words, a value of 100 is the highest search volume in a specified time period and a value of 50 stands for half of searches compared to the highest volume. For instance, the search interest for "Brexit" peaked in June 2016 with a value of 100 (Fig. 2.1). In June 2016, the highest search interest was registered on 24th June, which corresponds to the day after the referendum on 23rd June 2016 (Fig. 2.2).

The overall number of searches increases over time. Search volume in 2004 was lower compared to today. Furthermore, search volume data obtained from *Google Trends* is based on a sample [108]. On the same day, the query results for the same term will be the same [5]. On a different day, the result for the same term can be different. This also applies to weekly or monthly data where the results can differ when search volume is obtained again in a different week or month.

Apart from reporting search volume for a given query, *Google Trends* also reports related search terms for each query. Relations between search terms are

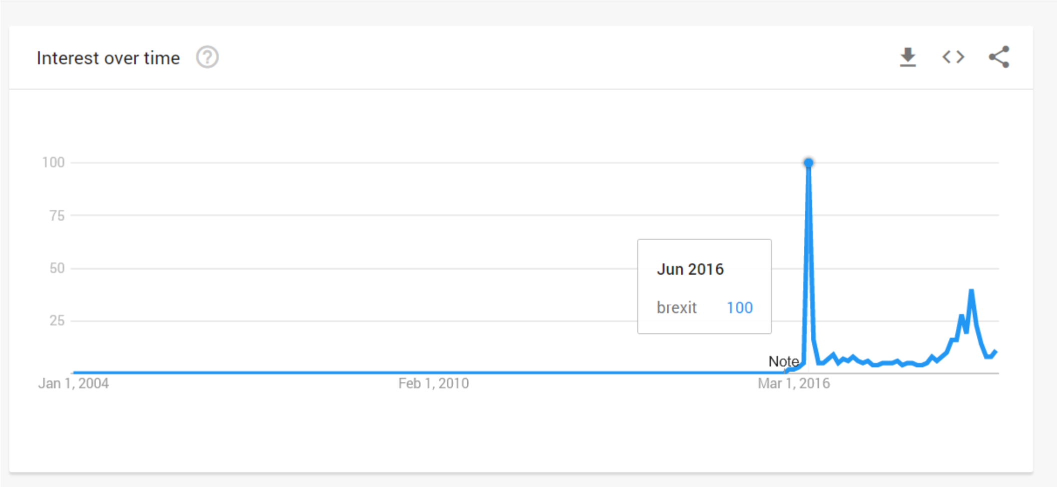

Figure 2.1: **Searching for "Brexit" since 2004.**

Figure taken from `https://trends.google.com/trends/explore?date=all&geo=GB&q=brexit`. Retrieved 20 August 2019.

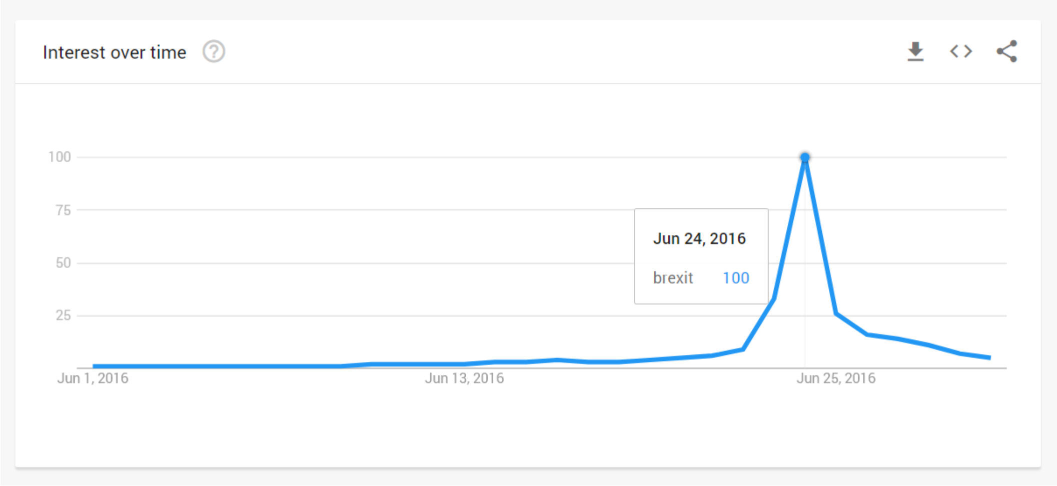

Figure 2.2: **Searching for "Brexit" in June 2016.**

Figure taken from `https://trends.google.com/trends/explore?date=2016-06-01 2016-06-30&geo=GB&q=brexit`. Retrieved 20 August 2019.

collected when a user enters a query which is then followed by another query. There are two main metrics of related queries that *Google Trends* reports: "Top" and "Rising" [109]. Top searches are the most frequent search terms used in the same search session within the specified country or region. Rising searches are keywords that have the highest growth in volume in the chosen time period.

Google search data has been used extensively by social scientists to gain a better understanding of collective human behaviour. One previous study discovered that search query volume can provide information on financial market moves within Eurozone countries to a certain extent [110]. A number of studies found the relationship between financial market movements and changes in search behaviour [111, 112]. The current transaction volumes of stock markets are correlated with changes in search volume [111]. The related study implied that search volume data could be used to provide insights before the stock market moves [112]. Moreover, previous research provides evidence that people within countries with a higher GDP are more likely to focus on the future as reflected in search behaviour [113]. In politics, "issue salience" refers to the importance of different political issues ranked by voters. It has been found that search data might be able to measure issue salience for some issues [114].

Nowcasting with Google search data

Having explained which information is available and accessible reflecting *Google* search behaviour, *Google* has recently been used as an online data source in many social science studies. This section will investigate previous studies that have used *Google* as data source specifically for nowcasting. Nowcasting has been a focus of several studies in areas as diverse as economics and health [4, 5, 13, 14, 16, 64].

Studies using Internet search data have investigated various aspects within the area of health, such as disease detection and monitoring. In the case of flu infections, traditional monitoring systems publish results with one or two weeks delay. A collaboration between a team of *Google* researchers and the *Centers for Disease Control and Prevention* resulted in a tool called *Google Flu Trends*, which monitors flu epidemics at different levels of geographical areas, specifically regional and state-level, in the US [13]. Other findings in disease detection and monitoring have underlined the importance of Internet search data [14, 115–121]. For instance, a recent study reveals that combining traditional data such as a time series of flu levels and online search data, can improve estimates of flu incidences [14]. In this study, nowcasting models including both *Google Flu Trends* data and historic flu levels are built to estimate current flu levels, which are only published with delay. An in-sample test reveals that such a nowcasting model can reduce the mean absolute error (MAE) by 14% in comparison with a baseline model that includes time series data on flu levels only. Furthermore, depending on the training windows size, out-of-sample nowcasting models can improve estimates by between 16% and 52%. In another study, an elastic net was used to improve nowcast estimates of influenza-like illness rates based on search query information from *Google Flu Trends* [122].

By using an elastic net regularisation to select important search terms, it can help reduce the mean absolute percentage error (MAPE) from 20% to 12%. It is compared with a nowcasting model including *Google Flu Trends* data only which aggregates all search terms. Apart from influenza, search query data from *Google Trends* has been discovered to be useful in monitoring dengue activity [123].

Numerous studies have recently investigated whether online search data can provide quicker estimates of the number of suicide incidents [124–129]. A possible link between the number of suicide occurrences and Internet search volume was examined for the USA between 2004 and 2007, suggesting a negative relationship [124]. Researchers found no evidence for a relationship between suicide-related search terms and suicide rates in Australia between 2004 and 2011 [125]. Using the search terms "suicide", "depression", and "suicide method" in Japanese, a previous study found that "depression" results in a positive correlation with the number of suicide incidents between 2004 to 2009 [126]. However, the number of suicides was found to be decreasing three months after and before the increase in search activity for "depression". In Taiwan, a study has shown that search volume of suicide-related search terms between 2004 to 2009 could be linked to specific age groups [127]. Another study focusing on Japanese individuals aged between 20 and 40 has found a positive relationship between online search activity and suicidal attempts between 2004 and 2010 [128]. A recent study estimated the number of suicide incidents using monthly *Google Trends* data, specifically focusing on the term "depression" and "suicide", restricted to England covering the period between 2004 and 2013 [129]. The authors found that incorporating *Google* search data and official data results in better estimates compared to using official data only.

In order to make a decisions in an economic context, policymakers are interested in obtaining instant and accurate estimates of key statistics such as unemployment rates. With the immediate availability of search volume data, numerous studies have investigated the use of Internet search volume data as economic indicators to date.

The unemployment rate - one of many factors measuring the economic health and a key indicator for the labour market - has been the focus of several studies in a wide range of countries. Researchers suggested that search query data could help estimate economic statistics, specifically the unemployment rate [130]. Moreover, researchers have examined the initial claims for unemployment in the United States [4, 131]. The study in 2009 by Choi and Varian reveals that the inclusion of search volume data on the "Jobs" and "Welfare & Unemployment" categories in autoregressive models can produce more accurate estimates of the initial claims than baseline models that exclude search data [131]. This study investigated initial claims and *Google Trends* data between 2004 and 2009 and performed out-of-sample tests, in which both long term and short term models with search data improved the MAE by 15% and 13% respectively. A later study by Choi and Varian in 2012, however, reveals that, through one-step-ahead out-of-sample forecasts, the baseline model fits slightly better than the *Google* model between 2004 and 2011 [4]. Nevertheless, they found that search query data in a short term model can reduce the MAE by 0.6% to 22%, when consider a turning point in the time series. Researchers have shown that *Google* search data can be used to nowcast monthly unemployment rates in Germany between 2004 and 2009 even under complex and rapid changing conditions by using certain groups of search terms in German language relating to unemployment [15]. As the unemployment rate in Germany is computed based on the period between the middle of the previous month and the middle of current month, the study also suggests that *Google* search data from the previous month has a better estimation performance for the current month than search data from the current month only. These results have drawn the attention of government institutions and central banks to the practical potential of using online search data for the estimation of various economic statistics [5, 18]. In the United Kingdom, the Bank of England has highlighted the usefulness of search volume data in nowcasting unemployment rates in the UK during 2004 and 2011 by choosing a keyword that is highly correlated with the official data amongst a group of keywords that was handpicked and related to unemployment. This study then compares its nowcasting performance with a baseline model and other traditional approaches such as survey data [5]. The chosen keyword "JSA" or Jobseeker's Allowance, which is an unemployment benefit in the UK, provides an improvement to the model fit based on in-sample goodness of fit measures; Akaike information criterion (AIC) and adjusted R-squared. Additionally, the one month ahead out-of-sample forecast reveals that the "JSA" model can reduce the root mean squared error (RMSE) by 13% compared to a baseline model with unemployment rates from official data only. Related findings reveal similar results for using *Google* search data for nowcasting unemployment rates in the United States and Eastern European countries; Czech Republic, Hungary, Poland and Slovakia [17, 132]. These studies underline the usefulness of search data. However, the methods used to select the keyword differ as well as the time period considered. Also, it remains unclear whether the inclusion of more relevant keywords can further improve nowcasting estimates.

The crucial first step for investigating search volume data is term selection. Numerous studies have used various approaches to select a keyword for their analyse.

For changes in unemployment, researchers have suggested "unemployment", "jobs", and "resume" as potential keywords when conducting initial surveys [4]. Other researchers considered "jobs" as an indicator because it is the most popular amongst job-related keywords [17]. Another study chose keywords based on the expected behaviour of people before being unemployed ("unemployment office") and during acquiring new jobs by searching for jobs via search engines [15]. The Bank of England's study employed a single term which is "JSA" (Job Seekers' Allowance) [5]. JSA is for people who are unemployed and search for jobs in the UK, who will receive monetary support. However, this is a specific keyword relating to the UK labour market only and might be less useful in the future. A recent study has claimed that "redundancy" is a more general term reflecting the flow into unemployment [6].

This section has summarised evidence that *Google* search data can be used as a potential online data source for obtaining early measurements in economics, disease detection and real-world events. However, the best method of choosing appropriate keywords and including them to improve nowcasting performance remains unclear. Although previous studies used a variety of search terms and approaches, they report similar results. Online search data was able to improve the nowcasting of unemployment rates by including search data into these models. However, the methods used to assemble nowcasting models differ. It is also unclear whether methods used in previous studies still hold at a later date, which then can rely on more data. This presents, therefore, an opportunity to consider nowcasting with longer time series and to include a combination of *Google* keywords, especially those relevant to unemployment.

2.2 Wikipedia

Wikipedia is the most commonly used knowledge repository worldwide. With more than 5 million English articles and approximately 40 million articles in other languages as of January 2019, *Wikipedia* is an active online community, in which the content can be viewed and edited by volunteers [133]. Data on *Wikipedia* page views and editing history is publicly available, allowing researchers to obtain data that reflects people interests similar to search data. The difference is that *Wikipedia* is limited to the pages that are available on the platform, whereas *Google* searches are open to any query a user might have entered. Aggregated information about these queries is then published via *Google Trends*.

Wikipedia is an online encyclopedia which offers definitions of various concepts as defined by human users. By exploiting this vast source of textual information written by humans, researchers have analysed the relatedness of words (e.g. how "cat" and "mouse" are related) [134, 135]. *Wikipedia* also allows users to insert links to other articles to create a semantic knowledge base of the relationship between concepts [136]. This, in turn, enables *Wikipedia* to enrich the articles with links that provide further explanation based on a piece of text using machine learning classification techniques [137]. Topic detection, information retrieval, and identifying related terms on *Wikipedia* are examples of major research areas [138–142]. For example, how to differentiate words that have the same spelling but differ in meaning such as the word "right"?. This area of research therefore aims to increase the accuracy for differentiating meanings. The history of editorial activity is also available on *Wikipedia*. A previous study investigated activity patterns of *Wikipedia* editors to estimate weekly activity patterns and the geographical distribution of editors' location across the world. Certain articles might be influenced by biases specific to a certain culture or society [143]. For instance, it has been found that English *Wikipedia* pages are largely edited by editors from European countries. Due to the open nature of *Wikipedia* articles, there are sometimes conflicts or disagreements in opinions between groups of editors. Researchers have developed an automated approach to detect such conflicts to study their social dynamical features. This includes the length of the discussion page [144–148]. The results reveal that the length of English discussion pages is correlated with the degree of controversiality. Other language discussion page lengths are weaker correlated with levels of controversiality [144]. This suggests that discussion pages can reflect editing conflicts when considering cultural differences.

Nowcasting with Wikipedia data

Researchers have investigated whether *Wikipedia* page view data, which reflects information on what people are interested in, can offer insights into current and subsequent collective human behaviour and decisions. A recent study has demonstrated that the number of page views on financial topics on *Wikipedia* can signal direction of stock market prices [149]. This implies that information on *Wikipedia* page views might provide crucial information for financial investors before making a decision.

Similar to *Google* search query data, information on page views and editor activities can be used to monitor events in the real world. For instance, researchers have used *Wikipedia* data to provide estimates of a movie's popularity before the movie is released [150]. When using *Twitter* data for event detection, a number of events might be identified inaccurately. *Wikipedia* page view data can therefore be used in a complementing way, filtering false events that were earlier detected using *Twitter* data [151]. Researchers have examined article updates (i.e. the number of edits and their timestamps) on *Wikipedia* and found a correlation with real-world events and in particular political events and natural disasters [152]. This suggests that real-world events could be automatically tracked and monitored through content extraction and aggregation from article updates on *Wikipedia*.

Previous research on *Wikipedia* has extensively demonstrated the usefulness of online data in studying social dynamics, conflict, topic detection and natural language processing. However, studies on *Wikipedia* usage and its potential in nowcasting real world events are limited, which presents an opportunity for social scientists.

2.3 Summary

We have presented previous research using online information gathering data.They provide evidence that underlines the opportunity of using online search data to nowcast various collective human behaviour patterns. *Google* search data, in particular, has been a focus for nowcasting economic figures. However, it is unclear whether including more relevant keywords can improve the nowcasting performance. Therefore, there is an opportunity to investigate online search data and nowcasting performance by including more keywords.

3 Limitations of online data

Even though online data provides large scale information of human behaviour, the biased nature of online data makes it difficult to replace traditional sources of data, such as surveys. Online data is therefore rather complementary and not a substitute. Generally, sampling can produce selection biases in which the selected sample does not represent the entire population equally. Although a large volume of data provides a number of advantages, it can produce less accurate results as the size can increase the errors caused by the sampling bias [153]. Internet users may not represent a whole population as age can be considered as a factor as younger people might be overrepresented.

Another concern when using online data sources is how data is being generated by online users. Most online data sources do not provide a transparent or detailed explanation of algorithms or mechanisms behind the data collection process. These can also be changed or updated without notice. Data availability is uncertain too. For instance, *Google Correlate* [154], a website that allows users to enter real-world data and find search volume that is closely correlated, is no longer updated as of 2018. An alternative approach for gathering data from the Internet that we have not examined in detail here and that can help address such transparency problems is crowdsourcing. Crowdsourcing approaches engage users directly in tasks actively rather than analysing data generated through the use of other services passively. For example, crowdsourcing approaches have made it possible to gather data on how beautiful Internet users consider different locations to be by asking them to rate photographs, opening up a stream of research assessing the visual beauty of our natural and built environment and how this beauty relates to our health and how happy we feel [155–158]. Crowdsourcing approaches are however also vulnerable to problems relating to the demographics of Internet users as described above.

4 Conclusion

This chapter has showcased how online data has helped in providing quicker estimates for the current state of society in various areas. It also has examined how social media data (e.g. *Instagram*, *Twitter*, *Facebook* and *Flickr*) and data from online information gathering platforms (e.g. *Google* and *Wikipedia*) can be used to investigate human activity patterns. Crucially, we highlighted the usefulness of online data in obtaining accurate and timely estimates of human activity patterns and real-world events before official figures are published.

Building on the literature review in this chapter, social media data can be used to provide early estimates and detect real-world events. Although much of the studies have focused on text-based social media sites (e.g. *Twitter*) and social networks (e.g. *Facebook*), other types of social media data sources, especially photo sharing platforms, e.g. *Instagram*, have only been used on a number of occasions. Moreover, there is a lack of detailed studies on language statistics across space. Although there is a growing number of studies on language distributions across areas using spatial data from other online media platforms, the question whether the spatial patterns of language usage can be estimated with data from a photo sharing website remains unanswered. It is also unclear whether the language usage on social media could be used to improve estimates of official language data. This opens an opportunity to investigate the usefulness of photo sharing data since *Instagram* provides spatial information. It provides a huge amount of geotagged photos which contain texts and captions that can be used for spatial analysis. Another opportunity is investigating tweets as *Twitter* also provides a large amount of geotagged information, which can complement the spatial analysis. Tackling these challenges could add further evidence for the usefulness of online data in estimating official statistics.

Furthermore, it was demonstrated that *Google* search data has been widely used to obtain quicker measurements of economic figures such as unemployment rates. However, the methods used to produce nowcasting models in each paper are different, as is the period of data considered. It is also unclear whether methods used in previous papers are still valid with later data. The question whether including more relevant keywords can improve nowcasting accuracy remains unclear too. This therefore presents an opportunity to focus on nowcasting with more recent search data and investigate the performance the corresponding models which include more keywords.

To summarise, this chapter has highlighted the potential of using online data Building on the opportunities mentioned above, this thesis will explore whether the distribution of language usage on *Instagram* and *Twitter* can help estimate language usage statistics. Lastly, this thesis will investigate whether recent online search data can be used to nowcast economic statistics, specifically unemployment rates, and whether including more relevant keywords can improve such estimates.

Chapter 3

Methodology

This chapter will outline the official statistics datasets that will be used in the following chapters. It will also describe the concepts and methods that will be applied for investigating the relationship between official statistics and online data in the following chapters.

To understand the concept of variable selection techniques, it will introduce a linear regression model as a background concept. The linear regression fundamentally model relationships between variables, estimating the coefficients to find a linear function that fits the observed data. The next section will explain the methods that will be used to quantify the relationship and to assess the accuracy of the fitted model. Specifically, R-Squared and various error metrics such as MAE and RMSE will be discussed. Afterwards, ridge, LASSO and elastic net, which are extensions of linear regression, will be introduced for variable selection. These techniques will be used to estimate the unemployment rates in Chapter 6.

Next, another type of regression – a logistic regression – that is designed for categorical dependent variables will be described. This method will be used to estimate how many people speak a given language, such as French, in different areas of British cities in Chapter 4 and Chapter 5. In addition, different error metrics specific to logistic regression models will be introduced.

The last section will discuss approaches to model selection and resampling techniques used for investigating model performance. A well-known technique is cross-validation, which will be used here for both analysis of statistics across space and statistics across time.

1 ONS datasets

In this thesis, we will investigate two different datasets obtained from the Office for National Statistics (ONS) in the UK: language usage and unemployment.

1.1 Language statistics data

In England and Wales, the ONS carries out a Census every ten years to collect valuable measurements of society. The information gathered via the Census is mainly used by public sectors and businesses. The public sector uses the information specifically for policy planning and resource allocation, while business uses the information to understand behaviour of their customer as well as identify new customers [159]. One of the measurements collected in the Census is the main language the respondent speaks. In the 2011 Census, people who were living in England and Wales were asked "What is your main language?". The statistics reveal that over 100 languages are spoken in England and Wales. Furthermore, over 7% of the population reported that their main language is not English [160]. Due to increasing concerns about the cost of the Census [2, 3], the ONS started to evaluate new possibilities to collect population statistics [7]. It has been suggested that the next Census in 2021 could be mainly conducted online [8].

1.2 Unemployment data

In the United Kingdom, employment, unemployment, and economic inactivity rates are calculated using a household survey called the Labour Force Survey (LFS), which is conducted and reported by the ONS. The survey interviews 90,000 randomly selected people face-to-face or via a phone interview every three months [161]. The ONS calculates unemployment figures for people who are 16 to 64 years old. The survey asks respondents whether they are either employed (starting from one hour a week), or unemployed and looking for a job, or economically inactive (unemployed and have no interest in employment) [162]. Specifically in the UK, people are counted as unemployed when they are not working, are available for work and have either looked for a job within the last four weeks or are waiting to start a new job. This is in line with the guidelines provided by the International Labour Organisation and the figures are therefore comparable with other countries [163]. The unemployment figures are published each month with one and a half months delay. For example, the April 2018 release reports the unemployment figure for February 2018, leaving a gap from March to April. For each month, the figures released reflect estimates for the UK population over a three-month period.

In the UK, the other key measure of unemployment is the unemployment benefits claimant count which is published monthly. This is derived from the number of Jobseeker's Allowance (JSA) claimants recorded by Jobcentre Plus. However, not every person who is unemployed is eligible for JSA. It is also not comparable with other countries' data, as not every country has a programme similar to JSA.

2 Linear regression

Linear regression is an approach for modelling the relationship between two quantitative variables. The output variable (Y) is also known as the dependent, response or outcome variable. The input variable (X) is also known as the independent, predictor, feature, or explanatory variable. Linear regression allows for important questions to be asked and resolved, for example, is there a relationship between x and y? (e.g. is there a relationship between the number of unemployment figures and the number of search queries for "jobs" in each city or country?). Linear regression can be used to study both single and multiple input variables. For p different variables, the input vector can be expressed as

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}. \tag{3.1}$$

The standard or simple linear model for the relationship between output variable Y and input variable X can be written as

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{3.2}$$

where $\epsilon \sim N(0, \sigma^2)$. The parameter or coefficient β_0 is an unknown constant that represents the intercept, and the slope is represented by β_1, while ϵ is a normally distributed error term with a mean of zero. The intercept is the expected value of Y when $X = 0$ and the slope is the average change in Y when X is changed by one unit. $\beta_0 + \beta_1 X$ forms the structural component and the error term ϵ is the random component. Since the structural or systematic component is unable to model the relationship perfectly, the error variable captures measurement errors and other discrepancies from the function [164].

Using training data to produce estimated coefficients $\hat{\beta_0}$ and $\hat{\beta_1}$, the model can estimate the value of the response $\hat{y}$ based on a predictor variable x. This is formulated in Eq. 3.3. The hat symbol indicates an estimated value from the data.

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1} x \tag{3.3}$$

2.1 Estimating the coefficients

Since β_0 and β_1 are unknown coefficients, training data is used to estimate the parameters. The goal is to find a line, represented by an intercept β_0 and a slope β_1, that is as close as possible to all training data points. One of the well-known methods to measure closeness is least squares, which chooses the intercept and slope to minimise the residual sum of squares (RSS) [165]. These coefficients become least square estimates.

Let $\hat{y}_i = \hat{\beta_0} + \hat{\beta_1} x_i$ be the estimate for y_i based on the ith value of X [165]. The ith residual is calculated from

$$e_i = y_i - \hat{y}_i. \tag{3.4}$$

The residual is the difference between the actual outcome and the predicted outcome. Alternatively, the residuals are the distance of each point from the line. Therefore, the residual sum of squares (RSS) are given by [165]

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2. \tag{3.5}$$

2.2 Linear mixed effects model

The linear mixed effects model is an another type of linear regression model which considers both fixed effects and random effects. In Eq.3.2, linear regression accounts for one or more fixed effects only. Linear model contains variation that is explained by the explanatory variables. The model also includes a random component which is the broad error term ϵ. Adding random effects provides structure to the error term to take into account variation that is not explained by the independent variables [166].

In repeated measures data, observations are not considered to be independently drawn from the population. Repeated observations are taken from the same unit of analysis (e.g. the same person or subject, or the same language). For example, a subject is asked to provide well being ratings every week. It is possible that some subjects tend to rate well being as high on average since they might be a happy person. This condition would violate the independence assumption in a linear model, which states that is each observation is independent. For this situation, adding random effects would resolve the non-independence by assuming a different random intercept value for each unit of analysis such as the subject.

Random effects are fundamentally consist of random intercepts and random slopes. Models that contain random intercepts assume different intercept for each level of the random-effect variable. For example, different subjects may have different means on well-being rating. On the other hand, random slopes in the model account for variation in the effect for each level of the random-effect variable. For instance, the rate of change in well-being score would be different for each subject. A simplified notation for linear mixed models is as follow [166]:

$$y_i = \alpha_i + \beta_i s, \tag{3.6}$$

where y is a dependent variable, α is an intercept, β is a slope, and s is a fixed effect variable. The index i represents the level of the random-effect variable. From our example of subjects, $i = 1$ would mean the first subject. This indicates that the intercept and slope will differ between each subject.

3 Assessing the accuracy of the model

After having investigated whether there exists a relationship between the dependent and independent variables, the next step is to quantify the relationship to determine how well the model fits the data. Generally, R-squared and various error metrics, such as mean absolute error (MAE) or root mean squared error (RMSE), are used to assess the accuracy of the model.

3.1 Correlation and R-squared

The Pearson correlation r provides the information about association between X and Y. The correlation is defined as

$$r = Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{3.7}$$

where $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ and $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

- r = 1 reflects a perfect positive correlation between X and Y.
- r = 0 reflect no correlation between X and Y.
- r = -1 reflects a perfect negative correlation between X and Y.

Since various error metrics (e.g. MAE and RMSE) provide an error measure in the units of the response variable (i.e. Y), it is unclear how to measure what number means a good model fit. Alternatively, R-squared is a goodness-of-fit measure in a form of a proportion. R-squared is the proportion of variance in Y that is explained by X. The formula of R-squared is

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}, \tag{3.8}$$

where residual sum of squares (RSS) is described in Eq.3.5, and TSS is the total sum of squares or $\Sigma(y_i - \bar{y})^2$. RSS computes the amount of unexplained variance from the regression model, while TSS calculates the total variance in the response variable (Y). Therefore, TSS - RSS can be interpreted as the amount of variance that is explained by the model. The value of the R-squared range is between 0 and 1. R-squared close to 1 implies a model with a better fit which explains a large proportion of the variability in Y. Conversely, a value of R-squared near 0 indicates the variability in Y is not explained much by the model. R-squared measures how closely the two variables are associated while p-value and t-statistic measure how strong the evidence is that there is a non-zero association. For simple linear regression, r^2 and R-squared are identical. In contrast, for a linear regression model with multiple variables, the alternative approach to calculate R-squared is Adjusted R-squared, which will be discussed in Section 6.

3.2 Error metrics

In order to evaluate model accuracy or compare prediction performance between different models, various error metrics are proposed in the literature [167–177].

Scale-dependent errors Scale-dependent errors simply measure the goodness of fit between the actual data and the prediction model. The forecast error or residual is given by

$$e_i = y_i - f_i, \tag{3.9}$$

where y_i is the actual value of ith observation and f_i is the forecast of y_i. The error metrics that are based on e_i, such as mean absolute error (MAE) and root mean squared error (RMSE), are scale-dependent [167]. This means they are on the same scale as the data. The MAE and RMSE are defined by the following formulas:

$$MAE = mean(|e_i|), \tag{3.10}$$

$$RMSE = \sqrt{mean(|e_i^2|)}. \tag{3.11}$$

For a fitted regression line, the average or the sum of the forecast errors e_i is equal to zero because the overestimates of some data points and the underestimates of other data points cancel each other out. Therefore, it is common to square or take the absolute difference of the residuals to indicate the magnitude of the errors.

It is suggested that MAE has the simplest interpretation as the absolute difference assigns equal weight to the spread of data [167]. This makes the MAE less sensitive or more robust to outliers. On the other hand, the RMSE is strongly affected by extreme outliers while minor differences are less significant [174]. In other words, in proportion to the total square error, the square of small errors is smaller whereas the square of large errors is bigger, such that they therefore have greater influence or weight. Calculating the square root of the MSE returns the error metrics to the original units of the data being measured.

There is no general consensus on whether the MAE or RMSE is more suitable to measure model accuracy [172, 173, 176, 177]. The limitation of scale-dependent errors is they are unsuitable for data sets that have different scales [167]. Furthermore, they disregard the direction of underestimate or overestimate in the model evaluation as it is removed by squaring or taking the absolute difference [177].

Percentage errors Mean absolute percentage error (MAPE) is a scale-independent measure of model performance. Unlike scale-dependent metrics, this metric is not expressed in the same units as the data [167]. MAPE is formulated as

$$MAPE = mean(|p_i|), \tag{3.12}$$

where p_i is the percentage error computed by $p_i = 100 \times \frac{e_i}{y_i}$. Although MAPE has the advantage of being scale-independent, it has four main disadvantages [167]. First, the percentage error could be infinite or undefined if y_i equals zero for any observation, known as the problem of division by zero. Second, if y_i is close to zero in conjunction with there being no upper bound and only a lower bound of zero for the absolute percentage error (APE), the MAPE could take on extremely high values, making the MAPE distribution positively skewed [170]. Thus, outliers can easily affect the MAPE [169]. Third, MAPE has an underlying assumption that in the data being modelled, zero is a meaningful value. However, this is problematic for some scales that have no meaningful zero, such as measuring temperature accuracy in Celsius [167, 168]. Fourth, MAPE penalises negative errors, or overestimates are more penalised than positive errors or underestimates although magnitude of the error is identical [169, 171, 175]. Negative errors are produced by forecasts that exceed the actual value, and positive errors are produced by forecasts which are lower than the actual value. For instance, let $y_i = 200$ and $f_i = 100$ — a positive error. The absolute percentage error is $|100 \times \frac{200-100}{200}| = 50$. In contrast, if $y_i = 100$ and $f_i = 200$ — a negative error — then the absolute percentage error is $|100 \times \frac{100-200}{100}| = 100$. In this way, while the absolute error for both forecasts is equal, the MAPE for the negative error is much higher. The MAPE is therefore systematically biased towards underestimates, making it asymmetric.

4 Ridge, LASSO, elastic net

To decide which input variable is important to include in the multiple linear regression model, there are several approaches for variable selection. By selecting or shrinking coefficients, the linear model can become more interpretable and provide more accurate estimates. To achieve this, there are 3 main types of variable selection techniques [178]:

- **Subset selection** computes the least squares fit for all possible subsets of the variables and then chooses the model based on the criterion that balances training error with the model size. However, this is computationally expensive when there are a large number of independent variables. For example, 20 variables will produce $2^{20} = 1,048,576$ models. Example techniques are *best subset selection*, and *backward and forward stepwise selection.*

- **Shrinkage or regularisation** select variables by introducing a penalty to penalise the model relative to least square estimates. This helps reduce variance and allows variable selection. Such techniques are *ridge*, *LASSO*, and *elastic net.*

- **Dimension reduction** finds combinations of variables, extracts important combination of variables and then uses them to fit a linear regression model. *Principal Component Regression (PCR)* and *partial least squares* are examples employing this technique.

In this thesis, we focus on shrinkage methods. Shrinkage methods generally fit a model with all available independent variables and constrain or regularise the coefficient estimates. In other words, they shrink the coefficient estimates towards zero. This helps determine which input variables are important in explaining output variables, and which are irrelevant. Ridge and LASSO are well-known techniques for shrinking the coefficient estimates. These methods control variability when there is a large number of variables by significantly reducing the variance of coefficient estimates, but they can introduce more bias afterwards. Bias is the difference between the model's average prediction and the actual value that the model is trying to predict. High bias generally can cause underfitting in which the model is oversimplified and produces inaccurate predictions on average. In other words, models with underfitting are unable to capture the pattern of the training data.

4.1 Ridge

In a least squares regression, the goal is to obtain coefficient estimates that minimise the residual sum of squares (RSS) as given by Eq.3.5. To penalise the coefficients in ridge regression that are considered insignificant, a shrinkage penalty term is added to the RSS. Specifically, the ridge regression [179] aims to minimise

$$\sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2, \tag{3.13}$$

where λ is a tuning parameter and the $\sum_{j=1}^{p}\beta_j^2$ after the λ is named L_2 normalisation. The shrinkage penalty causes coefficient estimates to be shrunk towards zero. The tuning parameter λ controls the amount of shrinkage. Increasing λ would reduce the coefficients' magnitudes. For instance, with large λ, ridge regression will reduce the size of coefficient estimates towards zero. However, if λ is zero, the result will be the same as least squares regression. Fundamentally, each λ will generate a different set of coefficient estimates. A search for an optimal λ parameter that minimises RSS on the testing data set can be conducted through cross-validation, which will be discussed in Section 6.1.

In ridge regression, adding a L_2 penalty term introduces bias into the model since all β coefficients are treated differently. On the other hand, it reduces the variance by shrinking coefficient estimates towards zero and RSS reduction on the testing data set.

4.2 LASSO

The disadvantage of ridge regression is it uses all predictors in the final model since it only shrinks the coefficients towards zero, but not exactly to zero. To tackle this problem, LASSO [180] was introduced as an alternative to ridge regression. The distinct feature is that the LASSO penalty is used instead, by using the absolute value ($|\beta_j|$) instead of the squared value on coefficients (β_j^2). The goal of LASSO is to minimise

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{3.14}$$

Instead of the L_2 penalty term ($\sum_{j=1}^{p} \beta_j^2$) used in ridge regression, the LASSO uses an L_1 penalty ($\sum_{j=1}^{p} |\beta_j|$). Similar to ridge regression, the LASSO has the effect of reducing the magnitudes of the coefficient towards zero. The difference is that LASSO selects variables to be included in the final model. Specifically, when λ is adequately large, the tuning parameter λ can set the coefficient estimates to be zero such that the variable will be excluded in the final model. Therefore, the L_1 penalty performs a sparse selection amongst the coefficients, resulting in sparse models which include relevant variables. This also allows the fitted model to be more interpretable and less complex. One limitation of the LASSO is that when the number of variables is greater than the number of observations, it can only select a number of variables less than or equal to the number of observations. Similar to ridge regression, the LASSO reduces variance at the cost of increase in bias. The bias-variance tradeoff is the balance between the two sources of errors: bias and variance. Less complex models often have low variance and high bias due to the simple structure of the model, whereas more complex models often have high variance and low bias since they have flexible structure. A model cannot be both less complex and more complex at the same time, therefore, a good balance between bias and variance is preferred for building an optimal model in which both errors are at the minimum level. Selecting an optimal λ can also be determined by cross-validation as described in Section 6.1.

For a multiple regression model, a predictor can be highly correlated with other independent variables. In other words, when X_j changes, other independent variables also change due to their correlations. However, correlation does not imply causation. It is difficult to isolate the effect separately from these correlated variables. This occurrence is known as multicollinearity. In the presence of multicollinearity or highly-correlated independent variables, ridge regression shrinks the correlated coefficients to a similar value. In contrast, the LASSO chooses the correlated coefficient that has a larger value [181].

It has been demonstrated that neither of these two shrinkage methods perform better than each other in general [178, 180, 182]. In theory, ridge regression would perform well when most of the variables are important with almost the same size of coefficients. When the number of observations is more than the number of variables and the predictors are correlated, ridge regression tends to provide better prediction performance than LASSO. In contrast, LASSO tends to perform well when there is a small quantity of important coefficients whereas the other predictors are close to zero. However, the true number of important predictors is not a known quantity when building the model. Cross-validation is therefore used to find the best fitted model, depending on the data set.

4.3 Elastic net

The elastic net is an alternative variable selection method introduced as a compromise between ridge regression and LASSO [181]. The elastic net includes both L_1 and L_2 penalties to achieve both shrinkage and automatic variable selection. Elastic net is defined to minimise

$$\begin{gathered}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha\lambda_2\sum_{j=1}^{p}\beta_j^2 + (1-\alpha)\lambda_1\sum_{j=1}^{p}|\beta_j| = \\ RSS + \alpha\lambda_2\sum_{j=1}^{p}\beta_j^2 + (1-\alpha)\lambda_1\sum_{j=1}^{p}|\beta_j|.\end{gathered} \tag{3.15}$$

α, ranging from zero to one, is the elastic net penalty determining how the ridge and LASSO penalties are combined. There are two special cases of the elastic net: When α=1, the elastic net is essentially a ridge regression model, while α=0 allows the elastic net to become a LASSO model. This method is flexible as it adapts the approach to the characteristics of the data by tuning the α parameter. Optimising the elastic net model can be performed by tuning an alpha value between 0 and 1 which will select some variables and shrink some coefficients.

5 Logistic regression

The linear regression model assumes that the output variable Y is quantitative. However, qualitative or categorical variables such as yes/no answers require logistic regression to estimate probabilities that are between 0 and 1. Using the linear equation in Eq.3.2 for a response variable that is not normally distributed, such as a binary variable that has a binomial distribution, might not be suitable, as linear regression can produce probabilities less than 0 or larger than 1. To address this issue, the generalised linear model (GLM) introduces logit and probit models which are appropriate for binary and categorical variables. Since the response variable is non-normal, the link function specifies the link between random and structural components. Logit and probit are example of link functions. The logit link for a binary response with the binomial distribution maps the structural component onto the interval between 0 and 1. Therefore, logistic regression is a binomial regression with the logistic link function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}, \tag{3.16}$$

where exp e 2.71828 is Euler's number and it is the base of the natural logarithm.

The equation 3.16 can be rearranged into the log odds or logit transformation of $p(X)$.

$$logit(p(X)) = log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X. \tag{3.17}$$

The term $\frac{p(x)}{1-p(X)}$ is called the odds which can take values between 0 and infinity. As the odds are close to 0, the indicated probabilities are very low. Odds near infinity implies higher probabilities. In contrast to the linear regression model which provides the average change in Y when X is changed by one unit, the effect of one unit increase in X changes the log odds by β_1 in a logistic regression model. Regardless of X's value, if β_1 is positive then increasing X will increase probabilities of X, while β_1 is negative then increasing X will decrease $p(X)$.

5.1 Estimating the coefficients

Although it is possible to use least squares from linear regression to fit the logistic regression model, the maximum likelihood method is preferred. Maximum likelihood tries to estimates β_0 and β_1 that results in a probability close to one for all observations that are success (or 1 in a binary variable) and a number close to zero for all observations that are opposite to success (or 0 in a binary variable). Mathematically, the likelihood function is formulated:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}. \tag{3.18}$$

The maximum likelihood chooses the β_0 and β_1 that maximise the likelihood function. The mathematical details of the fitting mechanism for this likelihood function are beyond the scope of this thesis. However, fitting logistic a regression model by maximum likelihood can be simply achieved via statistical software packages such as R.

5.2 Nagelkerke R-squared

When analysing data with a logistic regression, R-squared as defined for linear regression is not an appropriate measure of model fit. The model estimates from a logistic regression are maximum likelihood estimates. To evaluate the goodness-of-fit of logistic models, several pseudo R-squareds have been developed based on the concept of likelihood. Two methods that are widely used traditionally are McFadden (1974) and Cox and Snell (1989).

- Cox and Snell is based on the likelihood for the model compared to the log likelihood for a baseline model. However, with categorical outcomes, a well-known problem is it has a theoretical maximum value that is less than 1.
- Nagelkerke (1991) is a corrected version of the Cox and Snell R-squared that adjusts the scale to cover the full range from 0 to 1.
- McFadden is based on the log likelihood for the null (intercept-only) model and the fitted model.

The "pseudo" R-squared is similar to R-squared in linear regression. The difference is that R-squared in linear regression ranges from 0 to 1 while some pseudo R-squareds cannot reach 0 or 1. Moreover, pseudo R-squared values are not directly comparable to the R-squared for least squares models.

6 Model selection and cross-validations

In order to determine which model has the highest accuracy, R-squared and various error metrics (introduced in Section 3) are generally considered. When estimating the same set of observations that were used during training, the training error can be calculated. In contrast, a model that estimates values on a new set of observations that was not used during training, the test error can be calculated.

However, the regression model that contains all predictors available is highly likely to have the smallest RSS and the largest R-squared, because these measurements are related to the training error. When adding variable into the linear model, RSS will decrease, resulting in increase in the R-squared. In practice, the model with the lowest test error is preferred, as the trained model estimates test data that has not been seen. Thus, comparing R-squareds between linear models with multiple independent variables are not suitable for selecting the best fitted model. To address this issue, one approach is to adjust the training error to account for the bias result from model overfitting. Measures of model fit that look to make such an adjustment include AIC, BIC, and adjusted R-squared.

- **Akaike Information Criterion (AIC)** estimates the model fit relative to other models by maximum likelihood. AIC can also be used for different model types such as logistic regression. A smaller AIC value implies a better fit. The general AIC form is

$$AIC = -2\,log\,L + 2\,d, \tag{3.19}$$

where L is the maximised value of the likelihood function for the fitted model, d is the total number of parameters.

- **Bayesian Information Criterion (BIC)** introduces a heavy penalty on models with a large number of variables. As a result, smaller models are preferred.

  $$BIC = \frac{1}{n}(RSS + log(n)d\sigma^2), \tag{3.20}$$

  where σ^2 is the estimate of the variance of the error ϵ in a linear model, and n is the number of data points.

- **Adjusted R-squared** is an alternative approach based on R-squared to measure the goodness-of-fit of a model that contains multiple variables. In the normal R-squared as defined in Eq.3.8, the residual sum of squares (RSS) usually decreases when the model has more variables, subsequently increasing the R-squared. Adjusted R-squared places a penalty on models with many variables so that models with different numbers of predictor variables are comparable. The value range of adjusted R-squared is between 0 and 1, where a higher number indicates a model that fits the data well. It is calculated as

  $$\text{Adjusted R-squared} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}, \tag{3.21}$$

  where RSS is the residual sum of squares, TSS is the total sum of squares as defined in Eq 3.8, n is the number of observations and d is the number of predictors.

The advantage of adjusted R-squared over AIC and BIC is interpretability. However, adjusted R-squared can not be generalised to other types of model such as logistic regression. Alternatively, the test error can be directly estimated using a cross-validation approach.

6.1 Cross-validation

Usually, the goal of modelling is to obtain estimates that are accurate. To estimate the accuracy performance, a model is given a training data to test against a testing data. However, one round of testing would provides a bias result. Resampling methods are used by drawing different samples from the entire dataset to refit a model and then averaging the results to provide the overall estimation performance. One well-known method in resampling for model assessment and model selection is cross-validation. Two well-known types of cross-validation are leave-one-out cross-validation (LOOCV) and k-fold cross-validation.

Leave-one-out cross-validation Fundamentally, LOOCV repeatedly splits a set of observations into a training set and a validation set. For each procedure, it uses one observation as the validation set, while the remaining data are used as the training set, to fit a model to estimate the excluded observation. The test error can be calculated by averaging error metrics such as MAE or RMSE.

K-fold cross-validation K-fold cross-validation is one of the most widely used approaches for evaluating model performance. K-fold divides training data into k subsets or folds in which each fold has almost equal size. One of the folds will be left out as the testing set to determine how well the model performs, while the remaining subsets, as the training set, are used for fitting a model to estimate the excluded fold. For each fold, the test error can be calculated by averaging the error metrics such as MAE or RMSE. This process is performed repeatedly k times and, for each time, an another fold is selected as a validation set.

Time series cross-validation K-fold cross-validation seeks to randomly partition the data into several smaller data sets regardless of time sequence. However, for sequential data like time series, models should not be trained with data that would have not been known or available to avoid look-ahead bias. For instance, to predict the current value, a model should only train with past values and have no information about future values. Another problem is that K-fold cross-validation assumes there is no relationship between the observations, that is they are independent and identically distributed. However, time series data often possess autocorrelation or temporal components, that is the current value often derives or depends on the past values, and cross-validation does not take this issue into account [183].

To address these problems for cross-validation with time series data, two main techniques have been suggested. First, cross-validation with expanding training window [168]. Using this approach, the training set contains only values that occur prior to the test set data. For each step, the point at which the forecast is made rolls forward. The training window also includes more data as time moves forward resulting in expanding window.

Another technique is the sliding window approach, which is a variation of the first technique [184]. Instead of expanding the training window, keeping the training window at a fixed length to consider only recent data might offer different results. For example, a fixed training window of 10 years of yearly data between 2005 and 2015 is used to train the model. After a forecast is made, the next iteration of model training moves the training window forward in time to include more recent data and remove the oldest data in order to keep the training window at fixed length. Using the same example, after a prediction for year 2016 is made, the training window will slide to 2006 and 2016 to estimate the value of year 2017. These two techniques enable model to predict current or future values without having known future values in the training model.

7 Conclusion

This chapter has introduced the official statistics datasets on language usage and unemployment rate that will be used for investigations of the relationship between official statistics and online data.

This thesis will use the underlying concepts described in this chapter to investigate statistics across space and statistics across time. It will explore messages exchanged on the photo sharing site *Instagram* in Chapter 4, and text-based short messages on *Twitter* in Chapter 5. Both chapters will use logistic linear regression to investigate the relationship between online data and data from the ONS. It will use Nagelkerke's R-squared and the mean absolute error (MAE) to quantify the usefulness of online data. The cross-validation technique will be used to investigate the potential of using online data to improve estimates of number of speakers of each language across different urban areas.

In Chapter 6, linear regression will be used to analyse the relationship between search volume data and unemployment rates published by the ONS. Error metrics such as adjusted R-squared, MAE and RMSE will be used to compare the performance between nowcasting models with online data and models using official data alone. Next, ridge, LASSO and elastic net – techniques for variable selection – will be applied to select relevant variables from the search volume data. Time series cross-validation techniques will be used to investigate the performance of these nowcasting models.

Chapter 4

Estimating language statistics using Instagram photos

1 Introduction

There are currently more than one thousand spoken languages across the globe. Measuring spoken language statistics across regions helps policymakers in planning local public services such as identifying ethnic groups to develop equal opportunities, including jobs and training policies [1]. Numerous approaches have been employed to collect the measurements of the society in order to gain a detailed snapshot of the population. However, they require a vast amount of time and cost to record the relevant measurements. For instance, the Office for National Statistics (ONS) carries out a Census every ten years to collect valuable measurements of the status of the society in England and Wales. One of the key measurements collected in the Census is languages spoken in England and Wales. In the 2011 Census, people who were living in England and Wales were asked "What is your main language?". The statistics reveal that over 100 languages are spoken in England and Wales. Furthermore, over 7% of the population reported that their main language is not English [160].

People now communicate online and use online social networking websites to share their stories in forms of text, photo, or video in any language. Through these online activities, people are generating data which are collected by the online social networking service providers. Evidence from various studies has demonstrated that large-scale online data from online service providers such as *Google*, *Yahoo*, *Flickr*, *Wikipedia*, and *Twitter* can provide new insights of collective human behaviour [4, 5, 13, 15, 17, 19, 20, 45, 56, 57, 96, 97, 111–113, 130, 144, 149, 185–194]. A growing number of studies in recent years have studied language statistics across space using data from online platforms [33, 50–53]. Using *Twitter* data, previous work in this field has focused on determining the language and geographic location of tweets [33], mapping languages across different geographic scales ranging from country to city level [50], discovering the most commonly used languages within the top ten countries that actively tweet the most in a particular year [51], investigating different communication patterns across eight popular *Twitter* languages [52], and studying the effect of language on online social ties [53]. These studies suggest that spatial online data, especially from *Twitter* which has been popular amongst scholars, has the potential to estimate language statistics.

However, there is another large social media platform, *Instagram*, with more than one billion users around the world as of June 2018 [195]. *Instagram* was founded in 2010 as an online platform which allows users to upload photos and videos with the option to include geographic location. Previous work on data from *Instagram* has investigated mobility patterns [23] and socio-cultural characteristics [25, 26] but the study of the distribution of languages across areas using spatial data is lacking.

To study language distribution across different areas, we first consider a large and diverse city where different ethnicities and cultures are integrated in one society. New York, Sydney, and London are good examples of cities that have a large number of languages spoken. At the time of conducting this research, the most recent Census data from the USA, Australia, and the UK was from 2011. However, the USA uses a different survey, the American Community Survey (ACS), to obtain language statistics. While the latest version of these statistics were released in 2015 [196]. the ACS data for language statistics reports statistical areas in which each area has more than 100,000 people [196]. In contrast, the unit of analysis in the UK Census data includes areas with around 300 people, and others with between 5 000 to 15 000 people [197, 198], hence facilitating the study of this relationship at much higher granularity. London is also a much larger city than Sydney, with a population of nearly twice the size, potentially offering more data for study in both the Census and the online data streams. For comparison, there were about 4.8 million people living in Sydney in 2016 [199] while almost 8.2 million people lived in Greater London in 2011 [200]. For these reasons, this research will focus on UK Census data to analyse language usage statistics.

This chapter will investigate whether language data on messages exchanged online on *Instagram* can inform estimates of the language usage statistics in different regions in the UK by focusing on London and Manchester as case studies. Here, we retrieve comments, when and where the photos were taken, specific to the geographic area of Greater London and Greater Manchester from *Instagram*. For each language, we then compare the correspondence between language usage in *Instagram* photos and the number of people who responded their main language to ONS Census. Lastly, we build an estimating model for each language using *Instagram* data and compare its prediction performance with the baseline model to determine whether models with *Instagram* data can provide improvement in estimates.

2 Data and methodology

2.1 Instagram data

We sample about 7.3 million (7,301,807) *Instagram* photos that were taken in Greater London and uploaded over a period of six months between September 2015 and February 2016. When a user uploads a photo to *Instagram*, the caption, time, and geographic location (if specified) are recorded. Moreover, other users can communicate with the photo owner by leaving comments under that photo. We assume that people who exchange photos and messages online are more likely to do so in their main language.

There are tools that are capable of automatically detecting the language in which a given piece of text has been written. One of these tools is called Chromium Language Detector (CLD) which is developed by Google. The Chromium Language Detector is an open-source library which is a part of Google Chrome web browser that provides the user an option to translate a web page into their preferred language. The CLD analyses the given text in any language and reports one or more detected languages including the corresponding percentage of the language that was detected in the original text. Each detected language has a corresponding percentage of that language that was detected. For instance, providing the text "You might have to read this multiple times for it to fully sink in" into CLD returns English with 98%. Another example of piece of text is "je vais lÃă bas presque chaque fois que je remonte", which CLD detects as French with 98%. A combination of languages used in the text is also possible to detect, such as "I love his book which depicts beautiful pictures that breathe truly craftmanship. Altijd leuk om niet zomaar plaatjes te kijken maar het verhaal achter" in which CLD reports as 51% Dutch and 48% English.

For each *Instagram* photo, we pass the text provided by the caption and all comments into CLD and record the language in which it reports the greatest proportion of the text has been written. We then do this for every *Instagram* photo in Greater London dataset. To analyse languages that well represent the population of the dataset, we select the top 20 most commonly spoken languages across Greater London based on *Instagram* language usage, which aggregate to about 5.8 million photos (5,853,009) or 80% of the dataset. A list of photo counts for top 20 languages are provided in Table 4.1. We find that *Instagram* photos that are classified as mostly containing English captions and comments make up about 73% (5,320,973) of the dataset, while other 19 languages contribute only about 7% (532,036) of the dataset.

Every ten years, the Census is conducted by the Office for National Statistics (ONS) to record valuable measurements of the status of the society in England and Wales. The same questions are asked for all people who are living in England and Wales in which they can respond by either online or in paper to the questions. To gather language statistics, people who are living in England and Wales were asked "What is your main language?".

Census data has been widely used as a source of demographics study in respect to spatial area, which allow researchers to find relationship between official figures and other online data sources [201–203]. Census data is made available at various granularities for reporting area statistics such as Lower Layer Super Output Area and Middle Layer Super Output Area. Each Lower Layer Super Output Area (LSOA) has a population between 1,000 and 3,000 whereas Middle Layer Super Output (MSOA) has a minimum population of 5,000 and maximum population of 15,000 [198]. In terms of households, there are between 400 and 1,200 households per LSOA while each MSOA has between 2,000 and 6,000 households. We extract data at MSOA level as its granularity capture at least one Instagram photo in each area, especially area in Outer London where data points are more sparse. In the 2011 Census data, Greater London comprises 983 MSOAs in total, the boundaries of which are depicted in Fig. 4.1a.

To obtain area statistics of *Instagram* photos within Greater London and be comparable with the 2011 Census data from the ONS, we analyse the *Instagram* data at the level of MSOA by allocating each photograph to the corresponding MSOA using the recorded location of the photo. There are 983 MSOAs, covering all of Greater London's MSOAs, in which at least one *Instagram* photo was taken in between September 2015 and February 2016. English is the most commonly spoken language in all Greater London's MSOAs.

Excluding *Instagram* photos classified as mostly containing English by CLD, we find that *Instagram* photos labelled as other 19 non-English languages are clustered around Inner London while there are fewer photos in the Outer London area, as depicted in Fig. 4.1a. Amongst these 19 languages, we then map the most com-

Table 4.1: Top 20 spoken languages across Greater London based on *Instagram* usage between September 2015 and February 2016.

Top 20 most commonly spoken languages across Greater London based on *Instagram* language usage as classified by Chromium Language Detector (CLD), which aggregate to about 5.8 million photos (5,853,009) or 80% of the dataset (7,301,807). *Instagram* photos that are classified as mostly containing English captions and comments make up about 73% (5,320,973) of the dataset, while other 19 languages contribute only about 7% (532,036) of the dataset.

Language	Number of photos classified	Percentage of the dataset
English	5,320,973	72.87
Korean	66,097	0.91
Arabic	60,863	0.83
Spanish	60,778	0.83
Portuguese	55,542	0.76
Russian	53,500	0.73
Chinese	40,874	0.56
Italian	30,515	0.41
French	24,131	0.33
Japanese	21,732	0.30
Thai	17,529	0.24
Turkish	16,357	0.22
Danish	16,024	0.22
Swedish	15,686	0.21
Indonesian	12,845	0.17
Polish	11,082	0.15
German	10,361	0.14
Hebrew	6,392	0.08
Dutch	5,924	0.08
Greek	5,804	0.08

monly spoken language other than English based on *Instagram* language usage per MSOA which is shown in Fig. 4.1b. In each MSOA, we show the language with the highest number of photo counts amongst 19 non-English languages.

In Greater Manchester, there are 346 MSOAs in which we perform the same analysis for comparison purpose. We sample the *Instagram* photos that are geotagged within Greater Manchester between August 2013 and December 2013. This dataset contains about 0.8 million *Instagram* photos (803,439). The top 20 languages across Greater Manchester based on *Instagram* usage aggregate to about 0.6 million photos (637,249) or 79% of the dataset. *Instagram* photos with English language mostly are about 76% (612,098) of the dataset where as other top 19 languages makes up only about 3% (25,151). The most commonly spoken language based on *Instagram* usage in each MSOA across Greater Manchester are depicted in Fig. 4.2.

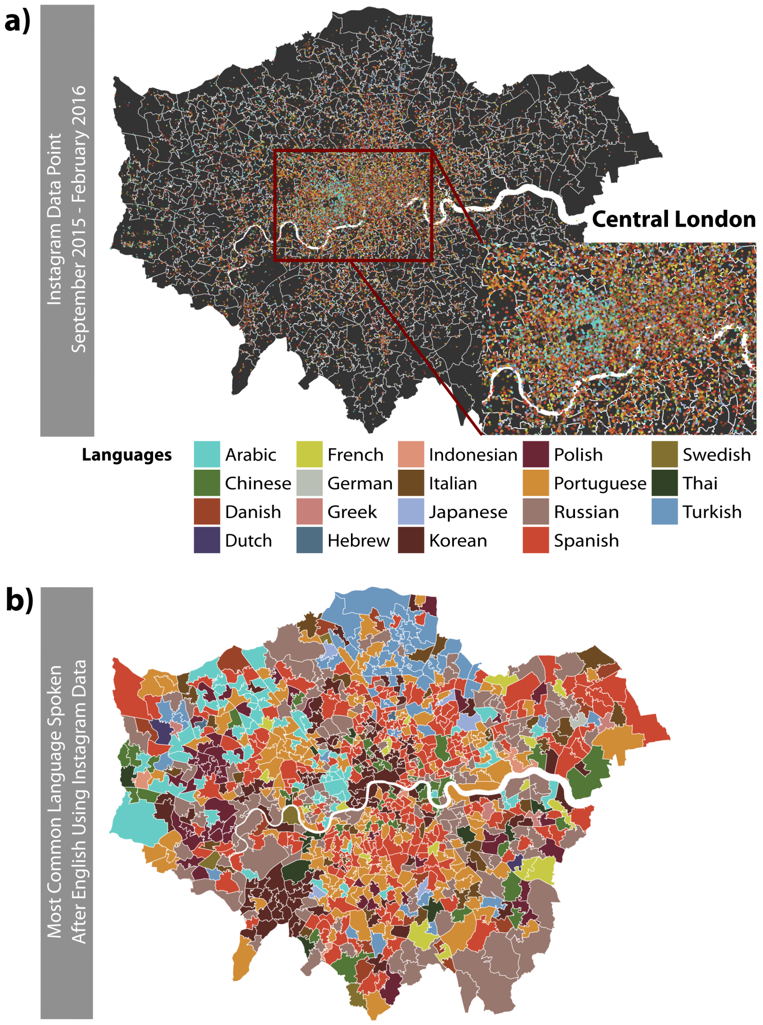

Figure 4.1: ***Instagram* measurements of language diversity in Greater London.**

(a) Locations of *Instagram* photos which were taken in Greater London within September 2015 and February 2016 filtered by using the top 20 most commonly spoken languages across Greater London from *Instagram* language usage as classified by Chromium Language Detector (CLD). Caption and comments from other users are submitted into CLD and record the language in which CLD reports the greatest proportion of the text has been written. There are 7.3 million (7,301,807) Instagram photos that were taken in Greater London area, of which there are 5.8 million (5,853,009) photos that are labelled amongst top 20 languages. Since 73% of photos are labelled as English, the map shows the locations of photographs for the other 19 languages. (b) Using all Instagram photos that are filtered earlier, we determine in which MSOA each Instagram photograph has been taken. Within September 2015 and February 2016, there are 983 MSOAs in which at least one Instagram photo was taken. We map the other 19 languages that are most commonly spoken in each MSOA other than English.

Table 4.2: Top 20 spoken languages across Greater Manchester based on *Instagram* usage between August 2013 and December 2013.

The top 20 languages across Greater Manchester based on *Instagram* usage aggregate to about 0.6 million photos (637,249) or 79% of the dataset (637,249). *Instagram* photos with English language mostly are about 76% (612,098) of the dataset where as other top 19 languages makes up only about 3% (25,151).

Language	Number of photos classified	Percentage of the dataset
English	612,098	76.18
Arabic	8,331	1.04
Danish	2,859	0.36
Chinese	2,406	0.30
Russian	2,178	0.27
Spanish	1,445	0.18
Thai	1,330	0.16
Indonesian	1,207	0.15
Portuguese	1,163	0.14
Afrikaans	657	0.08
Swedish	603	0.07
Italian	497	0.06
Polish	488	0.06
Japanese	430	0.05
Turkish	395	0.05
French	362	0.04
Manx	215	0.03
German	212	0.03
Dutch	194	0.02
Korean	179	0.02

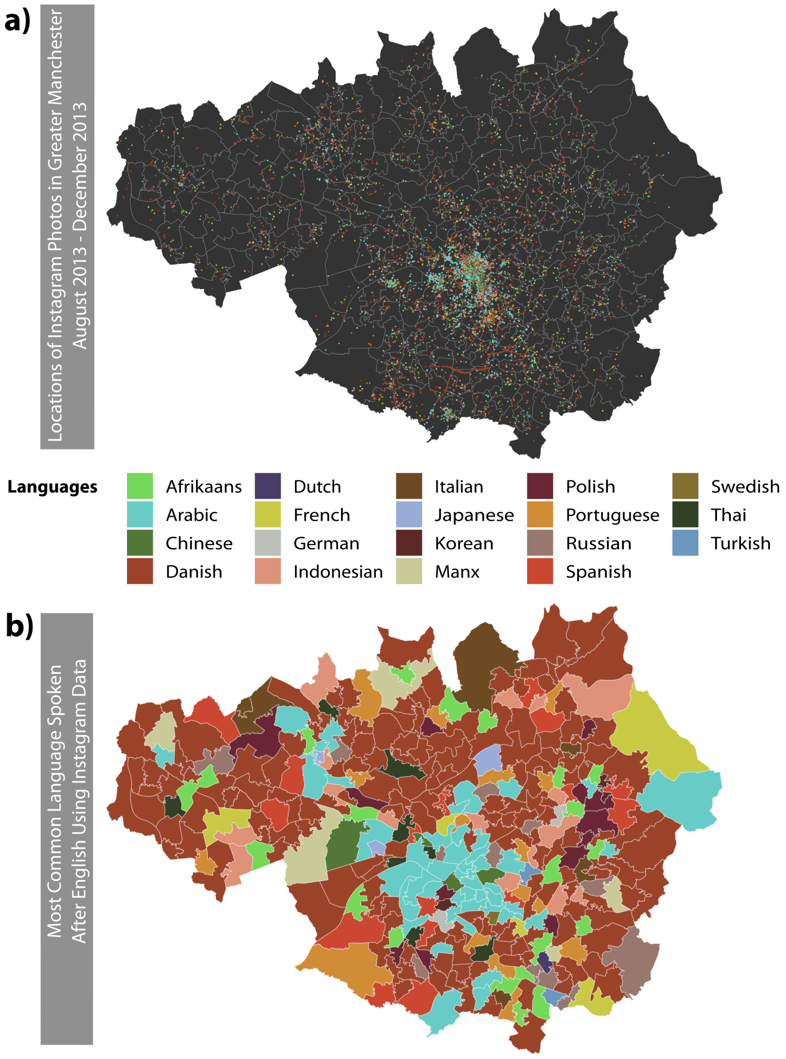

Figure 4.2: ***Instagram* measurements of language diversity in Greater Manchester.**

(a) Locations of *Instagram* photos which were taken in Greater Manchester within August 2013 and December 2013 filtered by using the top 20 most commonly spoken languages across Greater Manchester from *Instagram* language usage as classified by CLD. There are 0.8 million (803,439) Instagram photos that were taken in Greater Manchester area, of which there are 0.6 million (637,249) photos that are labelled amongst top 20 languages. Since 76% of photos are labelled as English, the map shows the locations of photographs for the other 19 languages that are most commonly spoken in each MSOA. (b) Using all Instagram photos that are filtered earlier, we determine in which MSOA each Instagram photograph has been taken. Within August 2013 and December 2013, there are 346 MSOAs in which at least one Instagram photo was taken. We map the other 19 languages that are most commonly spoken language in each MSOA other than English.

2.2 ONS Census data

In the 2011 Census data, there are 983 MSOAs across Greater London, and there are about 7.8 million (7,809,942) inhabitants of London who provided an answer to the main language question [204]. For Manchester, there are 346 MSOAs and about 2.6 million (2,572,731) respondents.

In all MSOAs for both areas, English is the most commonly spoken language. To be comparable with the *Instagram* data, we filter ONS data based on the top 20 most commonly spoken languages across London from *Instagram* language usage. Since English is the most commonly spoken language in all Greater London's MSOAs, Fig. 4.3 shows a map of the most commonly spoken language other than English in each MSOA across Greater London using ONS's Census data, and Fig. 4.4 depicts a similar map for the Greater Manchester area. It can be seen from the London map that some languages are clustered around neighbouring areas in which they have the same language that is the most commonly spoken. For example, a group of MSOAs where Turkish is the most commonly spoken language is clustered around the north of Greater London.

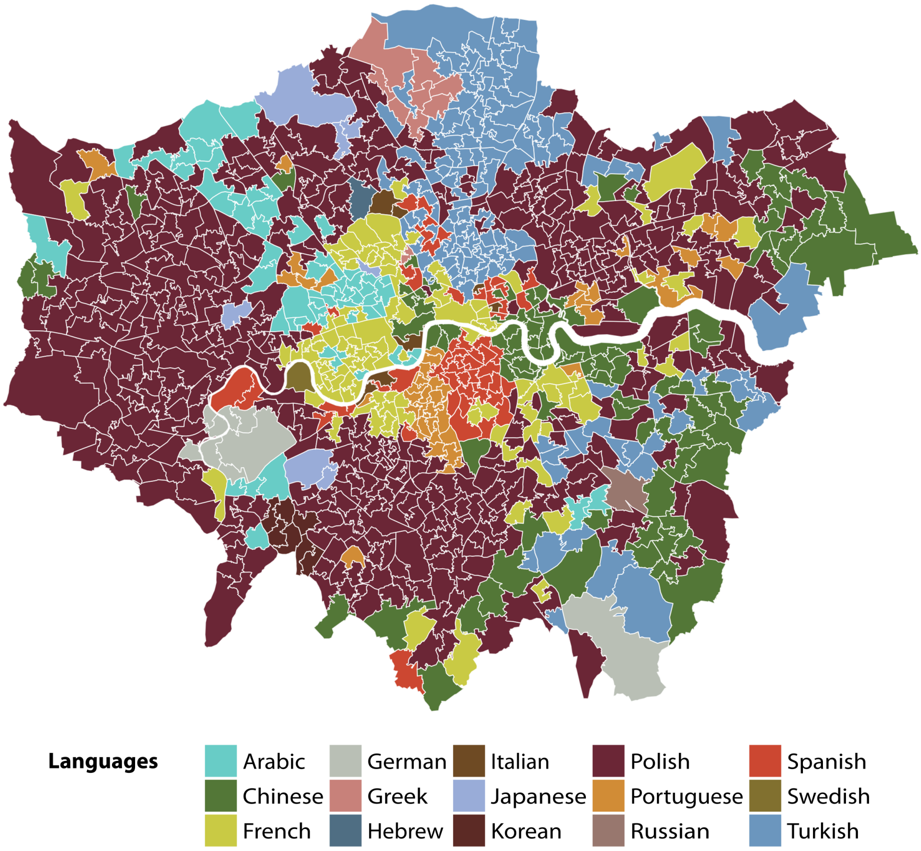

Figure 4.3: **Census measurements of language diversity in Greater London.**

The ONS Census was carried out in year 2011. Data from the Census is made available at various levels of spatial granularity. Here, we analyse data at the level of MSOAs or 'Middle Layer Super Output Areas'. An MSOA has a minimum population of 5,000 and maximum population of 15,000. Greater London comprises 983 MSOAs, the boundaries of which are depicted here. Our analysis investigates the top 20 most commonly spoken languages across Greater London according to *Instagram* usage. Since English is the most commonly spoken language in all Greater London's MSOAs, we produce a map of the top 19 languages other than English in each MSOA.

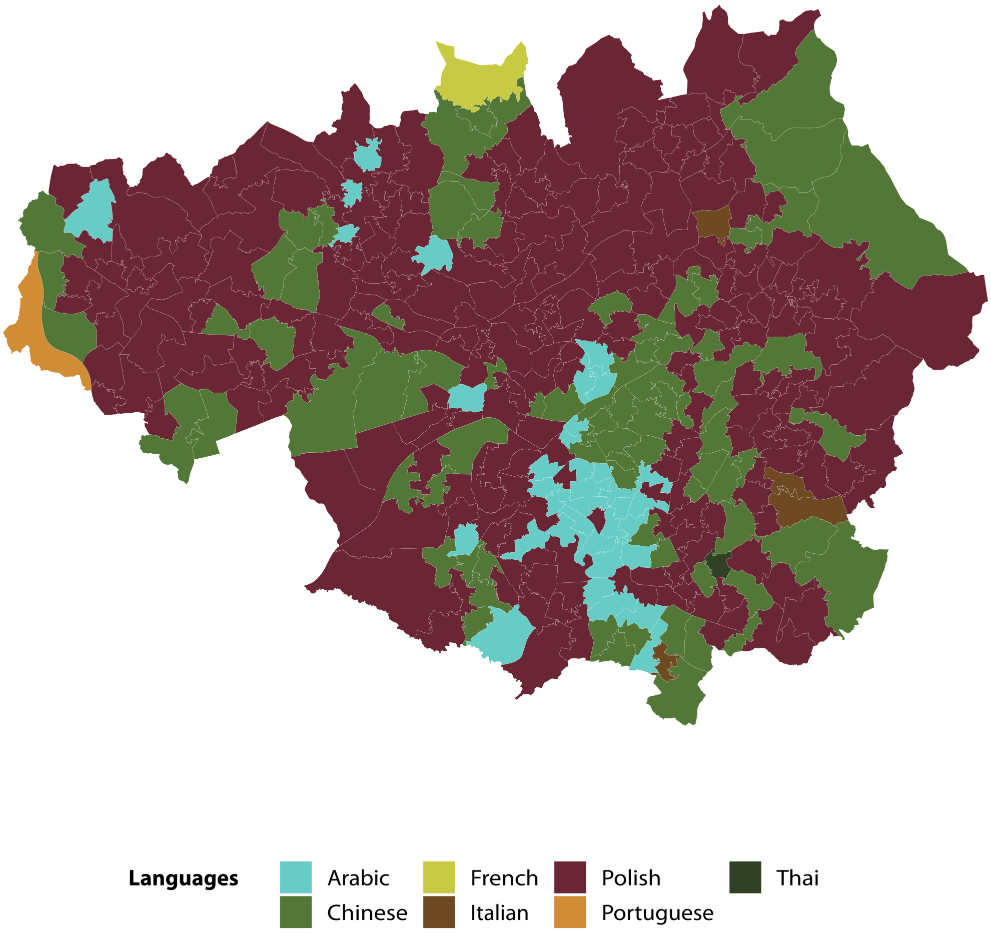

Figure 4.4: **Census measurements of language diversity in Greater Manchester.**

There are 346 MSOAs in Greater Manchester, the boundaries of which are depicted here. Our analysis investigates the top 20 most commonly spoken languages across Greater Manchester according to *Instagram* usage. Similarly to Greater London, English is the most commonly spoken language in all Greater Manchester's MSOAs. For this reason, we then map the top 19 most commonly spoken language other than English in each MSOA. According to the map, there are only seven out of 19 languages. Languages that are most commonly spoken and cover majority of MSOAs as can be seen from the map are Polish, Chinese, and Arabic. This suggests that Greater Manchester are less diverse in terms of language usage compared to Greater London.

3 Results

3.1 Greater London

The maps in Fig. 4.1b and Fig. 4.3 depict that London is a diverse city in terms of language spoken. English is excluded in the figures due to English is the most commonly spoken language in all of Greater London's MSOAs. Visual inspection reveals there is some correspondence between data retrieval from *Instagram* photos and ONS data. We then highlight MSOAs where the most common language spoken other than English according to ONS data is the same as the most common language spoken other than English in *Instagram* (Fig. 4.5). We find that the estimates of the most commonly spoken language other than English in 154 MSOAs out of 983 MSOAs are matched, such as Arabic, Polish, Spanish, and Turkish.

To investigate whether the language spoken in *Instagram* photos can be used to estimate the number of people who speak a particular language in each MSOA based on the top 20 languages found on *Instagram*, we first identify whether there is a hint of relationship between the number of *Instagram* photos and the number of respondents in ONS data. We then explore the extent to which *Instagram* photos can help estimate the percentage of spoken languages in different parts of Greater London and Greater Manchester. Finally, we investigate whether *Instagram* data would improve estimates generated by the baseline model which contains Census data only. We use the mean absolute error (MAE) to compare the predictive performance of these models. Then we perform Wilcoxon Signed Rank Test to find the evidence to support our findings of difference in performance between these models.

Across all MSOAs, the number of *Instagram* photos in each language, as recorded in our dataset after using the CLD to report the detected language with the largest proportion of the given text, is normalised by the total number of *Instagram* photos that were taken in Greater London (7,301,807). We then compare this with the number of main language usage for each language obtained from respondents of the ONS Census, which is normalised by the total number of respondents in the Census in Greater London area (7,809,942) (Fig. 4.6). However, with only 19 languages excluding English, we find no evidence of the statistical relationship, based on ranks, between the number of *Instagram* photos and the number of people who responded to the ONS Census ($\tau = 0.24$, $N = 19$, $p > 0.05$, Kendall's rank correlation). English is excluded in this statistical test due to huge difference, in terms of number of *Instagram* photos and main language usage obtained from the ONS, compared to other 19 languages, which would produce bias.

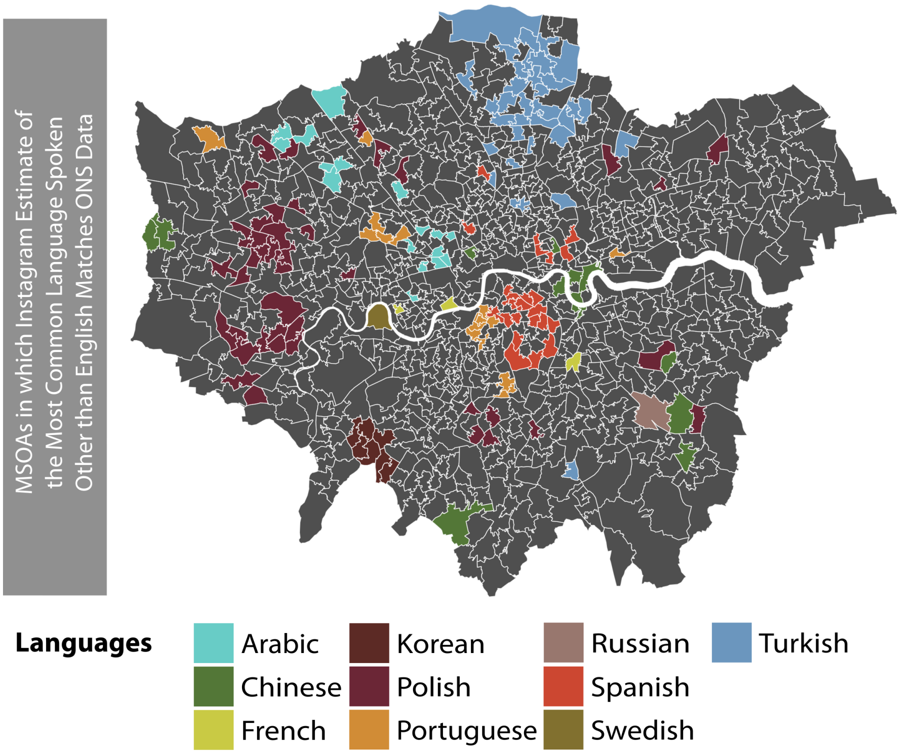
MSOAs in which Instagram Estimate of the Most Common Language Spoken Other than English Matches ONS Data
Languages
Arabic
Chinese
French
Korean
Polish
Portuguese
Russian
Spanish
Swedish
Turkish

Figure 4.5: **Identifying language usage relationship between *Instagram* photos and ONS Census data in Greater London.**

Based on the top 20 most commonly spoken languages across Greater London as determined from *Instagram* language usage, we identify and highlight MSOAs where the most common language spoken other than English according to ONS data is the same as the most common language spoken other than English according to *Instagram* data. There are 154 out of 983 MSOAs or 16% that are highlighted. The map reveals there are ten languages that MSOAs are matched and the majority of which are Arabic, Polish, Spanish and Turkish. This suggests that there is some correspondence between *Instagram* photos and ONS data.

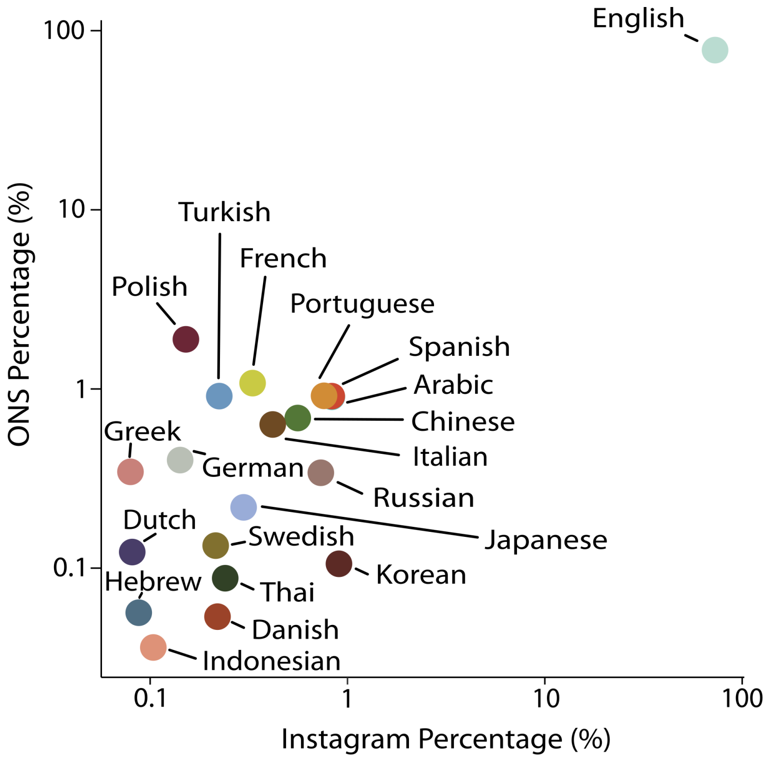

Figure 4.6: **Language usage percentages between *Instagram* photos and ONS Census data in Greater London.**

Across all MSOAs, we aggregate the number of *Instagram* photos as classified by CLD, normalise by the total number of *Instagram* photos in Greater London area (7,301,807) for each language. We then do the same with the number of Census respondents for each language and normalise by the total number of Census respondents in Greater London area (7,809,942). To compare percentages, we use log scales for both axis to produce the map. Excluding English language, we find no evidence for a correlation in terms of overall language usage percentages between Instagram and ONS data with only 19 languages ($\tau = 0.24$, $N = 19$, $p > 0.05$, Kendall's rank correlation).

For each language, for the MSOAs in which there is at least one *Instagram* photo, we investigate whether *Instagram* data can help in estimating the percentage of people in each MSOA who speak that language. As an example, we begin by building a model that provides an estimate for Arabic using logistic regression with the proportion of Arabic speakers across MSOAs as response variable and the percentage of *Instagram* photos reported as Arabic in each MSOA as explanatory variable (Fig. 4.7). To measure the amount of variance explained in each model, we use Nagelkerke R^2 [205]. It ranges from 0 to 1, where a value of 0 indicates that no variance is explained by the model, and a value of 1 means the all variance is explained by the model. Nagelkerke R^2 is a pseudo R^2, which is generalised from R^2 in linear regression, to measure the explained variation of the logistic regression model. A normal R^2, which is designed for linear regression, is non-applicable for logistic regression due to difference in parameter estimation approaches.

For an estimative model with Arabic *Instagram* photos, the Nagelkerke R^2 is 0.99 and this implies that including *Instagram* data helps explaining more variation and would generate more accurate estimates ($\beta = 0.311$, $N = 983$, $p < 0.001$, see Table 4.3). We conduct the same analysis for all top 20 languages and we find that including *Instagram* data offer value in estimating the percentages of people who speak in these languages in different parts of London as they help explain variations in the model (Fig. 4.8). To support our findings, we find strong evidence that all top 20 languages have significant relationship between the *Instagram* and the ONS Census data (*all* $\beta s > 0.04$, *all* $Ns = 983$, *all* $ps < 0.001$, see Table 4.3).

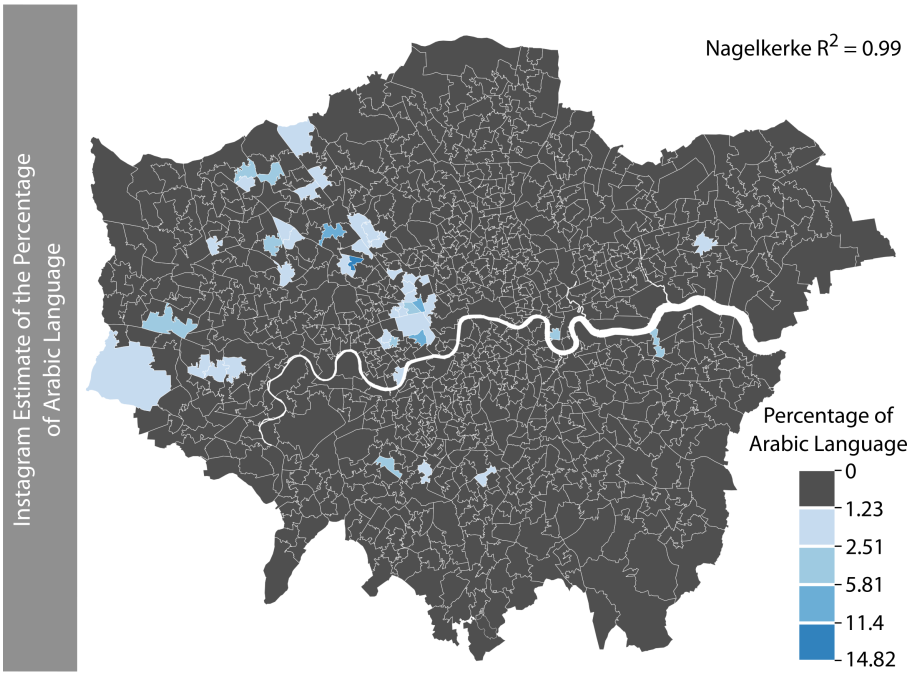
Instagram Estimate of the Percentage of Arabic Language
Nagelkerke R2 = 0.99
Percentage of Arabic Language
0
1.23
2.51
5.81
11.4
14.82

Figure 4.7: **Estimating the percentage of language spoken per MSOA in Greater London using *Instagram* photos.**

For each language, we investigate whether *Instagram* data can predict the percentage of people in each MSOA who speak that language. For example, we build an Arabic model using binomial logistic regression with the percentage of *Instagram* photos recorded as Arabic in each MSOA as an input to produce an estimated percentage of people who speak Arabic in each MSOA. We use Nagelkerke R^2 [205] to measure the amount of variance explained in each model. It ranges from 0 to 1, where a value of 0 indicates that no variance is explained by the model, and a value of 1 means the all variance is explained by the model. For an Arabic model, the Nagelkerke R^2 is 0.99 and this implies the Arabic model that includes *Instagram* data helps explaining the variance and would generate more accurate estimates ($\beta > 0.311$, $N = 983$, $p < 0.001$).

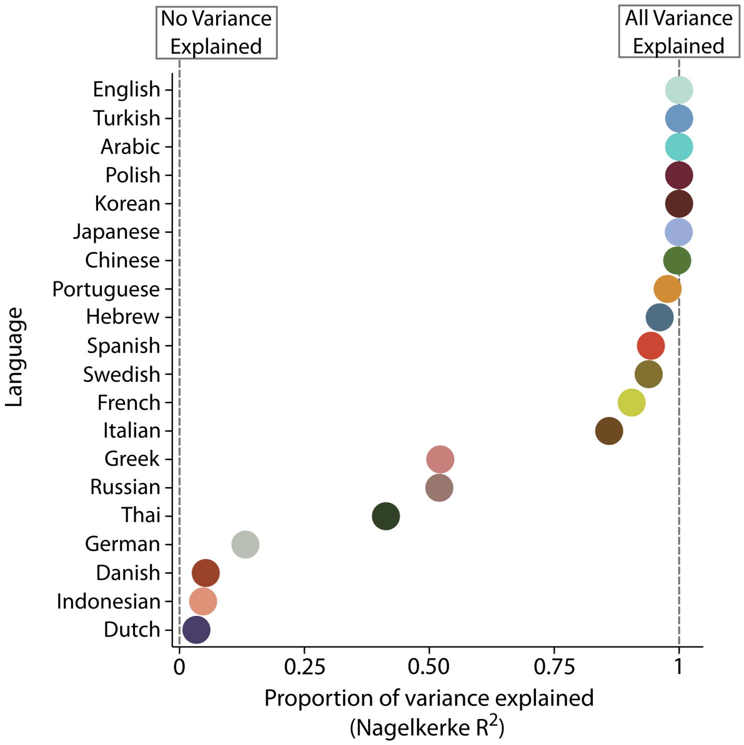

Figure 4.8: **Proportion of variance explained for each language model that include *Instagram* photos.**

We carry out the same analysis for all 20 languages and investigate the performance of *Instagram* data in estimating the ONS statistics. We find that language models including *Instagram* data offer value in producing accurate estimates of the language usage percentages in different parts of London. All top 20 languages have significant relationship between *Instagram* and ONS Census data ($all\ \beta s > 0.04$, $all\ Ns = 983$, $all\ ps < 0.001$, see Table 4.3).

Table 4.3: Results generated by logistic regression models on top 20 spoken languages across Greater London based on *Instagram* usage for the period between September 2015 and February 2016.

Each language model uses the percentage of *Instagram* photos in each MSOA as predictor while the proportion of ONS respondents is the dependent variable. The table is ordered by higher Nagelkerke R^2. We find that all top 20 languages have significant relationship between *Instagram* and ONS data. The higher number of Nagelkerke R^2 or the amount of variance explained by the model suggests the value of *Instagram* data in improving the percentage estimates of spoken language.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Language	Beta/ Coefficient	Nagelkerke R^2	Number of photos classified	Percentage of the *Instagram* dataset	Number of speakers according to ONS	Percentage of the ONS dataset
English	0.041***	1	5,320,973	72.87	6,083,420	77.89
Turkish	0.534***	1	16,357	0.22	71,242	0.91
Arabic	0.311***	0.999	60,863	0.83	70,602	0.90
Polish	0.268***	0.999	11,082	0.15	147,816	1.89
Korean	0.395***	0.999	66,097	0.91	8,257	0.11
Japanese	1.857***	0.999	21,732	0.30	17,050	0.22
Chinese	0.380***	0.996	40,874	0.56	53,759	0.69
Portuguese	0.137***	0.977	55,542	0.76	71,525	0.92
Hebrew	3.161***	0.961	6,392	0.08	4,403	0.06
Spanish	0.276***	0.943	60,778	0.83	71,192	0.91
Swedish	2.328***	0.939	15,686	0.21	10,428	0.13
French	0.489**	0.905	24,131	0.33	84,191	1.08
Italian	0.403***	0.860	30,515	0.41	49,484	0.63
Greek	0.684***	0.522	5,804	0.08	26,924	0.34
Russian	0.458***	0.520	53,500	0.73	26,603	0.34
Thai	0.693***	0.413	17,529	0.24	6,859	0.09
German	0.329***	0.132	10,361	0.14	31,306	0.40
Danish	0.513***	0.053	16,024	0.22	4,185	0.05
Indonesian	0.553***	0.047	12,845	0.17	2,817	0.04
Dutch	0.432***	0.034	5,924	0.08	9,603	0.12

To investigate the performance of the model using *Instagram* data, we use 20-fold cross-validation [206, 207] and compare the baseline models which include Census data only, and models that include *Instagram* data. Specifically, we randomly divide all MSOAs into 20 subsets, such that each subset contains 5% of total MSOAs or between 48 and 51 MSOAs. For each subset, we use that particular observations subset as the test set, while the other 19 subsets are combined as the training set. The training set are then used to build the model to estimates the values of remaining MSOAs, which are the test set. To compare and verify the results, the difference in values for each MSOA between the model estimates and the testing set is then recorded. This process, thus, is repeated 20 times, using each subset only once as the testing set. For each language, we fit a baseline logistic model on the training data using Census data only, which contains the percentage of language speakers across MSOAs. The baseline model estimates the mean of the percentage of language speakers across all MSOAs of the training data. The estimated mean is then used to represent the remaining MSOAs. We also build an *Instagram* model including both Census and *Instagram* data in which *Instagram* data contain the percentage of language speaker across MSOAs based on *Instagram* usage. Since we are interested in the top 20 languages, we repeat this process for each language. In each process, baseline and *Instagram* models are fitted to estimate the percentage of language speakers across MSOAs on the testing data. We then record errors or differences in values between estimates from these models and actual percentage from the testing set for each language and across all MSOAs.

We then compare the performance in terms of accuracy between the baseline and *Instagram* model for 20 individual languages by looking at which model has lower error rates, as measured by the Mean Absolute Error (MAE). Our findings show that 14 *Instagram* models provide better accurate estimates than baseline models in estimating the percentage of language spoken per MSOA (see Table 4.4). For 20 different languages, we then perform the Wilcoxon Signed Rank test to determine the significance of differences, in terms of rank, between errors generated by the *Instagram* model and the baseline model. We find evidence that there are 14 languages from *Instagram* models which improve estimates and are more accurate than the baseline models (see Table 4.4). English, Korean, Arabic, Chinese, Turkish, and Swedish are languages that have better performance than the others. These languages are well-represented in both Census and *Instagram* data in which *Instagram* data complements the Census data. Across the top 20 languages, the improvement as measured by the reduction in MAE using the *Instagram* models ranges from -9.9% to 11.3% compared to baseline models. Conversely, our results provides an evidence that there are 6 languages in which *Instagram* models show no improvements. This suggests that *Instagram* data can help estimate language usage to certain languages.

We build a linear mixed-effects model [208] to verify the overall performance between baseline and *Instagram* model across 20 languages by estimating absolute errors obtained from both models. To account for variation, the model assumes different random intercepts for languages and MSOAs. To do this, we specify model type (Instagram or baseline) as a fixed effect to estimate absolute errors, controlling for by-language and by-MSOA variability which are random effects. Our findings reveal that the *Instagram* model generates more accurate estimates than baseline models (β baseline = 0.008, β Instagram = -0.0005, N = 39320, p baseline = 0.07, p Instagram $<$ 0.001). Our results suggest that including *Instagram* data improve estimates within Greater London.

Table 4.4: 20-fold cross-validation results of logistic regression models on top 20 languages across Greater London.

All MSOAs are randomly split into 20 subsets in which each subset contains 5% of total MSOAs or between 48 and 51 MSOAs. We fit 20 logistic regression models for individual languages on the training data (19 subsets) using Census data only. Another 20 individual models are built by including *Instagram* data: the percentage of language usage per MSOA. The error rates are measured by using the Mean Absolute Error (MAE) and the units are in percentage of language speakers. This process is carried out repeatedly 20 times until every subset is used once as the test set and then, for each language, we average the MAE from 20 MAEs. By performing the Wilcoxon Signed Rank test, we find evidence that *Instagram* models improve estimates for 14 languages. The improvement as measured by reduction in MAE using *Instagram* models ranges from -9.9% to 11.3%. However, there are 6 languages that *Instagram* data provides no improvement to the estimates. The table is ordered by *Instagram* language usage. In the Wilcoxon Signed Rank test column, the star means the errors are significantly different ($p < 0.05$). The p-values are corrected using false discovery rate (FDR) correction to control the proportion of false positives.

Language	Baseline Model	Instagram Model	% Difference	Wilcoxon Test V
English	9.589	**8.656**	9.73	314635*
Korean	0.129	**0.114**	11.30	415481*
Arabic	0.720	**0.667**	7.19	395630*
Spanish	0.654	**0.638**	2.49	316439*
Portuguese	**0.575**	0.609	-6.00	348226*
Russian	**0.204**	0.225	-9.98	337959*
Chinese	0.432	**0.411**	4.66	374732*
Italian	0.447	**0.438**	2.03	321634*
French	0.699	**0.694**	0.70	310750*
Japanese	0.216	**0.209**	3.00	359132*
Thai	0.06	**0.059**	1.85	323604*
Turkish	0.992	**0.893**	9.97	404770*
Danish	0.0473	**0.0472**	0.03	280130*
Swedish	0.124	**0.116**	6.57	349335*
Indonesian	**0.035**	0.036	-0.27	305347*
Polish	1.329	**1.288**	3.09	373612*
German	**0.299**	0.2996	-0.19	301842*
Hebrew	0.076	**0.073**	3.79	400000*
Dutch	**0.0847**	0.0848	-0.14	311211*
Greek	**0.317**	0.318	-0.44	357653*

3.2 Greater Manchester

We investigate whether results from other large cities such as Manchester, are in line with Greater London's results by using *Instagram* data. Using the top 20 most commonly spoken languages across Greater Manchester, we identify and highlight MSOAs where the most common language spoken other than English according to the Census is the same as in *Instagram* data (Fig. 4.9). English is also excluded in the figures due to English is the most commonly spoken language in all Greater Manchester's MSOAs. The map reveals there is some correpondence between the number of *Instagram* photos and ONS data respondents in each MSOA although the number of correspondents MSOAs are less than Greater London. For Greater Manchester, there are only 7% of all MSOAs that are matched (24 out of 346 MSOAs). Also, there are only three languages, which are Arabic, Chinese, and Polish, that are identified as matched MSOAs.

Covering all MSOAs in Greater Manchester, the number of *Instagram* photos for each language as classified by CLD is aggregated and normalised by the total number of *Instagram* photos in Greater Manchester area (803,439). For each language, the number of Census respondents is normalised by the total number of Census respondents in Greater Manchester area (2,572,731) (Fig. 4.10). Similarly, with only 19 languages excluding English, we find no evidence in terms of statistical relationship between the number of *Instagram* photos and the number of people who responded to the ONS Census ($\tau = 0.17$, $N = 19$, $p > 0.05$, Kendall's rank correlation).

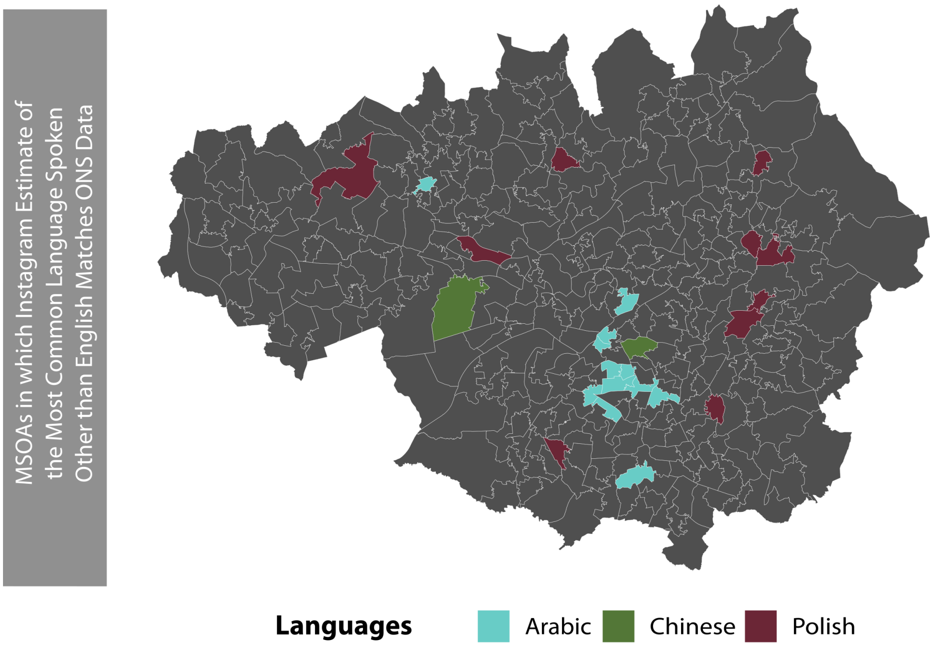
MSOAs in which Instagram Estimate of the Most Common Language Spoken Other than English Matches ONS Data
Languages
Arabic
Chinese
Polish

Figure 4.9: **Identifying language usage relationship between *Instagram* photos and ONS Census data in Greater Manchester.**

Based on the top 20 most commonly spoken languages across Greater Manchester as determined from *Instagram* language usage, we identify and highlight MSOAs where the most common language spoken other than English according to ONS data is the same as the most common language spoken other than English in *Instagram*. There are 24 out of 346 MSOAs or 7% that are highlighted. Arabic, Chinese and Polish are three languages that are identified as matched MSOAs. The map reveals there is some correspondence between *Instagram* photos and ONS data although the proportion of matched MSOAs is smaller than Greater London.

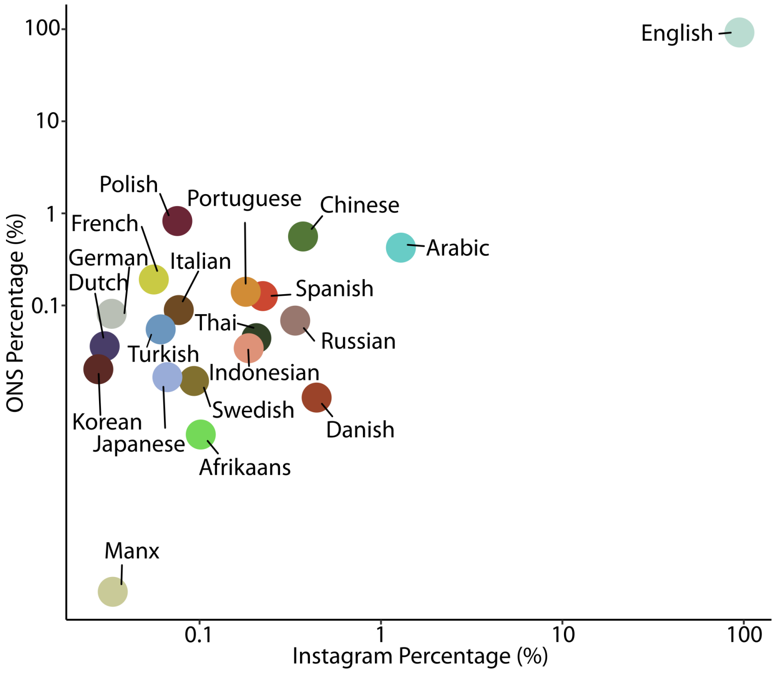

Figure 4.10: **Language usage percentages between *Instagram* photos and ONS Census data in Greater Manchester.**

Across all MSOAs and for each language, we aggregate the number of *Instagram* photos as classified by CLD, normalise by the total number of *Instagram* photos in Greater Manchester area (803,439). We then do the same with the number of Census respondents for each language and normalise by the total number of Census respondents in Greater Manchester area (2,572,731). To compare percentages, we use log scales for both axis to produce the map. Excluding English language, we find no evidence for a correlation in terms of overall language usage percentages between Instagram and ONS data with only 19 languages ($\tau = 0.17$, $N = 19$, $p > 0.05$, Kendall's rank correlation).

To investigate whether *Instagram* photos can also improve estimates on other large cities such as Manchester, we explore the extent to which *Instagram* photos can help estimate the percentage of spoken languages in different parts of Greater Manchester. For each language, we build logistic regression with the percentage of *Instagram* photos in each MSOA to estimate the percentage of language usage in each MSOA. Since we are also interested in top 20 languages based on *Instagram* usage in Greater Manchester, we repeat this step for top 20 languages. By comparing the amount of variance explained using Nagelkerke R^2 as we did with Greater London, we find that there are some languages in which *Instagram* data provide value in improving the percentage estimates of spoken language (Fig. 4.11). There are 16 languages out of the top 20 commonly spoken languages across Greater Manchester based on *Instagram* usage that have significant relationship (see Table 4.5).

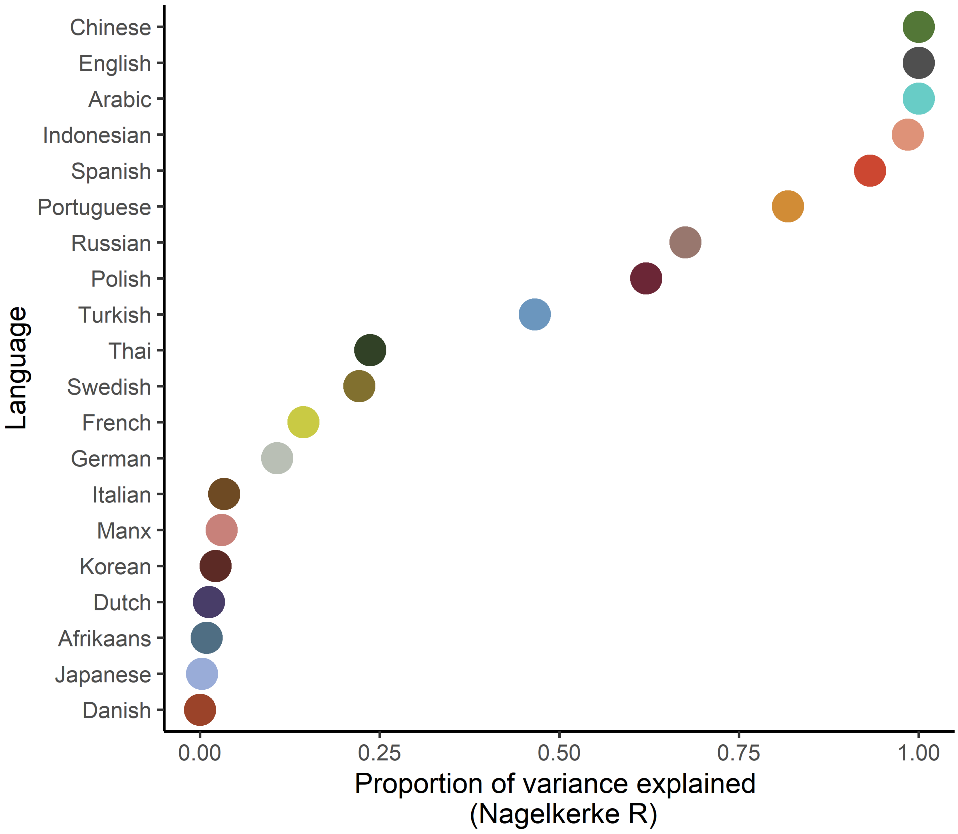

Figure 4.11: **Estimating the percentage of language spoken per MSOA in Greater Manchester using *Instagram* photos.**

The Nagelkerke R^2 are calculated for all 20 languages to investigate the potential of *Instagram* data in estimating the ONS statistics. We find that language models including *Instagram* data offer values in producing a good estimate of the percentages in different parts of Manchester. All top 20 languages have significant relationship ($all\ \beta s > 0.04$, $all\ Ns = 983$, $all\ ps < 0.001$, see Table 4.5).

Table 4.5: Results generated by logistic regression models on top 20 spoken languages across Greater Manchester based on *Instagram* usage for the period between August 2013 and December 2013.

Each language model uses the percentage of *Instagram* photos in each MSOA as predictor while the proportion of ONS respondents is the dependent variable. The table is ordered by higher Nagelkerke R^2. By using logistic regression models as we did with Greater London, we find that there are 16 languages out of the top 20 commonly spoken languages across Greater Manchester based on *Instagram* usage that have significant relationship. Also, the higher number of Nagelkerke R^2 or the amount of variance explained by the model suggests the value of *Instagram* data in improving the percentage estimates of spoken language.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Language	Beta/ Coefficient	Nagelkerke R^2	Number of photos classified	Percentage of the *Instagram* dataset	Number of speakers according to ONS	Percentage of the ONS dataset
English	0.104***	1	612,098	76.18	2,370,094	92.12
Chinese	1.213***	1	2,406	0.30	14,431	0.56
Arabic	0.389***	1	8,331	1.04	10,914	0.42
Indonesian	1.041***	0.984	1,207	0.15	883	0.03
Spanish	1.706***	0.932	1,445	0.18	3,252	0.12
Portuguese	0.986***	0.818	1,163	0.14	3,620	0.14
Russian	1.043***	0.675	2,178	0.27	1,761	0.07
Polish	0.279***	0.621	488	0.06	21,231	0.82
Turkish	3.069***	0.465	395	0.05	1,414	0.05
Thai	0.472***	0.236	1,330	0.16	1,136	0.04
Swedish	3.116***	0.221	603	0.07	389	0.01
French	0.965***	0.144	362	0.04	4,924	0.19
German	2.809***	0.107	212	0.03	2,088	0.08
Italian	0.544***	0.033	497	0.06	2,292	0.09
Manx	4.481	0.031	215	0.03	2	0.01
Korean	0.644**	0.022	179	0.02	522	0.02
Dutch	1.316*	0.013	194	0.02	927	0.04
Afrikaans	-1.342	0.009	657	0.08	102	0.01
Japanese	0.101	0.003	430	0.05	428	0.02
Danish	0.011	0.001	2,859	0.36	257	0.01

We investigate the performance of the model in Greater Manchester area as we did with Greater London. The 20-fold cross-validation is used to compare estimates between baseline models that include Census data only and models that incorporate *Instagram* language usage data. In the case of Greater Manchester, we also randomly divide all 346 MSOAs that are in Greater Manchester into 20 subsets such that each subset contains 5% of total MSOAs. Each subset contains between 327 and 330 MSOAs to train the model and estimate the percentage of language usage in between 16 and 19 MSOAs. By comparing error rates using the MAE, we find that there are 11 languages out of the top 20 commonly spoken languages based on *Instagram* usage in which models that include *Instagram* data provide better estimates in terms of accuracy. Furthermore, our results provide evidence that there are 10 *Instagram* models generating better estimates than baseline models. *Instagram* models across the top 20 languages provide reduction in MAE ranging from -104% to 23% compared to baseline models (see Table 4.6). On the contrary, there are 9 languages that have no improvement in estimates from *Instagram* models. Our findings reveal evidence that there are 7 languages in which *Instagram* models provide no improvement.

Similar to the Greater London analysis, we build a linear mixed-effects model to verify the overall estimating performance between baseline and *Instagram* model across 20 languages. However, our analysis do not provide evidence that the *Instagram* model generates more accurate estimates than baseline models (β baseline = 0.004, β Instagram = -0.0001, N = 13840, p baseline = 0.16, p Instagram = 0.47).

Table 4.6: 20-fold cross-validation results of logistic regression models on top 20 languages across Greater Manchester.

We randomly split all Greater Manchester's MSOAs into 20 subsets. One subset contains 5% (16-19) of total MSOAs and it is used as the test set. Similar to the Greater London analysis, we used the same process to fit 20 logistic regression models using Census data only and another 20 models that include *Instagram* data. 20 MAEs are then averaged to compare the performance between baseline and *Instagram* models, including Wilcoxon Signed Rank test. We find some evidence that *Instagram* models improve estimates for ten languages. Amongst the top 20 languages, *Instagram* models provide reduction in MAE ranging from -104% to 23%. On the other hand, our results show evidence that *Instagram* models provide no improvements on 7 languages. The table is ordered by *Instagram* language usage. The star in Wilcoxon Signed Rank test column indicates the errors are significantly different in terms of rank ($p < 0.05$). To control the proportion of false positives, FDR correction is used to adjust p-values.

Language	Baseline Model	Instagram Model	% Difference	Wilcoxon Test V
English	6.365	**5.964**	6.29	47062*
Arabic	0.454	**0.379**	16.33	53024*
Danish	**0.012**	0.013	-2.41	37508*
Chinese	0.498	**0.383**	23.03	50390*
Russian	0.063	**0.059**	5.87	43601*
Spanish	0.119	**0.114**	4.15	46140*
Thai	**0.032**	0.033	-1.94	39378*
Indonesian	**0.047**	0.054	-15.47	53322*
Portuguese	0.133	**0.130**	2.29	43252*
Afrikaans	0.007	**0.006**	0.56	30895
Swedish	0.018	**0.017**	0.74	41504*
Italian	**0.062**	0.065	-4.61	33451
Polish	**0.615**	0.839	-36.51	41980*
Japanese	**0.021**	0.042	-104.39	43597*
Turkish	0.046	**0.045**	1.92	40941*
French	**0.165**	0.167	-1.18	36577*
Manx	**0.001**	0.002	-6.36	44827*
German	0.061	**0.060**	1.17	36225*
Dutch	**0.033**	0.034	-0.38	31951
Korean	0.029	**0.028**	3.46	47453*

4 Discussion

In England and Wales, measuring language spoken is so costly and time consuming that it is only conducted once every ten years. Although there is a growing number of studies of distribution of language usage across areas using spatial data from other online media platforms, the number of such research on a photo sharing website is lacking. Also, the potential in estimating the spatial patterns of language usage with data from a photo sharing website is unclear. To answer this question, we focus on the ONS Census and *Instagram* data on two large cities as case study which are Greater London and Greater Manchester.

Looking at Greater London, we compare the measurements of language diversity from the ONS Census in year 2011 and *Instagram* photos taken between September 2015 and February 2016. We select the top 20 most commonly spoken languages across Greater London from the *Instagram* language usage as they represent the major population and their estimates would be more stable than less common languages. Across the top 20 languages, we use Nagelkerke R^2 to evaluate the relationship between the *Instagram* and ONS data. Afterwards, to validate the estimation performance, we use 20-fold cross-validation tests on both the baseline model, which contains ONS data only, and the *Instagram* model in order to compare the error rates. The 20-fold cross-validation test assesses model performance in estimating the spatial distribution of languages on different subsets of London areas (MSOAs). Based on Nagelkerke R^2 and the significance of coefficients, our analysis shows that all top 20 most commonly spoken languages, as labelled in *Instagram* photos, offer values in producing an estimate of the percentage of people who speak in these particular languages in different parts of London. However, we find no evidence of relationship, in terms of rank, between the number of *Instagram* photos and the number of people who responded their main language to the ONS Census. Based on cross-validation test and mixed-effect analysis, our findings provide evidence that there are 14 languages in which models that include *Instagram* data can generate more accurate estimates across London areas (MSOA) than baseline models consisting Census data only. Similarly, we perform the same analysis on another large city which is Manchester. Comparing to London, Census data shows that Manchester is less diverse in terms of language diversity while *Instagram* data reveals more language diversity. Our out-of-sample results reveal that there are ten languages out of top 20 commonly spoken languages based on *Instagram* usage that *Instagram* models have potential to generate better estimates, which is in line with the London analysis. However, based on mixed-effect analysis, we find no evidence that *Instagram* data improve estimates within Greater Manchester.

We note the time gap between the 2011 Census data and the *Instagram* data which spans from September 2015 to February 2016 where there is no official data to compared to. Here, we have only been able to investigate the use of recent *Instagram* data for estimating the spatial distribution of languages. This provides an opportunity for future work to investigate the potential for improving estimates across time, once new Census data becomes available after 2021. Nonetheless, our findings underline the potential of rapidly available online data that people are generating, such as *Instagram* in this case, to measure key population statistics at low cost by complementing official data.

Chapter 5

Estimating language statistics using Twitter data

1 Introduction

This thesis has investigated estimating language usage statistics across space using *Instagram* data in the previous chapter. However, it remains unclear whether other online data sources would provide results that are in line with the results in the *Instagram* chapter. Apart from *Instagram*, *Twitter* is a large-scale online platform where people mainly exchange short messages. *Instagram* allows users to upload photos and videos to interact with other users, while *Twitter*'s main communication method is the uploading of short messages or "tweets" to the online platform. For *Instagram*, geographic information (latitude and longitude) can be attached to the photographs, while, on *Twitter*, this information is attached to the messages. Also, a tool is required, such as Chromium Language Detector (CLD) as used in the previous chapter, to obtain language usage information on messages exchanged on *Instagram* photographs. On the other hand, *Twitter* provides language usage information with individual messages. This chapter will investigate whether another types of data, specifically messages exchanged on and language information provided by the *Twitter* platform would help in estimating language usage statistics. This would help policymakers or stakeholders to understand whether different types of online data sources will provide similar results or different perspectives of the current state of the society. It will also complement the investigation of the usefulness of online data by considering multiple data sources. In addition to MSOAs analysis, this chapter will consider a higher level of granularity using borough-level tagged places information as provided by *Twitter* for estimating language statistics.

Recent studies have investigated language statistics across space using data from *Twitter* by looking at the large picture of language and geography which suggests the potential to estimate language statistics using online data [33, 50–53]. However, the specific aspects of estimating distribution of languages across areas are lacking.

Here, we use six months of *Twitter* data between January and June 2018 in order to understand the relationship between tweets data with specific geolocations and official language statistics and to determine whether tweets data have potential to estimate language statistics across different areas in a city level by exploring Greater London. In order to quantify the extent of improvement in estimates from tweets, we compare estimation performance on the proportion of spoken languages across London by building and comparing baseline model with models that includes tweets data. Lastly, we include additional six months of *Twitter* data from July 2017 to December 2017 and investigating whether one year of *Twitter* data on tagged places can be used to estimate language statistics in London boroughs.

2 Materials and methods

We retrieve 0.9 million (932,662) *Twitter* messages known as tweets that were uploaded in Greater London from 1 January 2018 to 30 June 2018 via the *Twitter* API. Each tweet contains metadata which provides when and where (if specified by users) the message was uploaded along with the language detected by the *Twitter* language identifier, which is a proprietary classifier owned by *Twitter*. English is the most commonly used language in the dataset with 89 percent (838,465) of tweets. Other 49 languages are detected by the *Twitter* language identifier, which contribute to seven percent of tweets (68,641). Similar to the *Instagram* analysis, we select the top 20 most commonly used languages based on *Twitter* data. As a result, there are 0.9 million tweets (903,583) across top 20 languages, which are 96% of the dataset. Here, we visualise the locations of tweets of top 19 most commonly used languages other than English across Greater London in our *Twitter* dataset as shown in Fig. 5.1a. The counts of tweets uploaded in each language are listed in Table 5.1. We note that on the same length of period to the *Instagram* dataset from the previous chapter, although with different months and years, the number of tweets (932,662) is significantly smaller than the number of *Instagram* photos (7,301,807) in our dataset. Since photographs are more likely to have geolocations attached due to GPS-enabled technology in mobile phones and cameras than simple short messages, this might explain the huge difference in terms of quantity.

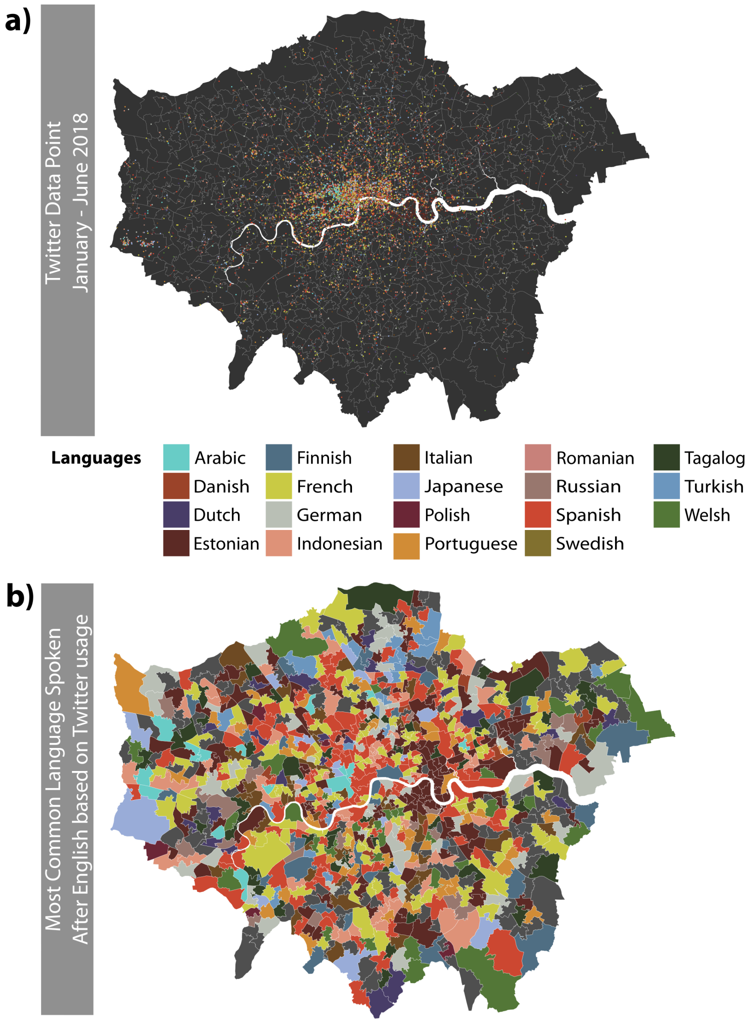

Figure 5.1: **Distribution of language diversity in Greater London using *Twitter* data.**

(a) Locations of tweets uploaded in Greater London between January and June 2018 filtered by top 20 most commonly used languages across London in the *Twitter* dataset. Each tweet contains metadata in which it provides when and where (if specified by users) the message was uploaded along with the language classified by *Twitter* language identifier. There are 0.9 million (932,662) tweets that were published in Greater London area, of which there are 0.9 million (903,583) tweets that are classified within top 20 languages. Since 89% of tweets are labelled as English, the map shows the locations of tweets for the remaining 19 languages. (b) For each tweet that are labelled within top 20 languages, we assign the MSOA corresponds to the location the tweet was uploaded or specified. For top 19 commonly used languages other than English, there are 851 MSOAs in which at least one tweet was published between January and June 2018. The map shows the top 19 most commonly used languages other than English in each MSOA.

Table 5.1: Top 20 commonly used languages based on *Twitter* language usage between January and June 2018.

We rank top languages based on *Twitter* language usage as classified by *Twitter* language identifier. We investigate on top 20 languages that have data available both in *Twitter* counts and ONS Census respondents. Hence, we remove Haitian, Norwegian, and Catalan from our analysis. The aggregated number of tweets across top 20 languages is 0.9 million (903,583) or 96% of the dataset (932,662). About 89 percent (838,465) of tweets are classified as English which is the largest proportion of the dataset whereas other 49 detected languages makes up only about seven percent (68,641).

Language	Number of tweets classified	Percentage of the *Twitter* dataset	Number of the 2011 Census respondents	Percentage of the ONS dataset
English	838,465	89.90	6,083,420	77.89
Spanish	10,689	1.14	71,192	0.91
French	8,219	0.88	84,191	1.07
Finnish	8,037	0.86	2,800	0.03
Portuguese	4,946	0.53	71,525	0.91
Italian	4,397	0.47	49,484	0.63
Estonian	3,526	0.37	1,192	0.01
Indonesian	3,417	0.36	2,817	0.03
German	3,338	0.35	31,306	0.40
Japanese	2,633	0.28	17,050	0.21
Arabic	2,182	0.23	70,602	0.90
Tagalog	1,715	0.18	25,869	0.33
Welsh	1,653	0.17	1,310	0.01
Turkish	1,626	0.17	71,242	0.91
Dutch	1,356	0.14	9,603	0.12
Haitian	1,198	0.12	-	-
Swedish	1,170	0.12	10,428	0.13
Danish	1,105	0.11	4,185	0.05
Norwegian	925	0.09	-	-
Russian	843	0.09	26,603	0.34
Catalan	783	0.08	-	-
Romanian	708	0.07	39,653	0.50
Polish	652	0.06	147,816	1.89

For the final analysis, there are 0.7 million tweets (731,631) that are automatically tagged with London borough names or tagged with locations within London boroughs by *Twitter* platform. We also include additional data of tweets from July 2017 to December 2017. This new data contains 0.9 million tweets (946,895) that are tagged as one of London boroughs or located within London. Therefore, the combined tweets dataset contains about 1.7 million messages (1,678,526) covering one year of data. Table 5.2 shows the number of tweets of combined dataset for each language.

The latest 2011 Census data reveals that about 7.8 million (7,809,942) inhabitants of Greater London have provided an answer to the main language question [204]. Census data reports area statistics at various granularities and we extract data at Middle Layer Super Output Area (MSOA) level and at Borough level. In Greater London, there are 983 MSOAs and the boundaries of which are depicted in Fig. 5.1a. Each MSOA has a population size of between 5,000 and 15,000 [198]. Greater London is divided into 32 boroughs and the City of London. The Census data reveals that English is the most commonly used language across all MSOAs and boroughs in Greater London.

We first analyse the tweets data at MSOA level by assigning corresponding MSOA to each tweet which contains geographical coordinates. Across 983 MSOAs in Greater London, English is the most commonly spoken language. For top 19 languages after English, there are 851 MSOAs in which at least one tweet was published between January and June 2018. The most commonly used language other than English per MSOA is depicted in Fig. 5.1b. Visual inspection reveals there are some MSOAs in which there is not any single tweet published between January and June 2018. Therefore, all MSOAs are not covered as we had hoped by using 6 months of *Twitter* data, only 851 MSOAs have language usage information. This suggests including more *Twitter* data would help fill the missing MSOAs.

In order to compare with *Twitter* data, we filter ONS data based on the top 20 most commonly spoken languages across London as determined by *Twitter* language usage. Since English is the most commonly spoken language in all Greater London's MSOAs, Fig. 5.2 shows a map of 19 most commonly used languages other than English in each MSOA using Census data. The map reveals that several languages are heavily clustered around neighboring areas, having the same language, such as Polish in South and West London.

Table 5.2: Top 20 commonly used languages based on *Twitter* language usage in tweets with tagged places between July 2017 and June 2018.

In total, there are 1.7 million tweets (1,678,526) with tagged places across one year of data. Top 20 languages contribute to 96% (1,606,516) of the dataset. We remove Haitian language from our analysis since it is not available in the ONS Census data.

Language	Number of tweets classified	Percentage of the *Twitter* dataset	Number of the 2011 Census respondents	Percentage of the ONS dataset
English	1,494,432	89.03	6,083,420	77.89
Spanish	17,195	1.02	71,192	0.91
French	11,776	0.70	84,191	1.07
Portuguese	9,395	0.55	71,525	0.91
Arabic	8,966	0.53	70,602	0.90
Indonesian	7,801	0.46	2,817	0.03
Italian	7,108	0.42	49,484	0.63
German	6,462	0.38	31,306	0.40
Estonian	6,421	0.38	1,192	0.01
Japanese	5,424	0.32	17,050	0.21
Turkish	5,306	0.31	71,242	0.91
Tagalog	4,629	0.27	25,869	0.33
Urdu	3,201	0.19	78,667	1.01
Welsh	3,086	0.18	1,310	0.01
Dutch	2,607	0.15	9,603	0.12
Polish	2,575	0.15	147,816	1.89
Haitian	2,539	0.15	-	-
Swedish	1,986	0.11	10,428	0.13
Danish	1,937	0.11	4,185	0.05
Finnish	1,913	0.11	2,800	0.03
Russian	1,757	0.10	26,603	0.34

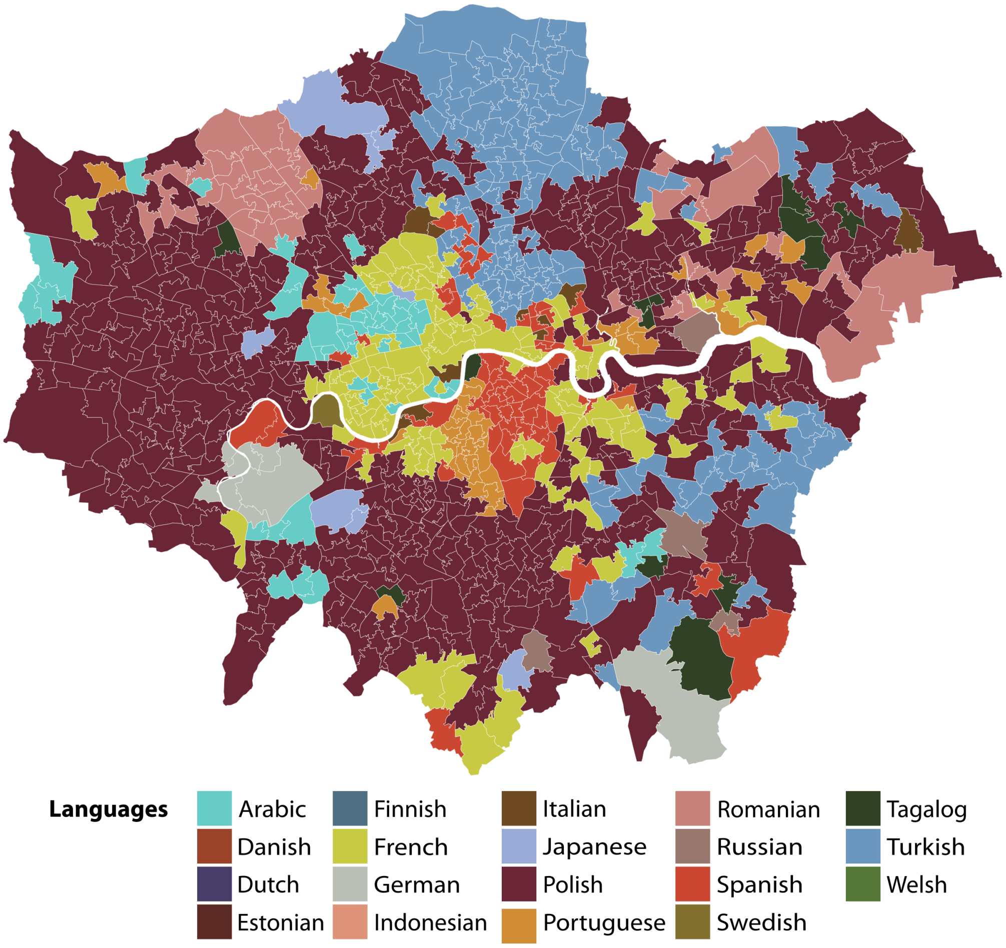

Figure 5.2: **Census measurements of language diversity in Greater London.**

The 2011 ONS Census data reports area statistics at various levels of spatial granularity. Here, we analyse data at the level of Middle Layer Super Output Areas (MSOA) which has the number of inhabitants between 5,000 and 15,000. Greater London comprises 983 MSOAs, the boundaries of which are depicted here. English is the most commonly spoken language in all Greater London's MSOAs. Here, our analysis investigates the top 19 most commonly spoken languages other than English across Greater London according to *Twitter* usage. The map shows 19 languages other than English that are most commonly spoken in each MSOA.

3 Results

We initially compare data between the ONS 2011 Census and tweets uploaded during January and June 2018 across all MSOAs through visual inspection. We highlight MSOAs where the most common language spoken other than English according to ONS data is the same as the most common language spoken other than English according to *Twitter* data (Fig. 5.3a). Out of 983 MSOAs, there are 50 MSOAs that are matched which reflects the relationship where some languages such as French, Spanish and Turkish have some correspondence. We note that the number of matched MSOAs and number of languages that are commonly used from *Twitter* data is smaller compared to the previous chapter analysing *Instagram* data.

To explore whether there is a relationship between the Census data and *Twitter* data, we analyse the tweets and the number of respondents in the ONS data for each language. We then investigate whether *Twitter* messages can be used to estimate the proportion of used languages across Greater London areas. Lastly, we compare the estimation performance between a model that contains Census data only and a model that incorporates *Twitter* data. To do this, we compare the mean absolute error (MAE) based on cross-validation techniques, and perform the Wilcoxon Signed Rank test to find evidence to support the differences in MAE reduction.

The number of tweets across all MSOAs for each language, as classified by the *Twitter* language identifier, is normalised by the total number of tweets that were published in Greater London (932,662). For each language, the number of language speakers according to the ONS is normalised by the total number of Census respondents (7,809,942). Figure 5.3b shows the relationship between percentages for each language between ONS and *Twitter* data. With only 19 languages other than English, our results suggest there is no relationship between the number of tweets and the number of Census respondents ($\tau = 0.02$, $N = 19$, $p > 0.05$, Kendall's rank correlation). However, visual inspection reveals that French, Spanish, and Turkish language might have some relationships as they tend to cluster around the neighbouring areas.

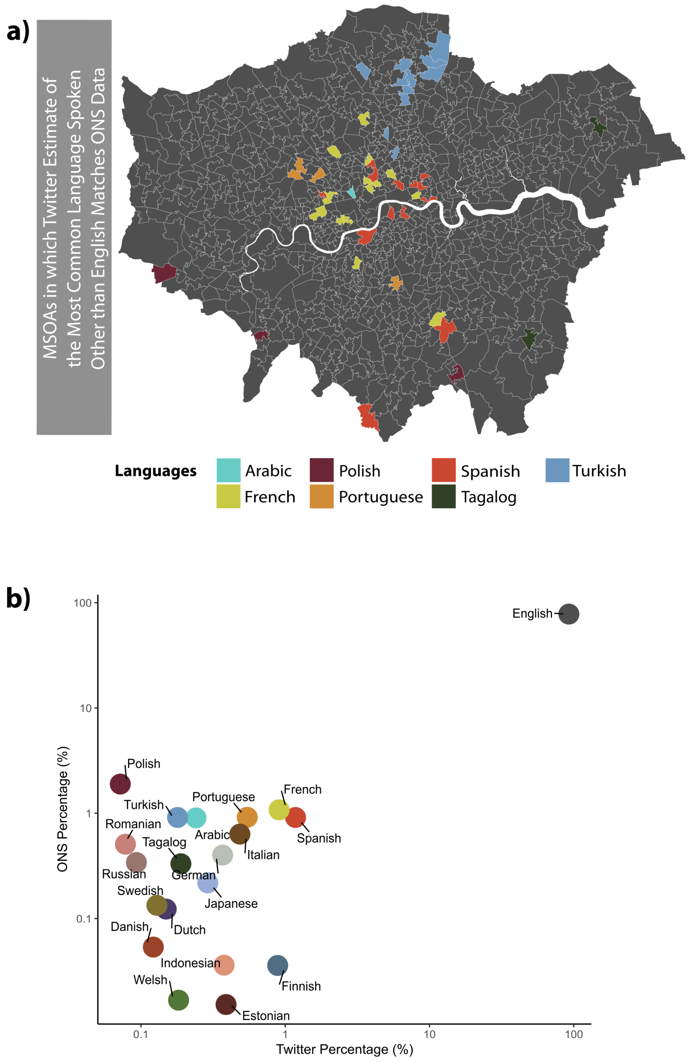

Figure 5.3: **Identifying relationship between *Twitter* and ONS Census data.**

(a) Based on the top 20 most commonly spoken languages across Greater London from *Twitter* language usage, we identify and highlight MSOAs where the most common language spoken other than English according to ONS data is the same as the most common language spoken other than English according to *Twitter* data. We find that there is there are 50 out of 983 MSOAs or 5% that are matched. The map reveals there are seven out of 20 languages that MSOAs are matched. (b) Across all MSOAs, the number of tweets across all MSOAs for each language, as classified by *Twitter* language identifier, is normalised by the total number of tweets that were published in Greater London (932,662). For each language, the number of language speakers according to the ONS is normalised by the total number of Census respondents (7,809,942). To compare percentages, we use log scales for both axis to produce the map. Excluding English language, we find no evidence for a correlation between the number of tweets and the number of Census respondents with only 19 languages ($\tau = 0.02$, $N = 19$, $p > 0.05$, Kendall's rank correlation).

Similar to Greater London analysis in the previous chapter, we investigate whether *Twitter* messages can help in estimating the proportion of spoken languages across MSOAs in which there is at least one tweet. For each language, we build a simple logistic regression model, which captures the proportions of spoken languages across MSOAs given by ONS data as response variable and we include the percentages of spoken languages obtained from tweets across MSOAs as predictor variable. We use Nagelkerke R^2 [205] to measure the amount of variance explained by the model. If the value is zero, no variance is explained while a value of one is equivalent to all variance explained by the model. Across the top 20 languages, we find that tweets help explain the variations and provide value in estimating spoken languages across areas in Greater London (Fig. 5.4). For example, English, Romanian, Arabic, and English have a higher Nagelkerke R^2 which implies *Twitter* data captures more variation in the models and have the potential to provide better estimates than other languages. To support our findings, 18 out of the top 20 languages show a significant relationship (Table 5.3).

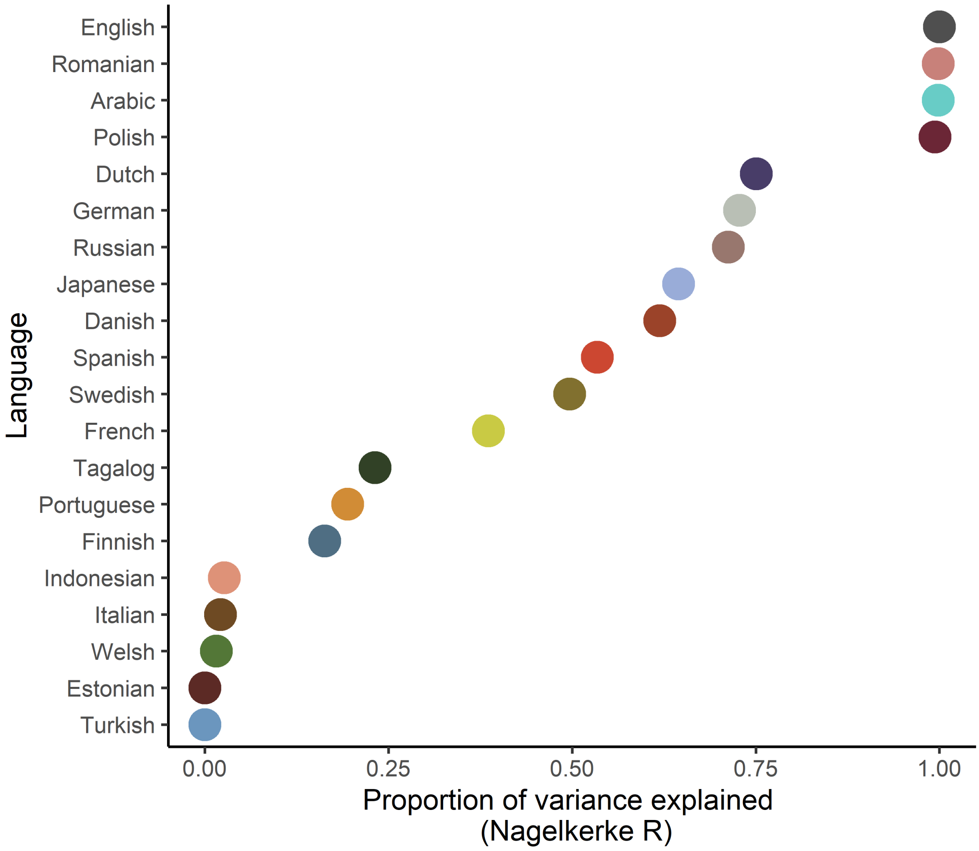

Figure 5.4: **Estimating the percentage of language spoken per MSOA using *Twitter* data.**

We investigate whether *Twitter* data can be used to estimate the percentage of language speakers in each MSOA. For each language, we build a simple logistic regression model which captures the proportions of spoken languages across MSOAs given by ONS data as response variable and we include the percentages of spoken languages obtained from tweets across MSOAs as predictor variable. We employ Nagelkerke R^2 to measure the amount of variance explained in each model. We carry out the same analysis for all 20 languages and investigate the performance of *Twitter* data in estimating the ONS statistics. We find that language models including *Twitter* data captures more variations and provides value in generating more accurate estimates of the percentages of language speakers in different parts of London. There are 18 out of top 20 languages that have a significant relationship between *Twitter* and ONS Census data (Table 5.3).

Table 5.3: Results generated by logistic regression models on top 20 commonly used languages across Greater London based on *Twitter* language usage for the period between January and June 2018.

Each language model uses the percentage of tweets in each MSOA as predictor while the proportion of ONS respondents is the dependent variable. The table is ordered by Nagelkerke R^2. The number of Nagelkerke R^2 represents the amount of variance explained by the model. A value of one means all variance explained while a value is zero means no variance explained. We find that there are 18 out of the top 20 languages that have a significant relationship between *Twitter* and ONS Census data. This suggests the values of *Instagram* data in improving the percentage estimates of spoken language.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Language	Beta/ Coefficient	Nagelkerke R^2	Number of tweets classified	Percentage of the *Twitter* dataset	Number of the 2011 Census respondents	Percentage of the ONS dataset
English	0.011***	0.999	838,465	89.90	6,083,420	77.89
Romanian	0.155***	0.998	708	0.07	39,653	0.50
Arabic	0.136***	0.998	2,182	0.23	70,602	0.90
Polish	0.193***	0.994	652	0.06	147,816	1.89
Dutch	-0.628***	0.750	1,356	0.14	9,603	0.12
German	-0.144***	0.727	3,338	0.35	31,306	0.40
Russian	-0.154***	0.712	843	0.09	26,603	0.34
Japanese	-0.336***	0.645	2,633	0.28	17,050	0.21
Danish	-0.614***	0.619	1,105	0.11	4,185	0.05
Spanish	-0.045***	0.534	10,689	1.14	71,192	0.91
Swedish	-0.164***	0.496	1,170	0.12	10,428	0.13
French	-0.025**	0.385	8,219	0.88	84,191	1.07
Tagalog	-0.039***	0.231	1,715	0.18	25,869	0.33
Portuguese	-0.019***	0.194	4,946	0.53	71,525	0.91
Finnish	-0.133***	0.163	8,037	0.86	2,800	0.03
Indonesian	-0.054**	0.026	3,417	0.36	2,817	0.03
Italian	-0.006*	0.021	4,397	0.47	49,484	0.63
Welsh	-0.078	0.015	1,653	0.17	1,310	0.01
Estonian	-0.001	0.001	3,526	0.37	1,192	0.01
Turkish	-0.001***	0.000	1,626	0.17	71,242	0.91

To investigate whether *Twitter* data can help complement ONS data, we compare a baseline model that has Census data only, containing the proportion of language speakers across MSOAs, and a *Twitter* model which includes the percentage of spoken language across MSOAs from *Twitter* data. We then perform a 20-fold cross-validation [206, 207] for each language, which is similar to the *Instagram* analysis, in which all MSOAs are randomly divided into 20 subsets of MSOAs, each comprising 5 percentage of all MSOAs or between 48 and 51 MSOAs. For each language, a baseline logistic regression model is fitted on the training data using Census data only, containing the percentage of language speakers across MSOAs. The baseline model estimates the mean of the percentage of language speakers across all MSOAs of the training data. The remaining MSOAs are represented by the estimated mean given by the baseline model. A *Twitter* model includes both Census and *Twitter* data. For both baseline and *Twitter* models and for each language, the cross-validation starts from training 19 subsets into the estimating model, leaving one subset out as test set for estimation and record the difference in values for each MSOA between the model estimates and the actual percentages from the testing set. This step is then repeated until every subset has been used once as test set. Then we record errors for both baseline and *Twitter* models. We do this process for each language out of the 20 languages. Since we are interested in the top 20 languages, we repeat this process for each language.

To determine whether *Twitter* data can generate more accurate estimates complementing the ONS data, we calculate the Mean Absolute Error (MAE), for each language, for both baseline and *Twitter* model, and determine which model has lower error rates. Our findings reveal that there are seven languages with MAE reductions which imply better estimation performance compared to the baseline models (Table 5.4). The Wilcoxon Signed Rank test supports the evidence of MAE reductions for these seven languages. Across the top 20 languages, the improvement measured by MAE reductions ranges from -32.6% to 0.92%. On the other hand, we find that *Twitter* data provides no improvement on 13 languages. This suggests that *Twitter* data does not provide as much information for such estimations as *Instagram*.

Table 5.4: 20-fold cross-validation across MSOAs results of out-of-sample tests on logistic regression models on top 20 languages for the period between January and June 2018.

We randomly split all 983 MSOAs in Greater London into 20 subsets. Each subset contains 5% of the total number of MSOAs. The training set consists of 19 subsets and one subset is used as the test set. For cross-validation, each subset is used once as the test set, therefore, we repeatedly perform this process 20 times. In each process, we fit 20 logistic models for individual languages on the training data using Census data only to estimate the mean percentage of language usage per MSOA. We build another 20 individual models including *Twitter* data. Mean Absolute Errors (MAE) are used as error measure and these are in units of percentage of language speakers. For each language and for each model type, we then calculate the average MAE based on 20 MAEs from 20 subsets. Afterwards, a Wilcoxon Signed Rank test is performed to determine whether the differences between the errors are significant. The table is ordered by *Twitter* language usage. For each language, model with better estimates are highlighted in bold. A star indicates significantly different errors ($p < 0.05$). The false discovery rate (FDR) correction is used on the p-values to control the proportion of false positives.

Language	Baseline Model	Twitter Model	% Difference	Wilcoxon Test V
English	9.589	**9.501**	0.92	299019*
Spanish	0.654	**0.649**	0.64	343963*
French	0.699	**0.698**	0.03	325409*
Finnish	**0.031**	0.031	-0.05	204309*
Portuguese	**0.575**	0.587	-2.23	317280*
Italian	**0.447**	0.545	-21.91	333635*
Estonian	**0.016**	0.016	-0.16	297123*
Indonesian	**0.035**	-0.14	-32.58	316196*
German	0.299	**0.298**	0.02	323970*
Japanese	0.216	**0.214**	0.54	345611*
Arabic	**0.718**	0.740	-2.98	368745*
Tagalog	0.251	**0.250**	0.01	179287*
Welsh	**0.015**	0.015	-0.06	303682*
Turkish	**0.992**	1.003	-1.17	278955*
Dutch	**0.085**	0.085	-0.08	169136*
Swedish	**0.124**	0.124	-0.05	341551*
Danish	**0.047**	0.047	-0.07	157961*
Russian	**0.204**	0.205	-0.13	322154*
Romanian	0.475	**0.472**	0.81	336104*
Polish	**1.329**	1.33	-0.02	299314*

We conduct a linear mixed-effects analysis [208] to determine whether the overall estimation performance across the top 20 languages is improved using *Twitter* data. We estimate absolute errors, as obtained from cross-validation results, by including the model type as fixed effect, controlling for by-language and by-MSOA variability as random effects. Our analysis suggests that models with *Twitter* data contribute to improvement in estimates within Greater London more than baseline model (β baseline $= 0.008$, β Twitter $= 0.001$, $N = 25890$, p baseline $= 0.085$, p Twitter < 0.001).

Moving to the other dataset, we analyse one year *Twitter* data about tagged places in Greater London starting from July 2017 to June 2018. Similar to the analysis above, we build logistic regression model for each language across all boroughs and use Nagelkerke R^2 to determine whether *Twitter* data can help improve estimates. Figure 5.5 shows the Nagelkerke R^2 for the top 20 languages of this dataset. With one year of data, we find that there are 14 languages that have a Nagelkerke R^2 of 1, suggesting that *Twitter* data on tagged places can also be used to improve estimates. Our findings reveal evidence that 17 out of the top 20 languages show a significant relationship between the *Twitter* and the ONS Census data although there is a negative relationship for two languages which are Tagalog and Danish (*all* $\beta s > 0.04$, *all* $Ns = 983$, *all* $ps < 0.001$, see Table 5.5).

For the boroughs analysis with *Twitter* data that are tagged with places, we perform a 33-fold cross-validation for each language to estimate language distributions for each borough and compare the performance. Our results suggest that there are six languages for which models with *Twitter* data generate more accurate estimates than the baseline models that include Census data only (Table 5.6). Amongst these six languages, we find evidence of MAE reductions for one language only, which is Turkish, while we find no evidence for the other 5 languages using Wilcoxon Signed Rank test. Across the top 20 languages, the range of MAE reductions are between -329.94% and 11.02%. The linear mixed-effects analysis is performed to investigate the overall estimation performance for the top 20 languages across boroughs in Greater London. We find no evidence there is an improvement in estimates in Greater London boroughs using *Twitter* data (β baseline $= 0.007$, β Twitter $= 0.002$, $N = 1320$, p baseline $= 0.07$, p Twitter $= 0.279$).

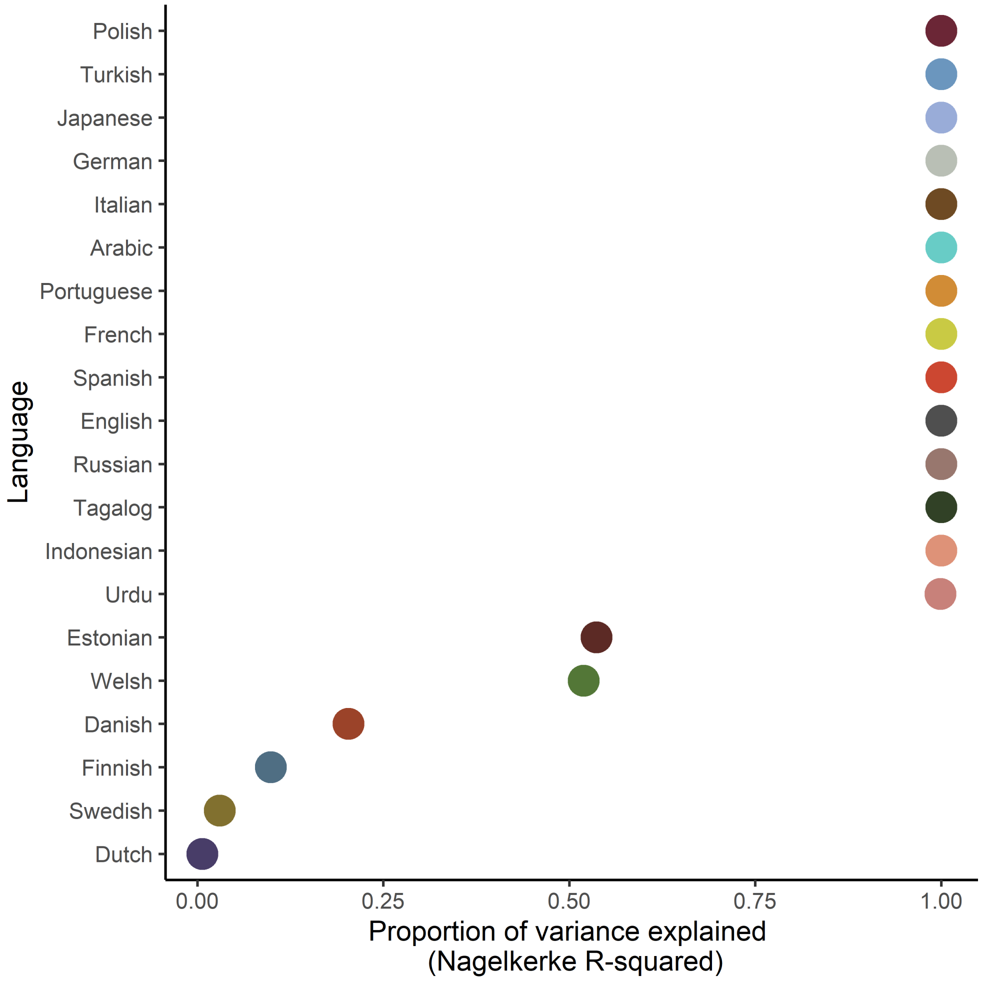

Figure 5.5: **Estimating the percentage of language spoken per borough using *Twitter* data on tagged places.**

We investigate whether *Twitter* data about tagged places can be used to estimate the percentage of language speakers in each borough. A logistic regression model is built on *Twitter* data on tagged places for each language to estimate the percentage of language usage across boroughs. A Nagelkerke R^2 is used to measure the amount of variance explained in the model. There are 17 out of the top 20 languages that show a significant relationship (Table 5.5).

Table 5.5: Results generated by logistic regression models on top 20 spoken languages across Greater London boroughs based on language usage on *Twitter* data on tagged places for the period between July 2017 and June 2018.

The logistic model for each language train on *Twitter* data in each borough to estimate the proportion of ONS respondents. The table is ordered by a Nagelkerke R^2. We find that 17 out of the top 20 languages have a significant relationship between the *Instagram* and the ONS data although two of which (Tagalog and Danish) have a negative relationship. The number of Nagelkerke R^2 refer to the amount of variance explained by the model which suggests the value of *Twitter* data in improving the estimates of language usage.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Language	Beta/ Coefficient	Nagelkerke R^2	Number of tweets classified	Percentage of the *Twitter* dataset	Number of the 2011 Census respondents	Percentage of the ONS dataset
English	0.098***	1	1,494,432	89.03	6,083,420	77.89
Spanish	0.656***	1	17,195	1.02	71,192	0.91
French	0.489***	1	11,776	0.70	84,191	1.07
Portuguese	1.144***	1	9,395	0.55	71,525	0.91
Arabic	1.016***	1	8,966	0.53	70,602	0.90
Italian	1.022***	1	7,108	0.42	49,484	0.63
German	0.514***	1	6,462	0.38	31,306	0.40
Japanese	2.921***	1	5,424	0.32	17,050	0.21
Turkish	1.820***	1	5,306	0.31	71,242	0.91
Polish	1.402***	1	2,575	0.15	147,816	1.89
Russian	2.718	1	1,757	0.10	26,603	0.34
Tagalog	-0.577***	0.999	4,629	0.27	25,869	0.33
Indonesian	2.814***	0.999	7,801	0.46	2,817	0.03
Urdu	0.344***	0.998	3,201	0.19	78,667	1.01
Estonian	0.383***	0.536	6,421	0.38	1,192	0.01
Welsh	0.514***	0.519	3,086	0.18	1,310	0.01
Danish	-0.307*	0.203	1,937	0.11	4,185	0.05
Finnish	0.413	0.098	1,913	0.11	2,800	0.03
Swedish	0.053	0.029	1,986	0.11	10,428	0.13
Dutch	-0.054	0.006	2,607	0.15	9,603	0.12

Table 5.6: 33-fold cross-validation across boroughs results of out-of-sample tests on logistic regression models on top 20 languages for the period between July 2017 and June 2018.

For the boroughs analysis, we find evidence for only one language (Turkish) out of six languages that models with *Twitter* data reduce MAE compared to baseline models. However, the MAEs across top 20 languages are reduced by starting from -329.94% to 11.02%. The table is ordered by language usage on *Twitter* data. The Wilcoxon Signed Rank test is used to determine whether the errors significantly differ between *Twitter* and baseline models ($p < 0.05$). To control the proportion of false positives, we use false discovery rate (FDR) correction to adjust the p-values.

Language	Baseline Model	Twitter Model	% Difference	Wilcoxon Test V
English	7.132	**7.06**	0.89	329
Spanish	**0.641**	0.677	-5.84	435*
French	**0.703**	0.724	-3.09	378
Portuguese	0.461	**0.411**	11.02	408
Arabic	**0.691**	0.880	-27.33	464*
Indonesian	0.024	**0.022**	6.79	360
Italian	**0.427**	0.669	-56.71	399
German	**0.296**	0.306	-3.41	345
Estonian	**0.006**	0.006	-1.89	259
Japanese	0.177	**0.166**	6.39	349
Turkish	0.972	**0.882**	9.24	442*
Tagalog	**0.180**	0.182	-1.07	248
Urdu	**0.903**	3.882	-329.94	321
Welsh	0.0084	**0.0083**	0.49	275
Dutch	**0.073**	0.074	-1.47	56*
Polish	**1.071**	1.272	-18.8	324
Swedish	**0.122**	0.131	-7.47	112*
Danish	**0.041**	0.118	-186.131	124*
Finnish	**0.024**	0.024	-1.377	221
Russian	**0.144**	0.158	-9.68	351

4 Discussion

Measuring language statistics with the Census is costly and time consuming and is undertaken only once a decade in England and Wales. Complementing the *Instagram* chapter, we investigate whether *Twitter* data can also be used for estimating spatial distribution of languages in cities. We analyse 0.9 million tweets uploaded in Greater London between January and June 2018. We focus on the top 20 most commonly used languages across areas in Greater London to determine a relationship between ONS Census data and *Twitter* data.

We find that *Twitter* data provides some value in estimating spoken languages in different parts of London as measured by Nagelkerke R^2, hinting at potential of tweets data. Our results suggest that, overall, incorporating *Twitter* data improve estimates more than baseline models. The cross-validation results of individual languages provides there are seven out of top 20 languages that can produce better estimates than the baseline model. Additionally, we conduct the analysis to improve the estimates at a borough level using place-tagged tweets for the period between July 2017 and June 2018. We found that there is a smaller number of languages that can provide better estimates than the baseline model although we find the evidence for only one language, which is Turkish.

We note that *Twitter* results provides smaller number of languages that improves the estimates of language speakers across MSOAs in Greater London compared to the *Instagram* analysis, which has 14 languages. The range of improvement in estimates in the *Twitter* analysis is also smaller. This suggests that *Twitter* data has a weaker effect on improving the estimates compared to the *Instagram* analysis. This might be due to the difference in terms of data quantity between *Instagram* data and *Twitter* data. For the same period of six months, our *Instagram* dataset have 7.3 million photos while there are about 0.9 million tweets in the *Twitter* dataset. It is possible that photographs are more attached to location than short messages. Qualitatively, the results of *Twitter* analysis are consistent with the *Instagram* chapter. As such, our findings underline the potential of text-based social media data, which is *Twitter* in this case, to estimate spatial patterns of language usage.

We note the time gap between the 2011 Census data and *Twitter* data which covers the first half of 2018 where there is no official data to compare to. Future work could also investigate the potential for improving estimates across time, once new Census data becomes available after 2021. Although our findings on *Twitter* data reveal a weaker effect compared to the *Instagram* analysis on the previous chapter, aggregating both *Instagram* and *Twitter* data or other online data sources could lead to improved estimates. Nevertheless, both *Twitter* and *Instagram* analysis highlight the usefulness of online data in estimating official statistics.

Chapter 6

Nowcasting unemployment figures using Google search data

1 Introduction

This thesis has investigated statistics across space using online data in the previous chapters. Statistics across time is also crucial for policymakers and stakeholders to understand the past and the present to make a decision on the future. In an economic context, they need to monitor the key statistics such as unemployment rates which represent current activity. This requires the most up to date figures however. The latest available official economic figures are published with a considerable time lag, such as one or two months delay, due to the time and effort which are required to gather the relevant data [4, 5]. This has become a challenge for nowcasting, which is forecasting the present rather than the future. It aims to predict current estimates where traditional methods are time-delayed.

Due to the advancement of the Internet, people are generating data via social online interactions, resulting in huge datasets. Thus, the Internet has turned into a new data source that captures human behaviour. Various studies [185–187, 209–212] have analysed human behaviour with the real world with data from online services such as *Google* [4, 5, 13, 111–113, 154, 188, 213–217], *Yahoo* [191], *Flickr* [19, 21, 97], *Wikipedia* [144, 149, 218], and Twitter [45, 189, 219, 220]. With the immediate availability of Internet search data, there are numerous studies now investigating whether online search data might be a good economic indicator [4, 5, 15–17, 130]. For instance, Choi and Varian [4] nowcasted the number of people who claimed for unemployment benefits in the US based on *Google Trends* data. *Google* is a search engine that dominates the United Kingdom's online search market. For example, it handled 90 percent of all queries made in 2018 [221]. It provides search volume data, which is publicly accessible via the *Google Trends* website.

Here, we investigate whether online search queries can be used to forecast current unemployment rates in a time period that goes beyond previous studies [4, 5, 15, 17, 130]. Specifically, we evaluate the usefulness of *Google Trends* data in nowcasting a recent dataset until February 2017 and investigate whether *Google Trends* queries can improve the current estimates of unemployment rates. Furthermore, while previous research has focused on small groups of search terms with various keyword selection approaches [4–6, 15, 17], we investigate whether including a broader relevant set of search terms provides additional value in nowcasting unemployment rates by using a variable selection technique, an elastic net.

2 Materials and methods

We retrieved online search data from *Google Trends* and obtained official unemployment data in the United Kingdom from the Office for National Statistics (ONS).

2.1 ONS data

In the United Kingdom, unemployment rates are calculated using the Labour Force Survey (LFS), which is conducted and reported by the Office for National Statistics (ONS). The seasonally unadjusted monthly results were obtained from the ONS website [222] on 5 May 2017. The rates are published with one and a half months delay. For example, the April 2017 release consists of the unemployment rates up to February 2017, leaving a gap from March to April. Since *Google* search volume data is available from January 2004, we used unemployment rates starting from January 2004. The top panel in Fig. 6.1 depicts the UK's unemployment rates.

2.2 Google Trends data

We retrieved Internet search volume data from *Google Trends*, previously known as *Google Insights* (https://www.google.co.uk/trends/) on 7 June 2017. The format of the data is a monthly time series and it is restricted to searches made in the United Kingdom. The *Google Trends* data we analyse spans from January 2004 to February 2017 and is not seasonally adjusted. We note that *Google Trends* reports search volume for a given query as an index relative to the highest search volume during the specified time period, rather than an absolute number of queries.

Considering search terms that are relevant to unemployment, three keywords

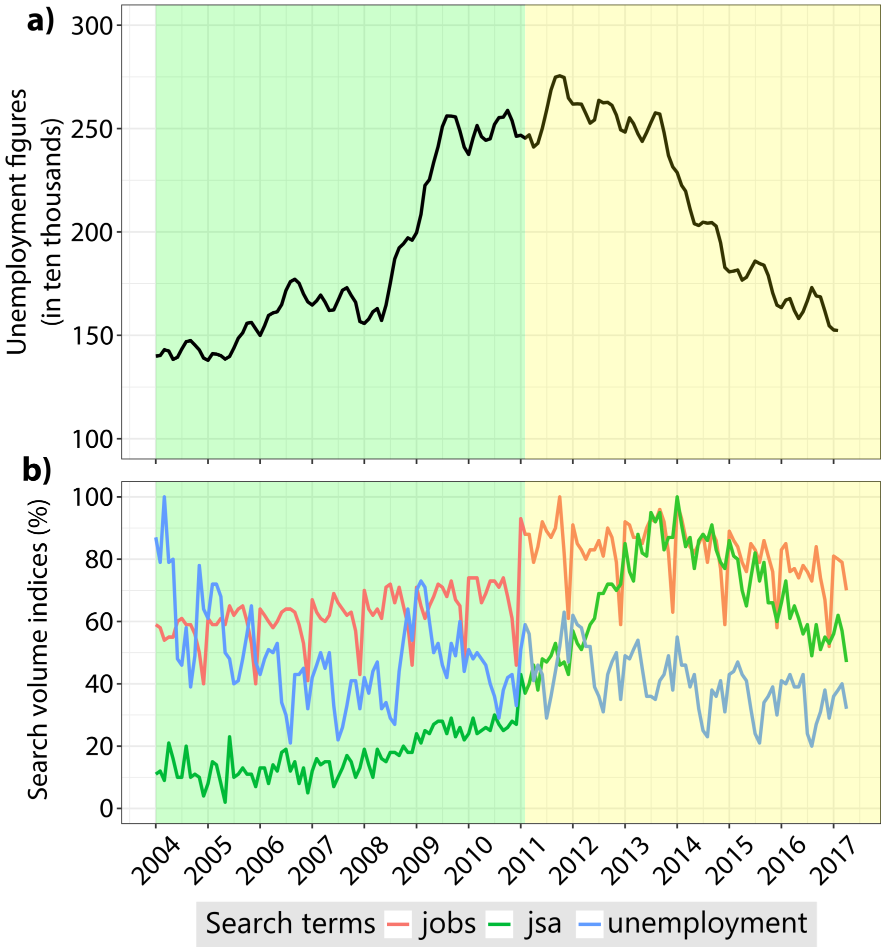

Figure 6.1: **UK unemployment figures and search volume data.**

(a) Time series of UK unemployment rates published by the ONS ranging from January 2004 to February 2017. (b) Time series of search volume data restricted to searches made in the United Kingdom between January 2004 and April 2017. McLaren and Shanbhogue's analysis [2] used the period from January 2004 until January 2011 as highlighted in green, while our new period of analysis is up to February 2017, which includes the yellow highlighted area.

are selected as suggested by previous studies [4, 5]. These are "jobs", "JSA", and "unemployment". The bottom panel in Fig. 6.1 shows *Google* search volumes for three different search terms.

We also explore more relevant search terms as identified by *Google Trends* and investigate whether alternative search terms can improve nowcasting current unemployment estimates. Therefore, we identify alternative search terms that are highly correlated with general search terms such as "jobs" and "unemployment" in the United Kingdom. In addition to search frequencies provided by *Google Trends* for a given query, it provides the top 25 search terms that are most highly correlated with the given query. Here, we obtained the top 25 search terms most highly correlated with "jobs" and the top 25 search terms whose search volume is highly correlated with the search volume for "unemployment" as listed in Table 6.1.

Table 6.1: Most highly correlated keywords for "jobs" and "unemployment" categories identified by *Google Trends*.

"jobs" category	"unemployment" category
jobs	unemployment
job	ww
indeed	welfare
jobcentre	office
careers	offices
job centre	benefits
cv	rsa
jobmatch	pole emploi
universal jobmatch	welfare rights
jobcentre plus	jsa
interview	uk unemployment
london jobs	disability
jobs indeed	workhouse
part time jobs	ww.
job vacancies	sign on
facebook	jsa claim
at	nav
job centre plus	unemployment benefit
hotmail	apl
gumtree	anpe
total jobs	dole
google	probation office
career	bony
interview questions	paro
reed	disability benefits

3 Results

3.1 Analysing the relationship

In order to investigate the relationship between *Google Trends* data and official unemployment rates beyond existing research [4, 5, 15, 17], we replicate and extend McLaren and Shanbhogue's [5] analysis, which covers the time period until January 2011. We expand this period of analysis up to February 2017. Based on these two different periods, we then build nowcasting models using a linear regression to quantify the extent to which *Google Trends* queries can improve current estimates of unemployment rates with a recent dataset.

To analyse the relationship between *Google Trends* data and the unemployment rates from the ONS, a baseline nowcasting model is first built using changes in unemployment rates in the previous three months only ($\Delta \mathrm{Y}_{\text{t-1}}$,$\Delta \mathrm{Y}_{\text{t-2}}$) to forecast change from the previous month ($\Delta \mathrm{Y}_{\mathrm{t}}$). This is a lag-2 autoregressive model, AR(2). Then, the changes in *Google Trends* queries are included as an external regressors ($\Delta \mathrm{X}_{\mathrm{t}}$) in order to create an advanced nowcasting model as follows:

$$\Delta \mathrm{Y}_{\mathrm{t}} = c + \beta_1 \Delta \mathrm{Y}_{\mathrm{t}-1} + \beta_2 \Delta \mathrm{Y}_{\mathrm{t}-2} + \Phi \Delta \mathrm{X}_{\mathrm{t}}, \tag{6.1}$$

where c is an intercept and β_1, β_2, and Φ are regression coefficients estimated from historical data.

The in-sample baseline model and advanced models with *Google* keywords are compared to evaluate the relationship. The in-sample test is conducted across five models: the baseline model, three advanced models that include three search terms ("jobs", "JSA", and "unemployment") separately as external regressors, and an advanced model, in which we incorporate all three search terms as external regressors. Then, we compare the results between these five models using performance and goodness-of-fit metrics such as the Mean absolute error (MAE) and the Akaike information criterion (AIC). We analyse the period from January 2004 until February 2017. This extends the time period analysed in McLaren and Shanbhogue [5].

The in-sample nowcasting results show that the model that includes three *Google* search terms slightly improves the overall goodness-of-fit (*AIC* baseline = 1572.491, *Adjusted* R^2 baseline = 0.3303, *AIC* advanced "three keywords" = 1557.65, *Adjusted* R^2 advanced "three keywords" = 0.4037). Details of the results are listed in Table 6.2. In Fig. 6.2b, we depict the in-sample nowcast errors between the baseline and model with three *Google* keywords. The error distribution of in-sample nowcast errors is illustrated in Fig. 6.3a, revealing the higher density of low error rates for the model with three *Google* keywords. We find that the model using three keywords from *Google Trends* generates more accurate estimates compared to the baseline model and the other three models in terms of MAE, RMSE and MAPE (MAE baseline = 31.64, MAE advanced "three keywords" = 29.36, $RMSE$ baseline = 40.20, $RMSE$ advanced "three keywords" = 37.55, $MAPE$ baseline = 154.61, $MAPE$ advanced "three keywords" = 146.60). In summary, the model with three keywords has smaller MAE, RMSE, and MAPE than the baseline model with 7.2%, 6.6% and 2.5% improvement respectively. This suggests that the best model to improve unemployment estimates is the model that include multiple keywords. Considering only models with single search term, the model with the search term "jobs" outperforms models with other individual search terms in terms of MAE and RMSE (MAE advanced "jobs" = 29.44, RMSE advanced "jobs" = 150.50). A model with "jsa" produces the smallest MAPE out of the three models with single search term (MAPE advanced "jsa" = 148.35).

Table 6.2: Comparison of in-sample unemployment nowcasting results between January 2004 and February 2017.

To evaluate the relationship between *Google* search data and UK unemployment rates, we used in-sample tests to compare baseline model and advanced models with *Google* keywords. Multiple performance measures are used for all five models. To prevent overfitting problem, we use adjusted R^2. AIC indicates the model fit relative to the other models given the data and balances the trade-off between model complexity and its goodness-of-fit. Our results suggest that including multiple keywords into the model improve estimates more than a single keyword. Stars indicate significance levels as follows: * p < 0.05; ** p < 0.01; *** p < 0.001.

Term	3 keywords	"jobs"	"unemployment"	"jsa"	Baseline
Intercept	0.2604	0.3896	0.7721	0.3207	0.6552
ΔY_{t-1}	0.7165***	0.7181***	0.6842***	0.6818***	0.6778***
ΔY_{t-2}	-0.2111**	-0.2245**	-0.2676***	-0.2440**	-0.2546**
$jobs_t$	1.4342***	1.3224***			
$unemployment_t$	-0.3249		0.405569		
jsa_t	0.0417			1.0772*	
AIC	1557.648	**1554.428**	1573.153	1570.321	1572.491
Adjusted R^2	0.4037	**0.4087**	0.3317	0.3439	0.3303
MAE	**29.3590**	29.4367	31.4660	30.4513	31.6409
RMSE	**37.5495**	37.6454	40.0210	39.6523	40.1963
MAPE	**146.5999**	150.4981	157.7204	148.3501	154.6145

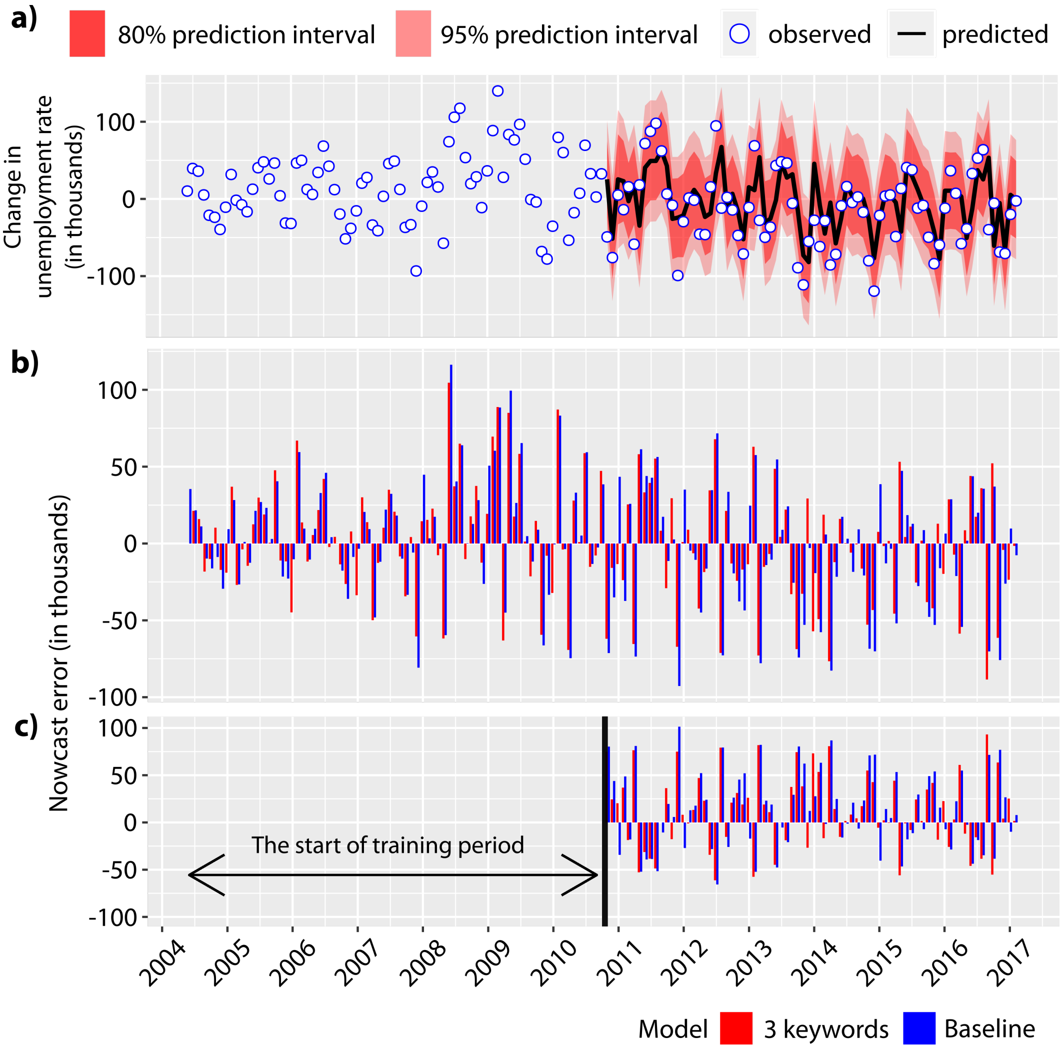

Figure 6.2: **Unemployment nowcasting results between January 2004 and February 2017 for the model with three *Google* keywords.**

(a) Out-of-sample nowcasts for the model with three keywords ("jobs", "jsa", "unemployment"). (b) In-sample nowcast errors for the baseline model, using only changes in unemployment rates from previous three months, and the model which includes change in *Google* search query data. (c) Comparison of out-of-sample nowcast errors between the baseline model and the model that includes three *Google* search query data. Our results in Table 6.3 show that the model with three search terms provides MAE and MAPE reduction greater than the other models for the longer time period between 2010 and 2017. On the other hand, for the shorter time period from 2008 to 2011, the model with a single keyword ("jobs") reduces MAE and RMSE the most compared to the other models with single search term.

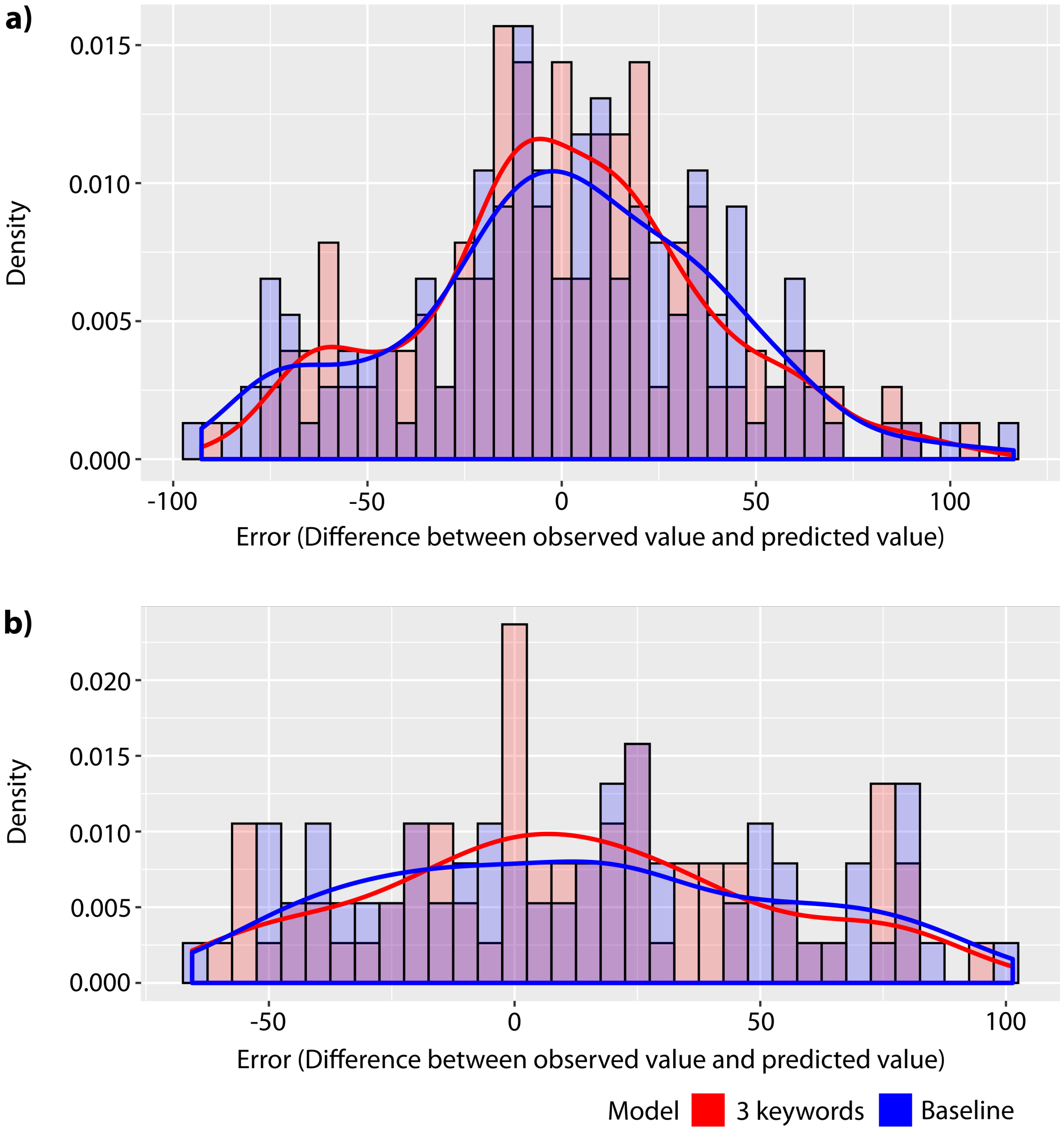

Figure 6.3: **Error distribution of unemployment nowcasts between baseline model and model with three *Google* keywords.**

(a) Error distribution of in-sample nowcasts between January 2004 and February 2017. (b) Out-of-sample error distribution for the time period between 2010 and 2017. Both distribution figures reveal that the model with three *Google* keywords produces a higher proportion of errors that are near zero or have low error rates compared to baseline model.

To evaluate the nowcasting performance of the advanced model which includes *Google Trends* data, an out-of-sample one month ahead forecast model is created as a baseline. This means that some of the sample data up to a specified month is used to estimate the model. Then, a prediction is calculated for the next month and compared to the actual observation to determine the accuracy of that model. We use training window as the first half of the dataset. This means that the model is estimating data from January 2004 to October 2010. Then, we nowcast the change in unemployment rates for November 2010. We then record the difference between the nowcast and actual change in the unemployment rate. Next, the model is re-estimated up to November 2010 and we nowcast the unemployment rate for December 2010. This is an iterative process until the end of the sample in February 2017. The coefficients of the model are recalibrated each time the training window increases. In Fig. 6.2, we visualise the estimates produced from the out-of-sample nowcasting model using *Google* three keywords. Furthermore, Fig. 6.3b depicts the error distribution obtained from out-of-sample test. In the first three columns of Table 6.3, we report the average MAE, RMSE, and MAPE for baseline and all four advanced models for both time periods considered.

Table 6.3: Comparison of out-of-sample unemployment nowcasting results on different periods.

The best performing model in each period is highlighted in bold. All estimations are based on changes in ONS's unemployment rates in comparison to the previous month.

	November 2010to February 2017			July 2008 to January 2011		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE
3 keywords	**31.9631**	40.2084	**151.4515**	39.6172	48.7793	243.5507
"jobs"	32.1627	**39.9473**	155.8729	**39.3390**	**48.0360**	242.1599
"unemployment"	35.6870	43.4078	182.0303	40.4665	49.9582	220.0004
"jsa"	33.7739	43.0845	157.69	39.5934	48.7523	218.4144
Baseline	35.9942	43.5630	186.6521	40.8144	49.5529	**215.4055**

For the longer time period of analysis, we find that the out-of-sample advanced model with three keywords reduces the MAE by 11.2% compared to the baseline model and the MAPE by 18.9% (MAE baseline = 35.99, MAE advanced "three keywords" = 31.96, $MAPE$ baseline = 186.65, $MAPE$ advanced "three keywords" = 151.45). Considering models with single search term only, the advanced model employing "jobs" search term provides the lowest RMSE amongst the other models with single search term, which is 10.7% lower than the baseline, and lowest MAE ($RMSE$ baseline = 43.56, $RMSE$ advanced "jobs" = 39.95, MAE advanced "jobs" = 32.16). On the other hand, we replicate the time period used in McLaren and Shanbhogue's [5] analysis for comparison, in which the training window of the first half of the dataset starts from January 2004 to June 2008. The performance of nowcasting models using data up until January 2011 reveals that models with *Google* search terms underperformed compared to the baseline model in terms of MAPE ($MAPE$ baseline = 215.41). The model with the "jobs" search term provides the smallest MAE and RMSE (MAE baseline = 40.81, MAE advanced "job" = 39.34, $RMSE$ baseline = 49.55, $RMSE$ advanced "job" = 48.04,). Furthermore, the model with three search terms produce slightly more accurate estimates than the baseline, with an improvement of 1.5%-3% (MAE advanced "three keywords" = 39.62, $RMSE$ advanced "three keywords" = 48.78). Our findings suggest that including more data and multiple keywords results in better nowcasting performance.

To investigate whether there is evidence of different levels of accuracy between nowcasting models, the Diebold-Mariano (DM) test is conducted across all five models by testing each pair of models as listed in Table 6.4. To control the proportion of false positives, we perform false discovery rate (FDR) correction. Although the out-of-sample analysis reveals that the advanced model with three keywords performs best in MAE reduction, we find no evidence in difference in accuracy between advanced model incorporating three keywords and the baseline model (DM "three keywords" = 2.07, N = 76, p = 0.07). Nevertheless, various error metric results from Table 6.3 suggest that including *Google* keywords help improve nowcast estimates of unemployment rates in the UK. Similarly, the shorter period of analysis results reveal that there is no evidence that model including *Google* search terms generates more accurate estimates than the baseline model (DM "three keywords" = 0.54, DM "jobs" = 1.00, N=31, p $>$ 0.05).

3.2 Nowcasting improvement with relevant search terms

To investigate whether including more online search queries provide extra value in nowcasting unemployment rates, we identify alternative search terms that are highly correlated with general search categories such as "jobs" and "unemployment" in the UK. Here, we include the top 25 search terms that are most highly correlated with "jobs" category and another top 25 highly correlated search terms with the search volume for "unemployment" in our nowcasting model. Since including all 25 or 50 search terms in a linear model would cause overfitting problem and complexity, we use a variable selection technique, which is elastic net, to select important keywords.

Table 6.4: Adjusted p-value and Diebold-Mariano (DM) test statistics of linear models.

The false discovery rate (FDR) correction is used to control the proportion of false positives and return adjusted p-values. Boldface highlights model that are statistically significant in terms of difference in level of accuracy. Positive values of the DM test means that the model on the row has smaller errors compared to the model on the column

a: Long time period: November 2010 to February 2017

	"jobs"	"unemployment"	"jsa"	Baseline
3 keywords				
Adjusted p-value	0.4223	0.0714	0.0714	0.0714
DM-statistic	-1.053	2.1655	2.0599	2.0737
"jobs"				
Adjusted p-value		0.0714	0.0714	0.0714
DM-statistic		2.4027	2.153009	2.2490
"unemployment"				
Adjusted p-value			0.7188	0.7188
DM-statistic			-0.3613	0.4139
"jsa"				
Adjusted p-value				0.7188
DM-statistic				0.5469

(To be continued)

Table 6.4 continued

b: Short time period: July 2008 to January 2011

	"jobs"	"unemployment"	"jsa"	Baseline
3 keywords				
Adjusted p-value	0.6479	0.6938	0.9858	0.7170
DM-statistic	-1.0766	0.8242	-0.0179	0.5423
"jobs"				
Adjusted p-value		0.6479	0.7170	0.6479
DM-statistic		1.0546	0.4649	1.0027
"unemployment"				
Adjusted p-value			0.6479	0.7170
DM-statistic			-1.2532	-0.5084
"jsa"				
Adjusted p-value				0.6479
DM-statistic				1.8565

For comparison of nowcasting performance, we build elastic net linear regression models in our out-of-sample one month ahead forecast test for each 25 keywords from two categories and we build another elastic net model which aggregates 50 keywords from both categories. We then compare the performance between five models: baseline AR(2), AR(2) with "jobs", "jsa", and "unemployment", AR(2) with 25 related keywords identified from "jobs" category, AR(2) with 25 related keywords identified from "unemployment" category, and AR(2) with 50 related keywords from both "jobs" and "unemployment" category. We use different training windows to explore the performance effect between short term and long term models. The training window used in this analysis ranges from 1 year, 5 year, 6 years and 5 months, and 10 years. 6 years and 5 months is a window of 50% of the whole period of the dataset. We consider 10 years as long term while less than 10 year periods (i.e. 1, 5, and 6 years) are considered as short term. For a short term period, the same starting training window from the out-of-sample analysis which starts from January 2004 to October 2010 is used. As for a longer period, the training window starts from January 2004 to May 2014. For each type of period, we consider two main types of training window: Growing and Sliding. Growing windows cover more data as time has passed, incorporating both old and current changes, while sliding window has a fixed number of data, covering only current changes.

For the elastic net model, we generate a range of tuning parameters using a random grid of 100 lambda and range of 100 alpha values. Then, for each alpha and lambda, we compute error metrics MAE, RMSE and MAPE. Next, we compare and select the tuning parameters that provide the best fitted model that has the smallest error according to MAE. We depict the nowcast errors of different models in Fig. 6.4 in which the details of the best performance models for each training window are listed in Table 6.5 and 6.6.

Starting from growing window size for short periods (1, 5, and 6 year and 5 months), we find that, on average the model with three search terms outperform other models in terms of MAE, RMSE and MAPE (*MAE* "three keywords" for 6 year and 5 months = 31.97, *RMSE* "three keywords" for 6 year and 5 months = 40.18, *MAPE* for 6 year and 5 months = 151.01). As for sliding window, we also find similar result with 1 year, 5 years, and 6 years and 5 months. While one year and five years of sliding training window reveal that the model with three keywords have smaller errors in terms of MAE and RMSE (*MAE* "three keywords" for five years = 32.35, *RMSE* "three keywords" for five years = 39.84), the model with 25 keywords from "unemployment" category has the smallest MAPE (*MAPE* "unemployment" five years = 156.221). This result is also similar for 1 year of sliding training window. Models with more than 25 keywords suffer from overfitting as they have more variables than the number of training data, they therefore have higher error rates than the baseline model. Using a sliding window of five years or more reveals that all models, except "unemployment" model, with relevant *Google* search terms identified by *Google Trends* have lower error rates than the baseline model in terms of MAE and RMSE.

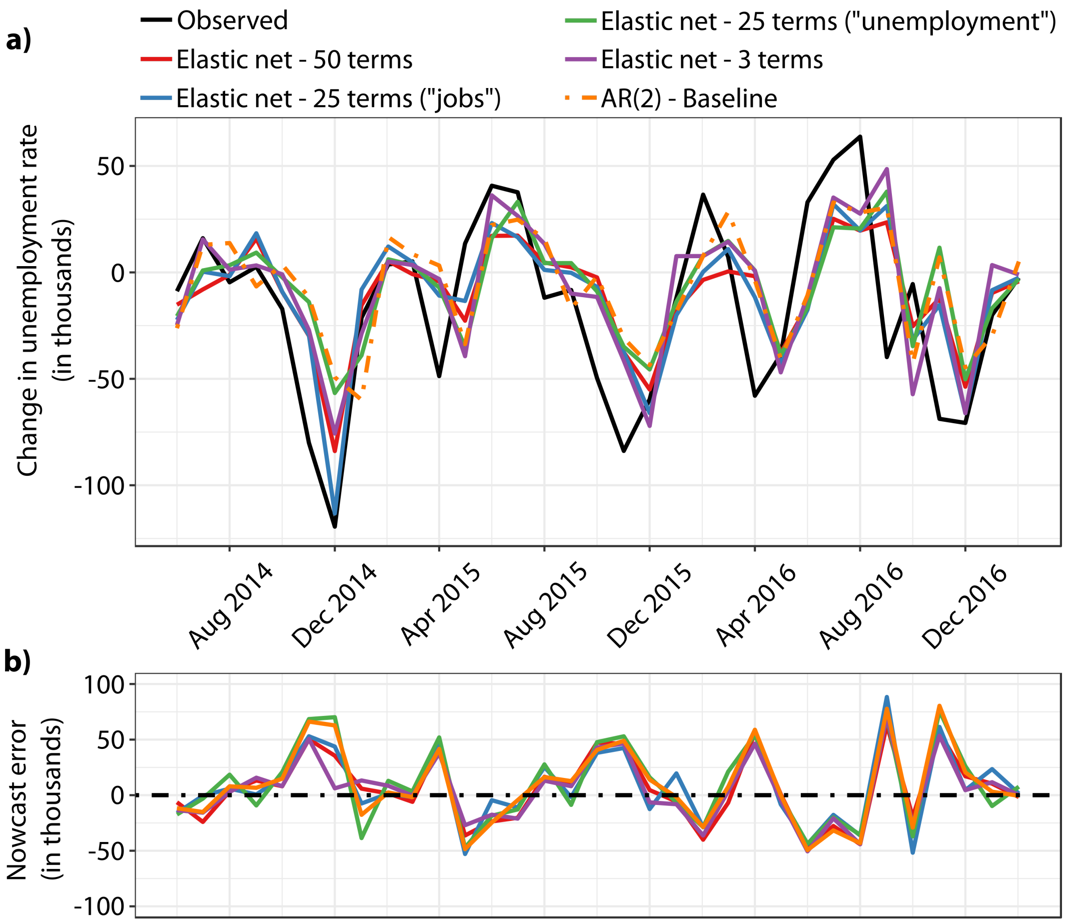

Figure 6.4: **Nowcast estimates of change in UK unemployment rates on different models using elastic net with 10 years of sliding window between June 2014 to February 2017.**

(a) The estimated change in UK rates of unemployment from elastic net incorporating relative volume changes in previous three months of ONS data and changes in *Google* search query data in previous month with relevant 50 keywords from both "jobs" and "unemployment" category (red), in comparison with the estimates from the ONS (black), elastic net models with relevant 25 search terms to "jobs" (blue) and "unemployment" (green), and the linear baseline model which includes relative volume changes in previous three months of ONS data only. (b) Comparison of one month ahead nowcast errors for the baseline model, using only changes in unemployment rates from previous three months, and elastic net models which include change in *Google* search query data from the current month. With a training window of 10 years, we find that the optimal model with 25 relevant keywords from "jobs" outperform the other models (MAE "jobs" $=$ 25.266, $RMSE$ "jobs" $=$ 34.225, $MAPE$ "jobs" $=$ 109.667)

Table 6.5: Comparison of nowcasting performance for change in unemployment rates on different models using elastic net between November 2010 and February 2017.

The best performed model in each window size is highlighted in bold. All comparison are based on change in ONS's unemployment rate from previous month.

	Growing window		Sliding window	
	1 year	5 years	1 year	5 years
MAE				
AR(2) + 50 terms	36.218	35.208	39.569	34.55
AR(2) + 25 terms ("jobs")	37.251	35.167	38.936	35.107
AR(2) + 25 terms ("unemployment")	35.281	34.987	39.1	34.093
AR(2) + 3 keywords	**34.201**	**31.948**	**37.667**	**32.35**
OLS AR(2)	36.273	36.008	38.61	35.711
RMSE				
AR(2) + 50 terms	44.781	42.488	47.851	42.781
AR(2) + 25 terms ("jobs")	46.892	43.27	47.053	42.276
AR(2) + 25 terms ("unemployment")	45.618	43.355	47.534	42.113
AR(2) + 3 keywords	**41.958**	**40.14**	**44.101**	**39.843**
OLS AR(2)	43.207	43.498	46.361	42.83
MAPE				
AR(2) + 50 terms	**150.41**	160.016	182.513	189.116
AR(2) + 25 terms ("jobs")	197.05	173.253	192.5	182.36
AR(2) + 25 terms ("unemployment")	167.018	160.983	**171.385**	**156.221**
AR(2) + 3 keywords	176.607	**150.851**	218.568	166.489
OLS AR(2)	187.182	185.365	237.724	184.451
AIC				
AR(2) + 50 terms	916.75	1406.801	954.0071	1404.542
AR(2) + 25 terms ("jobs")	917.699	1403.12	943.226	1405.091
AR(2) + 25 terms ("unemployment")	913.862	1405.117	935.6629	1421.066
AR(2) + 3 keywords	**901.875**	**1396.52**	916.1187	**1396.533**
OLS AR(2)	913.032	1409.991	**913.0321**	1409.991
Adjusted R^2				
AR(2) + 50 terms	0.3656	**0.4205**	0.3046	**0.4546**
AR(2) + 25 terms ("jobs")	0.3584	0.4103	0.30453	0.4001
AR(2) + 25 terms ("unemployment")	**0.4112**	0.4053	0.1943	0.3745
AR(2) + 3 keywords	0.3731	0.399	**0.3461**	0.3987
OLS AR(2)	0.2803	0.3271	0.2803	0.3271

Table 6.6: Comparison of nowcasting performance for change in unemployment rates on different models using elastic net.

	November 2010 to February 2017		June 2014 to February 2017	
	Growing	Sliding	Growing	Sliding
	6 years and 5 months		10 years	
MAE				
AR(2) + 50 terms	35.055	34.922	26.729	27.478
AR(2) + 25 terms ("jobs")	35.048	35.592	**24.957**	**25.266**
AR(2) + 25 terms ("unemployment")	34.268	34.334	27.548	28.495
AR(2) + 3 keywords	**31.972**	**32.096**	25.736	25.827
OLS AR(2)	35.994	35.73	30.248	30.215
RMSE				
AR(2) + 50 terms	42.465	42.018	**34.466**	35.120
AR(2) + 25 terms ("jobs")	42.874	42.756	34.484	**34.225**
AR(2) + 25 terms ("unemployment")	42.808	41.885	37.888	38.694
AR(2) + 3 keywords	**40.184**	**39.487**	34.689	34.689
OLS AR(2)	43.563	43.09	37.539	37.369
MAPE				
AR(2) + 50 terms	166.818	190.04	114.653	126.208
AR(2) + 25 terms ("jobs")	167.306	189.69	**111.357**	**109.667**
AR(2) + 25 terms ("unemployment")	154.047	176.254	112.059	125.738
AR(2) + 3 keywords	**151.005**	**160.923**	114.685	121.911
OLS AR(2)	186.652	179	152.593	149.848
AIC				
AR(2) + 50 terms	1563.632	1562.704	1564.902	1570.578
AR(2) + 25 terms ("jobs")	1564.224	1566.359	1563.127	1575.651
AR(2) + 25 terms ("unemployment")	1571.647	1568.963	1568.097	1570.931
AR(2) + 3 keywords	**1557.653**	**1555.78**	**1557.861**	**1557.653**
OLS AR(2)	1572.491	1572.491	1572.491	1572.491
Adjusted R^2				
AR(2) + 50 terms	**0.4499**	**0.4457**	**0.4693**	**0.4606**
AR(2) + 25 terms ("jobs")	0.4123	0.4048	0.4293	0.4030
AR(2) + 25 terms ("unemployment")	0.3937	0.3999	0.4010	0.3923
AR(2) + 3 keywords	0.3996	0.4034	0.3994	0.3996
OLS AR(2)	0.3303	0.3303	0.3303	0.3253

For long period, with ten years of fixed data length in the training model, we find that the optimal model with relevant keywords from "jobs" outperform other models (*MAE* "jobs" = 25.266, *RMSE* "jobs" = 34.225, *MAPE* "jobs" = 109.667). The growing window size also reveal that the optimal model with relevant keywords from "jobs" generate more accurate estimates than other models as measured by MAE and MAPE (*MAE* "jobs" = 24.957, *MAPE* "jobs" = 111.357). However, the optimal model with 50 relevant keywords from both categories provides smallest RMSE (*RMSE* "50 terms" = 34.466). This implies that including more relevant search terms with a long time period of training data can help improve nowcast estimates.

After we obtain the best tuning parameters from each model, we re-fit all training data to inspect the goodness-of-fit. We find that the optimal model with three keywords enhances the overall goodness-of-fit better than models that include more keywords. This implies that one or more of these three search terms are more relevant in improving nowcasting models. Amongst the optimal model with relevant keywords using 10 years growing window, the selected coefficients with stronger effect from each elastic net model are visualised in Fig. 6.5. The figure shows top keywords that have a strong effect on the model such as "hotmail", "jobmatch", "career", "part time jobs" and "interview questions".

We investigate whether there is an evidence of different levels of accuracy between elastic net nowcasting models by comparing the DM test across all five models that use 10 years of training data (Table 6.7). Our findings reveal no evidence that all four models including *Google* keywords produce more accurate estimates than the baseline model (*DM* "three keywords" = -2.5154, *DM* "25 terms unemployment" = 0.9943, *DM* "25 terms jobs" = -0.9884, *DM* "50 terms" = -2.0277, N = 33, $p < 0.05$). Nevertheless, the results given by the error metrics imply the potential of including more *Google* search terms to improve nowcast estimates.

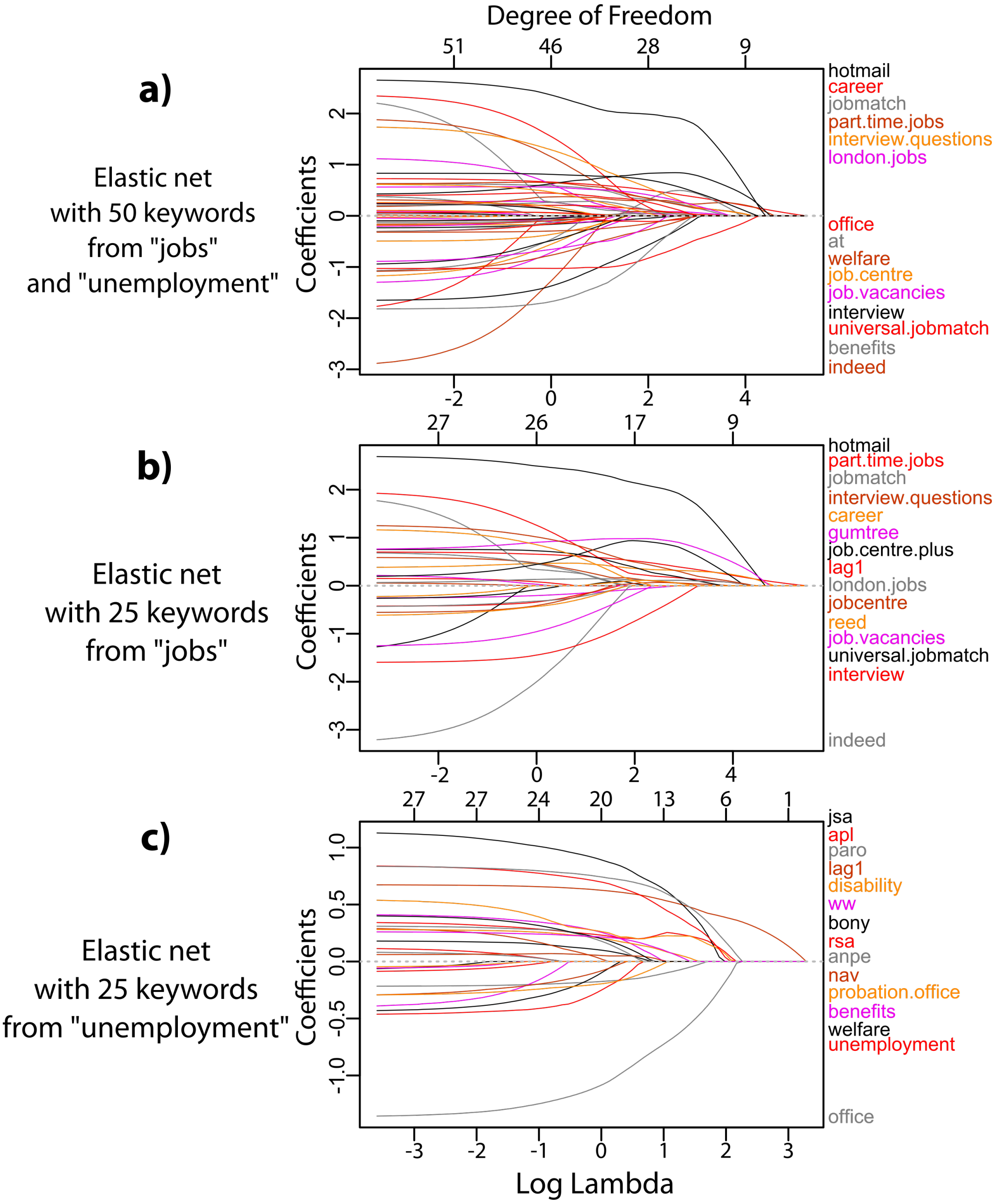

Figure 6.5: **The coefficient paths labelled with the top 15 largest coefficients.**

(a) The selected coefficients with stronger effect from an elastic net model with 50 relevant keywords from "jobs" and "unemployment" categories as identified by *Google Trends*. (b) The selected coefficients with stronger effect from an elastic net model with 25 relevant keywords from "jobs". (c) The coefficient paths of elastic net model incorporating 25 keywords from "unemployment". Visual inspection reveals that there are several top keywords that have a strong effect on the model such as "hotmail", "jobmatch", "career", "part time jobs" and "interview questions".

Table 6.7: Adjusted p-value and DM test statistics of elastic net models with 10 years of training data.

We use FDR correction to control the proportion of false positives, resulting in adjusted p-values. Models that significantly differ in terms of accuracy are highlighted in bold. Negative values of the DM test indicates that the model on the row has smaller errors compared to the model on the column.

	25 terms "jobs"	25 terms "unemployment"	3 keywords	Baseline
50 keywords				
Adjusted p-value	0.2315	0.0626	0.2013	0.1019
DM-statistic	-1.3538	-2.4751	1.5937	-2.0277
25 terms "jobs"				
Adjusted p-value		0.2315	0.0656	0.3303
DM-statistic		-1.3677	2.3302	-0.9884
25 terms "unemployment"				
Adjusted p-value			0.0626	0.3303
DM-statistic			2.7354	0.9943
3 keywords				
Adjusted p-value				0.0626
DM-statistic				-2.5154

4 Discussion

We investigate whether current online datasets hold the similar relationship and may provide new insights when the period is beyond the previous studies [4, 5]. To verify this relationship between *Google Trends* data and unemployment rates from ONS, we analyse on the period of January 2004 to February 2017. The results are in line with McLaren and Shanbhogue [5] analysis in which *Google Trends* data improves estimates of unemployment rates compared to the forecasts generated by model that incorporates the official data only. With the period extends to February 2017, the effect of *Google* search data is increasing based on the error metric results. We find that the linear model including three search terms produce more accurate estimates on both in-sample and out-of-sample tests than the baseline model.

We extend the study to validate whether the nowcasting model can be improved further by including more related *Google* search terms as identified by *Google Trends.* We employ shrinkage or regularisation approach to reduce overfitting problem occurred with linear regression. In particular, we employ elastic net technique to help remove irrelevant variables. Different periods of training data is used to investigate between short term and long term performance. Our results suggest that including more relevant search terms and using elastic net can improve nowcasting performance when use with a certain longer period of training data such as 10 years. On the other hand, for a shorter period of training data, including more search terms help reduce variance at the cost of increasing bias. Thus, we find that model with fewer search terms generally perform better than models that include more search terms.

Qualitatively, the results are in the same direction with previous studies in which models with *Google* search terms provide more accurate nowcast estimates than the baseline model. Moreover, this study shows that variable selection technique such as elastic net can help select the optimal model to obtain the estimates. This technique allow a broader sets of search term to be considered in order to select a number of keywords that are important since fitting all possible keywords would introduce overfitting problem in the nowcasting model. However, this does not mean it will provide the best model since the error of these machine learning techniques are broad and this is not the focus of this thesis. Building on our study, the future research could investigate whether the analysis could be improved using other models or techniques such as ARIMA and alternative versions of LASSO or ridge regression that account for temporal characteristics. Future work could also investigate whether other relevant search terms or other online search data sources, such as Bing, would have potential to improve estimates of current unemployment rate.

Chapter 7

Conclusion

People are constantly communicating online and interacting with social media services via mobile devices. Through these online activities, data is being generated and collected by a range of online service providers. This data can contain information on collective human behaviour and can be used to capture the current state of society. A quicker understanding of the current state of society could help inform decision making in both policy and business.

Previous research has exploited vast amounts of online data to investigate the relationship between online behaviour on the Internet and real world behaviour. In Chapter 2, this thesis has provided and described a wide range of previous studies gaining new insights into human behaviour using online data.

In this thesis, different online data sources - *Instagram* photo data, *Twitter* data and *Google* search data - have been used to obtain quicker estimates of key measurements of society. We started by estimating national statistics across space, specifically language statistics in Greater London and Greater Manchester, using *Instagram* and *Twitter* data and ended by estimating national statistics across time, specifically unemployment rates in the United Kingdom, using *Google Trends* data.

Photos, videos, and messages that are uploaded on social media platforms contain information about human behaviour. In England and Wales, measuring the number of people speaking a particular language across urban areas officially is conducted every ten years by the Census which requires human effort and time. For this reason, we focused on investigating the use of online data to complement the 2011 ONS Census, when estimating the spatial distribution of languages. In Chapter 4, we generated language usage statistics from *Instagram* photos that were uploaded and tagged with locations in Greater London and Greater Manchester. We selected the top 20 most commonly spoken languages across Greater London and Greater

Manchester from *Instagram* language usage. These represent the majority of spoken languages detected in messages exchanged on *Instagram* and their estimates would be more stable than less common languages. Instead of the top 20 languages based on the ONS Census, we chose the top 20 languages based on *Instagram* usage because some top languages in ONS census are not well-represented in the *Instagram* dataset. Our findings show that *Instagram* data has potential to help generate estimates of language usage in different areas of Greater London although we found no such evidence in Greater Manchester.

However, there are also other data sources which are available to investigate. Chapter 5 builds on the results from the previous chapter. We found that *Twitter* data has a smaller number of languages that improves the estimates compared to the *Instagram* analysis, suggesting a weaker effect of *Twitter* data on improving the estimates. This may be due to lower availability of posts with geotagged coordinates on *Twitter* than on *Instagram*. Nevertheless, our findings suggest that, overall, including *Twitter* data improve estimates more than baseline models. We explored *Twitter* data further on borough-level analysis in Greater London. At borough-level, we found that *Twitter* data that are automatically tagged with places provide a smaller number of languages that can help generate more accurate estimates compared to MSOA-level analysis on *Twitter*. However, our results revealed evidence of improvement in estimates for only one language.

On the other hand, statistics across time can capture the current trends of economic activities. In Chapter 6, we investigated the potential of using online data to estimate national statistics across time, specifically unemployment rates in the United Kingdom. We used UK unemployment rates from the ONS and *Google* search data from January 2004 to February 2017. Our results reveal that *Google* search data can help generate current estimates before the official data is released. Previous research has focused on single or small groups of *Google* search terms. The common approach has been subjective by manually selecting relevant keyword(s). In order to objectively select a number of keywords that are important, we have developed the nowcasting model further by including more relevant *Google* search terms, using regularisation techniques, specifically elastic net, employing different periods for the model to consider. Then, the out-of-sample and cross-validation techniques are used to validate the performance. Our results suggest that variable selection via an elastic net provides a better improvement in nowcasting performance when incorporating long periods of data while models with fewer search terms generally perform better in a shorter period. The study can be extended further by using other variable selection techniques or complex models to improve the results, including obtaining the optimal choice of keywords and the optimal time period considered for the model.

We noted the time gap between 2011 Census data and *Instagram* data which spans from September 2015 to February 2016 for Greater London and August 2013 to December 2013 for Greater Manchester where there is no official data to compare the accuracy of language usage estimates across areas. Future work could also investigate potential for improving estimates across time, once new Census data becomes available after 2021. This work has also focused on two large cities, Greater London and Greater Manchester, and it is sensible to assume that these cities will have greater levels of social media usage compared to smaller cities. However, there are various UK cities that have different characteristics such as size, income level, age demographics, and population density. Future research could group cities according to the scale of population density (i.e. classifying cities into small, medium, and large based on the number of people living in the city) and select one city in each group as a sample. Therefore, investigating whether language usage on *Instagram* or *Twitter* on cities with different characteristics (e.g. size and population density) could help in estimating language statistics and reveal a complete picture of the usefulness of online data.

Moreover, it is highly likely that there is a difference in the age distribution across language speakers for both the *Instagram* and the *Twitter* platform. The numbers of *Instagram* or *Twitter* users and the language variation that is reflected in *Instagram* and *Twitter* data might affect this analysis to some extent [50]. For example, older people might have less interest in using social media platforms than younger people. Therefore, some of the languages that are represented in the census from the ONS but not in *Instagram* and *Twitter* might be due to the difference in age population and their preference to use social media. This work has focused on top 20 common languages on *Instagram* and *Twitter* usage which are majority of these datasets. Future research could look into different criteria of common languages used for the analysis (e.g. top 20 languages from ONS Census), which might reveal other perspectives of online data.

This work has focused on only unemployment rates in the UK whereas there are also other countries, such as the US and Thailand, which could be further investigated to validate whether variable selection techniques or elastic net could improve nowcasting performance. This thesis has examined one example of economic indicators on a country-level analysis, whereas city level analysis could produce a clearer picture of the current state of the city. However, search volume data at a city level are currently unavailable from various search engines.

Apart from people searching online for jobs, one of the other possibilities is that people are looking for information online to support their decisions in buying consumer products. Therefore, using variable selection techniques to include relevant *Google* keywords for nowcasting other areas and key economic measurements such as retail sales statistics could improve nowcasting performance and provide more insight into the usefulness of online search data. There are also other widely used search engines such as *Bing*, and *Yahoo* that receive less attention. Merging search volume data from further search engines could reveal whether they can help provide better and quicker estimates of key economic measurements before the official figures are published.

There will be increasing challenges in the future for research that looks into online data, especially social media platforms. Recently, online data have been exploited for commercial or personal purposes, for instance, using online data to support political campaigns. There is also an increasing trend of leaked personal data from online service providers. Therefore, people in the public are more aware of online privacy when using both social media platforms and search engines. It is possible that these online data sources might restrict data access in the future, resulting lower volumes of data that are publicly available for research purposes.

This thesis has demonstrated the usefulness of online data in estimating statistics across space and statistics across time. It has investigated the relationship between online behaviour and real world behaviour. By exploiting spatial data and time series data that are publicly available from online platforms, researchers can gain new insight into human activity patterns and obtain key measurements of society at low cost. Obtaining quicker and cheaper estimates using online data could reveal crucial information about human behaviour at a collective level, for policymakers and businesses alike, helping illuminate new opportunities for society.

Bibliography

[1] ONS. Why we have a census - Office for National Statistics, n.d.. URL `https://www.ons.gov.uk/census/2011census/whywehaveacensus`.

[2] BBC. National Census could be scrapped, July 2010. URL `https://www.bbc.com/news/10584385`.

[3] ONS. Frequently asked questions - Office for National Statistics, n.d.. URL `https://www.ons.gov.uk/census/2011census/2011censusdata/2011censususerguide/frequentlyaskedquestions`.

[4] Hyunyoung Choi and Hal Varian. Predicting the present with Google Trends. *Economic Record*, 88(s1):2–9, June 2012. ISSN 1475-4932. doi: 10.1111/j.1475-4932.2012.00809.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-4932.2012.00809.x`.

[5] Nick McLaren and Rachana Shanbhogue. Using Internet search data as economic indicators. *Bank of England Quarterly Bulletin*, 51(2):134–140, 2011. URL `https://EconPapers.repec.org/RePEc:boe:qbullt:0052`.

[6] Paul Smith. Google's MIDAS touch: Predicting UK unemployment with Internet search data. *Journal of Forecasting*, 35(3):263–284, April 2016. ISSN 1099-131X. doi: 10.1002/for.2391. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2391`.

[7] ONS. Background to the Census transformation programme, August 2013. URL `https://webarchive.nationalarchives.gov.uk/20160111070955/http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/background-to-beyond-2011/index.html`.

[8] Stuart McPherson. The Census and future provision of population statistics in England and Wales: Recommendation from the national statistician and chief executive of the UK statistics authority, February 2014. URL

https://webarchive.nationalarchives.gov.uk/20160108193324/http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/beyond-2011-report-on-autumn-2013-consultation--and-recommendations/index.html.

[9] US Census Bureau. 2020 Census operational plan v4.0. Technical report, US Department of Commerce, December 2018. URL https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-oper-plan4.pdf.

[10] Cabinet Office. *Help Shape Our Future: The 2021 Census of Population and Housing in England and Wales*. 2018. ISBN 978-1-5286-0840-4. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/765089/Census2021WhitePaper.pdf. OCLC: 1090683406.

[11] Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676, May 2008. ISSN 0304-3932. doi: 10.1016/j.jmoneco.2008.05.010. URL http://www.sciencedirect.com/science/article/pii/S0304393208000652.

[12] Sam Miller, Helen Susannah Moat, and Tobias Preis. Using aircraft location data to estimate current economic activity. *Scientific Reports*, 2:7576, May 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-63734-w. URL https://www.nature.com/articles/s41598-020-63734-w.

[13] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, February 2009. ISSN 1476-4687. doi: 10.1038/nature07634. URL https://www.nature.com/articles/nature07634.

[14] Tobias Preis and Helen Susannah Moat. Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society Open Science*, 1(2):140095, October 2014. ISSN 2054-5703. doi: 10.1098/rsos.140095. URL http://rsos.royalsocietypublishing.org/content/1/2/140095.

[15] Nikolaos Askitas and Klaus F Zimmermann. Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(2):107–120, April

2009. ISSN 1611-6607. doi: 10.3790/aeq.55.2.107. URL https://ejournals.duncker-humblot.de/doi/abs/10.3790/aeq.55.2.107.

[16] Nikolaos Askitas and Klaus F. Zimmermann. The internet as a data source for advancement in social sciences. *International Journal of Manpower*, 36 (1):2–12, April 2015. ISSN 0143-7720. doi: 10.1108/IJM-02-2015-0029. URL http://www.emeraldinsight.com/doi/10.1108/IJM-02-2015-0029.

[17] Francesco D'Amuri and Juri Marcucci. The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, 33 (4):801–816, 2017. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2017.03.004. URL http://www.sciencedirect.com/science/article/pii/S0169207017300389.

[18] Tanya Suhoy. Query indices and a 2008 downturn: Israeli data. Technical Report, Bank of Israel, 2009. URL https://www.boi.org.il/deptdata/mehkar/papers/dp0906e.pdf.

[19] Daniele Barchiesi, Helen Susannah Moat, Christian Alis, Steven Bishop, and Tobias Preis. Quantifying international travel flows using Flickr. *PLOS ONE*, 10(7):e0128470, July 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0128470. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0128470.

[20] Daniele Barchiesi, Tobias Preis, Steven Bishop, and Helen Susannah Moat. Modelling human mobility patterns using photographic data shared online. *Royal Society Open Science*, 2(8):150046, August 2015. ISSN 2054-5703. doi: 10.1098/rsos.150046. URL http://rsos.royalsocietypublishing.org/content/2/8/150046.

[21] Tobias Preis, Federico Botta, and Helen Susannah Moat. Sensing global tourism numbers with millions of publicly shared online photographs. *Environment and Planning A: Economy and Space*, pages 471–477, 2020. doi: 10.1177/0308518X19872772. URL https://doi.org/10.1177/0308518X19872772.

[22] Lydia Manikonda, Yuheng Hu, and Subbarao Kambhampati. Analyzing user activities, demographics, social network structure and user-generated content on Instagram. *arXiv:1410.8099 [physics]*, October 2014. URL http://arxiv.org/abs/1410.8099.

[23] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, Juliana Salles, and Antonio A. F. Loureiro. A Comparison of Foursquare and Instagram

to the study of city dynamics and urban social behavior. In *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*, UrbComp '13, pages 4:1–4:8, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2331-4. doi: 10.1145/2505821.2505836. URL http://doi.acm.org/10.1145/2505821.2505836.

[24] Federico Botta, Helen Susannah Moat, and Tobias Preis. Measuring the size of a crowd using Instagram. *Environment and Planning B: Urban Analytics and City Science*, 2019. doi: 10.1177/2399808319841615. URL https://doi.org/10.1177/2399808319841615.

[25] Nadav Hochman and Lev Manovich. Zooming into an Instagram city: Reading the local through social media. *First Monday*, 18(7), June 2013. ISSN 13960466. doi: 10.5210/fm.v18i7.4711. URL https://journals.uic.edu/ojs/index.php/fm/article/view/4711.

[26] Nadav Hochman and Raz Schwartz. Visualizing Instagram: Tracing cultural visual rhythms. In *Proceedings of the Workshop on Social Media Visualization (SocMedVis) in Conjunction with the Sixth International AAAI Conference on Weblogs and Social Media*, pages 6–9, Dublin, Ireland, 2012. URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4782.

[27] Saeideh Bakhshi, David A. Shamma, and Eric Gilbert. Faces engage us: Photos with faces attract more likes and comments on Instagram. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 965–974, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557403. URL http://doi.acm.org/10.1145/2556288.2557403.

[28] Michele Zappavigna. Social media photography: Construing subjectivity in Instagram images. *Visual Communication*, 15(3):271–292, August 2016. ISSN 1470-3572. doi: 10.1177/1470357216643220. URL https://doi.org/10.1177/1470357216643220.

[29] Sara Santarossa, Paige Coyne, Carly Lisinski, and Sarah J Woodruff. #fitspo on Instagram: A mixed-methods approach using Netlytic and photo analysis, uncovering the online discussion and author/image characteristics. *Journal of Health Psychology*, page 1359105316676334, November 2016. ISSN 1359-1053. doi: 10.1177/1359105316676334. URL https://doi.org/10.1177/1359105316676334.

[30] Paulina Guerrero, Maja Steen Møller, Anton Stahl Olafsson, and Bernhard Snizek. Revealing cultural ecosystem services through Instagram images: The potential of social media volunteered geographic information for urban green infrastructure planning and governance. *Urban Planning*, 1(2):1–17, 2016. ISSN 2183-7635. doi: 10.17645/up.v1i2.609.

[31] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Analyzing labeled cyberbullying incidents on the Instagram social network. In Tie-Yan Liu, Christie Napa Scollon, and Wenwu Zhu, editors, *Social Informatics*, Lecture Notes in Computer Science, pages 49–66. Springer International Publishing, 2015. ISBN 978-3-319-27433-1.

[32] Twitter. Tweet geospatial metadata, n.d.. URL `https://developer.twitter.com/en/docs/tutorials/tweet-geo-metadata.html`.

[33] Mark Graham, Scott A. Hale, and Devin Gaffney. Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, 66(4):568–578, October 2014. ISSN 0033-0124, 1467-9272. doi: 10.1080/00330124.2014.907699. URL `http://arxiv.org/abs/1308.0683`.

[34] Twitter. How to add your location to a tweet, n.d.. URL `https://help.twitter.com/en/using-twitter/tweet-location`.

[35] Bingdong Li, Esra Erdin, Mehmet Hadi Gunes, George Bebis, and Todd Shipley. An overview of anonymity technology usage. *Computer Communications*, 36(12):1269–1283, July 2013. ISSN 0140-3664. doi: 10.1016/j.comcom.2013.04.009. URL `http://www.sciencedirect.com/science/article/pii/S0140366413001096`.

[36] M. Romaryo Molana Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. Population bias in geotagged tweets. In *Proceedings of the Ninth International AAAI Conference on Weblogs and Social Media*, Oxford, England, 2015. URL `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10662`.

[37] Simon Carter, Manos Tsagkias, and Wouter Weerkamp. Semi-supervised priors for microblog language identification. In *Dutch-Belgian Information Retrieval Workshop*, DIR, pages 12–15, Amsterdam, Netherlands, 2011.

[38] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Proceedings of the*

2010 IEEE Second International Conference on Social Computing, pages 177–184, Minneapolis, Minnesota, USA, August 2010. doi: 10.1109/SocialCom.2010.33.

[39] Scott A. Hale. Net increase? Cross-lingual linking in the Blogosphere. *Journal of Computer-Mediated Communication*, 17(2):135–151, January 2012. ISSN 1083-6101. doi: 10.1111/j.1083-6101.2011.01568.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1083-6101.2011.01568.x`.

[40] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on Twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0493-1. doi: 10.1145/1935826.1935845. URL `http://doi.acm.org/10.1145/1935826.1935845`.

[41] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, May 2012. ISSN 1947-4040. doi: 10.2200/S00416ED1V01Y201204HLT016. URL `https://www.morganclaypool.com/doi/abs/10.2200/s00416ed1v01y201204hlt016`.

[42] Vaibhavi N Patodkar and Sheikh I.R. Twitter as a corpus for sentiment analysis and opinion mining. *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)*, 5(12):320–322, December 2016. ISSN 22781021. doi: 10.17148/IJARCCE.2016.51274. URL `http://ijarcce.com/upload/2016/december-16/IJARCCE%2074.pdf`.

[43] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 538–541, Barcelona, Spain, 2011.

[44] Rushabh Mehta, Dhaval Mehta, Disha Chheda, Charmi Shah, and Pramila M. Chawan. Sentiment analysis and influence tracking using Twitter. *International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE)*, 1(2):pp:72–79–79, April 2012. ISSN 2277–9043. URL `http://www.ijarcsee.org/index.php/IJARCSEE/article/view/32`.

[45] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, March 2011. ISSN 1877-

7503. doi: 10.1016/j.jocs.2010.12.007. URL http://www.sciencedirect.com/science/article/pii/S187775031100007X.

[46] Pete Burnap, Matthew L. Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, 4(1):206, June 2014. ISSN 1869-5469. doi: 10.1007/s13278-014-0206-4. URL https://doi.org/10.1007/s13278-014-0206-4.

[47] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Alessandro Flammini, and Filippo Menczer. Political polarization on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 89–96, Barcelona, Spain, 2011. URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847.

[48] Sounman Hong and Sun Hyoung Kim. Political polarization on twitter: Implications for the use of social media in digital governments. *Government Information Quarterly*, 33(4):777–782, October 2016. ISSN 0740-624X. doi: 10.1016/j.giq.2016.04.007. URL http://www.sciencedirect.com/science/article/pii/S0740624X16300375.

[49] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185, Washington, D.C., USA, 2010.

[50] Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. The Twitter of Babel: Mapping world languages through microblogging platforms. *PLOS ONE*, 8(4):e61981, April 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0061981. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061981.

[51] Barbara Poblete, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. Do all birds tweet the same?: Characterizing Twitter around the world. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1025–1030, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063724. URL http://doi.acm.org/10.1145/2063576.2063724.

[52] Wouter Weerkamp, Simon Carter, and Manos Tsagkias. How people use twitter in different languages. In *ACM Web Science 2011*, Koblenz, Germany, 2011.

[53] Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman. Geography of Twitter networks. *Social Networks*, 34(1):73–81, January 2012. ISSN 0378-8733. doi: 10.1016/j.socnet.2011.05.006. URL `http://www.sciencedirect.com/science/article/pii/S0378873311000359`.

[54] Christian M. Alis, May T. Lim, Helen Susannah Moat, Daniele Barchiesi, Tobias Preis, and Steven R. Bishop. Quantifying regional differences in the length of Twitter messages. *PLOS ONE*, 10(4):e0122278, April 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0122278. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0122278`.

[55] Przemyslaw A. Grabowicz, José J. Ramasco, Bruno Gonçalves, and Víctor M. Eguíluz. Entangling mobility and interactions in social media. *PLOS ONE*, 9(3):e92196, March 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0092196. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0092196`.

[56] L. Alessandretti, P. Sapiezynski, S. Lehmann, and A. Baronchelli. Multi-scale spatio-temporal analysis of human mobility. *PLOS One*, 12:e0171686., February 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0171686. URL `http://openaccess.city.ac.uk/16791/`.

[57] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on Twitter networks: Validation of Dunbar's number. *PLOS ONE*, 6(8):e22656, August 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0022656. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0022656`.

[58] Vasileios Lampos and Nello Cristianini. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology*, 3(4):72:1–72:22, September 2012. ISSN 2157-6904. doi: 10.1145/2337542.2337557. URL `http://0.doi.acm.org/10.1145/2337542.2337557`.

[59] Giovanini Evelim Coelho, Priscila Leite Leal, Matheus de Paula Cerroni, Ana Cristina Rocha Simplicio, and João Bosco Siqueira Jr. Sensitivity of the dengue surveillance system in Brazil for detecting hospitalized cases. *PLOS Neglected Tropical Diseases*, 10(5):e0004705, May 2016. ISSN 1935-

2735. doi: 10.1371/journal.pntd.0004705. URL https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0004705.

[60] Janaína Gomide, Adriano Veloso, Wagner Meira, Jr., Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In *Proceedings of the 3rd International Web Science Conference*, WebSci '11, pages 3:1–3:8, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0855-7. doi: 10.1145/2527031.2527049. URL http://doi.acm.org/10.1145/2527031.2527049.

[61] Greg Miller. Social scientists wade into the tweet stream. *Science*, 333(6051): 1814–1815, September 2011. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.333.6051.1814. URL http://science.sciencemag.org/content/333/6051/1814.

[62] Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLOS ONE*, 6(5):e19467, May 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0019467. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0019467.

[63] Claudia Codeco, Flavio C. Coelho, Oswaldo Cruz, S. Oliveira, T. Castro, and Leonardo Bastos. Infodengue: A nowcasting system for the surveillance of arboviruses in Brazil. *Revue d'Épidémiologie et de Santé Publique*, 66:S386, 2018. ISSN 0398-7620. doi: https://doi.org/10.1016/j.respe.2018.05.408. URL http://www.sciencedirect.com/science/article/pii/S0398762018311088.

[64] Cecilia de Almeida Marques-Toledo, Carolin Marlen Degener, Livia Vinhal, Giovanini Coelho, Wagner Meira, Claudia Torres Codeço, and Mauro Martins Teixeira. Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level. *PLOS Neglected Tropical Diseases*, 11(7):e0005729, July 2017. ISSN 1935-2735. doi: 10.1371/journal.pntd.0005729. URL https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0005729.

[65] BBC. Facebook hits two billion users, June 2017. URL https://www.bbc.com/news/business-40424769.

[66] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Michelangelo Puliga, Antonio Scala, Guido Caldarelli, Brian Uzzi, and Walter Quattrocioc-

chi. Users polarization on Facebook and Youtube. *PLOS ONE*, 11(8): e0159641, August 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0159641. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0159641.

[67] Ana Lucía Schmidt, Fabiana Zollo, Michela Del Vicario, Alessandro Bessi, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. Anatomy of news consumption on Facebook. *Proceedings of the National Academy of Sciences*, 114(12):3035–3039, March 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1617052114. URL https://www.pnas.org/content/114/12/3035.

[68] Alessandro Bessi, Fabio Petroni, Michela Del Vicario, Fabiana Zollo, Aris Anagnostopoulos, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Viral misinformation: The role of homophily and polarization. In *Proceedings of the 24th International Conference on World Wide Web*, pages 355–356, Florence, Italy, May 2015. ACM. ISBN 978-1-4503-3473-0. doi: 10.1145/2740908.2745939. URL http://0-dl.acm.org.pugwash.lib.warwick.ac.uk/citation.cfm?id=2740908.2745939.

[69] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, May 2017. ISSN 0895-3309. doi: 10.1257/jep.31.2.211. URL https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211.

[70] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, June 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaa1160. URL http://science.sciencemag.org/content/348/6239/1130.

[71] Marco Della Vedova, Eugenio Tacchini, Stefano Moret, Gabriele Ballarin, Massimo DiPierro, and Luca de Alfaro. Automatic online fake news detection combining content and social signals. In *Proceedings of the 22st Conference of Open Innovations Association FRUCT*, FRUCT'22, pages 38:272–38:279, Jyvaskyla, Finland, 2018. FRUCT Oy. URL http://dl.acm.org/citation.cfm?id=3266365.3266403.

[72] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Science vs conspiracy: Collective narratives in the age of misinformation. *PLOS ONE*, 10

(2):e0118093, February 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0118093. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118093.

[73] Fabiana Zollo, Petra Kralj Novak, Michela Del Vicario, Alessandro Bessi, Igor Mozetič, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Emotional dynamics in the age of misinformation. *PLOS ONE*, 10 (9):e0138740, September 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0138740. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0138740.

[74] Delia Mocanu, Luca Rossi, Qian Zhang, Marton Karsai, and Walter Quattrociocchi. Collective attention in the age of (mis)information. *Computers in Human Behavior*, 51:1198–1204, October 2015. ISSN 0747-5632. doi: 10.1016/j.chb.2015.01.024. URL http://www.sciencedirect.com/science/article/pii/S0747563215000382.

[75] Alessandro Bessi, Antonio Scala, Luca Rossi, Qian Zhang, and Walter Quattrociocchi. The economy of attention in the age of (mis)information. *Journal of Trust Management*, 1(1):12, December 2014. ISSN 2196-064X. doi: 10.1186/s40493-014-0012-y. URL https://doi.org/10.1186/s40493-014-0012-y.

[76] Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Echo chambers: Emotional contagion and group polarization on Facebook. *Scientific Reports*, 6(1):37825, December 2016. ISSN 2045-2322. doi: 10.1038/srep37825. URL https://doi.org/10.1038/srep37825.

[77] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, January 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1517441113. URL https://www.pnas.org/content/113/3/554.

[78] Fabiana Zollo, Alessandro Bessi, Michela Del Vicario, Antonio Scala, Guido Caldarelli, Louis Shekhtman, Shlomo Havlin, and Walter Quattrociocchi. Debunking in a world of tribes. *PLOS ONE*, 12(7):e0181821, July 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0181821. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181821.

[79] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Trend of narratives in the age of misinformation. *PLOS ONE*, 10(8):e0134641, August 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0134641. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0134641`.

[80] Jessica Vitak, Paul Zube, Andrew Smock, Caleb T. Carr, Nicole Ellison, and Cliff Lampe. It's complicated: Facebook users' political participation in the 2008 election. *Cyberpsychology, Behavior, and Social Networking*, 14(3):107–114, March 2011. ISSN 2152-2715. doi: 10.1089/cyber.2009.0226. URL `https://www.liebertpub.com/doi/abs/10.1089/cyber.2009.0226`.

[81] Julia K. Woolley, Anthony M. Limperos, and Mary Beth Oliver. The 2008 presidential election, 2.0: A content analysis of user-generated political Facebook groups. *Mass Communication and Society*, 13(5):631–652, October 2010. ISSN 1520-5436. doi: 10.1080/15205436.2010.516864. URL `https://doi.org/10.1080/15205436.2010.516864`.

[82] Daniela V. Dimitrova and Dianne Bystrom. The effects of social media on political participation and candidate image evaluations in the 2012 Iowa caucuses:. *American Behavioral Scientist*, May 2013. doi: 10.1177/0002764213489011. URL `https://0-journals-sagepub-com.pugwash.lib.warwick.ac.uk/doi/abs/10.1177/0002764213489011`.

[83] Meredith Conroy, Jessica T. Feezell, and Mario Guerrero. Facebook and political engagement: A study of online political group membership and offline political engagement. *Computers in Human Behavior*, 28(5):1535–1546, September 2012. ISSN 0747-5632. doi: 10.1016/j.chb.2012.03.012. URL `http://www.sciencedirect.com/science/article/pii/S0747563212000787`.

[84] Gary Tang and Francis L. F. Lee. Facebook use and political participation: The impact of exposure to shared political information, connections with public political actors, and network structural heterogeneity. *Social Science Computer Review*, 31(6):763–773, December 2013. ISSN 0894-4393. doi: 10.1177/0894439313490625. URL `https://doi.org/10.1177/0894439313490625`.

[85] Terri L. Towner. Campaigns and elections in a web 2.0 world: Uses, effects, and implications for democracy. In Christopher G. Reddick and Stephen K. Aikins, editors, *Web 2.0 Technologies and Democratic Governance: Political, Policy and Management Implications*, Public Administration and Information

Technology, pages 185–199. Springer New York, New York, NY, 2012. ISBN 978-1-4614-1448-3. doi: 10.1007/978-1-4614-1448-3_12. URL `https://doi.org/10.1007/978-1-4614-1448-3_12`.

[86] Leticia Bode, Emily K. Vraga, Porismita Borah, and Dhavan V. Shah. A new space for political nehavior: Political social networking and its democratic consequences. *Journal of Computer-Mediated Communication*, 19(3):414–429, April 2014. doi: 10.1111/jcc4.12048. URL `https://academic.oup.com/jcmc/article/19/3/414/4067544`.

[87] Federica Liberini, Michela Redoano, Antonio Russo, Angel Cuevas, and Ruben Cuevas. Politics in the Facebook era evidence from the 2016 US presidential elections. Technical Report 389, Competitive Advantage in the Global Economy (CAGE), 2018. URL `https://ideas.repec.org/p/cge/wacage/389.html`.

[88] Martin Moore. Facebook, the Conservatives and the risk to fair and open elections in the UK. *The Political Quarterly*, 87(3):424–430, July 2016. ISSN 0032-3179. doi: 10.1111/1467-923X.12291. URL `https://0-onlinelibrary-wiley-com.pugwash.lib.warwick.ac.uk/doi/abs/10.1111/1467-923X.12291`.

[89] Porismita Borah. Political Facebook use: Campaign strategies used in 2008 and 2012 presidential elections. *Journal of Information Technology & Politics*, 13(4):326–338, October 2016. ISSN 1933-1681. doi: 10.1080/19331681.2016.1163519. URL `https://doi.org/10.1080/19331681.2016.1163519`.

[90] Tobias Moers, Florian Krebs, and Gerasimos Spanakis. SEMTec: Social emotion mining techniques for analysis and prediction of Facebook post reactions. In Jaap van den Herik and Ana Paula Rocha, editors, *Agents and Artificial Intelligence*, Lecture Notes in Computer Science, pages 361–382. Springer International Publishing, 2019. ISBN 978-3-030-05453-3.

[91] Ye Tian, Thiago Galery, Giulio Dulcinati, Emilia Molimpakis, and Chao Sun. Facebook sentiment: Reactions and emojis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 11–16, Valencia, Spain, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1102. URL `http://aclweb.org/anthology/W17-1102`.

[92] Chris Pool and Malvina Nissim. Distant supervision for emotion detection using Facebook reactions. In *Proceedings of the Workshop on Computational*

Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES), pages 30–39, Osaka, Japan, 2016. The COLING 2016 Organizing Committee. URL `https://www.aclweb.org/anthology/W16-4304`.

[93] Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. Quantifying the link between art and property prices in urban neighbourhoods. *Royal Society Open Science*, 3(4):160146, April 2016. doi: 10.1098/rsos.160146. URL `https://royalsocietypublishing.org/doi/full/10.1098/rsos.160146`.

[94] Aiello Luca Maria, Schifanella Rossano, Quercia Daniele, and Aletta Francesco. Chatty maps: Constructing sound maps of urban areas from social media data. *Royal Society Open Science*, 3(3):150690, March 2016. doi: 10.1098/rsos.150690. URL `https://royalsocietypublishing.org/doi/full/10.1098/rsos.150690`.

[95] Daniele Quercia, Luca Maria Aiello, and Rossano Schifanella. The emotional and chromatic layers of urban smells. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, Cologne, Germany, March 2016. URL `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13092`.

[96] Merve Alanyali, Tobias Preis, and Helen Susannah Moat. Tracking protests using geotagged Flickr photographs. *PLOS ONE*, 11(3):e0150466, March 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0150466. URL `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0150466`.

[97] Tobias Preis, Helen Susannah Moat, Steven R. Bishop, Philip Treleaven, and H. Eugene Stanley. Quantifying the digital traces of Hurricane Sandy on Flickr. *Scientific Reports*, 3:3141, November 2013. ISSN 2045-2322. doi: 10.1038/srep03141. URL `https://www.nature.com/articles/srep03141`.

[98] Huy Quan Vu, Gang Li, Rob Law, and Yanchun Zhang. Travel diaries analysis by sequential rule mining. *Journal of Travel Research*, 57(3):399–413, March 2018. ISSN 0047-2875. doi: 10.1177/0047287517692446. URL `https://doi.org/10.1177/0047287517692446`.

[99] Spencer A. Wood, Anne D. Guerry, Jessica M. Silver, and Martin Lacayo. Using social media to quantify nature-based tourism and recreation. *Scientific Reports*, 3:2976, October 2013. ISSN 2045-2322. doi: 10.1038/srep02976. URL `https://www.nature.com/articles/srep02976/`.

[100] Irem Önder, Wolfgang Koerbitz, and Alexander Hubmann-Haidvogel. Tracing tourists by their digital footprints: The case of Austria. *Journal of Travel Research*, 55(5):566–573, May 2016. ISSN 0047-2875. doi: 10.1177/0047287514563985. URL https://doi.org/10.1177/0047287514563985.

[101] Yihong Yuan and Monica Medel. Characterizing international travel behavior from geotagged photos: A case study of Flickr. *PLOS ONE*, 11 (5):e0154885, May 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0154885. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0154885.

[102] J. Clement. Instagram: Active users 2018 | Statista, 2019. URL https://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users/.

[103] Flickr. Jobs, n.d. URL https://www.flickr.com/jobs.

[104] Statista. Global Instagram user age & gender distribution 2019 | Statista, 2019. URL https://www.statista.com/statistics/248769/age-distribution-of-worldwide-instagram-users/.

[105] Paige Worthy. Top Instagram demographics that matter to social media marketers, September 2018. URL https://blog.hootsuite.com/instagram-demographics/.

[106] StatCounter. Search engine market share worldwide, n.d.. URL http://gs.statcounter.com/search-engine-market-share.

[107] Statista. Search engine market share worldwide, n.d. URL https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/.

[108] Google. How Trends data is adjusted - Trends help, n.d.. URL https://support.google.com/trends/answer/4365533?hl=en.

[109] Google. Find related searches - Trends help, n.d.. URL https://support.google.com/trends/answer/4355000?hl=en.

[110] Theologos Dergiades, Costas Milas, and Theodore Panagiotidis. Tweets, Google trends, and sovereign spreads in the GIIPS. *Oxford Economic Papers*, 67(2):406–432, April 2015. ISSN 0030-7653. doi: 10.1093/oep/gpu046. URL https://academic.oup.com/oep/article/67/2/406/2362293.

[111] Tobias Preis, Daniel Reith, and H. Eugene Stanley. Complex dynamics of our economic life on different scales: Insights from search engine query data. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1933):5707–5719, December 2010. ISSN 1364-503X, 1471-2962. doi: 10.1098/rsta.2010.0284. URL `http://rsta.royalsocietypublishing.org/content/368/1933/5707`.

[112] Tobias Preis, Helen Susannah Moat, and H. Eugene Stanley. Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports*, 3:1684, April 2013. ISSN 2045-2322. doi: 10.1038/srep01684. URL `https://www.nature.com/articles/srep01684`.

[113] Tobias Preis, Helen Susannah Moat, H. Eugene Stanley, and Steven R. Bishop. Quantifying the advantage of looking forward. *Scientific Reports*, 2:350, April 2012. ISSN 2045-2322. doi: 10.1038/srep00350. URL `https://www.nature.com/articles/srep00350`.

[114] Jonathan Mellon. Internet search data and issue salience: The properties of Google Trends as a measure of issue salience. *Journal of Elections, Public Opinion and Parties*, 24(1):45–72, January 2014. ISSN 1745-7289. doi: 10.1080/17457289.2013.846346. URL `https://doi.org/10.1080/17457289.2013.846346`.

[115] John S. Brownstein, Clark C. Freifeld, and Lawrence C. Madoff. Digital disease detection — Harnessing the web for public health surveillance. *New England Journal of Medicine*, 360(21):2153–2157, May 2009. ISSN 0028-4793. doi: 10.1056/NEJMp0900702. URL `https://doi.org/10.1056/NEJMp0900702`.

[116] Courtney D. Corley, Diane J. Cook, Armin R. Mikler, and Karan P. Singh. Text and structural data mining of influenza mentions in web and social media. *International Journal of Environmental Research and Public Health*, 7(2):596–615, 2010. ISSN 1660-4601. doi: 10.3390/ijerph7020596. URL `https://www.mdpi.com/1660-4601/7/2/596`.

[117] Benjamin M. Althouse, Samuel V. Scarpino, Lauren Ancel Meyers, John W. Ayers, Marisa Bargsten, Joan Baumbach, John S. Brownstein, Lauren Castro, Hannah Clapham, Derek AT Cummings, Sara Del Valle, Stephen Eubank, Geoffrey Fairchild, Lyn Finelli, Nicholas Generous, Dylan George, David R. Harper, Laurent Hébert-Dufresne, Michael A. Johansson, Kevin Konty, Marc Lipsitch, Gabriel Milinovich, Joseph D. Miller, Elaine O. Nsoesie, Donald R.

Olson, Michael Paul, Philip M. Polgreen, Reid Priedhorsky, Jonathan M. Read, Isabel Rodríguez-Barraquer, Derek J. Smith, Christian Stefansen, David L. Swerdlow, Deborah Thompson, Alessandro Vespignani, and Amy Wesolowski. Enhancing disease surveillance with novel data streams: Challenges and opportunities. *EPJ Data Science*, 4(1):17, October 2015. ISSN 2193-1127. doi: 10.1140/epjds/s13688-015-0054-0. URL `https://doi.org/10.1140/epjds/s13688-015-0054-0`.

[118] A. Valdivia, J. López-Alcalde, M. Vicente, M. Pichiule, M. Ruiz, and M. Ordobas. Monitoring influenza activity in Europe with Google Flu Trends: Comparison with the findings of sentinel physician networks – results for 2009-10. *Eurosurveillance*, 15(29):19621, July 2010. ISSN 1560-7917. doi: 10.2807/ese.15.29.19621-en. URL `https://www.eurosurveillance.org/content/10.2807/ese.15.29.19621-en`.

[119] Sungjin Cho, Chang Hwan Sohn, Min Woo Jo, Soo-Yong Shin, Jae Ho Lee, Seoung Mok Ryoo, Won Young Kim, and Dong-Woo Seo. Correlation between national influenza surveillance data and Google Trends in South Korea. *PLOS ONE*, 8(12):e81422, December 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0081422. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0081422`.

[120] Justin R. Ortiz, Hong Zhou, David K. Shay, Kathleen M. Neuzil, Ashley L. Fowlkes, and Christopher H. Goss. Monitoring influenza activity in the United States: A comparison of traditional surveillance systems with Google Flu Trends. *PLOS ONE*, 6(4):e18687, April 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0018687. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018687`.

[121] Min Kang, Haojie Zhong, Jianfeng He, Shannon Rutherford, and Fen Yang. Using Google Trends for influenza surveillance in South China. *PLOS ONE*, 8(1):e55205, January 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0055205. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0055205`.

[122] Vasileios Lampos, Andrew C. Miller, Steve Crossan, and Christian Stefansen. Advances in nowcasting influenza-like illness rates using search query logs. *Scientific Reports*, 5:12760, August 2015. ISSN 2045-2322. doi: 10.1038/srep12760. URL `https://www.nature.com/articles/srep12760`.

[123] Emily H. Chan, Vikram Sahai, Corrie Conrad, and John S. Brownstein. Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLOS Neglected Tropical Diseases*, 5(5):e1206, May 2011. ISSN 1935-2735. doi: 10.1371/journal.pntd.0001206. URL `https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0001206`.

[124] Michael J. McCarthy. Internet monitoring of suicide risk in the population. *Journal of Affective Disorders*, 122(3):277–279, May 2010. ISSN 0165-0327. doi: 10.1016/j.jad.2009.08.015. URL `http://www.sciencedirect.com/science/article/pii/S0165032709003978`.

[125] Andrew Page, Shu-Sen Chang, and David Gunnell. Surveillance of Australian suicidal behaviour using the Internet? *Australian & New Zealand Journal of Psychiatry*, 45(12):1020–1022, December 2011. ISSN 0004-8674. doi: 10.3109/00048674.2011.623660. URL `https://doi.org/10.3109/00048674.2011.623660`.

[126] Hajime Sueki. Does the volume of Internet searches using suicide-related search terms influence the suicide death rate: Data from 2004 to 2009 in Japan. *Psychiatry and Clinical Neurosciences*, 65(4):392–394, 2011. ISSN 1440-1819. doi: 10.1111/j.1440-1819.2011.02216.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1440-1819.2011.02216.x`.

[127] Albert C. Yang, Shi-Jen Tsai, Norden E. Huang, and Chung-Kang Peng. Association of Internet search trends with suicide death in Taipei City, Taiwan, 2004–2009. *Journal of Affective Disorders*, 132(1):179–184, July 2011. ISSN 0165-0327. doi: 10.1016/j.jad.2011.01.019. URL `http://www.sciencedirect.com/science/article/pii/S0165032711000528`.

[128] Akihito Hagihara, Shogo Miyazaki, and Takeru Abe. Internet suicide searches and the incidence of suicide in young people in Japan. *European Archives of Psychiatry and Clinical Neuroscience*, 262(1):39–46, February 2012. ISSN 1433-8491. doi: 10.1007/s00406-011-0212-8. URL `https://doi.org/10.1007/s00406-011-0212-8`.

[129] Ladislav Kristoufek, Helen Susannah Moat, and Tobias Preis. Estimating suicide occurrence statistics using Google Trends. *EPJ Data Science*, 5(1):32, November 2016. ISSN 2193-1127. doi: 10.1140/epjds/s13688-016-0094-0. URL `https://doi.org/10.1140/epjds/s13688-016-0094-0`.

[130] Michael Ettredge, John Gerdes, and Gilbert Karuga. Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, 48 (11):87–92, November 2005. ISSN 0001-0782. doi: 10.1145/1096000.1096010. URL `http://doi.acm.org/10.1145/1096000.1096010`.

[131] Hyunyoung Choi and Hal Varian. Predicting initial claims for unemployment benefits. Technical Report, Google, 2009. URL `https://static.googleusercontent.com/media/research.google.com/en//archive/papers/initialclaimsUS.pdf`.

[132] Jaroslav Pavlicek and Ladislav Kristoufek. Nowcasting unemployment rates with Google searches: Evidence from the Visegrad group countries. *PLOS ONE*, 10(5):e0127084, May 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0127084. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0127084`.

[133] Wikipedia. Wikipedia:About, December 2018. URL `https://en.wikipedia.org/w/index.php?title=Wikipedia:About&oldid=875800275`. Page Version ID: 875800275.

[134] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc. URL `http://dl.acm.org/citation.cfm?id=1625275.1625535`.

[135] Michael Strube and Simone Paolo Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pages 1419–1424, Boston, Massachusetts, 2006. AAAI Press. ISBN 978-1-57735-281-5. URL `http://dl.acm.org/citation.cfm?id=1597348.1597414`.

[136] Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. Semantic Wikipedia. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 585–594, New York, NY, USA, 2006. ACM. ISBN 978-1-59593-323-2. doi: 10.1145/1135777.1135863. URL `http://doi.acm.org/10.1145/1135777.1135863`.

[137] David Milne and Ian H. Witten. Learning to link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*,

CIKM '08, pages 509–518, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458150. URL http://doi.acm.org/10.1145/1458082.1458150.

[138] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. Clustering short texts using Wikipedia. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 787–788, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277909. URL http://doi.acm.org/10.1145/1277741.1277909.

[139] Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, 2007. Association for Computational Linguistics. URL http://aclweb.org/anthology/D07-1074.

[140] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A large ontology from Wikipedia and WordNet. *Journal of Web Semantics*, 6(3):203–217, September 2008. ISSN 1570-8268. doi: 10.1016/j.websem.2008.06.001. URL http://www.sciencedirect.com/science/article/pii/S1570826808000437.

[141] Simone Paolo Ponzetto and Michael Strube. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, pages 1440–1445, Vancouver, British Columbia, Canada, 2007. AAAI Press. ISBN 978-1-57735-323-2. URL http://dl.acm.org/citation.cfm?id=1619797.1619876.

[142] Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754, September 2009. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2009.05.004. URL http://www.sciencedirect.com/science/article/pii/S1071581909000561.

[143] Taha Yasseri, Robert Sumi, and János Kertész. Circadian patterns of Wikipedia editorial activity: A demographic analysis. *PLOS ONE*, 7(1):e30091, January 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0030091. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0030091.

[144] Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. Dynamics of conflicts in Wikipedia. *PLOS ONE*, 7(6):e38869, June 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0038869. URL `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0038869`.

[145] Róbert Sumi, Taha Yasseri, András Rung, András Kornai, and János Kertész. Characterization and prediction of Wikipedia edit wars. In *Proceedings of the ACM WebSci'11*, page 58, Koblenz, Germany, 2011. URL `http://eprints.sztaki.hu/8233/`.

[146] R. Sumi, T. Yasseri, A. Rung, A. Kornai, and J. Kertesz. Edit wars in Wikipedia. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 724–727, Massachusetts, Boston, USA, October 2011. doi: 10.1109/PASSAT/SocialCom.2011.47.

[147] Gerardo Iñiguez, János Török, Taha Yasseri, Kimmo Kaski, and János Kertész. Modeling social dynamics in a collaborative environment. *EPJ Data Science*, 3 (1):7, September 2014. ISSN 2193-1127. doi: 10.1140/epjds/s13688-014-0007-z. URL `https://doi.org/10.1140/epjds/s13688-014-0007-z`.

[148] János Török, Gerardo Iñiguez, Taha Yasseri, Maxi San Miguel, Kimmo Kaski, and János Kertész. Opinions, conflicts, and consensus: Modeling social dynamics in a collaborative environment. *Physical Review Letters*, 110(8):088701, February 2013. doi: 10.1103/PhysRevLett.110.088701. URL `https://link.aps.org/doi/10.1103/PhysRevLett.110.088701`.

[149] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis. Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports*, 3:1801, May 2013. ISSN 2045-2322. doi: 10.1038/srep01801. URL `https://www.nature.com/articles/srep01801`.

[150] Márton Mestyán, Taha Yasseri, and János Kertész. Early prediction of movie box office success based on Wikipedia activity big data. *PLOS ONE*, 8(8):e71226, August 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0071226. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0071226`.

[151] Miles Osborne, Saša Petrović, Richard Mccreadie, Craig Macdonald, and Iadh Ounis. Bieber no more: First story detection using Twitter and Wikipedia.

In *In SIGIR 2012 Workshop on Time-Aware Information Access*, Portland, Oregon, USA, 2012.

[152] Mihai Georgescu, Nattiya Kanhabua, Daniel Krause, Wolfgang Nejdl, and Stefan Siersdorfer. Extracting event-related information from article updates in Wikipedia. In Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 254–266. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-36973-5.

[153] Robert M. Kaplan, David A. Chambers, and Russell E. Glasgow. Big data and large sample size: A cautionary note on the potential for bias. *Clinical and Translational Science*, 7(4):342–346, 2014. doi: 10.1111/cts.12178. URL `https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1111/cts.12178`.

[154] Adrian Letchford, Tobias Preis, and Helen Susannah Moat. Quantifying the search behaviour of different demographics using Google Correlate. *PLOS ONE*, 11(2):1–11, February 2016. doi: 10.1371/journal.pone.0149025. URL `https://doi.org/10.1371/journal.pone.0149025`.

[155] Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. Quantifying the impact of scenic environments on health. *Scientific Reports*, 5 (1):16899, November 2015. ISSN 2045-2322. doi: 10.1038/srep16899. URL `https://doi.org/10.1038/srep16899`.

[156] Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. Using deep learning to quantify the beauty of outdoor places. *Royal Society Open Science*, 4(7):170170, 2017. doi: 10.1098/rsos.170170. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rsos.170170`.

[157] Chanuki Illushka Seresinhe, Helen Susannah Moat, and Tobias Preis. Quantifying scenic areas using crowdsourced data. *Environment and Planning B: Urban Analytics and City Science*, 45(3):567–582, 2018. doi: 10.1177/0265813516687302. URL `https://doi.org/10.1177/0265813516687302`.

[158] Chanuki Illushka Seresinhe, Tobias Preis, George MacKerron, and Helen Susannah Moat. Happiness is greater in more scenic locations. *Scientific Reports*, 9(1):4498, March 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-40854-6. URL `https://doi.org/10.1038/s41598-019-40854-6`.

[159] ONS. About the census - Office for National Statistics, n.d.. URL https://www.ons.gov.uk/census/censustransformationprogramme/aboutthecensus.

[160] ONS. Language in England and Wales - Office for National Statistics, n.d.. URL https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/language/articles/languageinenglandandwales/2013-03-04.

[161] Jonathan Athow. Working 9 to 5? – How we count unemployment and what the numbers show National Statistical, November 2018. URL https://blog.ons.gov.uk/2018/11/12/working-9-to-5-how-we-count-unemployment-and-what-the-numbers-show/.

[162] ONS. UK labour market - Office for National Statistics, n.d.. URL https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/uklabourmarket/december2018.

[163] International Labour Organization. Resolution concerning statistics of work, employment and labour underutilization, November 2013. URL http://www.ilo.org/global/statistics-and-databases/standards-and-guidelines/resolutions-adopted-by-international-conferences-of-labour-statisticians/WCMS_230304/lang--en/index.htm.

[164] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Statistical learning. In Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, editors, *An Introduction to Statistical Learning: With Applications in R*, Springer Texts in Statistics, pages 15–57. Springer New York, New York, NY, 2013. ISBN 978-1-4614-7138-7. doi: 10.1007/978-1-4614-7138-7_2. URL https://doi.org/10.1007/978-1-4614-7138-7_2.

[165] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Linear regression. In Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, editors, *An Introduction to Statistical Learning: With Applications in R*, Springer Texts in Statistics, pages 59–126. Springer New York, New York, NY, 2013. ISBN 978-1-4614-7138-7. doi: 10.1007/978-1-4614-7138-7_3. URL https://doi.org/10.1007/978-1-4614-7138-7_3.

[166] Bodo Winter and Martijn Wieling. How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling.

Journal of Language Evolution, 1(1):7–18, February 2016. ISSN 2058-4571. doi: 10.1093/jole/lzv003. URL `https://doi.org/10.1093/jole/lzv003`.

[167] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, October 2006. ISSN 01692070. doi: 10.1016/j.ijforecast.2006.03.001. URL `https://linkinghub.elsevier.com/retrieve/pii/S0169207006000239`.

[168] Rob J. Hyndman. *3.4 Evaluating Forecast Accuracy | Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 2nd edition edition, 2018. URL `https://Otexts.com/fpp2/`.

[169] Chris Tofallis. A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, 66(8):1352–1362, August 2015. ISSN 1476-9360. doi: 10.1057/jors.2014.103. URL `https://doi.org/10.1057/jors.2014.103`.

[170] David A. Swanson, Jeff Tayman, and Charles F. Barr. A note on the measurement of accuracy for subnational demographic estimates. *Demography*, 37(2):193–201, May 2000. ISSN 1533-7790. doi: 10.2307/2648121. URL `https://doi.org/10.2307/2648121`.

[171] Spyros Makridakis. Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, 9(4):527–529, December 1993. ISSN 0169-2070. doi: 10.1016/0169-2070(93)90079-3. URL `http://www.sciencedirect.com/science/article/pii/0169207093900793`.

[172] Cort J. Willmott and Kenji Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1):79–82, 2005. ISSN 0936577X, 16161572. URL `www.jstor.org/stable/24869236`.

[173] Cort J. Willmott, Kenji Matsuura, and Scott M. Robeson. Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment*, 43(3):749–752, January 2009. ISSN 1352-2310. doi: 10.1016/j.atmosenv.2008.10.005. URL `http://www.sciencedirect.com/science/article/pii/S1352231008009564`.

[174] Neil J Salkind. *Encyclopedia of Research Design*. SAGE Publications, Inc., Thousand Oaks, CA, 2010. doi: 10.4135/9781412961288. URL `http://sk.sagepub.com/reference/researchdesign`.

[175] Rob J. Hyndman. Errors on percentage errors | Rob J Hyndman, 2014. URL `https://robjhyndman.com/hyndsight/smape/`.

[176] J. Scott Armstrong. Evaluating forecasting methods. In J. Scott Armstrong, editor, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pages 443–472. Springer US, Boston, MA, 2001. ISBN 978-0-306-47630-3. doi: 10.1007/978-0-306-47630-3_20. URL `https://doi.org/10.1007/978-0-306-47630-3_20`.

[177] T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250, June 2014. ISSN 1991-9603. doi: 10.5194/gmd-7-1247-2014. URL `https://www.geosci-model-dev.net/7/1247/2014/`.

[178] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Linear model selection and regularization. In *An Introduction to Statistical Learning: With Applications in R*, pages 203–264. Springer New York, New York, NY, 2013. ISBN 978-1-4614-7138-7. doi: 10.1007/978-1-4614-7138-7_6. URL `https://doi.org/10.1007/978-1-4614-7138-7_6`.

[179] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, February 1970. ISSN 0040-1706. doi: 10.1080/00401706.1970.10488634. URL `https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634`.

[180] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, January 1996. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL `https://doi.org/10.1111/j.2517-6161.1996.tb02080.x`.

[181] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x. URL `https://doi.org/10.1111/j.1467-9868.2005.00503.x`.

[182] Wenjiang J. Fu. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, September 1998. ISSN 1061-8600. doi: 10.1080/10618600.1998.10474784. URL `https://amstat.tandfonline.com/doi/abs/10.1080/10618600.1998.10474784`.

[183] Christoph Bergmeir and José M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, May 2012. ISSN 0020-0255. doi: 10.1016/j.ins.2011.12.028. URL `http://www.sciencedirect.com/science/article/pii/S0020025511006773`.

[184] Rob J. Hyndman. Time series cross-validation: An R example | Rob J Hyndman, n.d. URL `https://robjhyndman.com/hyndsight/tscvexample/`.

[185] G. King. Ensuring the data-rich future of the social sciences. *Science*, 331(6018):719–721, February 2011. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1197872. URL `http://www.sciencemag.org/cgi/doi/10.1126/science.1197872`.

[186] Alessandro Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428, July 2009. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1171990. URL `http://science.sciencemag.org/content/325/5939/425`.

[187] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323(5915): 721–723, February 2009. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1167742. URL `http://science.sciencemag.org/content/323/5915/721`.

[188] Helen Susannah Moat, Tobias Preis, Christopher Y. Olivola, Chengwei Liu, and Nick Chater. Using big data to predict collective behavior in the real world. *Behavioral and Brain Sciences*, 37(1):92–93, February 2014. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X13001817. URL `https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/using-big-data-to-predict-collective-behavior-in-the-real-world1/419945C852EAC07BB49C5A943EC7E5BE`.

[189] Federico Botta, Helen Susannah Moat, and Tobias Preis. Quantifying crowd size with mobile phone and Twitter data. *Royal Society Open Science*, 2(5): 150162, May 2015. ISSN 2054-5703. doi: 10.1098/rsos.150162. URL `http://rsos.royalsocietypublishing.org/content/2/5/150162`.

[190] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of*

the 19th International Conference on World Wide Web, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772777. URL http://doi.acm.org/10.1145/1772690.1772777.

[191] Ilaria Bordino, Stefano Battiston, Guido Caldarelli, Matthieu Cristelli, Antti Ukkonen, and Ingmar Weber. Web search queries can predict stock market volumes. *PLOS ONE*, 7(7):e40014, July 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0040014. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0040014.

[192] Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts. Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences*, 107(41):17486–17490, October 2010. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1005962107. URL http://www.pnas.org/content/107/41/17486.

[193] Liang Liu, Bo Qu, Bin Chen, Alan Hanjalic, and Huijuan Wang. Modelling of information diffusion on social networks with applications to WeChat. *Physica A: Statistical Mechanics and its Applications*, 496:318–329, April 2018. ISSN 0378-4371. doi: 10.1016/j.physa.2017.12.026. URL http://www.sciencedirect.com/science/article/pii/S0378437117312785.

[194] Nicoló Musmeci, Tomaso Aste, and T. Di Matteo. Relation between financial market structure and the real economy: Comparison between clustering methods. *PLOS ONE*, 10(3):e0116201, March 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0116201. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0116201.

[195] Avery Hartmans and Rob Price Insider, Business. Instagram just reached 1 billion users, 2018. URL http://uk.businessinsider.com/instagram-monthly-active-users-1-billion-2018-6.

[196] US Census Bureau. Detailed Languages Spoken at Home and Ability to Speak English, n.d. URL https://www.census.gov/data/tables/2013/demo/2009-2013-lang-tables.html.

[197] ONS. Output Area (OA), December 2011. URL https://webarchive.nationalarchives.gov.uk/20160107193025/http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/output-area--oas-/index.html.

[198] ONS. Super Output Area (SOA), December 2011. URL http://webarchive.nationalarchives.gov.uk/20160106001702/http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/super-output-areas--soas-/index.html.

[199] ABS. 2016 Census QuickStats: Greater Sydney, 2017. URL https://quickstats.censusdata.abs.gov.au/census_services/getproduct/census/2016/quickstat/1GSYD?opendocument.

[200] London Councils. Population and census | London Councils, n.d. URL https://www.londoncouncils.gov.uk/our-key-themes/local-government-finance/population-and-census.

[201] Elsa Arcaute, Carlos Molinero, Erez Hatna, Roberto Murcio, Camilo Vargas-Ruiz, A. Paolo Masucci, and Michael Batty. Cities and regions in Britain through hierarchical percolation. *Royal Society Open Science*, 3(4):150691, April 2016. ISSN 2054-5703. doi: 10.1098/rsos.150691. URL https://doi.org/10.1098/rsos.150691.

[202] Elsa Arcaute, Erez Hatna, Peter Ferguson, Hyejin Youn, Anders Johansson, and Michael Batty. Constructing cities, deconstructing scaling laws. *Journal of The Royal Society Interface*, 12(102):20140745, January 2015. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2014.0745. URL http://rsif.royalsocietypublishing.org/content/12/102/20140745.

[203] Clémentine Cottineau, Erez Hatna, Elsa Arcaute, and Michael Batty. Diverse cities or the systematic paradox of Urban Scaling Laws. *Computers, Environment and Urban Systems*, 63:80–94, May 2017. ISSN 0198-9715. doi: 10.1016/j.compenvurbsys.2016.04.006. URL http://www.sciencedirect.com/science/article/pii/S0198971516300448.

[204] ONS. Data viewer - Nomis - Official labour market statistics, n.d.. URL https://www.nomisweb.co.uk/census/2011/QS204EW/view/2013265927?rows=cell&cols=rural_urban.

[205] Nico J D Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692, 1991. ISSN 0006-3444. doi: 10.1093/biomet/78.3.691. URL https://doi.org/10.1093/biomet/78.3.691.

[206] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Linear model selection and regularization. In *An Introduction to Statistical Learning*,

Springer Texts in Statistics, pages 203–264. Springer, New York, NY, 2013. ISBN 978-1-4614-7137-0 978-1-4614-7138-7. doi: 10.1007/978-1-4614-7138-7_6. URL https://link.springer.com/chapter/10.1007/978-1-4614-7138-7_6.

[207] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Resampling methods. In *An Introduction to Statistical Learning*, Springer Texts in Statistics, pages 175–201. Springer, New York, NY, 2013. ISBN 978-1-4614-7137-0 978-1-4614-7138-7. doi: 10.1007/978-1-4614-7138-7_5. URL https://link.springer.com/chapter/10.1007/978-1-4614-7138-7_5.

[208] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software; Vol 1, Issue 1 (2015)*, 67(1):1–48, October 2015. ISSN 1548-7660. doi: 10.18637/jss.v067.i01. URL https://www.jstatsoft.org/v067/i01.

[209] Zhi-Qiang Jiang, Wen-Jie Xie, Ming-Xia Li, Boris Podobnik, Wei-Xing Zhou, and H. Eugene Stanley. Calling patterns in human communication dynamics. *Proceedings of the National Academy of Sciences*, 110(5):1600–1605, January 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1220433110. URL http://www.pnas.org/content/110/5/1600.

[210] Leonidas Sandoval and Italo De Paula Franca. Correlation of financial markets in times of crisis. *Physica A: Statistical Mechanics and its Applications*, 391(1):187–208, January 2012. ISSN 0378-4371. doi: 10.1016/j.physa.2011.07.023. URL http://www.sciencedirect.com/science/article/pii/S037843711100570X.

[211] Marco A. Javarone, Roberto Interdonato, and Andrea Tagarelli. Modeling evolutionary dynamics of lurking in social networks. In Hocine Cherifi, Bruno Gonçalves, Ronaldo Menezes, and Roberta Sinatra, editors, *Complex Networks VII: Proceedings of the 7th Workshop on Complex Networks CompleNet 2016*, Studies in Computational Intelligence, pages 227–239. Springer International Publishing, Cham, 2016. ISBN 978-3-319-30569-1. doi: 10.1007/978-3-319-30569-1_17. URL https://doi.org/10.1007/978-3-319-30569-1_17.

[212] R. Conte, N. Gilbert, G. Bonelli, C. Cioffi-Revilla, G. Deffuant, J. Kertesz, V. Loreto, S. Moat, J. P. Nadal, A. Sanchez, A. Nowak, A. Flache, M. San Miguel, and D. Helbing. Manifesto of computational social science. *The European Physical Journal Special Topics*, 214(1):325–346, Novem-

ber 2012. ISSN 1951-6401. doi: 10.1140/epjst/e2012-01697-8. URL https://doi.org/10.1140/epjst/e2012-01697-8.

[213] Nikolaos Askitas. Predicting road conditions with Internet search. *PLOS ONE*, 11(8):e0162080, August 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0162080. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0162080.

[214] Helen Susannah Moat, Christopher Y. Olivola, Nick Chater, and Tobias Preis. Searching choices: Quantifying decision-making processes using search engine data. *Topics in Cognitive Science*, 8(3):685–696, July 2016. ISSN 1756-8757. doi: 10.1111/tops.12207. URL https://doi.org/10.1111/tops.12207.

[215] Tobias Preis and Helen Susannah Moat. Early signs of financial market moves reflected by Google searches. In Bruno Gonçalves and Nicola Perra, editors, *Social Phenomena: From Data Analysis to Models*, pages 85–97. Springer International Publishing, Cham, Switzerland, 2015. ISBN 978-3-319-14011-7. doi: 10.1007/978-3-319-14011-7_5. URL https://doi.org/10.1007/978-3-319-14011-7_5.

[216] Chester Curme, Tobias Preis, H. Eugene Stanley, and Helen Susannah Moat. Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences*, 111(32):11600–11605, August 2014. doi: 10.1073/pnas.1324054111. URL http://www.pnas.org/content/111/32/11600.abstract.

[217] Takao Noguchi, Neil Stewart, Christopher Y. Olivola, Helen Susannah Moat, and Tobias Preis. Characterizing the time-perspective of nations with search engine query data. *PLOS ONE*, 9(4):e95209, April 2014. doi: 10.1371/journal.pone.0095209. URL https://doi.org/10.1371/journal.pone.0095209.

[218] Scott A. Hale. Multilinguals and Wikipedia editing. In *Proceedings of the 2014 ACM Conference on Web Science - WebSci '14*, pages 99–108, Bloomington, Indiana, USA, 2014. ACM Press. ISBN 978-1-4503-2622-3. doi: 10.1145/2615569.2615684. URL http://dl.acm.org/citation.cfm?doid=2615569.2615684.

[219] Zeyu Zheng, Huancheng Yang, Yang Fu, Dianzheng Fu, Boris Podobnik, and H. Eugene Stanley. Factors influencing message dissemination through social media. *Physical Review E*, 97(6):062306, June 2018. doi: 10.1103/PhysRevE.97.062306. URL https://link.aps.org/doi/10.1103/PhysRevE.97.062306.

[220] Scott A. Hale. Global connectivity and multilinguals in the Twitter network. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14*, pages 833–842, Toronto, Ontario, Canada, 2014. ACM Press. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557203. URL `http://dl.acm.org/citation.cfm?doid=2556288.2557203`.

[221] StatCounter. Search engine market share United Kingdom, n.d.. URL `http://gs.statcounter.com/search-engine-market-share/all/united-kingdom/2016`.

[222] ONS. Unemployment by age and duration (not seasonally adjusted): UNEM01 NSA - Office for National Statistics, n.d.. URL `https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/unemploymentbyageanddurationnotseasonallyadjustedunem01nsa/current`.