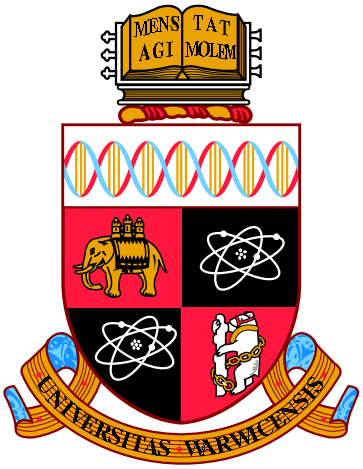

Faster socieoeconomic indicators using novel data sources

by

Sam Miller

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy in

Business and Management

Warwick Business School

June 2022

Contents

Acknowledgments

I first thank my supervisors, Suzy Moat and Tobias Preis, for supporting me every step of this journey. In particular, I appreciate their tolerance of my slightly unorthodox approach to tackling a PhD. I also thank my friends who helped me through what have been both the best and worst years of my life. You know who you are. Finally, and most importantly, I thank my mother and father. They would both have been so proud to see this work published.

This thesis was funded by The Alan Turing Institute, via EPSRC grant EP/N510129/1, and the Office for National Statistics Data Science Campus. The Alan Turing Institute provided cloud computing resources through a Microsoft Azure for Research Award.

Declarations

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

Parts of this thesis have been published, as detailed below. In all cases, I am the first author and primary contributor:

1. Results from Chapter 3 were published in May 2020 in *Scientific Reports* [Miller et al., 2020b].

2. Results from Chapter 5 were published in April 2020 in *The Proceedings of the Web Conference 2020* [Miller et al., 2020a].

3. Results from Chapter 6 were published in June 2022 in *PLOS Neglected Tropical Diseases* [Miller et al., 2022].

Abstract

Policymakers require up-to-date statistics to make good decisions. Most official statistics in economics and public health are released only after a significant delay. The goal of "nowcasting" (a combination of the words "now" and "forecasting") is to estimate these statistics before their official release. Better nowcasts would help policymakers respond to rapidly developing crises in a range of domains. The recent Covid-19 crisis has highlighted this issue: it is extremely challenging to make decisions in crisis without knowledge of either the current state of the economy or the incidence of disease in the population.

Recent technological advances mean we now generate real-time data simply by going about our lives. This thesis shows how we can use novel data sources to improve nowcasts in economics and public health. We highlight how we are no longer constrained to traditional data sources, such as surveys.

We first investigate whether high-frequency aircraft location data can generate faster GDP estimates. We also show that this dataset can help improve estimates of airport performance, particularly at the onset of the Covid-19 crisis. We next use a novel combination of Wikipedia page views and data scraped from online "darknet" drug markets to nowcast illicit drug demand. Better statistics on drug markets would be highly valuable to policymakers in both economics and public health. Finally, we use data from Google Trends to nowcast the incidence of chikungunya in Rio de Janeiro. Official disease data is delivered with long and variable delays. We show that including real-time Google Trends data allows earlier detection of epidemics.

This thesis finds evidence that novel data sources can improve the speed and accuracy of official statistics in a range of domains. As the variety of novel data sources keeps growing, these may give policymakers more complete, real-time information when making crucial decisions.

Chapter 1

Introduction

Policymakers require fast and accurate statistics to make effective decisions across a range of domains. Currently, there is usually a significant delay with releasing official statistics for both economics and public health. For example: the first official publication of UK GDP for March 2020, the beginning of the Covid-19 crisis, was not available until two months later in May 2020. Accurate statistics for public health can take even longer to publish. In Brazil, policymakers tracking dengue fever often still do not have complete data even after three months. Given the potential lag between important policy decisions, such as economic stimulus, and their impact on society, it is important for policymakers to act as quickly as possible in response to crises. Detecting these crises quickly requires up-to-date information on the world.

"Nowcasting" (a combination of the terms "now" and "forecasting") refers to the practice of estimating the current value of statistics before their official release. There is growing recognition among policymakers of the importance of better nowcasts. The UK Office for National Statistics (ONS) now has a program dedicated to providing faster indicators of economic activity.[1] Similarly, both the Bank of England[2] and Federal Reserve[3] have teams for estimating the current state of macroeconomic variables such as Gross Domestic Product (GDP) and inflation. In public health, the US Center for Disease Control (CDC) has invested in rapid estimates of flu incidence at weekly frequency[4]. The Covid-19 crisis has underlined the importance of better nowcasting in both economics and public health: policymakers are often making critical decisions, such as economic stimulus or lockdown timing, without accurate knowledge of either the current state of the economy or the spread

[1] https://www.ons.gov.uk/economy/economicoutputandproductivity/output/bulletins/fasterindicatorsofeconomicactivityuk/february2020

[2] https://www.bankofengland.co.uk/quarterly-bulletin/2018/2018-q3/

[3] https://www.frbatlanta.org/cqer/research/gdpnow

[4] https://www.cdc.gov/flu/weekly/index.htm

of Covid-19 among the population.

The new field of computational social science holds much promise for improving nowcasting. We now generate far more data than ever before simply by going about our lives. Activities ranging from searching on the internet to travelling on aircraft leave streams of data behind that may have value for nowcasting. Simultaneously, researchers have developed computational techniques for handling this rapidly growing volume and variety of new data sources. For example, our high-frequency aircraft location data, used in Chapters 3 and 4 of this thesis, contains over 30 billion observations. Without modern computational power and techniques, such as MapReduce, it would be infeasible to analyse data at such scale. New techniques from machine learning for handling unstructured data, such as text and images, allow for analysis of a greater variety of data sources than ever before. This is particularly relevant for nowcasting, as traditional data capture methods such as surveys rarely yield data in real-time.

In Chapter 2, we first review recent developments in computational social science that have paved the way for a greater variety of data sources to feed into better nowcasts. We then explore nowcasting specifically for economics, where a rich methodological literature has evolved but is somewhat limited in the variety of data sources considered. We next assess nowcasting for public health, where there is a particularly well-developed literature on using internet search data in epidemiology. The chapter finishes with a review of literature using the specific novel data sources featured in this thesis: aircraft location data and illicit drug sales data scraped from darknet markets.

Nowcasting GDP is a well-studied problem in economics. Most economic policy tools, such as interest rates, only fully impact the economy after a significant delay. Recessions, such as the 2008 financial crisis, usually occur quickly and without prior warning. Policymakers must therefore respond as quickly as possible to recessions in order to limit their damage. In Chapter 3, we analyse the potential for novel aircraft location data to improve nowcasts of airline performance and GDP. Aviation is a major component of GDP, directly contributing at least 3% in both the UK and USA, alongside supporting other economic sectors. The Covid-19 crisis shows how rapidly the aviation sector can change, so accurate data on its current performance is particularly important. However, current aviation statistics are released with a delay of several months. Aircraft now broadcast their location in real-time at high frequency using the Automated Dependent Surveillance Broadcast (ADS-B) system. We show that ADS-B data can be transformed to accurately estimate airline flight volumes in real-time, which is a crucial indicator of their

performance. We further show that real-time knowledge of airline flight volumes provides a real-time indicator for aviation's contribution to GDP. This indicator is particularly valuable during volatile economic times, such as the 2008 financial crisis and 2012 Eurozone crisis. These results were published in Nature Publishing Group's *Scientific Reports* [Miller et al., 2020b].

In Chapter 4, we extend the application of ADS-B data to nowcasting airport performance. Flight volumes are a key determinant of airport performance. We exploit the geographic dimension of ADS-B data to show that it can significantly improve nowcasts of airport flight volumes. Chapter 4 also extends the time series of ADS-B data to April 2020, which allows us to show that ADS-B data is particularly valuable at the start of the Covid-19 crisis. Future economic applications of this data could include nowcasting the international trade network. UK and US authorities currently publish granular data on trade flows through airports, but its publication is delayed by at least three months.

Black markets, such as illicit drugs, have always been difficult for economists to monitor. Currently, policymakers mostly rely on surveys that are infrequent - sometimes as low as annual frequency - and prone to response bias. This has also been problematic for public health policymakers, who sometimes fail to detect rapid changes in drug use, such as during the US Fentanyl crisis, until they have become a fully fledged epidemic. In Chapter 5, we explore the potential for a novel combination of drug sales data scraped from online markets and Wikipedia page views to help nowcast illicit drug demand. Owing to recent developments in cryptocurrencies and web browsing, there are now online "darknet" markets that enable anonymous trade in illicit goods. We collect a transaction-level dataset of illicit drug trades by scraping the buyer feedback on these markets, which is available in real-time. We then show that we can construct a global, high frequency measure of illicit drug demand from the transaction-level data. However, darknet markets are difficult to scrape reliably so we cannot always assume that data will be available from them when needed. Therefore, we further show that Wikipedia page views for each drug can accurately nowcast darknet demand for that drug. We provide evidence that darknet demand is generally representative of broader demand. Monitoring Wikipedia page views could therefore allow early detection of rapid changes in drug use before they become public health crises. These results were published in *Proceedings of The Web Conference 2020* [Miller et al., 2020a].

A more common application of nowcasting in public health is in epidemiology. Accurate, real-time data on disease incidence is critical for policymakers to limit, or respond effectively to, the spread of disease. Chikungunya, a mosquito-

borne disease or *arbovirus*, is a growing public health problem in Brazil. There were over 130,000 cases in 2019 alone. Chikungunya can lead to complications such as paralysis and workplace absence. These complications are catastrophic for the low-income households usually affected by chikungunya. In Chapter 6, we analyse whether internet search data from Google Trends can improve nowcasts of chikungunya incidence in Rio de Janeiro. Current statistics on chikungunya case counts are prone to long delays, as there is a long and varied lag between a patient reporting symptoms and entry of the case into the official monitoring system. However, data on Google searches for chikungunya-related terms is reliably available in real-time. We build on a Bayesian approach designed for data that is subject to long and varied delays. This differs from models used in previous chapters that require regular and complete data on previous periods, and are therefore unsuitable for nowcasting chikungunya. We find that including Google search data improves both nowcast accuracy and precision. Furthermore, these improvements are largest during epidemics, which are particularly important periods for policymakers to have accurate current data. Including Google search data in surveillance systems may therefore help policymakers respond more quickly to future chikungunya epidemics. These results are due to be published in *PLOS Neglected Tropical Diseases* [Miller et al., 2022].

In the final chapter, we review our main results, their limitations and opportunities for future research to build on this thesis. We conclude that there is great potential for new data sources to improve nowcasts in both economics and public health. However, we emphasise the need for caution and a gradual approach when using novel data sources. In all chapters, we include novel data sources as a complement to a baseline model based on traditional methods, rather than replacing traditional methods entirely. We then analyse the marginal predictive value of novel data sources. When deployed in this way, we show that novel data sources can be a powerful tool to augment existing nowcasting methods. In turn, they may help improve the speed and accuracy of official statistics.

Chapter 2

Literature review

2.1 Introduction

Fast and accurate statistics are important for policymakers. First used in economics [Giannone et al., 2008], nowcast combines the words "now" and "forecast" to refer to the estimation of the current value of statistics that are only officially released after a lag. Without fast and accurate statistics, it is much more difficult to make critical decisions in many areas of policy. In the midst of a recession, economic policymakers deciding how large to make an economic stimulus package would benefit greatly from precise estimates of the recession's depth. Similarly, governments considering a lockdown in response to Covid-19 would benefit from up-to-date knowledge of the current number of infections. While nowcasting Covid-19 infections is out of scope for this thesis, the crisis has underlined the importance of fast and accurate statistics. This thesis focuses on novel data and methods that could provide such statistics in two key policy areas: economics and public health.

This chapter reviews existing literature on nowcasting. We first review the developments in the field of computational social science. We focus particularly on the use of "unstructured" data, such as images and text, which initially come in a non-tabular format. We document how new data and methods have recently enabled large improvements in nowcasting, which could be very valuable for policy.

We then review the development of nowcasting in economics. Much of the literature in economics has focused on improving the econometric methodology for incorporating higher frequency economic data into nowcasting low frequency official statistics. For example: Giannone et al. [2008] show how incorporating monthly indicators, such as purchasing manager surveys, improves nowcasts of quarterly GDP releases in the USA. Generally, this literature has taken the breadth of available

data sources as a given, rather than trying to engineer new data sources. There is now some body of research into including real-time, online indicators, and growing awareness of potential for using broader data sources in policy. For example, the UK Office for National Statistics now has a dedicated "faster indicators" program for improving the speed of its economic statistics [Nolan, 2019]. However, studies of the inclusion of a wider range of novel data sources and use of machine learning techniques are less well developed. This justifies the focus of this thesis on expanding the sources of data available for nowcasting.

Next, we review developments in nowcasting for public health. There has been little research so far into nowcasting drug demand, where faster statistics would benefit both economics and public health policymakers. This is partially due to a lack of high-frequency drug demand data so far, which is why we highlight darknet markets as an important source of data in this thesis. The literature on nowcasting in epidemiology is much better developed; there have been many studies using online search data since the seminal paper on using Google Trends to nowcast the flu, known as Google Flu Trends, in 2009 [Ginsberg et al., 2009]. However, nowcasting the spread of disease has proven difficult – the original Google Flu Trends method led to some high-profile overestimates of flu incidence in later flu seasons [Lazer et al., 2014]. Much of the literature so far has been restricted to documenting a correlation between a data source and the spread of a disease, rather than a fuller out-of-sample evaluation of whether a new data source is useful for nowcasting. We evaluate all models in this thesis out-of-sample against a relevant baseline. We also retrain our models across time as new data becomes available, which was a key missing component in the original Google Flu Trends study [Preis and Moat, 2014]. Most nowcasting models also assume complete, regular data on disease prevalence in previous time periods, which is often not the case when nowcasting diseases in practice. In Chapter 6 of this thesis, we highlight that our proposed models are always operationally feasible in that they use only data available at the time of the nowcast, without needing to aggregate to long time periods such as months.

The final section reviews the nascent literature using aircraft location data and data from darknet markets, which are the two novel data sources contributed by this thesis. There have been some applications of darknet data in economics, but none so far in public health or nowcasting of statistics from any field. Similarly, location data from vehicles has rarely been used in economics, with only some preliminary studies documenting possible uses as an early indicator for economic variables. Aircraft location data in particular has barely been used outside at all outside its originally intended application in aviation. This thesis therefore marks

the first use of both data sources in nowcasting, while paying heed to methodological developments in the literature to ensure rigorous evaluation of their nowcasting value.

2.2 Recent advances in computational social science

2.2.1 Use of text data

We now generate much more real-time data than ever before. When we search for information on Google, this leaves a record of our search terms, time of search and location we searched from [Choi and Varian, 2009b]. When we review a product on Amazon, we similarly create a record of our review time, content and evidence of demand for the product [See-To and Ngai, 2018]. When we purchase anything with a credit card, we create a record of the price, time of purchase and account transaction, potentially along with further personal data on ourselves [de Montjoye et al., 2015; Galbraith and Tkacz, 2018]. Even our decisions to travel, whether by road [Rowland, 2019], sea [Bonham et al., 2018] or air [Strohmeier et al., 2018], can now create data from the location tracks left by the vehicles we have travelled in. Unlike traditional data sources such as surveys, much of this new data is generated incidentally – we create it as an unintentional by-product of our regular actions. The new field of "computational social science" focuses on generating insights from this vast volume of incidental data [Conte et al., 2012; Moat et al., 2014]. This section reviews how the development of computational social science has enabled large potential improvements in nowcasting.

Lazer et al. [2009] were the first to identify computational social science as a new academic field. Even before the internet became so deeply embedded in our daily lives, there was a rapidly developing literature on how to exploit the first sources of internet data, such as email. Social network analysis using email, which previously relied on smaller scale survey data, provides some of the earliest examples. Eckmann et al. [2004] are able to infer the social network of an office by its email traffic. Kossinets and Watts [2006] find similar results when analysing the email network of a university, and are able to analyse network robustness at a large scale. These studies were among the first to show the potential for using internet data to analyse problems at a scale not previously feasible.

The computational social science literature began to expand more quickly with high profile studies using search engine data. Ginsberg et al. [2009] show that Google searches for flu symptoms could provide an accurate leading indicator for weekly flu data published by the CDC. This was not the first paper to hypothe-

sise a relationship between internet searches and influenza case volume, Polgreen et al. [2008] show a correlation with Yahoo searches. However, Ginsberg et al. [2009] demonstrate a very concrete policy case for computational social science with publicly available internet data. Despite later controversy over the performance of "Google Flu Trends" [Lazer et al., 2014], there is now a rich literature on using online search data to better model the spread of disease, as well as other public health concerns such as suicide [Kristoufek et al., 2016]. We discuss this literature further in Chapter 6, as the final chapter of this thesis builds on it directly by using Google search data to improve nowcasts of chikungunya incidence.

Researchers very soon started using Google search data in wider applications across both public health and economics. Tkachenko et al. [2017a] show how Google Trends, the volume of Google searches for a given topic, can improve surveillance of type 2 diabetes in the UK. Choi and Varian [2009b] show that Google Trends (the volume of Google searches for a given topic) could help nowcast over several economic indicators, such as consumer confidence, car sales and travel patterns (see also Botta et al. [2020b]). They later extend this analysis to claims for unemployment benefits [Choi and Varian, 2009a]. While Choi and Varian [2009b] focused primarily on the US, other studies confirmed these results held in Israel [Suhoy, 2009], Germany [Askitas and Zimmermann, 2009] and the UK [Chamberlin, 2010]. These analyses quickly attracted interest from policymakers, with the Bank of England publishing its own report on the potential use of internet search data for macroeconomic analysis [McLaren and Shanbhogue, 2011].

Google Trends has also been used widely in a closely related field – finance – for more granular analyses. Preis et al. [2010] show that the volume of Google searches for companies in the S&P 500 is predictive over trading volume for their stocks. Da et al. [2011] similarly find that Google searches are predictive over stock prices for Russell 3000 companies between 2004 and 2008. In a later analysis, Preis et al. [2013b] find that including Google Trends data for key terms in finance in a trading strategy can outperform market returns. The early, and continued, popularity of Google Trends may be due to ease of access, with Google providing a public API. Furthermore, search engine data is relatively easy to engineer into a useful format, without needing novel methods from machine learning (although they may be necessary to help identify the most relevant terms as in Curme et al. [2014]). The final two chapters of this thesis show that search engine data continues to provide useful insights, when analysing both illegal drug market trends and the spread of disease.

Computational social science soon expanded to include other forms of text

data from the internet, such as Twitter. The literature on using Twitter data has benefited greatly from developments in computer science techniques for analysing text, referred to as natural language processing [Manning and Schutze, 1999]. Paul and Dredze [2011] show that the volume of flu-related tweets can help nowcast the CDC's weekly flu estimates. Aramaki et al. [2011] quickly build on this study by showing that extracting the sentiment of the tweet, using techniques from natural language processing, adds further value in nowcasting the flu. Gomide et al. [2011] also show value from adding Twitter sentiment to estimates of dengue fever. Similarly, Bollen et al. [2011] show that analysing the sentiment of large volumes of tweets in the US may be predictive over the level of the Dow Jones Industrial Average (DJIA), with more negative sentiment on Twitter foreshadowing falls in the DJIA. Procter et al. [2013] show the potential for using Twitter data to track the 2011 London riots. However Zubiaga et al. [2018] show the potential for rumours to spread quickly across Twitter, which may reduce its value for nowcasting. Nevertheless, analysis of the actual text in tweets is not always necessary to extract useful information. In a particularly innovative study, Botta et al. [2015] exploit the location metadata from tweets to show that the volume of tweets from a football stadium can help with instant estimates of crowd sizes.

Wikipedia provides another form of text data that can be relatively simple to include in modelling. There is a public API for the daily volume of Wikipedia views for each page [Wikipedia API, 2019], so researchers can collect useful data without needing to analyse the text of the pages themselves. Moat et al. [2013] find a negative correlation between Wikipedia page views for DJIA companies and their share prices. They tie this finding to loss aversion: traders are more likely to search for information when making a decision that would book a loss (see also Moat et al. [2016]). Curme et al. [2014] build on this analysis by using a combination of Google and Wikipedia data to show a link between searches for events in business and politics with subsequent moves in stock markets. In another trading application, El-Bahrawy et al. [2019] show that Wikipedia page views for cryptocurrencies, such as Bitcoin, can be used to improve the profitability of trading strategies. Wikipedia data has also been used to predict economic performance on a more granular level. Mestyan et al. [2013] use Wikipedia data to predict box office performance for newly released films. They find that the daily volume of Wikipedia page views for each film is predictive of box office revenues, over and above traditional indicators such as number of theatres distributing the film, for up to a month before the film's release.

An issue with some of the previous literature is that the value from adding Wikipedia data may not be independent of similar Google data. McMahon et al.

[2017] document that Wikipedia page landings are often the direct result of a Google search, as the Wikipedia page is often the first search result. For example, Google searches for a given film may be highly correlated with views for its Wikipedia page. If so, there may be limited predictive value from adding Wikipedia page values to a model over and above simply adding the Google Trends data for its topic. In Chapter 5 of this thesis, on predicting drug demand with Wikipedia page views, we are mindful of these issues and present analysis using both Wikipedia page views and Google Trends.

An advantage of using Wikipedia views for some topics, relative to Google searches or Twitter data, may be the relative lack of language ambiguity. For example there may be multiple Google searches related to certain diseases, depending on their symptoms, but there is only one Wikipedia page for the disease. Generous et al. [2014] show that Wikipedia views can be used to improve surveillance of a range of diseases. They successfully use the language of each page as a location proxy for the viewer to improve the geographic granularity of their analysis. For example, they can assume that Portuguese language page views for dengue fever are from Brazil, whereas Thai page views are from Thailand. In Chapter 5 of this thesis, we similarly exploit the language of each Wikipedia page to add a geographic dimension to our analysis of drug markets on the darknet.

Computational social science has also been able to extract text data at scale from the migration of traditional print media online. Alanyali et al. [2013] find a positive relationship between daily mentions of a company in the *Financial Times* and its daily stock market trading volumes. This provides novel empirical support that financial news is a primary driver of financial markets. Curme et al. [2017] find that the diversity of news in the *Financial Times* around a company is predictive over trading volumes. Piškorec et al. [2014] find a significant correlation between the cohesiveness of financial news and volatility of national stock indices. Souma et al. [2019] show the that stock price reactions to news can be used to train sentiment analysis models, which can then quantify the sentiment of future news articles. Previous studies that tested hypotheses on the effect of news, such as Chan [2003], had to use manually assembled datasets. Computational social science now allows us to evaluate these issues at a much greater scale. However, many studies do not go further than documenting correlations, which is not always sufficient to show a new data source has predictive value in practice. In this thesis, novel data sources are evaluated in a framework designed to assess practical nowcasting value against appropriate baseline models.

As well as traditional media, researchers have made extensive use of data

from social media platforms. Facebook is the most widely used social network in the world, with 68% of US adults regularly using the platform as of 2018 [Smith and Anderson, 2018]. Much of the research using Facebook data has been in political science. Aral and Walker [2012] obtained data on 1.3 million Facebook users to analyse which demographics are more susceptible to peer pressure. They find that younger users and males are more easily influenced, which may have implications for policies aimed at countering the spread of negative behaviours online such as extremism. Bakshy et al. [2015] find that Facebook leads to users being less exposed to diverse sources of news. Bond et al. [2012] show that voters were directly influenced in the 2010 US congressional elections by messages they received on Facebook. However, Facebook now lacks a public API which may limit its use to researchers in computational social science. It may instead be more useful as a way to recruit participants for more traditional data collection [Kosinski et al., 2015].

2.2.2 Use of image data

Other social networks have, at least temporarily, had public APIs, making their data easier to collect at scale. The photo-sharing site *Flickr* has a public API, which has led to a large volume of research using its data. Preis et al. [2013a] show that the volume of Flickr images uploaded around the time of Hurricane Sandy and tagged with the name of the hurricane may provide a proxy for the intensity of Hurricane Sandy in New Jersey. In turn, they suggest that Flickr data may be useful for rapid crisis response, given the real-time availability of Flickr images. Tkachenko et al. [2017b] build on this result by showing Flickr data improves predictions of floods. Barchiesi et al. [2015a] turn to data on the location of Flickr users over time, inferred from photo metadata, and show a positive correlation with international travel flows from the UK's Office for National Statistics (ONS). Barchiesi et al. [2015b] demonstrate a similar relationship for mobility patterns within the UK, and Preis et al. [2020] show that the methodology can be generalised to all of the G7 countries.

Beyond mobility, Aiello et al. [2016] are able to infer the level of noise pollution in London and Barcelona at high geographic resolution, by using the text in tags from Flickr photos. Seresinhe et al. [2016] exploit text data attached to Flickr photos to estimate the presence of street art. They show that, in London, house prices rise more in areas in which more art is detected. More broadly, Seresinhe et al. [2018] demonstrate that data from Flickr photos can be used to make conclusions about the aesthetic appeal of a neighbourhood.

Prior to 2016, *Instagram* also had a public API that researchers used. Sim-

ilarly to the Barchiesi et al. [2015a] analysis of international travel flows, Weilenmann et al. [2013] show that Instagram uploads from museums can be used to analyse attendance. Botta et al. [2020a] demonstrate that the volume of Instagram photographs from a given location can provide an instant indicator of crowd size. These studies are remarkable as they generate useful data from images without actually using any of the image content, relying only on metadata such as timestamp, location and tags.

There has been rapid growth recently in studies classifying the content of the images themselves, due to key developments in an area of machine learning called "deep learning" [LeCun et al., 2015]. In 2012, Krizhevsky et al. [2012] made a major breakthrough in the speed of training more complex image classification models called "convolutional neural networks" (CNNs). This led to a dramatic expansion of research using image data. For example, Alanyali [2018] is able to train a CNN to detect protests from Flickr image uploads. This built on previous work tracking protests using only the tags on images, rather than the content of the images themselves [Alanyali et al., 2016].

In one particularly innovative paper, Seresinhe et al. [2017] combine two novel data sources – imagery and crowdsourced reviews of image "scenicness" – to analyse, at scale, what makes landscapes attractive. Previous research using images had been mostly restricted to analysis of image colour. For example, landscape studies found that images with green or blue areas tend to be more attractive, as they usually represented forests or bodies of water [Ward Thompson et al., 2012; Triguero-Mas et al., 2015]. Seresinhe et al. [2017] were able to go further, using CNNs to extract specific features of images and therefore go beyond pixel colour. For example, they find that areas with man-made features, such as castles, were often highly attractive, despite being less green and blue. On the other hand, people often rated bland grassy scenes unattractive, despite being heavily green. These findings have direct policy relevance, given the evidence for a connection between people's wellbeing and their environment [Arriaza et al., 2004; Real et al., 2000; Seresinhe et al., 2015, 2019], and related policy trade-offs around the placement of onshore wind turbines [McKenna et al., 2021]. In Chapter 5 of this thesis, we also combine two novel sources of data – darknet market reviews and Wikipedia page views – to shed light on an important policy issue.

Advances in computational social science have led to a well-developed literature on using satellite imagery. Satellite imagery has been popular among researchers, particularly in economics, because it is available at high geographic resolution, in real-time and at low cost [Donaldson and Storeygard, 2016]. This has

allowed researchers to measure variables at a scale that would previously have been prohibitively costly. Henderson et al. [2012] is arguably the seminal study in the field. Using satellite imagery, they measure the amount of light emitted at night time across a range of countries. They show that including the light data improves measurement of GDP growth for countries with poor data quality, consistent with earlier initial findings on light data from Chen and Nordhaus [2011]. The high geographic resolution of their light data allows them to construct a higher geographic resolution measure of GDP growth. They show this can provide new evidence on outstanding development questions, such as the impact of urbanization on growth and the impact of malaria prevalence on growth. While previous studies had used satellite data [Battese et al., 1988; Sutton et al., 2007; Sutton and Costanza, 2002], this was the first study to formally incorporate satellite data into a statistical framework for economic measurement.

The satellite data literature has since developed to include faster measures of other indicators, particularly in developing countries where data is more likely to be poor. Harari [2020] uses satellite imagery to measure the geometric properties of Indian cities, such as compactness and transport accessibility, for which large-scale data did not previously exist. They are able to provide evidence for a causal link between these properties and differing growth rates across cities. Jean et al. [2016] go further in complexity, by applying deep learning to satellite imagery in order to extract features for predicting poverty at high geographic resolution. Other studies have used satellite data to provide rapid measures of other variables that are important in economic development, such as deforestation [Burgess et al., 2012, 2019] and pollution [Jayachandran, 2009].

There is a nascent research deploying more complex techniques to derive measures from satellite images that are useful for developed countries. Law et al. [2019] analyse the issue of measuring house prices in London. They use a convolutional neural network (CNN) to extract important features, such as housing density and proximity to transport. They find that models measuring prices that included features extracted from satellite imagery outperformed those that used only traditional data, such as local crime rates and school quality. Law et al. [2019] also use images from Google Street View, building on previous work documenting the value of these images for statistical measurement [Gebru et al., 2017]. In Chapters 3 and 4 of this thesis, our methods to extract useful features from unstructured high-dimensional aircraft location data, build on the efforts made to work with satellite data so far.

This thesis uses similarly rapid developments in other areas of computer science for dealing with big data. For example, the aircraft location dataset we

use in Chapters 3 and 4 is approximately 50 terabytes. Analysing data at this scale is only now possible due to recent advances in compute power [Moore, 1998; Waldrop, 2016] and the MapReduce technique for distributing workloads across many cores [Dean and Ghemawat, 2008]. We are also able to use new data collection methods, such as web scraping, to leverage new sources of data such as feedback from darknet markets, where the original format is raw HTML [Dittus et al., 2018]. Over the following chapters, we show the potential for leveraging these cutting-edge techniques, along with the above developments in computational social science, to improve nowcasting in economics and public health.

2.3 Nowcasting in economics

2.3.1 Methodological developments

While now used in many disciplines, the term "nowcasting" was first coined in economics. In a seminal paper, Giannone et al. [2008] study the problem of GDP releases. This important statistic was released only quarterly, despite the increasing availability of more frequent releases of other data sources. Giannone et al. [2008] develop a framework for updating estimates of current GDP using other higher frequency series. Despite there being some limited, earlier literature on using higher frequency data [Evans, 2005], Giannone et al. [2008] was the first study to propose a formal nowcasting framework that could integrate a general number of higher frequency series at an arbitrary frequency. Giannone et al. [2008] find that integrating higher frequency economic data, such as monthly business sentiment surveys, significantly reduces nowcast errors. Soon after, real-time estimates for Chinese GDP were also derived from directly applying this method [Yiu and Chow, 2010].

There is now an extensive literature in macroeconomic nowcasting, which is mostly focused on how to integrate higher frequency conventional data [Kapetanios and Papailias, 2018]. Many papers have used Mixed Data Sampling (MIDAS) methods, developed by Ghysels et al. [2004] for models including time series at different frequencies. This technique has been extensively applied to Euro area GDP, which is released at monthly frequency, with a particular focus on how to incorporate surveys of business and consumer confidence [Angelini et al., 2008; Giannone et al., 2009]. Another strand of the literature focuses on how to use asset price data, which is even higher frequency. Andreou et al. [2013] integrate daily financial asset prices into US GDP estimates. Aastveit and Trovik [2012] find that daily Oslo Stock Exchange prices, aggregated to monthly frequency, are the most important feature in their model for nowcasting Norwegian GDP. The above literature has seen

much development statistically, but is still generally focused on inclusion of existing economic indicators, rather than newer sources of data.

Central banks, recognising the value of accurate nowcasts, have also developed their own nowcasting models. The Bank of England nowcasts both UK [Bell et al., 2014; Anesti et al., 2017] and global GDP growth [Kindberg-Hanlon and Sokol, 2018]. Several branches of the US Federal Reserve independently nowcast US GDP growth [Doh and Bae, 2019; Atlanta, 2021] There is therefore a clear policy demand for better nowcasting models. We include even higher frequency data to improve GDP nowcasts – our aircraft location dataset often receives multiple observations for a given aircraft within a second.

2.3.2 Use of online data in economics

There have been more recent studies that begin to integrate unconventional data sources into economic nowcasts. This literature arguably begins with Ettredge et al. [2005]. They analyse the US unemployment rate, published monthly by the Bureau of Labor Statistics (BLS). They show that an aggregate of weekly online job search activity was significantly positively correlated with the unemployment rate. However, they were limited by the short time series of their data to documenting an in-sample correlation, rather than a fuller statistical analysis of whether web search data can predict unemployment out-of-sample.

The use of online data became more widespread in economics after the publication of Choi and Varian [2009b]. They use Google Trends data to improve nowcasts for several variables outside of GDP. For example, they show that Google searches for motor vehicles are indicative over monthly sales of motor vehicles data, as published by the US census bureau. Their results extend to services, too. Hong Kong publishes monthly statistics on visitors by country of origin. Choi and Varian [2009b], exploiting the geographic dimension of Google Trends, show that Google searches for vacations to Hong Kong from a given country are indicative of arrivals from that country. Consistent with the initial findings of Ettredge et al. [2005], they also show that searches for jobs and unemployment benefits are indicative of monthly unemployment claims. Choi and Varian [2009b] have enough data to evaluate models out-of-sample against a sensible baseline autoregressive model, rather than simply documenting a correlation, so this is arguably the seminal study in nowcasting with novel data for economics. Goel et al. [2010], when using Google search data to predict consumer behaviour, are similarly careful to evaluate model performance against a sensible baseline. For example, when predicting revenue of major film releases, the baseline model includes the budget of the film and the num-

ber of screens in which the film opened in. In this thesis, we always evaluate model performance out-of-sample in comparison to relevant baseline models.

Many studies since Choi and Varian [2009b] have followed suit in using Google data to nowcast unemployment. Unemployment may be a particularly good choice of macroeconomic variable to nowcast as, given its personal nature, there is likely a stronger link with internet search. Askitas and Zimmermann [2009] find similar results for Germany, although they are restricted by their sample size to documenting a positive correlation rather than an out-of-sample performance analysis. Pavlicek and Kristoufek [2015] find mixed results for the Visegrad Group, which they attribute to differing degrees of internet usage across the countries. Fondeur and Karamé [2013] find that Google searches are a significant indicator of youth unemployment in France, and including them in a nowcasting model reduces out-of-sample errors relative to the prediction based on past unemployment data alone. Their model may perform particularly well because the young are most likely to use the internet as a job search tool [D'Amuri, 2009]. D'Amuri and Marcucci [2017] focus on US unemployment during the great recession. They find that Google-based models generally outperform professional forecasters out-of-sample, and their performance is particularly strong at the beginning of the recession. In Chapter 3 of this thesis, we also focus specifically on model performance during times of high volatility, which is when rapid estimates of macroeconomic variables are likely of most value to policymakers.

The literature has extended to further macroeconomic variables, such as inflation. One innovative strand of this research is web-scraping online prices, which are useful for nowcasting given their real-time availability and widespread coverage. The Billion Prices Project [Cavallo and Rigobon, 2016] demonstrates the potential of web-scraping for generating useful, new data at scale. The authors scraped 50 million online prices across 30 different countries at a daily frequency for eight years. This proved effective for nowcasting prices at high frequency, and making international comparisons. Cavallo [2017] finds that online prices tend to mirror offline prices, therefore they serve as an accurate proxy for nowcasting the overall price level. In separate studies, statistics authorities in both the UK [Breton et al., 2015] and Netherlands [Griffioen et al., 2014] have also found that clothing, a major component of their Consumer Price Index (CPI), can be nowcast using clothing prices scraped from the internet.

Further studies used scraped data from the Billion Prices Project in a range of applications, both nowcasting and to analyse other issues in economics. Cavallo et al. [2018] show that online prices from similar goods across different countries

can yield rapid estimates of changes in Purchasing Power Parity, therefore helping to nowcast changes in real consumption across countries. Cavallo [2016] uses online price data to provide new evidence that price changes tend to be infrequent, and previous results documenting frequent changes are driven by measurement bias. Data from the Billion Prices project has been further used for policy evaluation. For example, Cavallo et al. [2019] shows the burden of US tariffs on Chinese goods fell almost entirely on US importers, rather than Chinese exporters. Another study showed that currency unions, such as the Eurozone, effectively reduce price differences across countries [Cavallo et al., 2014]. In Chapter 4, we make recommendations for how our similarly novel aircraft location data may be applied to further topics in economics.

There is growing appreciation among economic policymakers for the value of online data. From May 2020, the Bank of England began including Google trends data in their quarterly Monetary Policy Report [Monetary Policy Report, 2020]. They document sharp drops in Google searches for hotels, cars and theatres as early indicators of the economic impact from Covid-19. The Chicago Federal Reserve have also considered Google data in their nowcast of unemployment claims for the current crisis [Brave, 2020]. Their initial results show that Google search data is both an economically and statistically significant indicator of the geographic distribution of unemployment. In this thesis, we show that inclusion of yet wider data sources, namely aircraft location data and darknet reviews, are useful to economic policymakers. We provide the first evidence that real-time location data from aircraft may be a leading indicator for aviation's contribution to GDP. We also show the value of novel online data for estimating market trends – specifically that drug sales scraped from darknet markets and Wikipedia page views for each drug may be useful for nowcasting trends in drug demand. We build on the previous methodological literature by ensuring models are evaluated out-of-sample against relevant baselines, rather than simply documenting a correlation between a novel data source and an economic statistic.

2.4 Nowcasting in public health

2.4.1 Use of online data for monitoring illicit drug consumption

As discussed in the previous section, Chapter 5 of this thesis nowcasts drug demand using a combination of different online data sources. As well as being relevant to economic policymakers, accurate knowledge of changes in drug demand are useful for public health authorities too. Rapid changes in drug use are often characterised

as "epidemics" in their own right, such as the US opioid crisis [Frank and Pollack, 2017; National Institute on Drug Abuse, 2019]. However, these rapid changes in drug use are often hard to monitor. Authorities have traditionally relied on annual surveys, for example the United Nations Office on Drugs and Crime (UNODC) World Drug Report [United Nations Office on Drugs and Crime, 2019]. The low frequency of these statistics may be insufficient for authorities to respond to rapid changes in drug use, such as the US Fentanyl epidemic, which arguably took off within a year [Higham et al., 2019].

There is a nascent literature in public health on better nowcasting drug use with online data, which is generally available much faster than the official surveys. Perdue et al. [2018] consider the issue of tracking growing popularity of new substances, such as cannabinoids and opioids, as measured by the annual *Monitoring the Future* (MTF) survey in the US. The lack of high frequency data is a particular issue for monitoring novel substances, as their growth can be especially fast and they may not show up on traditional drug surveys at all. Perdue et al. [2018] therefore hypothesise that there may be value in using Google Trends data to improve monitoring of these substances. They extract Google Trends data related to several categories of drugs and aggregate to annual frequency to match the MTF surveys. They find significant, positive correlations between Google Trends and drug consumption in most cases. This provides evidence that Google Trends may be a useful fast indicator of changing drug consumption. However, the study is limited by the short time series of annual frequency data available to the authors. For example, they estimate the correlation for the "Bath salts" drug category with just five data points. This limits their ability to go beyond correlation in their analysis.

Balsamo et al. [2019] analyse the problem of monitoring changing levels of opioid abuse across US states. The US CDC publishes annual survey data from each state monitoring this issue. Balsamo et al. [2019] generate a novel data source from comments about opioids on the large online forum, Reddit. These comments have location metadata, so the authors are able to identify which state a comment is from. They find a significant positive correlations between the volume of Reddit comments about opioids from a given state, and the level of opioid abuse in that state. This suggests online data may be useful for monitoring geographic differences in drug use over time. While the sample size is larger in this study, as they have data for 50 states, they too are limited to documenting a correlation by having just two years of data available. The short time series of their data means they are unable to build a nowcasting model to assess the value of their online data.

Zheluk et al. [2014] analyse the issue of demand for the Russian opioid

Krokodil. They collect data on court appearances for Krokodil possession as a proxy for demand between 2010 and 2012. They find a strong positive correlation between Google searches for Krokodil-related topics and court appearances. While having access to data from eight Russian regions, they too are limited to documenting a correlation by the short time series of the court appearance dataset. In an innovative study, Kapitány-Fövény and Demetrovics [2017] use web-search data to analyse cross-elasticity between traditional drugs and novel substances, which may act as substitutes. They analyse Hungarian interest in Mephedrone, a stimulant that was briefly legal but banned in 2011. By assuming monthly web search data is a proxy for demand, they find evidence that Mephedrone was being used as a substitute for MDMA. However, they concede that more evidence is required that web search is a proxy for demand, noting that searches for Mephedrone spiked after it was banned despite a fall in demand.

In Chapter 5 of this thesis, we construct a high-frequency time series of drug demand with data scraped from the darknet. This allows out-of-sample evaluation of a nowcasting model with online data against a autoregressive baseline. No previous papers had access to a long enough time series of data for such an evaluation. We therefore build on the previous literature by providing the first evidence that online data can help nowcast drug demand out-of-sample. Furthermore, this thesis is the first study, to our knowledge, focusing on nowcasting demand on the darknet more generally.

2.4.2 Use of online data in epidemiology

Epidemiology is one of the academic fields with the most well-developed nowcasting literature. Researchers have used time series methods for predicting the spread of disease for over 20 years [Allard, 1998], which can be effective independent of the underlying causal process for why the disease spreads. Better nowcasts of disease spread are highly valuable, as there is a very clear policy case for having the most up-to-date statistics on disease spread. As recently underlined by the Covid-19 crisis, policymakers often have to make crucial decisions, such as whether to impose stricter lockdown measures, without accurate, current data on the prevalence of disease [Mavragani, 2020]. Earlier, more effective interventions have already been shown to reduce mortality rates across a range of diseases, such as arboviruses [Olliaro et al., 2018], influenza [Longini et al., 2005; Ferguson et al., 2005] and Covid-19 [Sun et al., 2020; Galea et al., 2020].

Ginsberg et al. [2009] is arguably the seminal study in nowcasting with online data. The CDC release weekly estimates of influenza incidence across regions in the

USA, with a lag of one week. Ginsberg et al. [2009] collect Google search data for symptoms of influenza, aggregated to weekly frequency. They show very strong positive correlations between the Google search data and influenza rates. Given the real-time available of Google search data, they conclude that their modelling approach can generate accurate estimates of influenza prevalence with a lag of just one day, as opposed to the one week lag in the CDC data. This modelling approach came to be known as "Google Flu Trends".

The Ginsberg et al. [2009] study was highly innovative, marking the first high-profile use of online data for disease nowcasting, but it suffered from some key methodological limitations [Lazer et al., 2014]. Goel et al. [2010] show that a simple autoregressive baseline model, where they estimate the current period's influenza incidence using the previous period, performs just as well as the Google Flu Trends model. Moreover, the Google Flu Trends model performance may have declined over time. Copeland et al. [2013] find the Google Flu Trends model performed relatively well until the 2013 flu season, when it dramatically overestimated the size of the flu outbreak in the USA. They attribute this to heightened media coverage of the flu that year, and that their model had not been updated since 2009. This highlights the need for online search data to be used as a complement, rather than a substitute, for official data, and that time-series models should be updated as new data becomes available. Preis and Moat [2014] incorporate both these lessons into their modelling approach. They first build a baseline autoregressive model that nowcasts current flu incidence using the past history of flu incidence. They then augment this model with real-time data on the current volume of Google search queries for flu-like symptoms. They refit their model each week as new data arrives, therefore allowing it to adapt over time. Using this approach, they show that including Google data alongside official data improves model accuracy by over 14%. Yang et al. [2015] further build on Google Flu Trends, with an approach called Autoregressive Google (ARGO) by explicitly modelling seasonality and allowing the weight on different search terms to vary over time. In Chapter 6 of this thesis on nowcasting chikungunya incidence, we similarly build a model that augments an autoregressive baseline, rather than replacing it altogether, and adapts over time as new data arrives.

There is now also an extensive literature on using online data to nowcast the incidence of dengue, the most prevalent arbovirus globally [Wilder-Smith et al., 2017]. Chan et al. [2011b] extend the Google Flu Trends methodology to tracking dengue fever across five countries, with a modelling approach called Google Dengue Trends (GDT). They find strong positive correlations, ranging from 0.82 to 0.99, in each country. They attribute periods of weaker performance, such as the 2005 Indian

dengue season, to a lack of internet access. While they validate results out-of-sample against a "hold-out" set, they do not check their models outperform a baseline autoregressive model using past official data on dengue prevalence. Yang et al. [2017] build on these results by extending their ARGO method from influenza to dengue. They find that appropriate inclusion of Google data generally improves model performance, although Google modelling is vulnerable to over-estimating outbreaks that attract significant media attention. Gluskin et al. [2014] also find that GDT performs well at the country-level. However, they find more variable performance when analysing the data at a higher spatial resolution across Mexican states. In particular, they find that local weather is an important predictor, as GDT performs worse in areas with climates less suited to dengue transmission.

There have been far fewer studies using online search data to nowcast chikungunya, a less-studied arbovirus, which is the focus of Chapter 6 of this thesis. Naveca et al. [2019] analyse chikungunya incidence across regions of the Brazilian Amazon from 2014 to 2018. They find a strong positive correlation between Google searches for chikungunya-related terms and chikungunya case counts over time in each region. The timing of the peak of the 2017 epidemic also aligns with the peak of Google searches. However, they are limited to documenting a correlation, which is not sufficient evidence alone that Google search data is valuable for nowcasting chikungunya over and above the history of past case counts alone. Chapter 6 of this thesis therefore builds on the previous literature by providing the first evidence of the value of online search data for monitoring chikungunya in a rigorous nowcasting framework.

2.4.3 Issues with incomplete data

The vast majority of the epidemiological nowcasting literature assumes complete data on previous periods is reliably available at the time of making the nowcast. This may be a reasonable assumption in some cases, such as nowcasting the weekly CDC influenza case counts, where data is entered very regularly. However for nowcasting arbovirus incidence in developing countries, such as in Yang et al. [2017], this assumption may not hold. There is substantial literature documenting that dengue fever is prone to delays and revisions in reporting [Madoff et al., 2011; Runge-Ranzinger et al., 2008]. This means that, when nowcasting the current period in practice, we also have to estimate disease incidence in previous periods, rather than conditioning on it. Therefore, models that condition on accurate knowledge of disease incidence in previous periods may not be usable in practice for nowcasting arboviruses in developing countries. In Chapter 6 of this thesis, we have access to

case-level data on both the date of diagnosis and date of entry into the monitoring system for arboviruses. This allows us to construct a model that is operationally usable for nowcasting chikungunya incidence.

There is now a nascent literature in nowcasting the spread of dengue in the presence of incomplete data. A research team based in Rio de Janeiro has developed a system called InfoDengue [Codeço et al., 2018]. The system, developed by researchers at Fiocruz (a research institute linked to the Brazilian Ministry of Health) and FGV, monitors the incidence of dengue, chikungunya and Zika across 792 Brazilian cities. InfoDengue has a nowcasting model which is updated every week as new case data is entered. The system collects data on weather conditions and the volume of dengue-related Twitter posts, given previous studies documenting the predictive potential of these variables [Lowe et al., 2014; Gomide et al., 2011; Marques-Toledo et al., 2017]. However, these variables are not currently used when making the nowcast [Codeço et al., 2018].

The original model was relatively simple: it considered the number of cases entered in a given week, estimates how many actual cases are missing based on historic data and applies an uplift to nowcast the actual current number of cases. Bastos et al. [2019] develop a new methodology for nowcasting dengue incidence, which is now being used in InfoDengue. Their model is based on Integrated Nested Laplacian Approximation (INLA – see Rue et al. [2009]). INLA is a Bayesian simulation approach assuming an underlying Gaussian process generating observed data. Bastos et al. [2019] show that their approach is much more accurate than the estimates produced by the original InfoDengue approach. They explicitly account for the missing data issue, using only data on past cases that had been entered at the time of the nowcast, rather than assuming complete availability of past data. However, the Bastos et al. [2019] model uses only historic case count data - they do not include any other variables such as online data or weather.

Mizzi et al. [2021] build directly on the Bastos et al. [2019] approach to improving dengue estimates in the presence of missing data. They first extend the INLA model to include the Twitter data that is provided as part of InfoDengue, and show this improves accuracy. They incorporate a further online data source - Google Trends for dengue symptoms - to show a further improvement in model accuracy. They also show that including online data improves model precision, measured by the width of the prediction intervals output by the INLA algorithm. Mizzi et al. [2021] find that the best model includes both Google and Twitter data, although the marginal improvement of including both is small compared to the improvement from having just one online data source. Finally, Mizzi [2019] further improves the

literature by extending the geographic scope of the analysis. Whereas Bastos et al. [2019] analysed only Rio de Janeiro, Mizzi [2019] shows that online data generally improves model accuracy and precision across 10 other Brazilian cities, although the improvement is largest in Rio de Janeiro.

In chapter 6 of this thesis, we build directly on Mizzi et al. [2021] by showing their approach to nowcasting in the presence of partial data, and the inclusion of Google Trends data, improves estimates for chikungunya, as well as dengue. This is important given the substantial overlap in symptoms between dengue and chikungunya can make it difficult to distinguish the two diseases [Hochedez et al., 2006]. Furthermore, we ensure all our models are evaluated out-of-sample against a rigorous baseline, rather than being limited to documenting a correlation.

2.5 Previous work using aircraft location data and data from darknet markets

2.5.1 Previous work on darknet data

Scraping of online reviews is a a particularly relevant strand of research, given the focus of Chapter 5 on data from darknet markets. The online review aggregator, Yelp, provides high-time frequency data at a high geographic resolution, and has therefore been used in many recent studies. Glaeser et al. [2017] analyse whether Yelp data can help nowcast local economic changes, as measured by the annual US ZIP code level County Business Patterns dataset. They find that growing numbers of Yelp reviews are indicative of faster local business growth, even after controlling for prior business growth in the region. Other studies have dug deeper into the content of the Yelp reviews themselves. Luca and Zervas [2016] find that restaurants are more likely to commit review fraud if facing increased competition, which may have implications for business propensity to break laws more generally in response to increased competition. At a more micro-level, Chong et al. [2017] scrape Amazon reviews to analyse whether they contain useful information for predicting product-level demand. They find that both the volume and positivity of the reviews are relevant, but the performance improvement relative to the model with no review data is small. See-To and Ngai [2018] find similar results when analysing reviews on Chinese e-commerce sites. This thesis builds on the scraping literature using online reviews by incorporating a data source used rarely before in economics and never in public health - reviews from darknet markets.

It is unsurprising the literature on darknet markets is sparse, given that

the markets themselves are a fairly recent innovation. Christin [2013] was the first study to use scraped data from the markets, and their study is mostly a descriptive analysis of how the first big darknet market, Silk Road, evolved over time. Soska and Christin [2015] then documents how the darknet markets that followed Silk Road evolved. Use of darknet data in economic research has been very rare, so far. Bhaskar et al. [2019] use sales data from the darknet to analyse the heightened importance of reputation in black markets, given the lack of legal enforcement to resolve disputes. Červený and van Ours [2019] use cannabis price data from darknet markets to gain a rare insight into differences in international drug pricing. They find that, despite international competition between sellers, prices are still higher in countries with higher incomes, suggesting buyers have a preference for their "home" country.

The most relevant darknet study for this thesis is Dittus et al. [2018]. They study the demographics of drug consumption on the darknet, and whether its existence has altered supply chains. They find that darknet demand for illicit drugs is geographically representative of traditional demand i.e. countries with high cannabis consumption traditionally, such as the UK, also have a relatively high share of cannabis consumption on the darknet. This is an important finding for Chapter 5 of this thesis, as it is more useful for policymakers to be able to accurately nowcast darknet drug demand if it is a good proxy for drug demand from traditional sources too. This thesis therefore contributes the first study into estimating drug demand with darknet data, which may be useful for policymakers in both economics and public health.

2.5.2 Previous work using vehicle location data

Location data has rarely been used in nowcasting, likely due to the difficulty of engineering such large datasets until recently. The UK Office for National Statistics (ONS) has published preliminary findings on using two such data sources for nowcasting. The first data source is UK road traffic data from Highways England dataset, which is available after a 3-week lag [Highways England, 2022]. Rowland [2019] measure the volume of road traffic from high-frequency sensor data around economically important locations, such as ports. They show evidence of a positive correlation between the volume of road activity for large vehicles and a variety of economic measures, such as GVA and international trade. They nevertheless advise caution with using these statistics alone to nowcast GDP, as the relationship seems to be weaker outside of large economic events such as recessions.

The second data source is freight shipping data from the Automatic Iden-

tification System (AIS - see Bonham et al. [2018]). The AIS records the location, speed and heading of ships in close proximity to UK ports at high frequency. The high dimensionality of their dataset (the raw data is reportedly one terabyte) means it requires extensive engineering to extract useful features, such as the ship's trajectory. Bonham et al. [2018] show that the AIS data, after such feature engineering, can help nowcast shipping delays. In turn, they speculate that these delays may be a useful early indicator for components of GDP that are related to shipping. However, they do not formally analyse this extension in the report, which is limited to predicting shipping delays with the AIS data.

This thesis contributes a novel data source to the nowcasting literature: aircraft location data from the ADS-B protocol. So far, the vast majority of literature using ADS-B data has focussed on its originally intended application in the field of aviation. For example, studies have used ADS-B data for improving trajectory prediction [Alligier and Gianazza, 2018] and delay estimation [Calvo-Palomino et al., 2018]. Some research has also shown that ADS-B data can help nowcast weather conditions at high geographic resolution [Kapoor et al., 2014; Trub et al., 2018]. However, there have been few studies using ADS-B data outside of aviation. The closest study to economics is arguably Strohmeier et al. [2018], who show that tracking the flights of corporate jets through ADS-B data may provide an early indicator of merger activity. Chapter 3 therefore makes a significant contribution as it marks the first use of location data from aircraft in nowcasting (published in Miller et al. [2020b]). More recently, partly due to the Covid-19 crisis, more research has begun using ADS-B data for economic measurement [Iacus et al., 2020].

Chapters 3 and 4 of this thesis build directly on the initial exploration of shipping location data in Bonham et al. [2018]. The ADS-B dataset also records the location, speed and heading of aircraft at high frequency. We similarly perform extensive feature engineering when faced with a high dimensional raw dataset (my raw data is approximately 50 terabytes) to extract useful features, such as takeoffs and landings. We also go further in economic analysis, by testing whether these features are useful for nowcasting economic variables, rather than leaving it to future research. The methods in this thesis could be applied to other location data, such as the roads data from Rowland [2019] or shipping data from Bonham et al. [2018].

2.6 Contributions of this thesis

This review has shown how recent developments in the field of computational social science have enabled large improvements in nowcasting. There have been many

studies into both nowcasting methodology and inclusion of new data sources, such as online search data, that have been applied to a range of statistics across economics and public health. Initial results have shown potential for speeding up key statistics, such as flu incidence, but many studies are limited to documenting a correlation between a novel data source and a socioeconomic statistic. A correlation alone is insufficient to show that a new data source is useful for nowcasting. In this thesis, we evaluate models by their out-of-sample nowcast performance, rather than correlation. We also ensure models are always evaluated against a relevant baseline, to show that the novel data source should add value when deployed in practice alongside existing methods to nowcast statistics.

In economic nowcasting, the methodological literature for integrating higher frequency data into lower frequency official releases, such as GDP, is already well-developed. However, there is scope to broaden the sources of data considered, as most of the literature so far has focused on higher frequency economic data such as surveys of purchasing managers and asset prices. This thesis is the first to analyse the potential for aircraft location data to improve the speed of estimates of GDP. We also contribute the first evidence that online data from darknet markets and Wikipedia page views may be useful for nowcasting drug demand, which would benefit policymakers monitoring black markets.

In nowcasting for public health, online data (particularly online search data) has already been widely used. Much of this literature assumes complete data is available on past incidence of a disease when nowcasting its current period incidence. This may hold for some applications, such as monitoring influenza in the USA, where the arrival of data in the central monitoring system is so quick that the gradual fashion in which case data usually arrives is somewhat hidden. However, it is a less accurate assumption for other applications, such as monitoring mosquito-borne diseases in Brazil, a situation where data arrival is slower, and hence the gradual delivery of the data and the inconsistent delays in reporting are much more evident. In this thesis, we build on previous literature developing models for situations when the arrival of data on previous disease incidence is inconsistent. We extend the scope of diseases considered to chikungunya, for which there are, to the best of our knowledge, no nowcasting studies yet.

Chapter 3

Estimating current economic activity with aircraft radar data

3.1 Introduction

Most economic statistics, such as GDP, are released with a significant delay. Estimating their current values before they are published is known as "nowcasting" [Giannone et al., 2008]. As discussed in Chapter 2, there have been previous studies using real-time internet data, from sources such as *Google* [Choi and Varian, 2009b; Carriere-Swallow and Labbe, 2013; Vosen and Schmidt, 2011; Preis, Moat, Stanley, and Bishop, 2012; Da, Engelberg, and Gao, 2011], *Twitter* [Bollen, Mao, and Zeng, 2011; Botta, Moat, and Preis, 2015] and *Wikipedia* [Moat, Curme, Avakian, Kenett, Stanley, and Preis, 2013; Mestyan, Yasseri, and Kertesz, 2013], to nowcast economic data. Better nowcasts are highly valuable, as major economic policy tools such as interest rates can take up to 20 months to fully impact the economy [Bernanke and Gertler, 1995]. When faced with shocks like the 2008 financial crisis, policymakers must therefore respond as quickly as possible, which requires accurate knowledge of the current state of the economy. Failure to do so has deepened past recessions [Auerbach and Gorodnichenko, 2012], leading to political instability across Europe following both the Great Depression and the 2008 financial crisis.

Aviation is a key economic sector, contributing at least 3% to GDP in the UK and the US [Federal Aviation Authority, 2016; Oxford Economics, 2014]. Aircraft now broadcast their location, among other data, in real-time using the Automated Dependent Surveillance Broadcast (ADS-B) system. So far, ADS-B data has rarely been used outside its intended application in the aviation sector. An exception is one study that showed that ADS-B data could track corporate jets, therefore providing

a leading indicator for certain mergers between firms [Strohmeier et al., 2018]. This chapter analyses whether ADS-B data can help nowcast airline performance and aviation's direct contribution to GDP. In contrast to real-time ADS-B data, the current statistics for these variables are published with a three month delay because they require surveys of businesses. Faster statistics could help policymakers respond more quickly to future economic shocks, thereby limiting their damage.

3.2 Methods

3.2.1 ADS-B data

We retrieve aircraft data from the *ADS-B Exchange* [ADS-B Exchange]. Commercial aircraft in Europe have been required to broadcast ADS-B data since 2017 [European Commission, 2011], and it has been mandatory for US aircraft since January 2020 [Federal Aviation Authority, 2010]. These broadcasts include the aircraft's speed and location, specified as their altitude, latitude and longitude alongside a timestamp. Each ADS-B message also includes a six-digit hex identification code assigned to the aircraft by the *International Civil Aviation Organisation* (ICAO). Amongst other things, the ICAO code makes it possible to link an aircraft to its operating airline by looking up the ICAO code in a corresponding database. This pre-processing is carried out by *ADS-B Exchange* and the operating airline is included in each of the resulting ADS-B records.

ADS-B messages are unencrypted, in order to be receivable by other aircraft, which means they are available to anyone with an ADS-B receiver. The *ADS-B Exchange* collects data from thousands of receivers [ADS-B Exchange]. The resulting database covers global flight activity (Fig. 3.1). We analyse the period from July 2016 to December 2018.

We note that coverage has improved over time, as the number of receivers feeding the database has grown. Figure 3.2 shows how Western Europe and the coastal USA have had good coverage since the exchange launched in June 2016. However, coverage has notably improved in the central USA, eastern Europe and Brazil. Much of Africa and Asia remains uncovered, limiting our ability to analyse these regions.

The raw data we analyse contains roughly 25 billion messages. First, we reduce this data from one row for each message to one row for each flight. Figure 3.3 shows how we identify take-offs and landings by analysing the altitude of an aircraft over time.

We use the altitude and timestamp fields from the ADS-B messages to es-

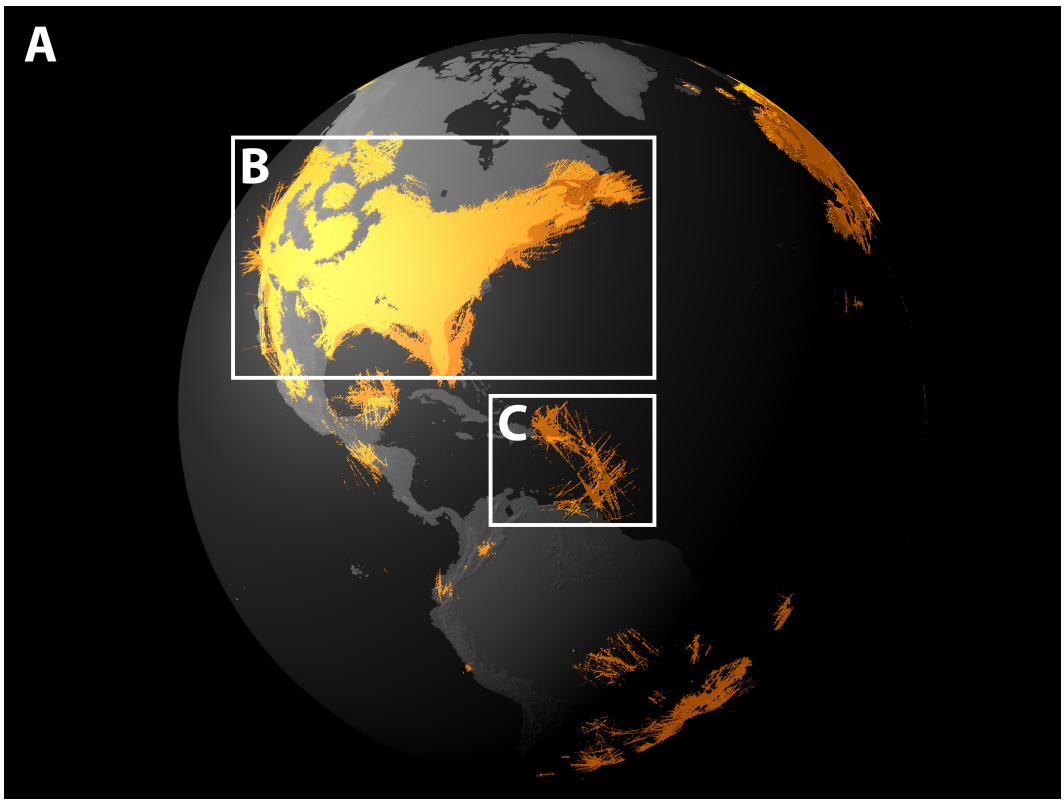

Figure 3.1: **Recorded flight paths over the western hemisphere on 30th September 2016.** (A) In orange, we depict the locations from which at least one ADS-B message was received by the network of receivers. (B) The network covers large parts of the United States and in particular their coastal regions. Visual inspection of the US east coast reveals that the land-based receivers are capable of tracking air traffic over coastal waters too. (C) In regions within coverage but with less dense air traffic flow, e.g. in the broad vicinity of the Caribbean Sea, distinct flight routes emerge. The increase in coverage from 2016 to 2018 is shown in Figure 3.2. The base layer of this map utilises the ALOS World 3D global digital surface model provided by the Japan Aerospace Exploration Agency (JAXA), which is available to use with no charge via https://www.eorc.jaxa.jp/ALOS/en/ (©JAXA).

timate how many flights each aircraft makes per month. Figure 3.4 shows that, for some aircraft, there are clear errors in altitude data. These errors could reduce the accuracy of our estimates of aircraft activity, so we clean the altitude data using median filtering. To median filter, we first set a window size $k = 5$. For each altitude observation a_t, we calculate the median of observations in the window $[a_{t-2}, a_{t-1}, a_t, a_{t+1}, a_{t+2}]$. If the altitude is different from the median then we replace it with the median.

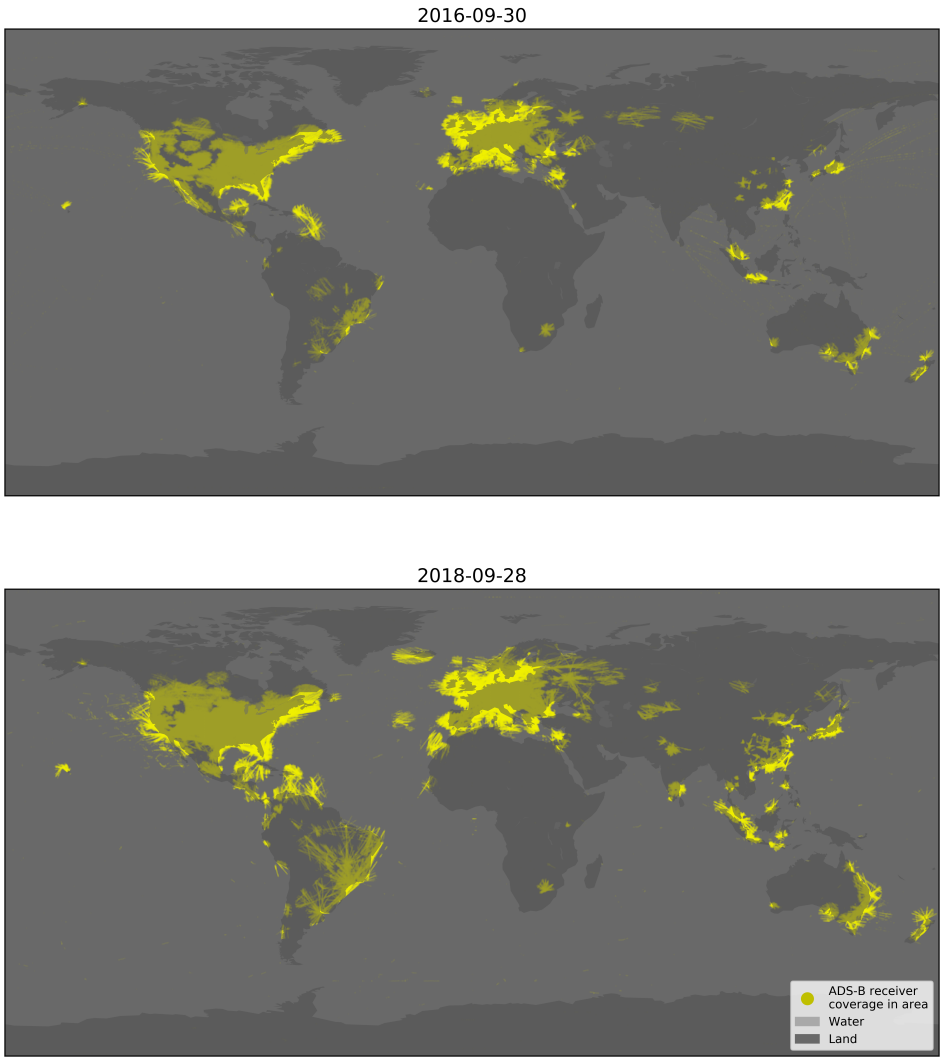

Figure 3.2: **ADS-B coverage over time.** This figure depicts coverage over time. An area is highlighted in yellow if an ADS-B message was received from that location on (A) 30th September 2016 or (B) 28th September 2018. The ADS-B Exchange has had good coverage in Western Europe and the coastal USA since launching in June 2016. Coverage has clearly improved across time. This improvement is most noticeable in the central USA, eastern Europe and Brazil. Much of Africa and Asia remains uncovered. The base layer of this map uses data from OpenStreetMap (©OSM contributors), which is available to use with no charge via https://www.openstreetmap.org/.

To give some numeric examples, suppose we observe the following altitude vector for an airraft $A_1 = [10000, 10100, 10200, 2000, 10400, 10500, 10600]$. There are 3 points that have enough neighbours to apply the median filter: a_3, a_4, a_5.

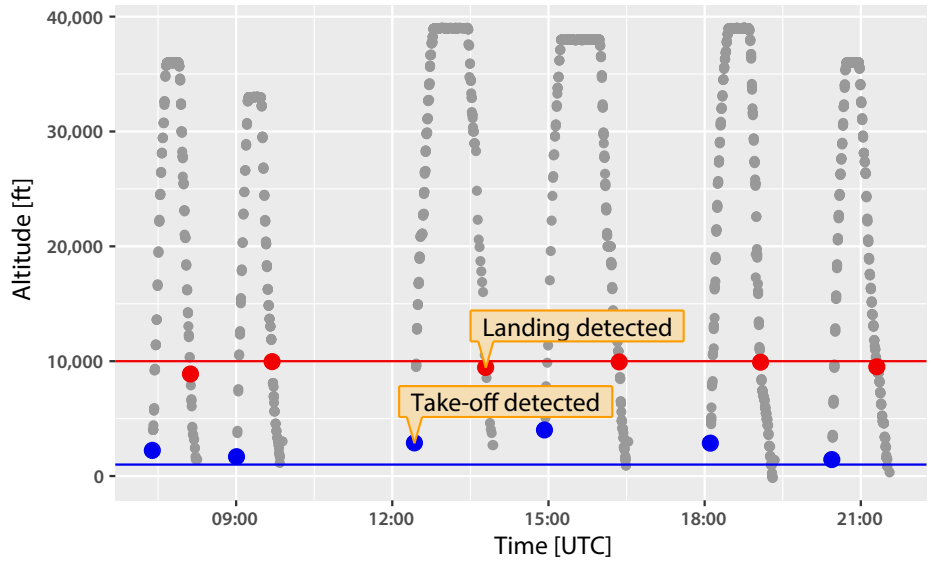

Figure 3.3: **The flight-counting algorithm.** An example of the flight-counting algorithm crawling through altitude data from real-time ADS-B messages. The algorithm identifies a take-off and landing of six separate flights for this aircraft over the course of a day. These form a new data structure, where we have one row for each flight rather than one row for each ADS-B observation. In total we identify 67 million separate flights from July 2016 to December 2018.

- $a_3 = 10200$: neighbours $[10000, 10100, 10200, 2000, 10500]$ has median 10100, so replace it.
- $a_4 = 2000$: neighbours $[10100, 10200, 2000, 10400, 10500]$ has median 10200, so replace it.
- $a_5 = 10400$: neighbours $[10200, 2000, 10400, 10500, 10600$ has median 10400, so do not replace it.

The final vector $A_1^{medfilt} = [10000, 10100, 10100, 10200, 10400, 10500, 10600]$ is a more realistic take-off trajectory. Figure 3.4 depicts the impact of median filtering on 3 aircraft over a given day. The upper panel is an aircraft with fairly clean data, so filtering changes only 0.8% of observations. The middle and lower panels show aircraft with less clean data, so filtering changes 1.2% and 2.8% of their observations respectively.

Initially, there is one row for each message containing, among other fields, the aircraft's latitude, longitude, altitude and timestamp. To estimate monthly airline performance, we do not need such a large volume of data. To facilitate our analysis, we therefore reduce the data to one row for each unique flight. This contains the

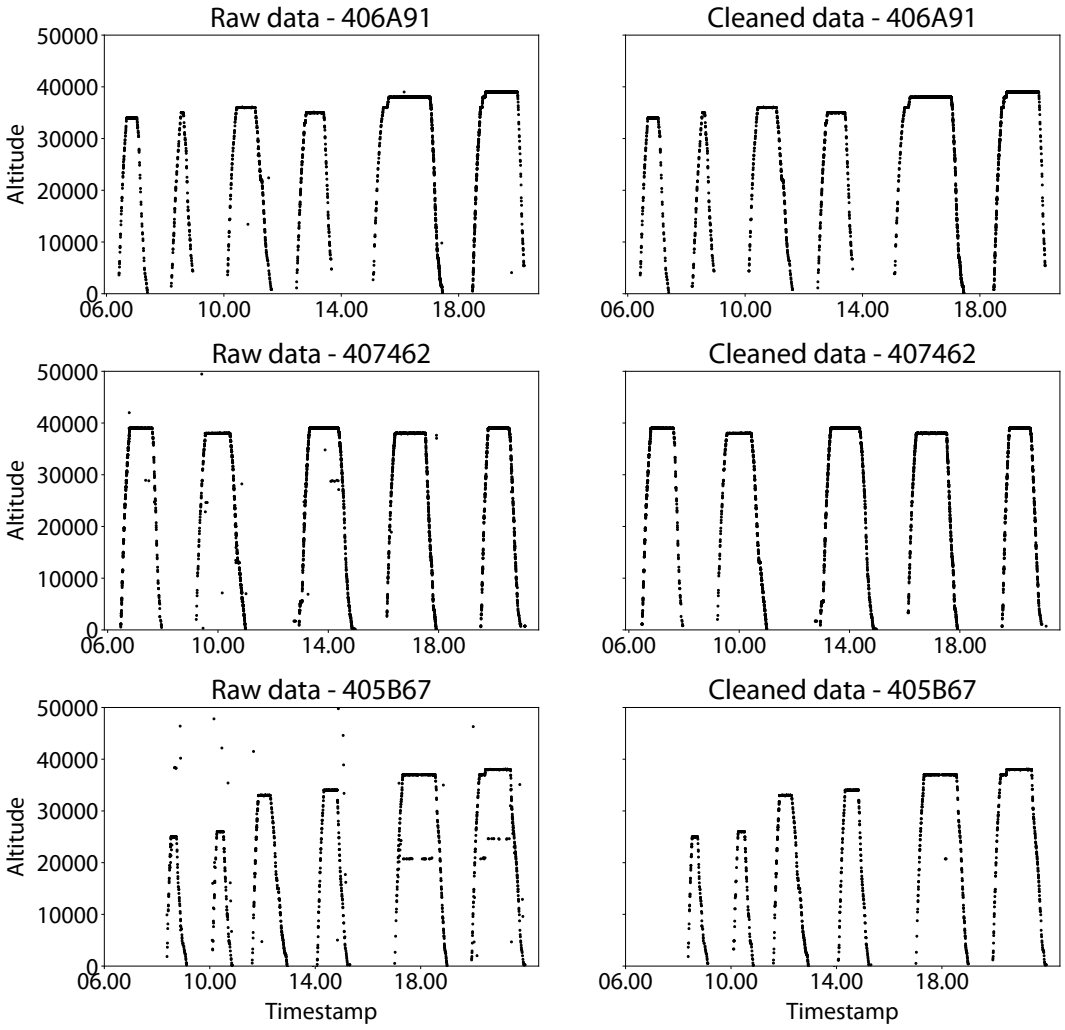

Figure 3.4: **Impact of median filtering on aircraft altitude profile.** The left show the raw data, and the right show data after median filtering. The raw data clearly contains trajectory errors, which median filtering is mostly able to correct.

aircraft's take-off time and location, landing time and location, aircraft ID and the airline.

Counting unique flights from the ADS-B messages is a non-trivial task. Firstly, each aircraft can make multiple flights per day, and there is a lot of variation depending on journey length. We show the distribution of daily flight counts per aircraft in Figure 3.5.

There is a field in the ADS-B message identifying the aircraft, but no reliable field to identify separate flights. We therefore created an algorithm to identify separate flights from the raw altitude data, henceforth referred to as the *flight-*

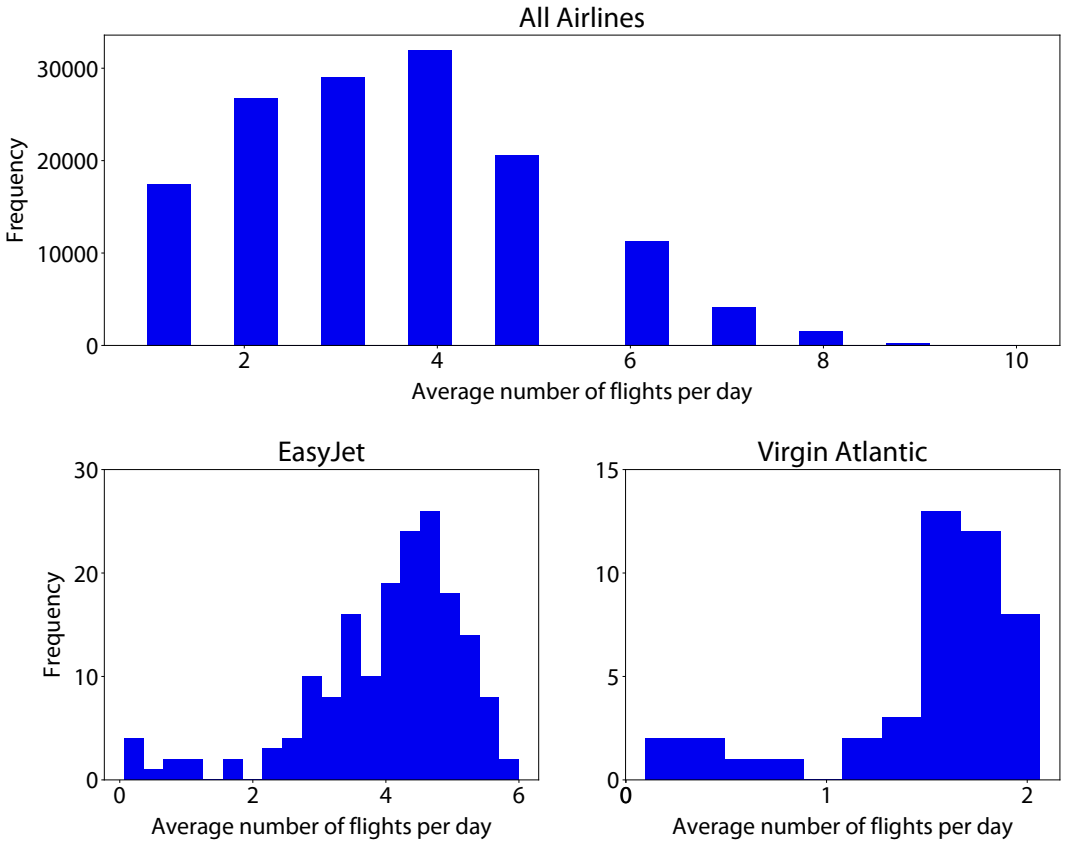

Figure 3.5: **Distributions of daily flight counts for each aircraft in September 2018.** The upper panel shows that the range is wide: many aircraft make just one flight, but some make up to ten flights. Most of this variation is explained by journey length, which varies systematically by airline. The lower panels show that Easyjet aircraft, which fly short haul, make far more flights on average than Virgin Atlantic, which also fly long haul.

counting algorithm. The flight-counting algorithm uses the aircraft's altitude to identify the take-off and landing of each flight. Figure 3.3 shows how we crawl through the time-ordered altitude observations and creates a take-off if the aircraft ascends above a certain threshold. Similarly, it creates a landing if the aircraft descends below a given threshold. The below figures shows how the algorithm identifies take-offs and landings for given aircraft over the course of a day. The counting algorithm reduces the data from having a row for each ADS-B message to a row for each flight. We identify over 67 million flights this way.

Second, the raw data is not very clean, as shown in Figure 3.4. This can reduce the flight-counting algorithm's accuracy. Figure 3.6 depicts the impact of cleaning the altitude data. The left panel shows counts carried out on the raw data,

and we can see that they are often wrong. The right panel shows more accurate counts carried out on the cleaned data.

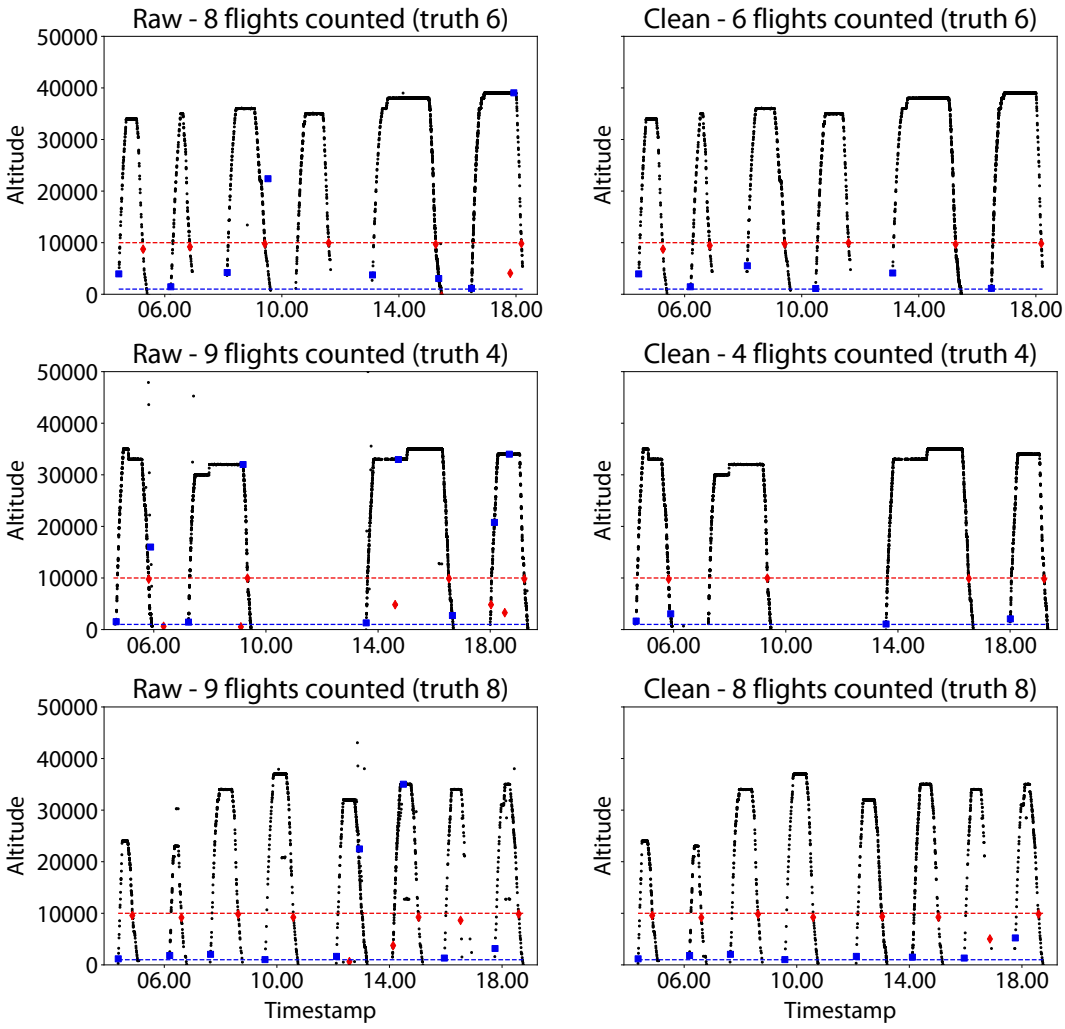

Figure 3.6: **Impact of data cleaning on accuracy of flight counting.** The left panel shows the counting algorithm on the raw data for given aircraft over a day. Noise in the data causes the algorithm to identify false positives, so it counts too many flights. The right panel shows the counting algorithm operating on the same data after median filtering. Visual inspection suggests that the algorithm is more accurate after cleaning.

Even after filtering, the altitude data still has occasional errors. To minimise their impact, we introduce some heuristics to the algorithm. If a take-off (landing) is recorded, we stipulate a lag of 30 minutes before the aircraft can land (take-off) again. We select this lag as a reasonable minimum journey time for commercial flights. Figure 3.7 illustrates how adding the lag heuristic makes the counting algo-

rithm more robust to any residual noise in the data.

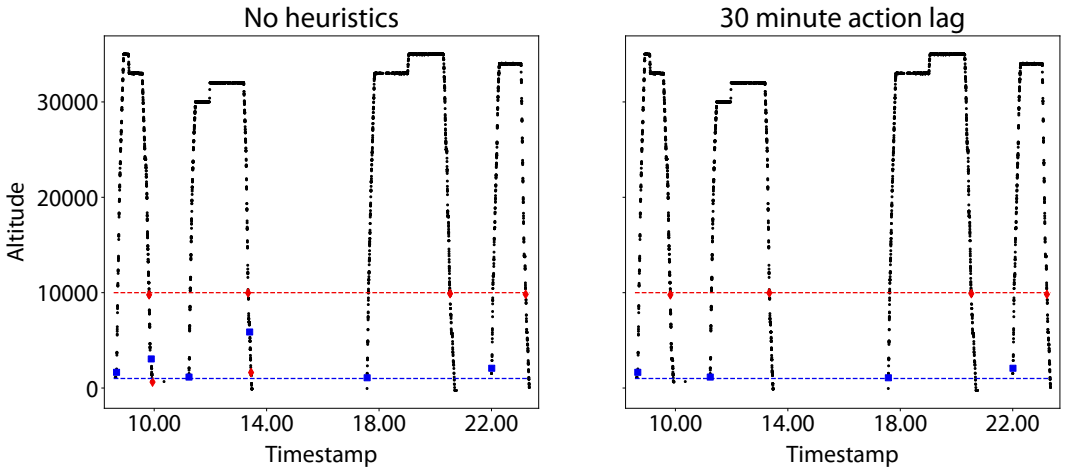

Figure 3.7: **Impact of adding a heuristic that enforces a lag between actions.** This figure demonstrates the importance of enforcing a lag between recording actions for an aircraft. Even after median filtering, there may be residual issues with data quality. The left figure shows these issues can lead to recording of multiple false positives, counting six flights versus a true count of four. The right figure shows that enforcing a lag of 30 minutes between actions can help deal with residual data quality issues, with the algorithm correctly counting four flights.

Finally, we set the landing altitude at 10,000 feet, which is much higher than the take-off altitude of 1,000 feet. This is to minimise the impact of missing data on the algorithm's accuracy. Figure 3.8 shows that higher landing thresholds are much less vulnerable to missing data, although they record landing times slightly too early. We set the take-off threshold to pick up as many take-offs as possible while accounting for the issue that some airports are above sea level. We would not want to record planes that are moving on the ground, while sending ADS-B messages, as take-offs.

Applying this method, we extract 67 million separate flights and record their take-off and landing time as well as corresponding locations. Next, we aggregate these flights by month and airline to generate estimates for published airline statistics. This further reduces our dataset to 303 monthly airline flight counts for the UK, and 405 for the US. We include the largest 13 UK and 15 US airlines, with the cut-off for airline inclusion set at 1% of total air traffic in each country. Overall, our ADS-B data captures 83% of UK flights and 41% of US flights since July 2016.

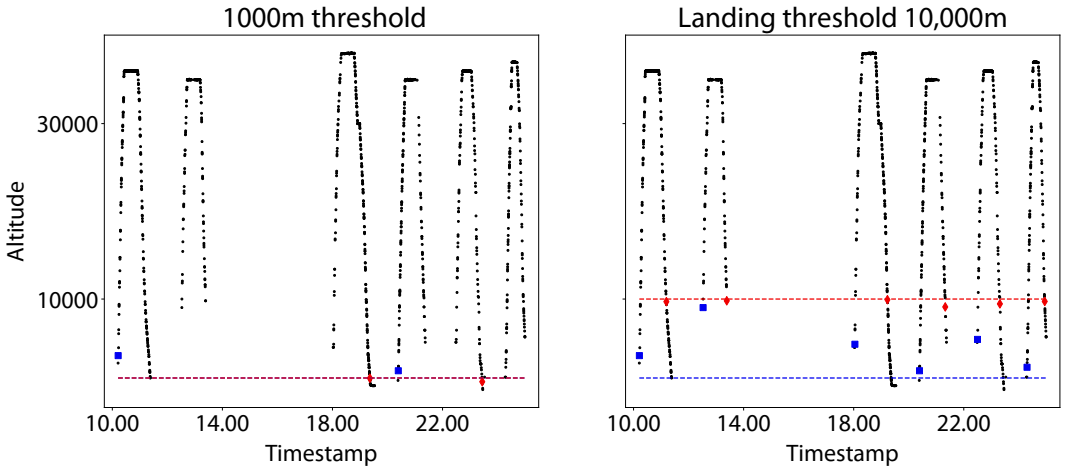

Figure 3.8: **Impact of raising the landing altitude threshold.** This figure demonstrates the importance of setting a landing altitude threshold well above sea level. The depicted aircraft makes six flights over the course of the day. The left panel shows that a low altitude threshold causes the algorithm to miss several landings, counting only two flights. The right panel shows that raising the altitude threshold helps correct this, as the algorithm now records six flights.

3.2.2 Published aviation statistics

Both the UK [Civil Aviation Authority, 2020] and US [Federal Aviation Authority, 2020] aviation authorities publish monthly airline statistics. They contain a range of performance indicators, such as flight volume and capacity utilisation, but are currently released with a three month delay. Figure 3.9 depicts the monthly percentage change in both the airline statistics and in flight volumes calculated using the ADS-B data. Visual inspection suggests there is a strong correlation.

However, there is seasonality for both countries. Figure 3.10 shows the level of both series. There are consistently more flights in summer, likely due to school holidays, than winter. Figure 3.11 shows that there is no obvious seasonality in annual percentage changes. We account for this seasonality by transforming longer time series into annual percentage changes. Figure 3.11 shows this effectively deseasonalises the data.

Finally, we collect economic data from the *UK Office for National Statistics* (Office for National Statistics) and *US Bureau of Economic Analysis* (United States Bureau of Economic Analysis). Both the ONS and BEA publish a GDP series that is split by industry, from which we consider air transport. The UK series is a monthly time series dating back to 1997, and the US series is a quarterly time series dating back to 2005.

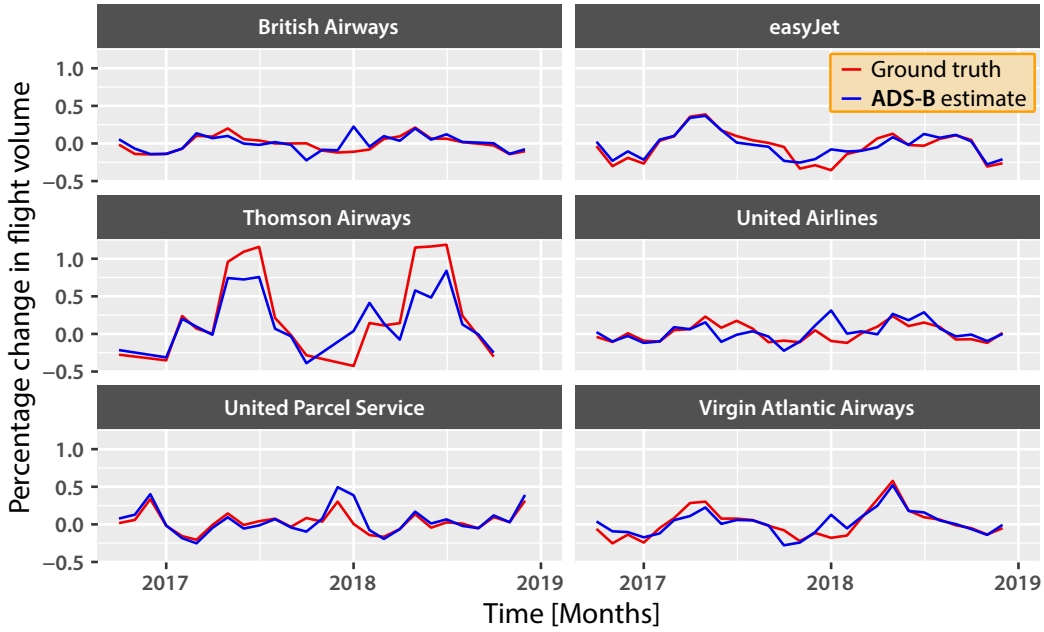

Figure 3.9: **Comparing official aviation statistics to measures derived from ADS-B data.** Plots of the percentage change in official flight count and ADS-B flight count for six UK and US airlines. The ADS-B estimate of the monthly change tracks the official statistics very closely, although there is some variation in accuracy across airlines. We suggest the two most likely sources of error are the database that maps aircraft to airline, and imperfect coverage of the ADS-B data.

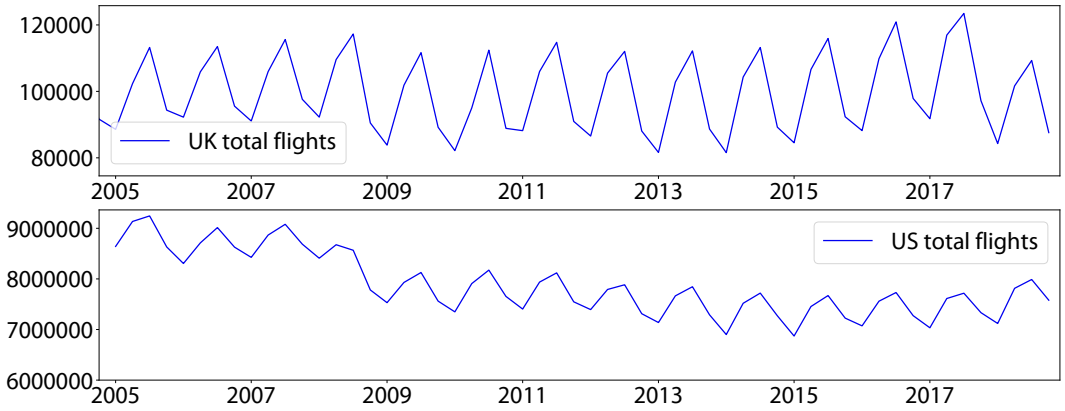

Figure 3.10: **Total flights series in levels.** This plots depicts aggregated flight volumes in the levels space. Data is from official statistics on flight volume, reported by airlines. The upper panel depicts UK aggregate flight volume over time. The lower panel depicts US aggregate flight volume over time. There is clear seasonality in the aggregate series, with consistently more flights in summer than winter.

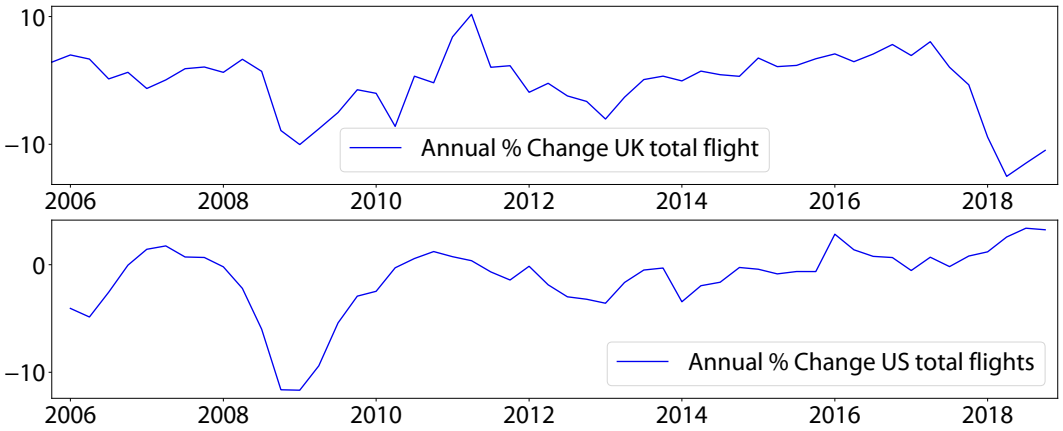

Figure 3.11: **Total flights series in annual percentage changes.** This plots depicts aggregated flight volumes after transforming to annual percentage changes. Data is from official statistics on flight volume, reported by airlines. The upper panel depicts the annual percentage change in UK aggregate flight volume over time. The lower panel depicts the annual percentage change in US aggregate flight volume over time. The transformation successfully de-seasonalises the data.

When analysing these series, we consider key time series properties, such as stationarity. Analysis of nonstationary series can lead to spurious conclusions, so it is important to transform series into a stationary space prior to analysis. Figure 3.12 depicts this visually, and these series both fail to reject the null of the ADF test. Figure 3.13 shows that converting these series to annual percentage changes makes them far more stationary. Both series now clearly reject the null of the ADF test, with p-values less than 0.01.

We considered using other data sources in the economic nowcasting model, such as the Purchasing Managers Index (PMI). These are released at higher frequency than GDP and have been shown to add value in prior nowcasting papers [Giannone et al., 2008]. However, here we are only nowcasting aviation's contribution to GDP, rather than GDP as a whole, and we cannot subset the PMI to just the aviation sector. The ADS-B data is a very strong domain match in this case, for which we did not feel there were any alternatives available in real-time. Therefore, we did not decide to include other data sources, such as the PMI, in our GDP model.

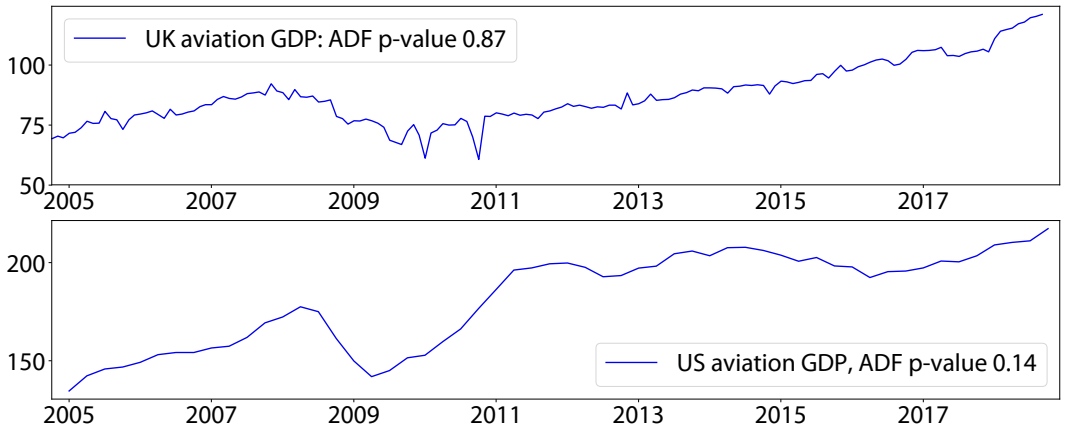

Figure 3.12: **GDP series in levels.** This plot shows aviation's contribution to GDP. The upper panel shows the UK's contribution, and the lower panel shows the US. We perform an ADF test for stationarity on both series. In neither case can we reject the null hypothesis of nonstationary data. Therefore, inference on these series in levels risks spurious results.

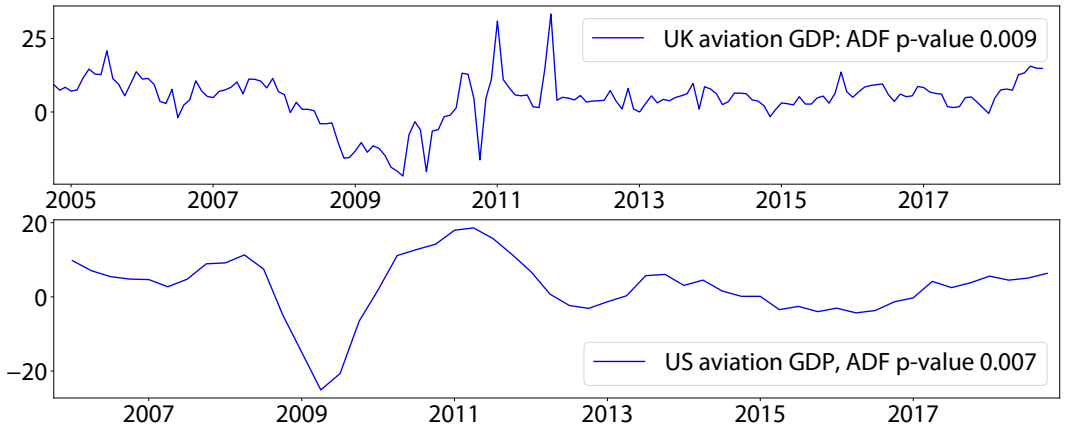

Figure 3.13: **GDP series in annual percentage changes.** This plot shows the annual percentage change in aviation's contribution to GDP. The upper panel shows the UK's contribution, and the lower panel shows the US. We perform an ADF test for stationarity on the annual percentage change in both series. In both cases, we find p-values of less than 0.01. This provides strong evidence of stationarity after transformation to annual percentage changes. This enables valid inference on these series.

3.3 Results

3.3.1 Estimating airline flight volume

For each airline, we aim to generate rapid estimates of flight volume across time. Some airlines are much larger than others. To ensure comparability between airlines,

we therefore normalise the flight volume data by indexing the first period to each airline for 100. We then re-scale subsequent periods so they are measured relative to the first. An airline whose original flight counts were (5000, 6000, ... 6500) would be normalised to (100, 120, ... 130). A normalised flight volume of 120 reflects a flight volume 20% higher than the first period.

A reasonable baseline model would be an autoregressive (AR) model where we estimate normalised airline flight volumes with their own history:

$$y_{i,t} = \alpha_i + \gamma_t + \beta y_{i,t-3} + \epsilon_{i,t}, \tag{3.1}$$

where $y_{i,t}$ is the number of flights and $\epsilon_{i,t}$ is a noise term for airline i in month t. Due to the three month publication lag for the official flight volume statistics, when nowcasting the flight volume for month t we only have official data from month $t-3$. The baseline therefore includes an AR(3) term, $y_{i,t-3}$ and β is the weight on the AR(3) term.

We also derive binary ("dummy") variables from the longitudinal data structure. γ_t are coefficients for dummy variables for each month (12 in total), which proxy for seasonality. A positive value for γ_t would reflect that flight volumes are usually higher than average in month t. α_i are coefficients for 28 airline-specific dummy variables, which capture the airline's average growth over time. A positive value of α_i would reflect an increase in the mean flight volume for airline i across the time period.

To measure the performance boost from ADS-B data, we add this data to the baseline model. Denoting $x_{i,t}$ as the ADS-B flight count for airline i in period t, and δ as the weight on the ADS-B term:

$$y_{i,t} = \alpha_i + \gamma_t + \delta x_{i,t} + \beta y_{i,t-3} + \epsilon_{i,t} \tag{3.2}$$

Table 3.1 shows that adding ADS-B data boosts the in-sample accuracy of all baseline models, regardless of whether month and airline dummies are included. This shows that ADS-B data can help estimate dynamic airline-specific changes in flight volume, and does not only proxy seasonality or differences in airline growth.

Table 3.2 provides a fuller breakdown of model performance by dummies included. We note that airline and month dummies both individually boost model performance substantially. However, adding ADS-B data still substantially boosts model performance in all specifications.

Our results so far suggest that ADS-B improves in-sample estimates of airline flight volume. However, in-sample scores may overstate true predictive accuracy. We

Table 3.1: **Estimating airline flight volume: in-sample results.** In-sample adjusted R^2 scores from models built to generate rapid estimates of airline flight volume. All models are unpenalised linear regression. The baseline model is autoregressive: it estimates the change in each airline's monthly flights using the most recently available flight count statistics. The ADS-B model additionally includes the ADS-B estimate of each airline's monthly flights as a predictor. The simple model does not include any further predictors. The complex model includes binary variables for each airline and month, to capture seasonality and differences in airline growth across the period. ADS-B data boosts performance across all model specifications, including against the more complex baseline. This shows that ADS-B data can help estimate dynamic airline-specific changes in flight volume, and does not only proxy seasonality or differences in airline growth.

	UK		US	
Model	Simple	Complex	Simple	Complex
Baseline adj R^2	0.06	0.66	0.36	0.75
ADS-B adj R^2	0.54	0.90	0.56	0.89
Number of parameters	2	27	2	29
Number of airlines	13	13	15	15

cannot use a random train-test split to assess out-of-sample performance because time series data is not independently and identically distributed (i.i.d). A random train-test split would put data in the training set that occurs after the data in the testing set. We would therefore use data from the future to fit a model that estimates the past, which would not be a valid measure of out-of-sample accuracy.

Instead, we use adaptive nowcasting [Preis and Moat, 2014] to measure out-of-sample accuracy. For each period t in our dataset, we use periods $\in [1, t-1]$ as our training set. The trained model then estimates the flight volumes for each airline in period t. We record the mean absolute error (MAE) across airlines, and that is the test score for period t. Each time we increase t, we re-fit the model to add new training data (which is why we call it "adaptive"). This procedure only uses past data to predict the present, so we know performance is out-of-sample.

The models with firm and time dummies have many parameters. This could lead to overfitting, which would reduce out-of-sample performance. We therefore regularise our adaptive nowcast models using LASSO regression. Let β be the vector of parameters in a linear forecasting model for T time periods and N airlines:

$$\beta_{LASSO} = min\{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t} - \beta X_{i,t})^2 + \lambda\|\beta\|\} \tag{3.3}$$

where $y_{i,t}$ is the flight count for airline i in period t, $X_{i,t}$ is their feature

Table 3.2: **Estimating airline performance: in-sample results with different dummy choices.** In-sample adjusted R^2 scores from predicting each airline's monthly flights. All models are unpenalised linear regression. The baseline model makes predictions without ADS-B data and the ADS-B score is after adding ADS-B data. Model 1 uses only airline's flights from the previous period as a predictor. Model 2 adds 12 month dummy variables to proxy for seasonality, model 3 adds airline dummy variables, and model 4 uses both.

	UK				US			
Model	1	2	3	4	1	2	3	4
Baseline R^2	0.06	0.37	0.39	0.66	0.36	0.53	0.61	0.75
ADS-B R^2	0.54	0.82	0.73	0.90	0.56	0.78	0.71	0.89
Dummies	None	Airlines	Months	All	None	Airlines	Months	All
N parameters	2	15	14	27	2	17	14	29

vector. LASSO applies a linear penalty λ to the magnitude of each coefficient, which punishes more complex models. It also allows automatic variable selection as the linear penalty results in many zero parameters. The weight to penalise complexity is determined by λ. We tune using 5-fold cross-validation across all data from period 1 to $t-1$ to find the optimal values of λ and β. Next we record the tuned model's predictions for period t, and measure the error for each of the N airlines. The performance score in period t is the mean absolute error (MAE) across the N airlines.

For both the UK and US, we construct a baseline adaptive nowcasting model which produces estimates for each airline. The baseline adaptive nowcasting model contains airline dummies. We made this choice because fitting month dummies with a short time dimension is difficult). Suppose we are nowcasting for December 2016. Our training data would include July to November 2016. As we had no observations for December, we would not be able to include the month dummies in the predictive model.

Now suppose we are instead nowcasting December 2017. We would have a month of prior airline observations of December to fit the month dummies with, so we would be able to include them. However, they would be very sensitive to any one-off events in December 2016 such as a global network disruption. Furthermore, the model would not be strictly comparable to December 2016 because it would include extra features. This would make comparing results across time more difficult.

The ADS-B model adds ADS-B data to the baseline adaptive nowcasting model. Figure 3.14 depicts our adaptive nowcasting results. Adding ADS-B data reduces the MAE by 29% for the UK (baseline MAE = 17.3, ADS-B MAE = 12.2)

and 18% for the US (baseline MAE = 7.4, ADS-B MAE = 6.1).

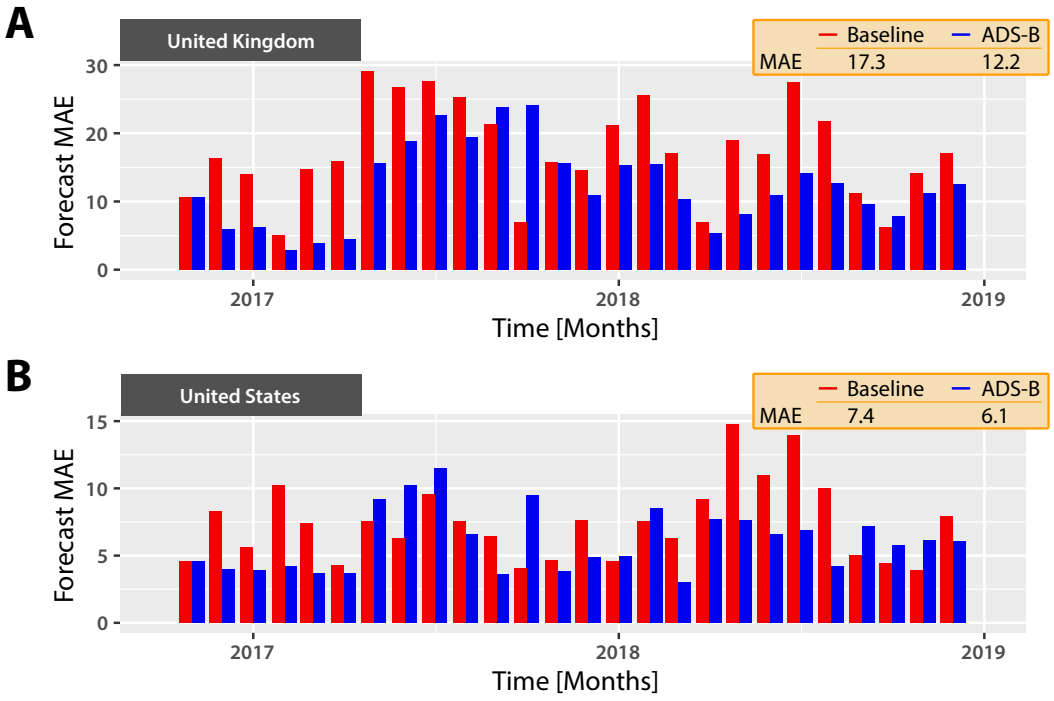

Figure 3.14: **Adaptive nowcasting of airline flight volume.** We build adaptive nowcasting models [Preis and Moat, 2014] to generate rapid estimates of flight volume for UK and US airlines. We investigate whether models enhanced with real-time ADS-B data deliver more accurate estimates than nowcasting models based on historic flight volume alone. For each month $t \in [5, T]$, we fit a model with all data up to $t-1$, then test performance in month t. The model is an autoregressive linear regression penalised using LASSO and tuned through 5-fold cross-validation. (A) Performance for the UK. (B) Performance for the US. The red series show the mean absolute error (MAE) from the baseline model with no ADS-B data, whereas the blue series show the MAE when adding ADS-B data. We find that including ADS-B data as a predictor improves rapid estimates of airline flight volume in both the UK and the US.

These results hold across a range of dummy specifications. Table 3.3 shows results from the other possible choices of dummies. Column 1 is the simplest model, with no dummies included. Column 2 adds month dummies, but not airline dummies. Colum 3 adds airline dummies, but not month dummies. Column 4 adds both airline and month dummies. Adding ADS-B data reduces MAE in all specifications. The reductions range from 20% to 26% for the UK, and 11% to 21% for the USA, which is consistent with our primary results using airline dummies only.

The means reported in this section are means across airlines. We also show

Table 3.3: **Adaptive nowcast results with varying dummies included.** Adaptive nowcasting results, as measured by MAE, with different dummy variable configurations. The model is LASSO. For each country, column 1 is the simplest model, with only the basic autoregressive predictor. Column 2 adds dummy variables for the airlines, which proxy for airline flight volume growth during the analysis period. Column 3 adds 12 month dummy variables to the model, which proxy for seasonality. Column 4 adds both month and airline dummy variables, which is the most complex specification.

	UK				US			
Baseline MAE	18.8	17.3	17.7	15.2	8.2	7.4	7.3	6.7
ADS-B MAE	14.0	12.2	13.9	12.3	6.9	6.1	6.5	5.3
Dummies	None	Airlines	Months	All	None	Airlines	Months	All

the distribution of errors, across airline, for each model over time in Figure 3.15. Our results are not noticeably being driven by outliers.

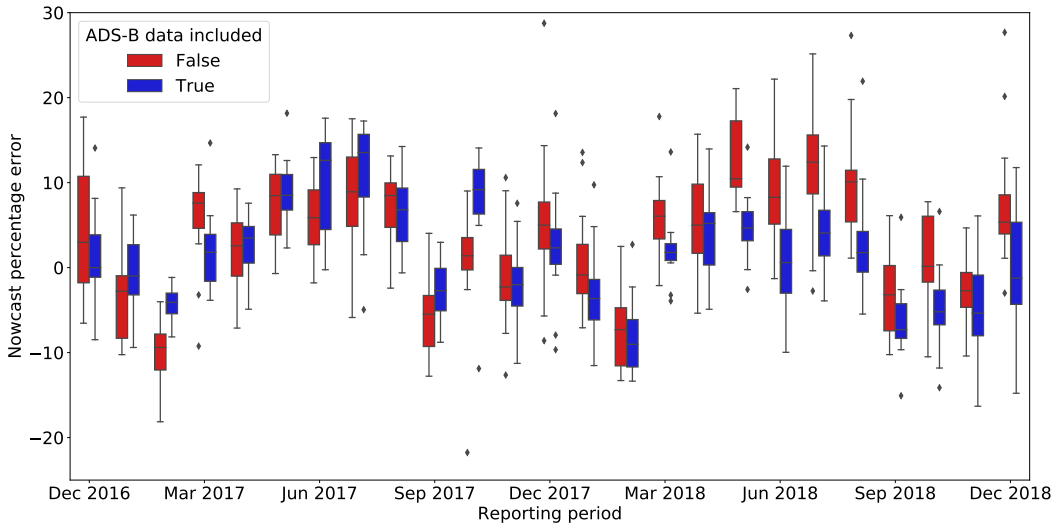

Figure 3.15: **Distribution of nowcast errors over time.** This plots show the distribution of errors, for each airline, over time. The scores we report in the main text are mean absolute errors (MAE), where we take the mean across airlines. ADS-B data reduces the mean of the distribution. Moreover, the lower means are not noticeably being driven by outliers.

We chose an expanding training window, rather than a fixed window. A fixed training window of w months means we lose the first $w + 1$ months from the sample. Given the short time dimension, our primary setup uses an expanding window instead. Each test period t, we train with all data up to $t-1$. This minimises data loss, but means later periods are trained on more data than earlier periods.

Table 3.4 reports results from a range of fixed training windows from 6 to 18 months. Adding ADS-B data substantially reduces nowcast MAEs in all cases and none of them qualitatively differ from the expanding window results. Therefore, our results are highly robust to varying the length of training window.

Table 3.4: **Adaptive nowcasting results with varying training windows.** This table shows adaptive nowcasting results with different training windows. The model is LASSO, with dummy variables for each airline. A fixed training window of w months means we lose the first $w+1$ months from the sample. Given the short time dimension, our primary setup uses an expanding window instead. Each test period t, we train with all data up to $t-1$. This minimises data loss, but means later periods are trained on more data than earlier periods. Increasing the training window reduces the size of the test set as the first test period becomes later. Here we report results from a range of fixed training windows from 6 to 18 months. None of them qualitatively differ from the expanding window results in the main text. With a 6 month window, the first period is May 2017, a 12 month window is November 2017 and an 18 month window is May 2018.

	UK			US		
Baseline MAE	21.4	14.7	17.0	7.6	7.3	8.2
Augmented MAE	14.8	11.6	12.0	6.6	5.7	6.2
Window Length	6	12	18	6	12	18

3.3.2 Estimating economic activity

We next analyse whether ADS-B data may help estimate aviation's direct contribution to GDP. Both the UK and US aviation GDP series are non-stationary based on Augmented Dickey-Fuller tests (UK: *Dickey-Fuller* = -1.4; US: *Dickey-Fuller* = -2.9; both ps > 0.05). Therefore their distributions are not constant over time, so we cannot use them for regression. Instead, we use the rolling annual percentage change in GDP, which deals effectively with both non-stationarity and seasonality (see Figure 3.13 for analysis of stationarity and Figure 3.11 for analysis of seasonality).

Our baseline specification for the annual percentage change in aviation's direct contribution to GDP Δz_t is

$$\Delta z_t = \alpha + \beta \Delta z_{t-j} + \epsilon_t, \tag{3.4}$$

where ϵ_t is a noise term. There is a two month publication lag for the first complete estimate of UK GDP, and a one quarter lag for the US. Therefore Δz_{t-j} is the most recent value known at month t, where $j = 2$ for the UK and $j = 1$ for the US. The

augmented model includes the rolling annual percentage change in ADS-B flight volume Δx_t:

$$\Delta z_t = \alpha + \beta \Delta z_{t-j} + \gamma \Delta x_t + \epsilon_t \tag{3.5}$$

The in-sample results are promising: adding ADS-B data boosts adjusted R^2 from 31% to 55% for the UK and 12% to 42% for the US. Table 3.5 shows further details of these results. There is a large boost in R^2 for both the UK and US. However, there are only 18 monthly periods for the UK and 6 quarterly periods for the US due to the limited time series of ADS-B data. These sample sizes are clearly too small to assess out-of-sample performance with an adaptive nowcasting model. We cannot construct longer time series as ADS-B data has only been available since July 2016.

Table 3.5: **In-sample results for nowcasting GDP.** This table shows full in-sample results for nowcasting GDP. All models are unpenalised linear regression. Baseline is estimation without ADS-B data and ADS-B score is after adding ADS-B data. There is a large boost in R^2 for both the UK and US. However the sample sizes are possibly too small for valid inference. There are only 18 monthly observations for the UK, and 6 quarterly observations for the USA. We cannot construct longer time series as ADS-B data has only been available since July 2016.

Country	UK	USA
Baseline R^2	0.31	0.12
ADS-B R^2	0.55	0.42
Sample size	18	6
Frequency	Monthly	Quarterly

We previously showed that ADS-B data could help estimate official flight volumes. To obtain greater insight into whether ADS-B data can improve nowcasts of aviation's direct contribution to GDP, we therefore substitute the official airline flight volume series in place of the ADS-B data. The official airline series are available for the full period for which we have aviation GDP data for both the UK (from 1997) and the US (from 2005). We again use adaptive nowcasting, but with fixed training window lengths of 60 months for the UK and 8 quarters for the US.

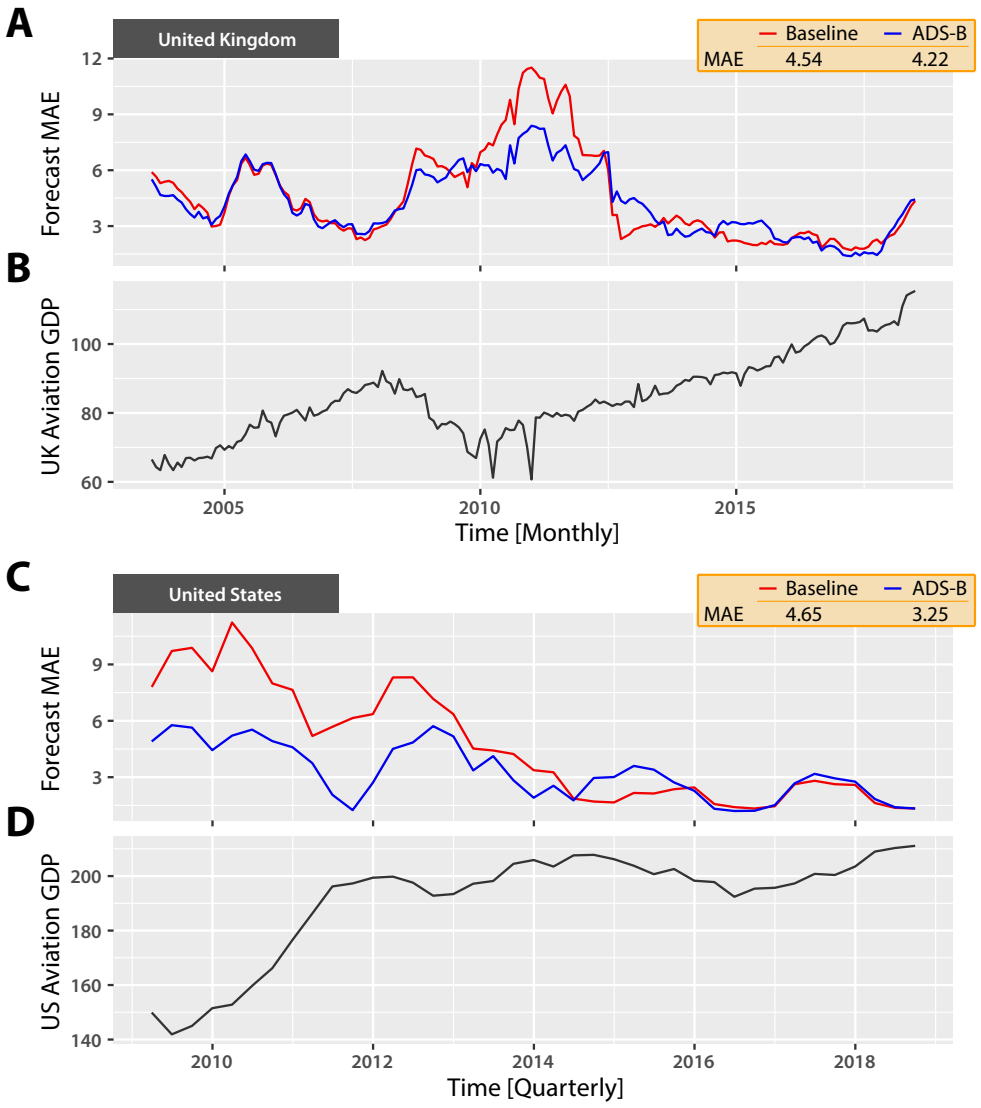

Figure 3.16: **Adaptive nowcasting of aviation's direct contribution to GDP.** We build adaptive nowcasting models [Preis and Moat, 2014] to generate rapid estimates of aviation's direct contribution to GDP in the UK and the US. We investigate whether models enhanced with real-time flight volume data would deliver more accurate estimates than nowcasting models based on historic GDP data alone. Given a training window w, for each period $t \in [w, T]$ we fit a model with all data $\in [t-w, t-1]$, then test performance in period t. (A) Adding flight volume data reduces the UK nowcast MAE by 7%. (B) Aviation's direct contribution to UK GDP. Visual inspection suggests that flight volume data delivers the greatest improvements in estimates during volatile economic periods, such as the period from 2008 to 2012. (C) Similarly, adding flight volume results in the MAE decreasing by 30% in the US. (D) Aviation's direct contribution to GDP in the US. Again, we see that the US model is most improved by flight volume data during the volatile economic period until 2012.

Figure 3.16 depicts out-of-sample results for GDP estimation. Adding real-time flight volume data reduces the MAE by 7% for the UK (baseline MAE = 4.54, ADS-B MAE = 4.22) and 30% for the US (baseline MAE = 4.65, ADS-B MAE = 3.25).

The improvement is greatest during the 2008-2010 financial crisis and the 2012 Euro crisis. Table 3.6 shows the performance split by time period. The baseline model is relatively strong outside the crisis period, as the autoregressive component is likely a stronger predictor. However during the crisis, when GDP is more volatile, the baseline model is much weaker. The augmented model, that adds real-time flight data, becomes relatively stronger. It reduces MAE by 20% for the UK and 46% for the US. This may be because the baseline AR model performs worse during volatile economic times.

Table 3.6: **GDP results: split by crisis against non-crisis.** This table shows the performance split by time period. We report that the augmented GDP model performs better relative to the baseline during volatile economic times. The baseline model is relatively strong outside the crisis period, as the autoregressive component is likely a stronger predictor. However during the crisis, when GDP is more volatile, the baseline model is much weaker. The augmented model, which includes real-time flight data as a predictor, becomes relatively stronger. It reduces MAE by 20% for the UK and 46% for the US.

	UK		US	
Time Period	Crisis	Non-crisis	Crisis	Non-crisis
Baseline MAE	8.0	3.2	8.1	2.3
Augmented MAE	6.4	3.3	4.4	2.4
Sample size	48	131	17	25
Frequency	Monthly	Monthly	Quarterly	Quarterly

For our main results, we chose a fixed training window of 60 months for the UK, and 8 quarters for the US. Longer training windows allow for more data to fit the model. However, we lose testing data as there is insufficient data to nowcast earlier time periods. Table 3.7 presents results from varying the training window. In each case, including data on real-time flight volume reduces nowcast MAEs. Our results are qualitatively unchanged by both shorter and longer training windows, across both the UK and US.

3.4 Discussion

We have assessed whether ADS-B data can help nowcast aviation statistics, which are currently published with a three month delay. We first show that ADS-B data

Table 3.7: **GDP nowcasting results with a varying training window.** The main text shows results from training windows of 60 months for the UK and 8 quarters for the US. This table presents results from varying the training window. For both countries, the second column is the training window length reported in the main text. In the first and third columns for each country, we show how the results change if the training window is increased or decreased by 6 months (2 quarters). Our results are qualitatively unchanged by both shorter and longer training windows, across both the UK and US. Therefore, our results are robust to varying the training window.

	UK (monthly)			US (quarterly)		
Window	54	60	66	6	8	10
Baseline MAE	4.63	4.54	4.52	4.90	4.65	4.03
Augmented MAE	4.41	4.22	4.13	4.42	3.25	3.10
Sample size	197	191	185	44	42	40

can accurately estimate airline performance, as measured by their flight volume. Second, we show that real-time knowledge of flight volume is a leading indicator for aviation's direct contribution to GDP. We find that this indicator is of greatest value during volatile periods, such as the crises between 2008 and 2012. Crisis periods are when rapid estimates for GDP are most crucial for policymakers, as they must take decisions quickly. In certain crises, such as disease outbreaks, real-time information on flight volumes may also be important beyond the economic domain.

The main limitation of our analysis comes from the novelty of ADS-B data. We do not have a long enough time series to determine whether ADS-B data, which we only had access to from July 2016 onward, can directly nowcast GDP out-of-sample. Future work will have access to a longer ADS-B time series and could therefore better evaluate out-of-sample performance. Chapter 4 benefits from a slightly longer time series, as we extend the ADS-B data out to April 2020 to cover the beginning of the Covid-19 crisis. Continued monitoring of ADS-B data will also be important in case its value as an economic indicator changes. For example, airlines who knew ADS-B data were being used to assess their performance may instruct pilots to fly differently. This however seems unlikely given the probable costs of flying more erratically. ADS-B data is therefore likely to be less prone to manipulation than other nowcasting data, such as internet search activity.

Finally, our analysis is restricted to aviation which comprises only 3-5% of GDP in total, including indirect contributions which are not analysed here. However, our methods could be extended to other sectors of the economy where data is shared at a similar level of granularity. For example: Chapter 4 analyses whether these

methods can be extended to estimating airport flight volumes, which are also an important part of aviation's infrastructure. Chapter 5 applies similar longitudinal methods to analyse whether we can nowcast illicit drug demand with real-time internet data. As the availability of real-time data increases, we could develop more accurate estimates of a large enough number of economic sectors to build a complete, real-time picture of the economy. In turn, policymakers would be able to respond more effectively to future crises.

Chapter 4

Nowcasting airport traffic with aircraft location data

4.1 Introduction

In Chapter 3, we analyse whether ADS-B data could be used to nowcast airline performance and aviation's contribution to GDP. This chapter extends the analysis to airports, as air traffic volumes are also crucial to airport financial performance. When publishing annual reports, airport's headline numbers are often their annual growth in passengers and flight volumes. Statistics on airport flight volumes are currently available in the UK, but are published with a 3 month lag. However, as discussed in Chapter 3, there has been a publicly accessible ADS-B database with global coverage since 2016. In this chapter, we exploit the geographic granularity of ADS-B data to show that it can help estimate airport traffic in real-time. We build on the analysis in Chapter 3 by using the latitude and longitude profiles of flights, as well as their altitude, to match them to airports. In turn, this could be a leading indicator for airport financial performance.

We also significantly extend the time series of the data relative to Chapter 3. We consider a longer time period, extending the end of our sample from December 2018 to April 2020. This covers the beginning of the Covid-19 crisis, which caused unprecedented volatility for the aviation sector. The longer time series allows us to build a real-time dataset containing some 117 million flights between July 2016 and April 2020, relative to the 67 million flights analysed in Chapter 3. Our dataset contains information on the precise location of both the takeoff and landing of each flight.

The geographic granularity of our dataset could be exploited in future re-

search on estimating trade volumes. Many countries publish granular trade data that is split by both airport of arrival and trading partner. Our real-time flight dataset may serve as a useful nowcasting indicator for changes in trade volumes.

In this chapter, we are conscious of the methodological issues with some of the nowcasting literature. Similarly to Chapter 3, we compare our ADS-B model to several plausible choices of baseline model in this chapter. We also allow the predictive model to change over time in a process called "adaptive nowcasting" [Preis and Moat, 2014], which is consistent with the methodology from Chapter 3. Taken together, these choices allow for a rigorous assessment of the out-of-sample predictive power of ADS-B data.

4.2 Data

4.2.1 Airport statistics

The UK requires airports to disclose their volume of flights each month. These disclosures are submitted to the Civil Aviation Authority [Authority], who publish monthly tables containing all UK airports. The publications have a 3 month lag: to receive January 2019 data, one must wait until April 2019.

The US does not require airports to make such disclosures. However, they require airlines to make detailed disclosures of their flight volumes to the Federal Aviation Authority [FAA, 2020]. Similarly to the UK, the FAA publish monthly tables with a 3 month lag. These must be split by origin and destination airport, so it is possible to construct airport flight statistics from the airline disclosures.

Finally, we manually assemble a dataset of major global airport locations. This is necessary for matching the ADS-B takeoffs and landings to an airport. ADS-B broadcasts do not contain reliable information on origin and destination airport, so we have to match manually. We describe this process in the next section.

ADS-B data

As in Chapter 3, we collect ADS-B data from the ADS-B Exchange. We provide a reminder of the structure of this data below.

In order to aid air traffic control, aircraft in the EU and USA must broadcast their position in real time under the ADS-B protocol. This has been mandatory for US commercial aircraft since 2020, and EU aircraft since 2017. Furthermore, other aircraft must be able to receive these broadcasts, so they are unencrypted. Thousands of aviation enthusiasts have also set up receivers to pick up the broadcasts.

They feed into a centralised database called the ADS-B Exchange, which has global coverage since July 2016.

The key fields from these broadcasts are the aircraft's location (altitude, latitude, longitude) and the timestamp. We extract these fields, along with a unique identifier for the aircraft and their operating airline, from around 20 billion broadcasts since July 2016. This information is sufficient to reduce the dataset from a row for each broadcast to a row for each journey. The journey contains both the aircraft's inferred takeoff and landing times, and takeoff and landing location. While there is a field in the ADS-B data indicating origin and destination airport, it is missing for many flights and the ADS-B Exchange warns that it is heavily prone to error.

The upper panel of Figure 4.1 depicts the reduction process for a given aircraft over a day of data. We record a takeoff if the aircraft clears 1,000 feet while ascending. Similarly, we record a landing if the aircraft crosses 10,000 feet while descending. When either a takeoff or landing occurs, we record the precise location of the aircraft. For this aircraft, we are able to reduce thousands of location broadcasts to just three flights. Chapter 3 describes this algorithm in more detail, along with relevant modelling choices. We run this algorithm over all the ADS-B data from July 2016 to April 2020, recording some 117 million flights globally. This extends the dataset from Chapter 3, where the end date for the analysis was December 2018 and we record some 67 million flights.

We further build on Chapter 3 by projecting this aircraft's flights into latitude and longitude space in the lower panel of Figure 4.1. The takeoff locations tend to be closer to the actual airport than the landing locations. This is because the landing altitude threshold is much higher, which is necessary due to missing data. Lower altitude thresholds therefore caused us to miss landings. We conclude that this is not worth a slight improvement in location accuracy.

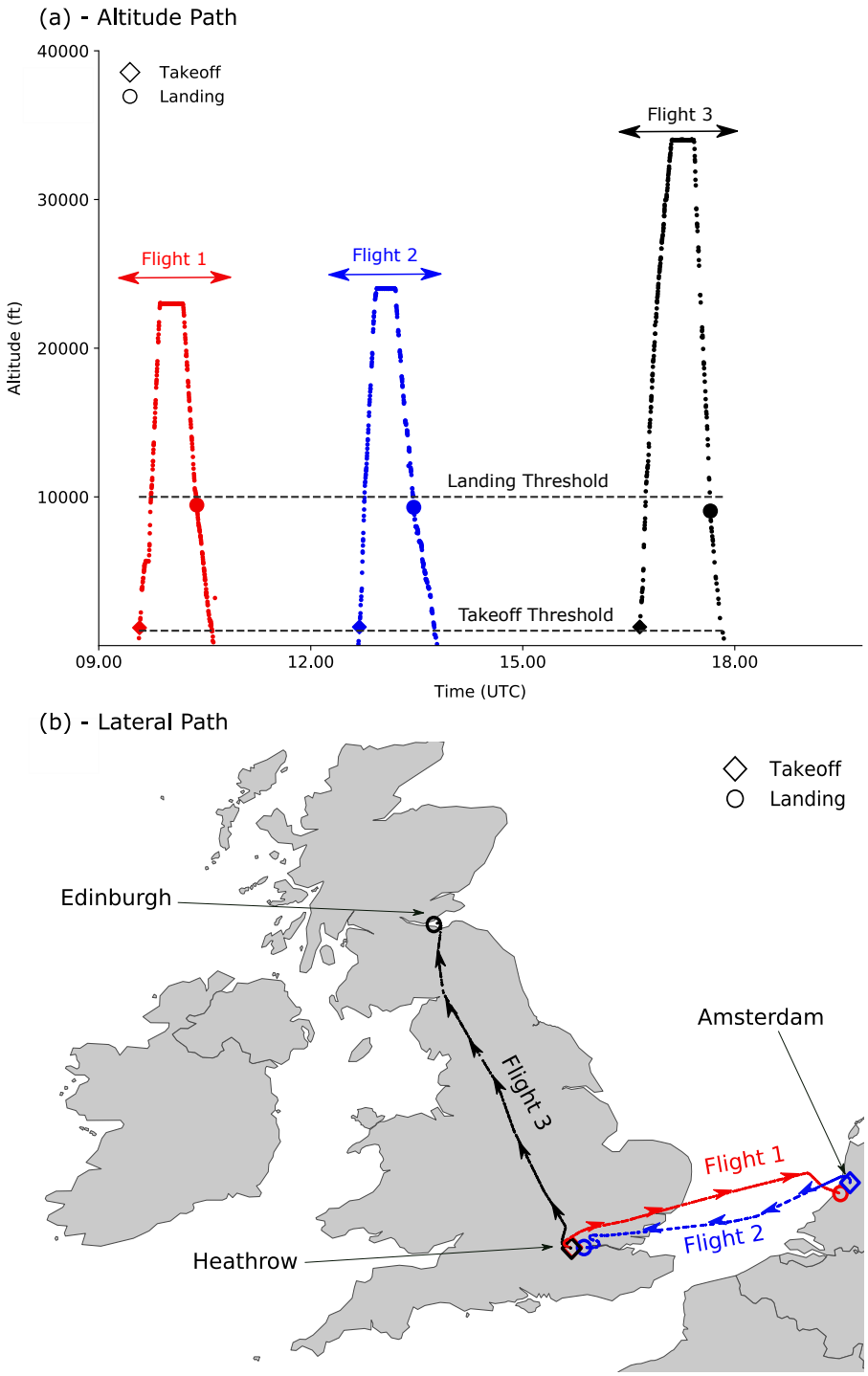

Figure 4.1: **Path of aircraft 40083B on 1 September 2018.** This figure shows how the flight counting algorithm tracks a given aircraft. Panel (a) shows the aircraft in altitude space, and panel (b) in latitude and longitude space. As in Chapter 3, the takeoff threshold is set to 1,000 feet, and the landing threshold is 10,000 feet. Unlike in Chapter 3, the recording of takeoff and landing location is critical for matching the flight to an airport. We record three takeoffs at 09.30, 13.00 and 17.00, and three landings at 10.30, 13.30 and 18.00. The first flight is from Heathrow to Amsterdam and the second flight is the return. The third flight takes off again at Heathrow, and lands in Edinburgh. The slight imprecision in landing locations is due to the higher landing threshold, shown in panel (a).

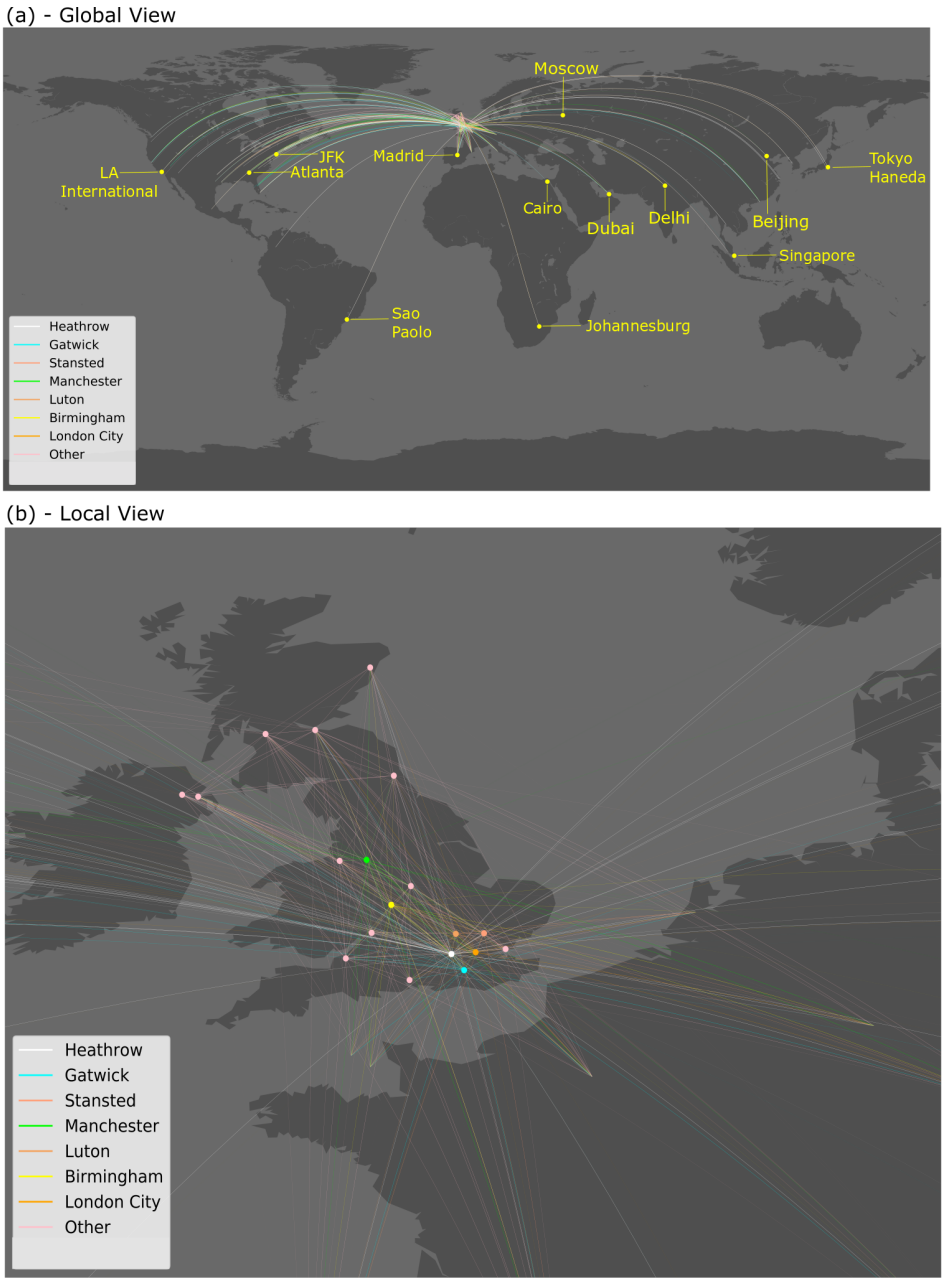

Figure 4.2: **Recorded flights during September 2018.** This figure shows flight paths originating in the UK during September 2018 that were picked up from ADS-B data. The upper panel is a global view, with some major hub airports such as New York and Dubai plotted as solid yellow circles. Different arc colours mark the major UK origin airports. We analyse the 18 largest airports, as measured by annual volume of flights in the UK airport statistics. Heathrow clearly dominates long haul traffic, but it is difficult to assess short haul due to the density of traffic around the UK. We therefore provide a closer view of the UK in the lower panel. The 18 UK airports are plotted as circles with colour matching their flight arc.

We next match each flight's takeoff and landing location to an airport. This is the airport with the shortest Euclidean distance, as measured by latitude and longitude. Figure 4.2 is a spatial snapshot of this process for flights originating in the UK during September 2018. The upper panel shows that long haul flight traffic is dominated by Heathrow, which is the UK's major hub, Gatwick and Manchester. Few other airports fly beyond Europe.

Because of the concentration in traffic, we provide a local view of UK flights in the lower panel of Figure 4.2. The ADS-B flight volumes do not always match the UK airport statistics well on aggregate. For example, smaller London airports seem to be overrepresented relative to some larger regional airports such as Manchester and Birmingham. This is likely because some airports are better covered by ADS-B receivers than others. Furthermore, some of the regional airports predominantly make intra-UK flights.

We restrict the matching process to only include airports with at least 1% of total traffic for their country. This leaves 18 UK airports and 21 US airports. This restriction is necessary because the takeoff and landing locations are recorded with some error. There are many very small airports and we may mistakenly attribute takeoffs to them that actually occurred at nearby major airports.

Variable coverage is a much bigger issue for the ADS-B data in the USA than the UK. Figure 4.3 depicts aggregate coverage for both regions. ADS-B coverage is much higher in the UK, and the aggregate series tracks the airport statistics more closely. The lack of coverage in the USA is problematic for comparing across airports. ADS-B data records fewer takeoffs at the largest airport in the USA (Atlanta) than the 6th largest (New York JFK), which has only half as many flights according to US airport statistics. Therefore we separate inference in Section 4.3 between the UK and US.

We have 18 UK airports and monthly data from July 2016 to April 2020. Given the 3 month delay in airport statistics, we drop the first 3 months of the sample. Four smaller airports had very poor ADS-B coverage so we also remove them from the sample. This leaves us with 14 UK airports and 43 time periods, for a UK sample size of 602 airport-months. In the US, after applying a similar procedure, we have 21 airports remaining. Given the same 43 time periods, we have a US sample size of 903 airport-months.

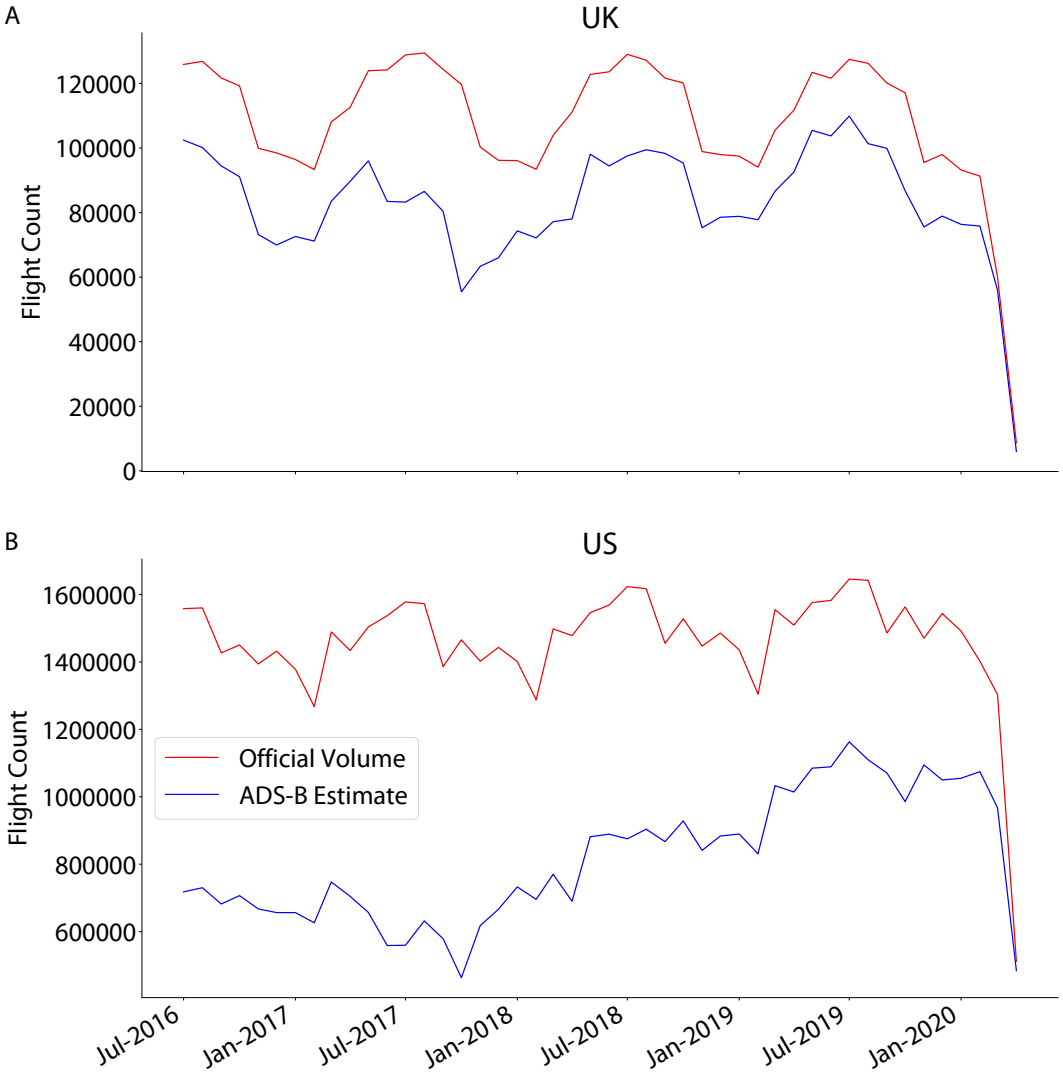

Figure 4.3: **Aggregate flight volumes for UK and USA.** This figure compares the aggregate flight volumes measured by ADS-B data and the airport statistics. The upper panel depicts the UK and the lower panel depicts the USA. In both cases, the blue series shows the airport statistics and the red series is the ADS-B estimate. The UK clearly has better coverage. Across the sample, the ADS-B estimate varies from 75% – 95% of the airport statistics and matches the seasonality closely. In contrast, US coverage varies from 40% – 60%. While it increases across the sample, it does not seem to match the aggregate seasonality well. Our sample runs from July 2016 to April 2020, which is longer than the sample in Chapter 3 that ended in December 2018. This allows us to cover the beginning of the Covid-19 crisis in April 2020, where we observe a sharp decline in volumes for both countries.

4.3 Results

4.3.1 Nowcasting airport flight volumes

In-sample results

Consistent with the methodology from Chapter 3, we first construct a full baseline model. The full baseline model is our best estimate of airport statistics, without using any ADS-B data:

$$y_{i,t} = \alpha_i + \gamma_t + \beta y_{i,t-3} + \epsilon_{i,t} \tag{4.1}$$

where $y_{i,t}$ denotes the volume of air traffic departing from airport i in month t. Due to differences in size across airports, we normalise $y_{i,t}$ by indexing the first period (July 2016) to 100 for all airports. The model is autoregressive (AR) - it includes the third lag of the airport statistics $y_{i,t-3}$ as this is known when estimating period t.

We engineer two sets of dummy variables using the longitudinal structure of the data. α_i are airport dummies, which estimate the average volume of flights departing from airport i across all time periods. γ_t are month dummies, which proxy for seasonal effects. As shown in Section 3.2.1 of Chapter 3, there is clear seasonality in the data with more flights in spring and summer.

The full model adds the ADS-B measure of airport traffic to the baseline model:

$$y_{i,t} = \alpha_i + \gamma_t + \beta_1 x_{i,t} + \beta_2 y_{i,t-3} + \epsilon_{i,t} \tag{4.2}$$

where $x_{i,t}$ denotes the ADS-B measure for airport i in period t. We first assess the value of ADS-B data by comparing accuracy between the baseline model and full model.

Table 4.1 presents a comparison between the baseline model and the full model, as measured by R^2. These are adjusted R^2 scores that penalise more complex models with more variables. Adding new variables can therefore reduce the R^2 score if the model does not become more accurate. The simplest model is the first column, which is an AR model with none of the firm or month dummies. The second column adds firm dummies and the third column adds month dummies. The final column is the most complex model, given by Equation 4.2, which includes both sets of dummies.

The ADS-B data raises the adjusted R^2 score of every specification. For the UK, the improvements are generally larger. The size of this boost varies considerably

Table 4.1: **Estimating flight volumes, in-sample results.** In-sample adjusted R^2 scores from predicting the number of monthly flights of each airport. All models are unpenalised linear regression. Baseline score is the prediction without ADS-B data and ADS-B score is from the same model after adding ADS-B data. Model 1 uses only flights from the previous period as a predictor. Model 2 adds month dummies to proxy for seasonality, model 3 adds firm dummies, and model 4 uses both sets of dummmies. Adding ADS-B data improves performance across all models, although the boost is much smaller when we include month dummies.

Model	Simple	Firm dummies	Time dummies	All dummies
UK				
Baseline R^2	0.06	0.53	0.08	0.55
ADS-B R^2	0.46	0.66	0.62	0.73
Number of features	2	45	16	59
USA				
Baseline R^2	0.01	0.24	0.17	0.41
ADS-B R^2	0.06	0.28	0.25	0.47
Number of features	2	45	23	66

across models. In the simplest setup, with no dummies, adding ADS-B data boosts R^2 from 6% to 46%. In the most complex setup, with both firm and month dummies, ADS-B data boosts R^2 from 55% to 73%. This provides evidence that ADS-B data is predictive even after controlling for firm-specific and seasonal variation.

The improvement is smaller for US airports. However, the improvement is fairly consistent across models. In the simplest setup, with no dummies, adding ADS-B data boosts R^2 from 1% to 6%. In the most complex setup, with both firm and month dummies, ADS-B data boosts R^2 from 41% to 47%. The smaller improvement may be due to poorer ADS-B coverage, as documented in Section 3.2.1.

A major caveat is that these are in-sample results. If the models overfits, the R^2 scores may overstate out-of-sample performance. This is a bigger issue for the more complex models with large numbers of dummies. We may be particularly concerned about the models with month dummies, given the short time series. A forecasting model would not be able to estimate these, because they are using data from the future to fit parameters estimating the present. Furthermore, longitudinal data is not usually independent and identically distributed (i.i.d) so we cannot assess out-of-sample performance with a random train-test-split. Instead, we use an adaptive nowcasting [Preis and Moat, 2014] to analyse whether ADS-B improves

out-of-sample performance.

Out-of-sample results

Consistent with Chapter 3, we use adaptive nowcasting to assess out-of-sample performance. We provide a reminder of this procedure below.

Adaptive nowcasting uses only data from the past to estimate the present. We first set a training window, w, which determines the length of past data in the training set. For each period t in our dataset, we use periods $\in [t-(w+1), t-1]$ as our training data. We then test the models by finding their mean absolute error (MAE) across airports in period t. This measures out of sample performance, because the testing set always occurs chronologically after the training set.

We re-train the model each time we increase t, because the training window moves and the training data changes. This adaptation is important as the relationship between our variables may change over time. For example, if ADS-B receiver coverage improves over time then it would become a stronger predictor. Optimally, the parameters would change to reflect that.

The length of the training window is a hyperparameter that we can tune. We choose a training window of 12 months. This does not yield the lowest error, but there is a trade-off between increasing the length of the training window and the size of the testing set. Results using training windows from 3 - 21 months are not qualitatively different.

Figure 4.4 presents adaptive nowcasting results. The baseline model is from the third column of Table 4.1, which includes firm dummies. However we do not include month dummies, because we can't fit them given the limited number of time periods in the training set. We need to be wary of overfitting, because there are a large number of firm dummies. Therefore we regularise our model using LASSO regression, and we tune using 5-fold cross validation on the training set. Chapter 3 provides a fuller description of this procedure.

Panel A of Figure 4.4 shows nowcast MAE for UK airports. ADS-B data reduces nowcast MAE, relative to the baseline model, in all periods. Across the whole sample, MAE falls by 45%. This is a bigger improvement than some of the in-sample models, suggesting that the baseline model overfit from the number of dummies. In turn, this may have limited the potential boost from adding ADS-B data. MAE also seems to be smaller in the second half of the sample, which may reflect improving ADS-B coverage.

Panel B of Figure 4.4 shows nowcast MAE for US airports. Adding the ADS-B data reduces nowcast MAE by 2.5% across the sample, which is much smaller than

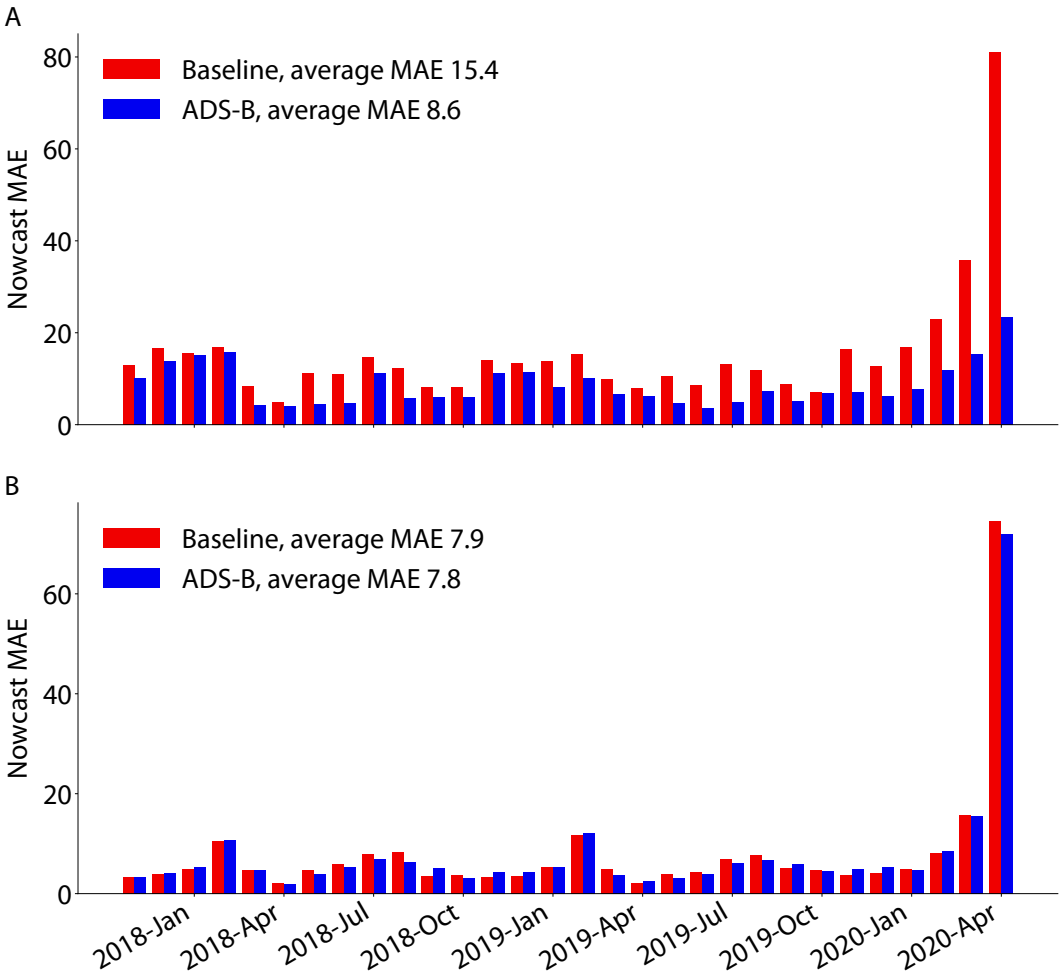

Figure 4.4: **Adaptive nowcasting results for UK and US airport flight volumes.** We use the adaptive nowcasting procedure with a 12 month training window. For each month t from in November 2016 to April 2020, the testing set is period t and the training set is months $\in [(t-13, t-1]$. Panel A shows nowcast mean absolute error (MAE) for UK airports. The red series show the MAE from the baseline model with no ADS-B data and the blue series shows MAE after adding ADS-B data. This reduces nowcast MAE in every time period, and by 45% across the sample on average. Panel B shows nowcast MAE for US airports. Adding the ADS-B data reduces nowcast MAE by 2.5% across the sample, which is much smaller than the improvement for UK airports. This is likely due to poorer US receiver coverage. For both countries, the reduction is greatest in April 2020, which is when the Covid crisis began.

the improvement for UK airports. This is likely due to poorer ADS-B coverage in the US, as ADS-B only became mandatory in 2020 for the US compared to 2017 in the UK.

The upper panel of Figure 4.5 shows US nowcast results if we train a model using only data from 2019 onwards. As shown in Figure 4.3, US coverage improves towards the end of our sample. The reduction in nowcast MAE is 32% in this subsample, which is much greater and comparable to the UK results. This provides further evidence that receiver coverage heavily impacts ADS-B model performance. We may expect continued improvement of the ADS-B model in future if receiver coverage continues to improve.

In both countries, the ADS-B model improves results the most in April 2020, when the Covid crisis sharply reduced air traffic. We depict an example of this in the lower panel of Figure 4.5, which shows flight volumes for JFK airport. The ADS-B model predictions are more variable than the baseline model, and generally track closer to the ground truth. This difference is particularly stark in April 2020. The baseline model completely misses the sharp contraction in true flight volumes, but the ADS-B model detects it. The ADS-B model is therefore able to effectively provide an early warning of declining traffic for JFK airport.

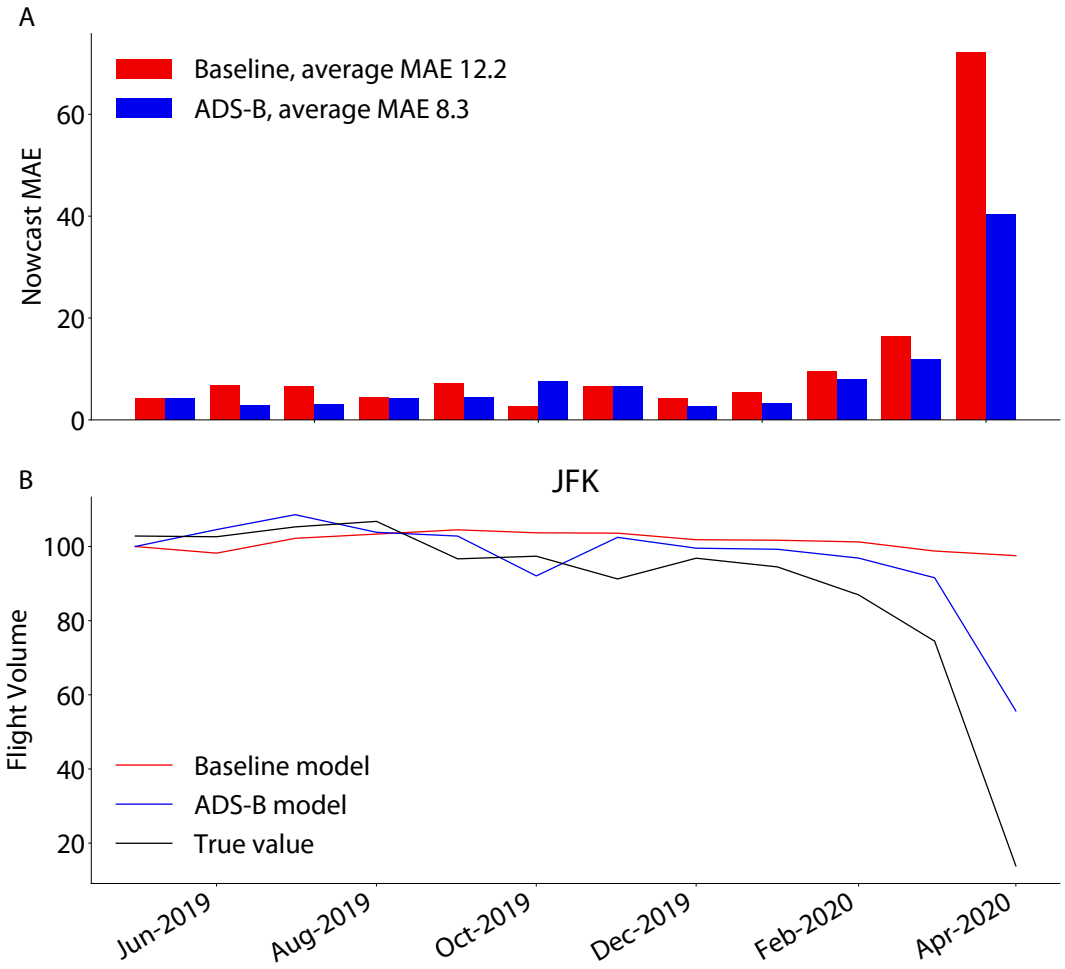

Figure 4.5: **Adaptive nowcasting results for US airport flight volumes after 2018.** Panel A shows adaptive nowcast results for the US when only using data after 2018. We use the adaptive nowcasting procedure: for each month t from May 2019 to April 2020, the testing set is month t and the training set is all months from January 2019 to $t-1$. Due to the short time series, we use an expanding training window. Later months therefore have more training data. The red series show the MAE from the baseline model with no ADS-B data and the blue series shows MAE after adding ADS-B data. This reduces nowcast MAE by 32% on average across the sample. The reduction is particularly large in April 2020, which is when the Covid crisis sharply reduced air traffic. Panel B depicts an example of nowcast output for JFK airport. The ADS-B model generally predicts closer to ground truth than the baseline model. This difference is particularly stark in April 2020, when the baseline model fails to depict the sharp fall in air traffic.

4.4 Discussion

In this chapter, we analyse whether real-time location data from ADS-B can improve estimates of UK and US airport flight volumes. We set up a range of baseline models, and find that ADS-B data boosts in-sample performance across all specifications. The size of this boost varies from 4% to 54%, depending on how many features we engineer from the longitudinal data structure. Adding seasonal effects in particular seems to make the baseline model much stronger, therefore reducing the gain from adding ADS-B data. These results build on similar findings in Chapter 3 which show that ADS-B is a in-sample strong predictor of airline flight volumes.

We find the in-sample improvements also apply out-of-sample. Using an adaptive nowcasting procedure, we find that ADS-B data reduces out-of-sample MAE by 45% for the UK. The improvement is smaller for the US, at just 2.5%. However, the performance boost grows over time. If we restrict the US data to 2018 onward, the improvement in MAE from including ADS-B data rises to 32%. Therefore ADS-B data may provide greater gains in the future.

The improvement in airport out-of-sample performance is not as large as the improvement from Chapter 3, which focused on airlines. This may be because the airport analysis is more vulnerable to varying geographic coverage. For example, suppose Manchester Airport was poorly covered with ADS-B receivers relative to London Heathrow. This would have a very direct impact on the accuracy of relative estimated flight volumes for these airports. However, the impact on relative airline flight volumes would be more limited, because the ADS-B data has a variable for matching flights directly to airlines.

Future research could further build on our results by finding applications for real-time knowledge of airport traffic. Some airports stocks are publicly traded, so ADS-B could be predictive over their stock prices. The data could also be useful in producing faster economic statistics. For example, real-time airport statistics would likely help to estimate the overall economic health of the aviation sector. We already showed in Chapter 3 how real-time data on airline flight volumes could be a leading indicator for aviation's contribution to GDP in the UK and USA.

Statistics authorities in both the UK and USA also publish granular data on international trade flows through each airport. These are further split by route e.g trade arriving in London Heathrow specifically from China. Future research could analyse whether changes in air traffic along each route, as measured by ADS-B, are a real-time indicator for changes in trade. This would provide powerful further evidence of the potential for real-time location data as a nowcasting indicator.

Chapter 5

Nowcasting drug demand with Wikipedia page views: evidence from darknet markets

5.1 Introduction

Chapters 3 and 4 find that real-time aircraft location data has economic value in nowcasting the aviation sector. We noted that extending these methods to other sectors may help build up a complete, real-time picture of the economy. This chapter analyses whether similar methods, applied to other real-time data from the internet, can be used to nowcast illicit drug demand. Given its black market status, this is a difficult sector to observe, so real-time information may be particularly valuable.

As well as being an economic issue, rapid changes in illicit drug use are also a major public health concern. In the USA, 30,000 people died from Fentanyl overdoses in 2018 alone [National Institute on Drug Abuse, 2019]. It has become harder for authorities to monitor illicit drug markets. This is partly caused by shifts in production and distribution channels [Demant et al., 2018]. Traditional global supply chains for organic drugs, such as cocaine and heroin, are being replaced by new supply routes for synthetic drugs, which can be produced anywhere [Dittus, Wright, and Graham, 2018; Smith and Garlich, 2013; Perdue, Hawdon, and Thames, 2018; Tracy, Wood, and Baumeister, 2017]. As a result, monitoring supply chains is now more challenging.

Changes in drug demand are also hard to monitor. Traditionally, authorities have relied on annual surveys, such as the United Nations Office on Drugs and Crime (UNODC) World Drug Report [United Nations Office on Drugs and Crime, 2019].

The low frequency of these statistics means that authorities may miss opportunities to intervene early in drug crises, such as the US Fentanyl epidemic [Higham et al., 2019]. New drug categories may not appear in such surveys at all [Smith and Garlich, 2013]. For example, there were at least 36 novel psychoactive substances discovered between January and August 2019 alone [EMCDDA, 2019]. A more frequent measure of drug use would enable public health authorities to intervene earlier, thereby using their limited resources more effectively.

To address these concerns, we present a novel method to nowcast drug demand based on high-frequency sales data from darknet markets. These are online markets that rely on encryption and digital currencies to enable anonymous trade of goods and services [Barratt and Aldridge, 2016]. We measure "nowcast" value by the out-of-sample nowcast errors of our models. As in previous chapters, "nowcasting" means estimating the current value of statistics that are usually released with a lag. As detailed in Chapter 2, this term was first coined for economic variables, such as GDP, inflation and migration [Giannone, Reichlin, and Small, 2008; Carriere-Swallow and Labbe, 2013; Lin, Cranshaw, and Counts, 2019]. We consider economic applications in Chapters 3 and 4, which nowcast the aviation sector using real-time aircraft location data. More recently, nowcasting has also been applied in epidemiology to estimate flu and dengue outbreaks [Ginsberg, Mohebbi, Patel, Brammer, Smolinski, and Brilliant, 2009; Preis and Moat, 2014; Olson, Konty, Paladini, Viboud, and Simonsen, 2013; Majumder, Santillana, Mekaru, McGinnis, Khan, and Brownstein, 2016; Mizzi, Preis, Bastos, Gomes, Codeço, and Moat, 2021]. Although we consider an epidemiology application in Chapter 6, accurate nowcasts of drug demand would also be useful, given the long lag between current annual surveys.

We find that nowcasting models based on historic sales alone cannot accurately nowcast drug demand. It is also difficult to scrape the markets and there are frequent outages [Molnar et al., 2010], so a measure based on darknet data alone could be unreliable. However, consumers may search for information on drugs before making a purchase [Su, He, Liu, Zhang, and Ma, 2018; Wightman, Perrone, and Nelson, 2017]. We therefore also collect data on Wikipedia page views for each drug, because these are reliably available in real-time [Yoshida et al., 2015]. We find that adding data on Wikipedia page views for these drugs dramatically improves the models' accuracy. Therefore, we can construct a more frequent measure of drug demand using Wikipedia data. Our Wikipedia model remains effective at weekly frequency, which is a substantial improvement relative to the current annual surveys. In turn, this could reduce public response times to future drug epidemics.

Nowcasting drug use with online behaviour is an emergent field with little directly comparable literature. One recent paper found a correlation between Google searches for novel psychoactive substances and their annual change in sales, as measured by the United Nations Office on Drugs and Crime surveys [Perdue et al., 2018]. However, this study could not assess whether this relationship holds out-of-sample due to the low time frequency of their sales data. Another study found a correlation between the volume of online comments about opioids, on the large forum Reddit, and the level of opioid abuse across US states [Balsamo et al., 2019]. This study was also limited to in-sample analysis by the low time-frequency of their data, and were geographically restricted to the USA. Our global drug demand data is available at higher time frequency, which allows us to evaluate model performance out-of-sample. Moreover, the darknet data is actual sales rather than drug user surveys, which are particularly vulnerable to response bias [Zhao et al., 2009]. This chapter is therefore the first evidence on predicting darknet drug sales, and also whether internet search can predict drug use out-of-sample.

Another growing strand of literature is using Wikipedia page views to predict economic variables. Previous papers have found relevance for traditional variables such as the stock market [Moat et al., 2013] and box office sales [Mestyan et al., 2013]. Some papers have also found Wikipedia data can predict cryptocurrency prices and Bitcoin trades [ElBahrawy et al., 2017; Kristoufek, 2013], which is particularly relevant given our use of darknet data. Finally, our results are consistent with prior findings that Wikipedia page views can predict other epidemics such as the flu [Generous et al., 2014].

5.2 Data

5.2.1 Darknet sales

When buying drugs on the darknet, it is common (although not always mandatory) to leave reviews. Each review has an associated vendor country, product and timestamp. Our data comes from scraping the reviews from the four largest markets (Alphabay, Hansa, Traderoute and Valhalla) during June and July 2017. These covered 80% of global trade at the time [Dittus et al., 2018].

Drug sales on the darknet have risen over time. This means the sales data may be nonstationary, which is problematic for assessing time series model performance [Cheung and Lai, 1995]. Figure 5.1 shows global darknet demand for MDMA. Panel A shows that sales over time are growing rapidly, so they may not be stationary. Panel B shows the percentage change in sales over time. While there is

volatility, the distribution looks more constant across time. We formally test for stationarity in Section 5.3.1.

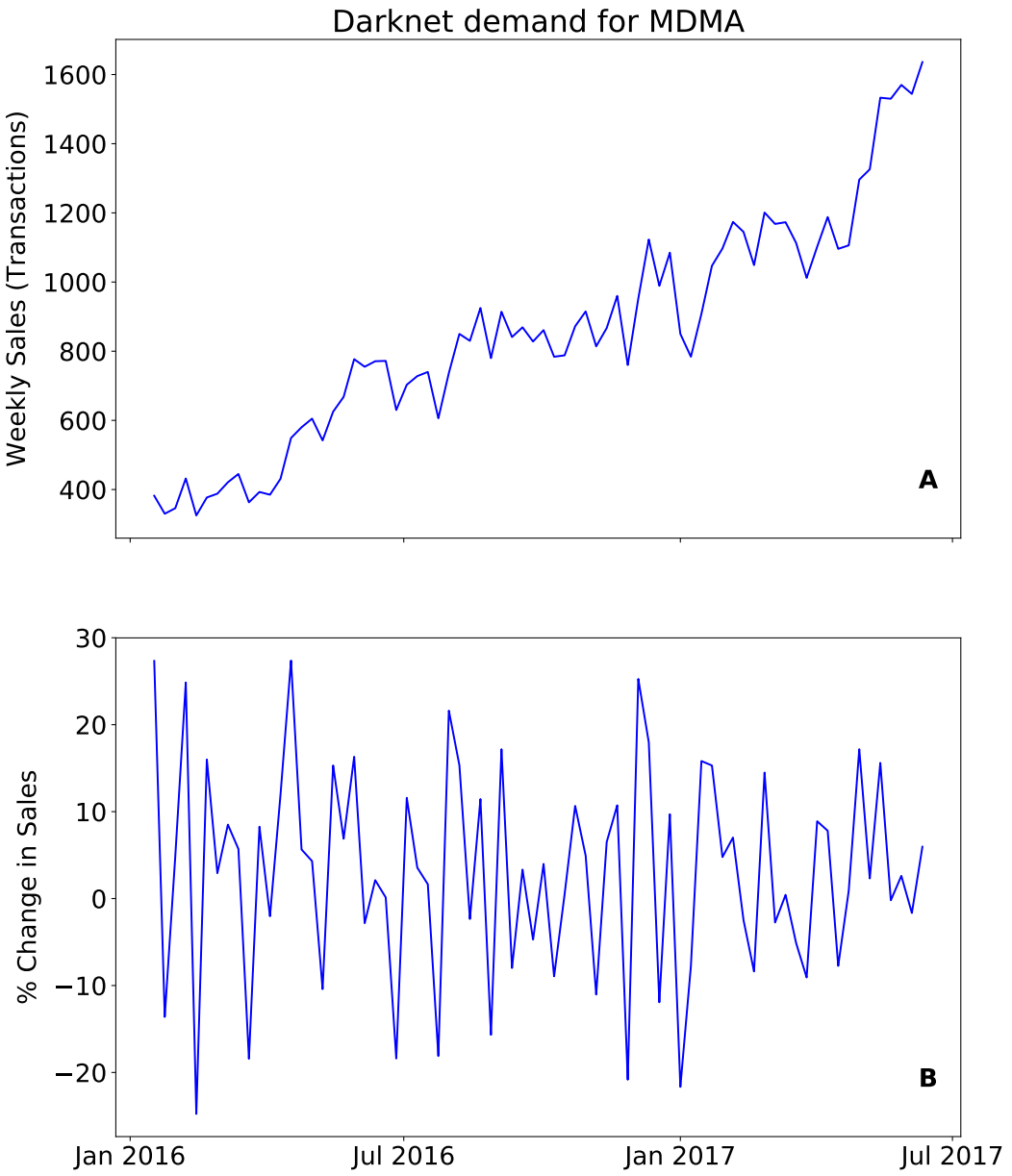

Figure 5.1: **Darknet MDMA sales over time.** This figure shows weekly sales of MDMA on the darknet over time. The unit is number of unique sales, so a given transaction is only one sale regardless of the volume of the transaction. Panel A shows that sales over time are growing rapidly, so sales may not be stationary. Panel B shows the percentage change in sales over time. While there is significant volatility, the distribution looks more constant across time. We formally test for stationarity in Section 5.3.1.

The sales data is transaction-level: for each review, we assume one sale. This is a lower bound on true sales, because it is not always mandatory for buyers to leave a review [Barratt and Aldridge, 2016]. The timestamps are continuous, so we could conduct our analysis at different levels of time aggregation. The higher frequency the aggregation, the faster the measure of drug demand would be.

However, higher frequencies make the sales data sparser with more zero observations. Figure 5.2 shows the distribution of percentage changes in drug sales, aggregated at different frequencies. The daily frequency distribution is clearly sparser than weekly, which in turn is sparser than monthly.

To manage this trade-off, we aggregate the sales data to monthly frequency, which is still much more frequent than the annual official surveys. We assess the robustness of our results to differing aggregation frequencies in Table 5.3.

A potential limitation with the scraped sales data is that it only captures drug listings that were still available from June to July 2017. If a vendor were to create a listing and remove it before June 2017, we would not observe any of the sales in the scrape. We could reduce the impact on our analysis by limiting our data to be as close to the scraping period as possible. For example, if we only consider sales from May to July 2017 then there would be far fewer removed listings. However this would also reduce our sample size. Instead, we use all available data for our analysis and assess the impact of restricting the sample period in Table 5.4.

5.2.2 Wikipedia views

We collect Wikipedia page views data through the Wikipedia API, which runs from July 2015 onward [Wikipedia API, 2019]. The raw data is available at daily frequency, but we aggregate to monthly frequency to match the sales data. We further split the data by language, and use that as a proxy for the country of the viewer. This is likely a reasonable assumption for some languages, as shown in [Generous et al., 2014]. For example, viewers of the Dutch language page are probably located in the Netherlands. However there may be measurement error, particularly for the English language page which we assume to cover several countries. We assess the results for each country in Section 5.3.3.

Figure 5.3 describes the Wikipedia data over the sample period. Panel A shows that total views are relatively stable over time. Panel B shows the distribution of views across languages, of which the English pages are unsurprisingly by far the most popular. Panel C shows the distribution of views across drugs, which is more evenly spread.

We considered using Google Trends as an alternative web search indicator.

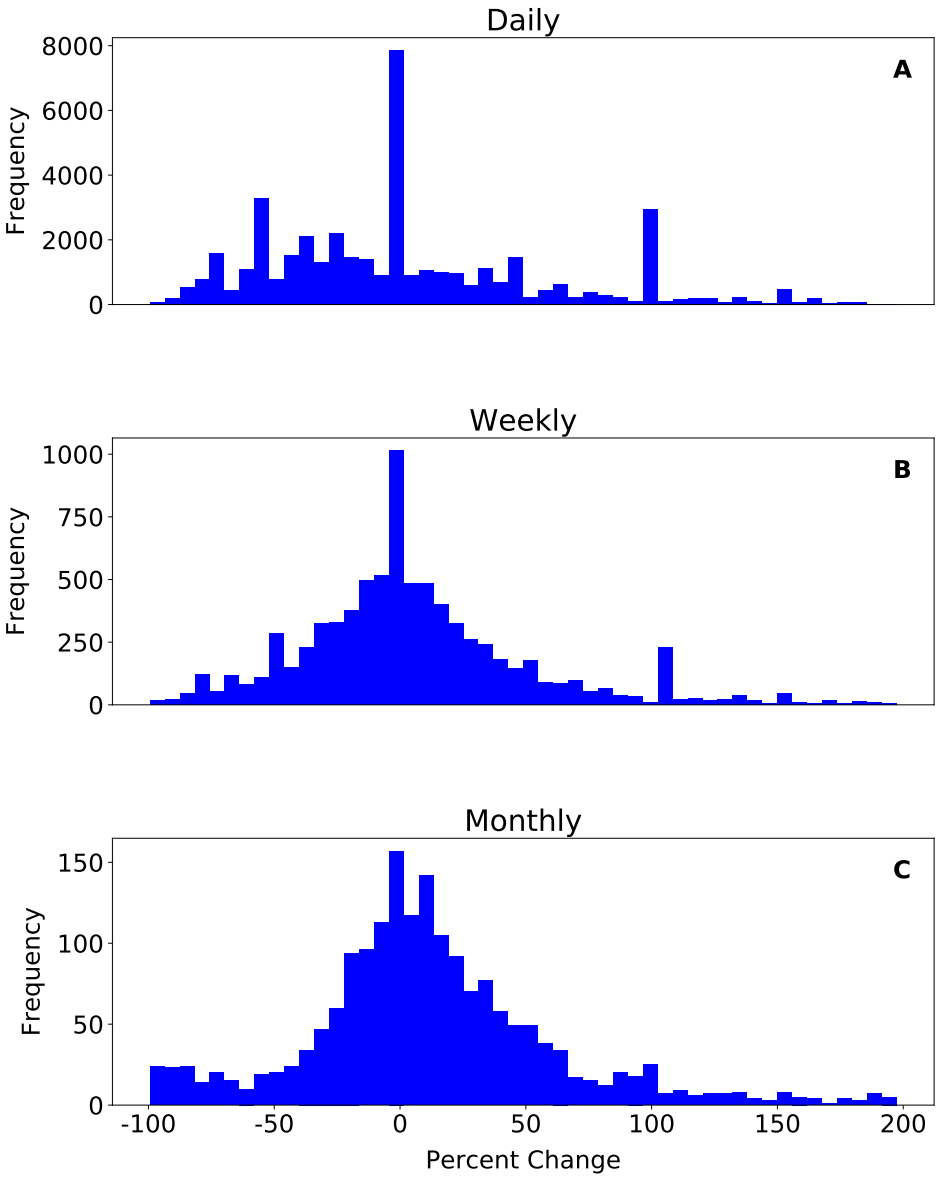

Figure 5.2: **Distribution of percentage changes in drug demand at differing aggregation frequencies.** These plots present distributions of the percentage changes in drug demand at daily, weekly and monthly frequencies. Higher time frequencies are beneficial for policymakers, and for raising the sample size for analysis. However, the higher time frequencies are problematic because the data is sparser. At daily frequency, roughly 18% of the percentage changes in sales are zeroes, and the distribution is not normal. We manage this trade-off by aggregating to monthly frequency, where only 3% of the percentage changes in sales are zeroes. The distribution is therefore much more normal.

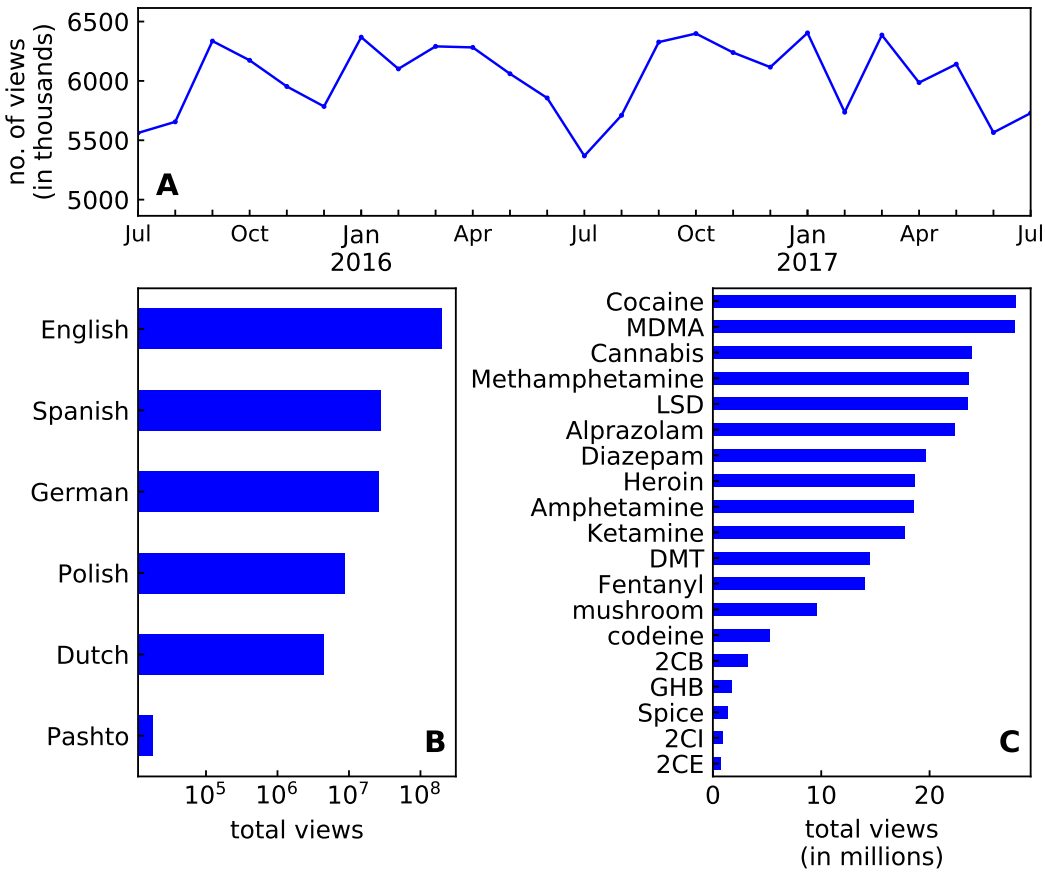

Figure 5.3: **Wikipedia data summary.** Panel A shows the total number of Wikipedia pages views for all languages and drugs, aggregated to monthly frequency. This seems stationary over time. Panel B shows total Wikipedia views over the entire period of study for each language, at a log scale. English is clearly dominant, with views being an order of magnitude greater than the next most common language (Spanish). Panel C shows total Wikipedia views over the entire period of study for each drug Wikipedia page. This is fairly evenly distributed across drugs, although the most popular drugs (e.g. Cannabis) are not necessarily the most viewed.

However Google Trends may be problematic due to language ambiguity for drugs. For example, a search for Magic Mushrooms could feasibly be expressed as "mushrooms", "shrooms", "magic shrooms", or "truffles". Wikipedia is much simpler as there is a set page for each drug [Tsurel et al., 2017]. We conducted an additional analysis where we augmented the Wikipedia data with Google Trends, but found little change in our results. This is consistent with previous research finding substantial correlation between Wikipedia and Google searches [McMahon et al., 2017], which suggests there may be limited value in using both.

5.3 Results

5.3.1 Pooled model

For valid time-series inference, we require the distribution of our data to be stationary across time [Cheung and Lai, 1995]. To formally test for stationarity, we conduct Augmented Dickey Fuller (ADF) tests on the sales data for each drug. We find monthly sales to be nonstationary, with ADF test statistics ranging from -0.8 to -2.6 (all p-values > 0.09). In contrast, the ADF statistics after conversion to monthly percentage changes range from -3.9 to -11.9 (all p-values < 0.01). We therefore conduct our analysis on the monthly percentage change in sales over time. Our data is longitudinal with 3 dimensions: drug, country and time. Let $y_{i,j,t}$ denote the percentage change in sales of drug i in country j and time period t. Our baseline model is:

$$y_{i,j,t} = \beta_0 y_{i,j,t-1} + \alpha_i + \delta_j + \gamma_t \tag{5.1}$$

where $y_{i,j,t-1}$ is an autoregressive term, in case of serial correlation. We also engineer binary variables ("dummies") from the longitudinal data structure. These features allow the baseline model's estimates to capture more variation in drug demand:

- α_i are dummies for each drug.
- δ_j are dummies for each country.
- γ_t are dummies for each month, in case of seasonality.

In this specification (the "pooled" model), we model all drugs jointly. The advantage is that we have more data to fit each of the pooled parameters, which makes overfitting less likely. For example, if we have N drugs and J countries then we have $N * J$ observations to fit each time dummy γ_t. Having fewer drug-specific parameters may also allow us to nowcast drugs that are not in our sample.

However, the disadvantage is that we restrict the model relative to modelling each drug separately. If we were to model each drug separately, this would allow for separate country and time dummies for each drug. Equation 5.2 demonstrates this approach, which we analyse later in Section 5.3.2.

$$y_{i,j,t} = \beta_0^i y_{i,j,t-1} + \delta_j^i + \gamma_t^i \tag{5.2}$$

To estimate the performance improvement from Wikipedia data, we add it to the baseline model. Letting $X_{i,j,t}$ be the percent change in Wikipedia views for drug i in country j and time period t, the "Wikipedia model" is:

$$y_{i,j,t} = \beta_0 y_{i,j,t-1} + \beta_1 X_{i,j,t} + \alpha_i + \delta_j + \gamma_t \tag{5.3}$$

Table 5.1 presents in-sample results comparing the pooled models. All models are unpenalised regression. Scores are adjusted R^2, which includes a penalty term for models with more features. The baseline score is the model's adjusted R^2 without including Wikipedia views. The Wikipedia model includes data on Wikipedia views. The models in the first column use only the autoregressive terms and Wikipedia views as predictors. The models in the second column add complexity with country, drug and month dummies.

The Wikipedia model outperforms the baseline by between 49 and 64 percentage points (pp), depending on the model choice. Therefore Wikipedia data is a strong in-sample indicator for drug demand. This effect is also much larger than the boost from adding the dummies, which we estimate at 7-22pp. However, in-sample performance may not reflect true nowcast potential because of possible overfitting.

Table 5.1: **Pooled model - in-sample performance**. This table compares in-sample accuracy, as measured by Adjusted R^2, with varying numbers of dummies. The first column is a simple model with no dummies included. The second column is a more complex model, including dummies for country, drug and month. Including Wikipedia data boosts model performance by 64pp for the simple model, and 49pp for the complex model. This performance boost is much larger than the boost from adding dummies alone, which we estimate at 22pp. Therefore Wikipedia data is a strong in-sample indicator for drug demand.

	Simple model	All dummies
Baseline Adj. R^2	0.003	0.22
Wikipedia model Adj. R^2	0.64	0.71
Sample Size	1918	1918
Number of features	2	35

We cannot evaluate out-of-sample performance with a random train test split, as time series data is not independent and identically distributed (i.i.d). A random split would put some data in the training set that occurs chronologically after some of the testing set. We would therefore be using data from the future to fit a model estimating the past. This is clearly not possible when performing an actual nowcast.

As in Chapters 3 and 4, we instead use a one-step-ahead nowcasting proce-

dure to measure out-of-sample performance. We first set a training window, w, that determines the size of the training set. Then for each period $t \in [w, T]$ in the data, the training set is data from periods $\in [t - w - 1, t - 1]$. To prevent overfitting, we penalise the model's coefficients using LASSO and 5-fold cross validation in the training set. The penalised model then nowcasts the test set from period t, which is completely held out from training. This procedure only nowcasts the present with data from the past, so it is truly out-of-sample.

We record the errors in period t and use the mean absolute error (MAE) to measure that period's accuracy. Each time we increase t, we slide the training window to update the data and re-fit the model. The model therefore "adapts" over time to new data, which helps maintain accuracy if the underlying relationship changes over time. We set a training window of 12 months, which allows the model to see each month in the training set and fit the seasonality dummies. The first period in our test set is therefore October 2016.

Figure 5.4 compares out-of-sample results from the pooled models. We include month, drug and country dummies in both models. As explained in Section 6.2.1, our analysis is in the monthly percentage changes space. For example, if the true monthly change in demand for a given drug were 0.3 (30%) and our model predicted 0.5 (50%), this would yield an error of 0.2 (20%).

Adding Wikipedia data to the model reduces nowcast MAE in almost every time period. The average reduction in error across the sample is 43% relative to the baseline MAE. Therefore, Wikipedia data is also a strong out-of-sample predictor for drug demand.

The US Fentanyl epidemic demonstrates how a more frequent measure of drug use could be highly valuable for providing early warnings to policymakers. The federal government only declared a national emergency in January 2017, which was arguably too late [Higham et al., 2019]. Figure 5.5 shows that the Wikipedia model's errors are 20% lower than the baseline for US Fentanyl demand. The Wikipedia model makes more variable estimates and is therefore more able to detect shifts in demand. For example, the Wikipedia model correctly nowcasts the big demand spikes in June 2016 and January 2017, whereas the baseline model does not. Therefore, the Wikipedia model may have been able to provide early warning of the US Fentanyl epidemic.

5.3.2 Modelling each drug separately

We have both country and time dimensions in the data, which increases the sample size for each individual drug. This allows us to fit separate models for each drug i:

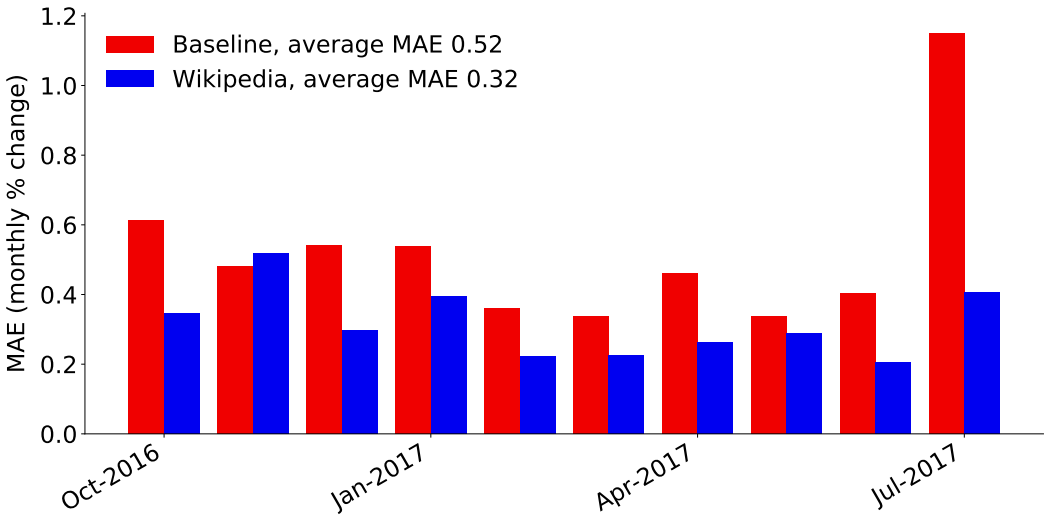

Figure 5.4: **Out-of-sample adaptive nowcasting results - pooled model.** This plot shows out-of-sample adaptive nowcast results using a model that pools data from all drugs. The red bars show mean absolute error (MAE) from the baseline model, where the error refers to the error in estimating the monthly percentage change in demand. The mean is the mean across drugs. The blue bars show MAE from the Wikipedia model. Adding Wikipedia data to the model reduces nowcast MAE in almost every time period. The average reduction in error across the sample is 43%. Therefore, Wikipedia data is also a strong out-of-sample predictor for drug demand.

$$y_{i,j,t} = \beta_0^i y_{i,j,t-1} + \beta_1^i X_{i,j,t} + \delta_j^i + \gamma_t^i \tag{5.4}$$

The parameters β_0^i, β_1^i, α_j^i and γ_t^i now vary by drug. This allows for the model estimates to vary more: for example we can estimate a separate feature weight on Wikipedia views for each drug. The Wikipedia model from Equation 5.4 now has 299 features, compared to 35 features for the model from Equation 5.3. The models may be more accurate, but also more prone to overfitting. This is particularly true for the dummies, as there will be far fewer data points to fit each of them. For example, there are on average only 5 data points to fit each time dummy in the Fentanyl model, which may not be enough to avoid overfitting. We therefore focus on the out-of-sample model performance.

We again assess out-of-sample performance using adaptive nowcasting, but we perform this separately for each drug. Figure 5.6 compares nowcast results between the baseline and Wikipedia models. The MAE is the mean across the entire nowcasting procedure for a given drug - we do not display results over time

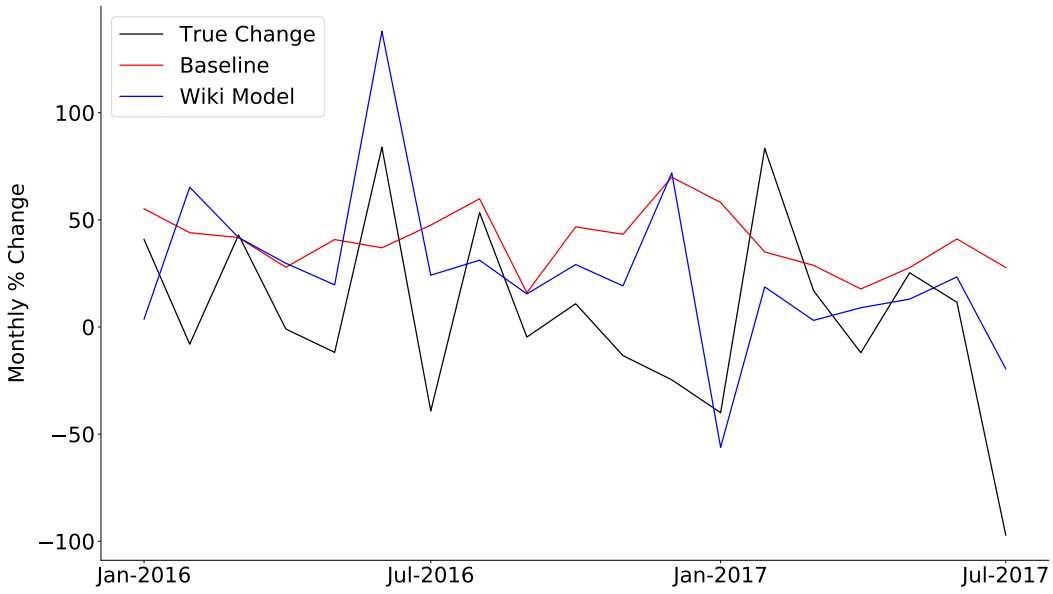

Figure 5.5: **Out-of-sample nowcast of US demand for Fentanyl in the USA.** This figure compares performance of the baseline model against the Wikipedia model for nowcasting US fentanyl demand. The black series is the true change in demand. The red series is the out-of-sample nowcast from the baseline model. The blue series is the out-of-sample nowcast from the Wikipedia model. The Wikipedia model makes more variable estimates and is therefore more able to detect shifts in demand. For example, the Wikipedia model correctly detects the big demand spikes in June 2016 and January 2017, whereas the baseline model does not. Therefore, the Wikipedia model may have been able to provide early warning of the US Fentanyl epidemic.

as in Figure 5.4. We do not include any dummies in these models due to the small sample size for some drugs.

Adding Wikipedia data reduces nowcast errors for every drug relative to the baseline. The average MAE reduction across drugs is 42%. Therefore, Wikipedia data remains a strong indicator of demand when modelling each drug separately.

5.3.3 Modelling each country separately

Similarly to Section 5.3.2, we can also fit a separate model for each country j:

$$y_{i,j,t} = \beta_0^j y_{i,j,t-1} + \beta_1^j X_{i,j,t} + \alpha_i^j + \gamma_t^j \tag{5.5}$$

The parameters β_0^j, β_1^j, α_j^j and γ_t^j now vary by country, which again allows for greater model complexity.

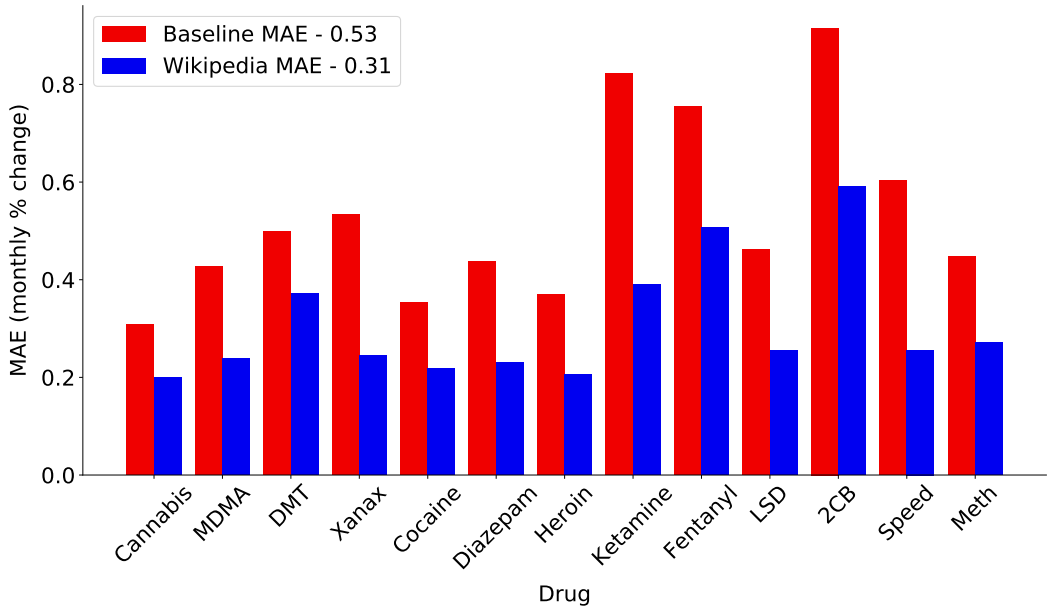

Figure 5.6: **Out-of-sample adaptive nowcast results, modelling each drug separately.** This plot show out-of-sample adaptive nowcast results when constructing a separate model for each drug. We do not include any dummies in these models due to the small sample size for some drugs. The red bars show mean absolute error (MAE) from the baseline model, where mean refers to mean across time periods for the entire nowcast period. The blue bars show MAE from the Wikipedia model. Adding Wikipedia data reduces nowcast errors for every drug relative to the baseline. The average MAE reduction across drugs is 42% relative to the baseline model. Therefore, Wikipedia data remains a strong indicator of demand when modelling each drug separately.

Figure 5.7 presents out-of-sample adaptive nowcast results when modelling each country separately. We fit a separate model for each country, so each country has its own feature weights on the autoregressive term and Wikipedia page views. MAE is the mean error across the entire nowcasting procedure, for a given country.

Adding Wikipedia data to the baseline improves accuracy in every country relative to the baseline, which is consistent with previous models. The average MAE reduction across countries is 40%. Therefore, Wikipedia data remains a strong out-of-sample predictor when modelling each country separately.

The results from the separate modelling approaches suggest that accuracy may be reduced at lower sample sizes, particularly for the baseline model. Figure 5.6 shows that MAE tends to be greater for drugs with fewer sales, such as 2CB and DMT, than more popular drugs such as MDMA and Cannabis. The correlation between the size of the drug category we are predicting, as measured by sales

volume, and MAE for that category is -.78 for the baseline model and -.73 for the Wikipedia model. The results from modelling each country separately, discussed later in Section 5.3.3, further support this conclusion. The correlation between the size of the country, as measured by sales volume, and MAE for that category is -.81 for the baseline model and -.68 for the Wikipedia model. Series with smaller sample sizes are likely to be more volatile, therefore harder for an autoregressive model to nowcast. Adding Wikipedia data is therefore likely to improve accuracy more where sales data is sparser, which could be particularly useful in detecting emerging drugs such as the novel psychoactive substances.

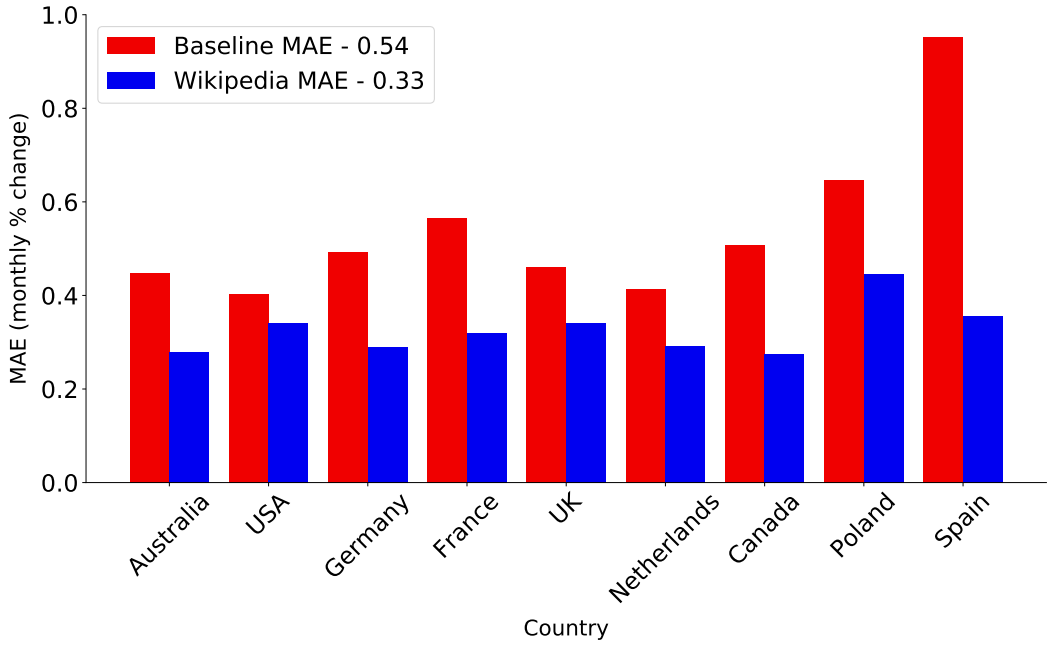

Figure 5.7: **Out-of-sample adaptive nowcast results, modelling each country separately.** This plots show out-of-sample adaptive nowcast results when constructing a separate model for each country. The red bars show mean absolute error (MAE) from the baseline model, where mean refers to mean across time periods for the entire nowcast period. The blue bars show MAE from the Wikipedia model. Adding Wikipedia data reduces nowcast errors for every country relative to the baseline. The average MAE reduction across drugs is 40%. Therefore, Wikipedia data remains a strong indicator of demand when modelling each country separately.

There may be an issue with the geographic link between Wikipedia page views and darknet sales. The Wikipedia data is split by language of the page (e.g. French), whereas the darknet sales are split by country of sale (e.g. France). The link between them is likely to be strong when the language is not widely spoken outside its origin country, such as Dutch. However there are languages where the country

of origin is less clear, such as English. We analyse this issue in Figure 5.7, which presents results from modelling each country separately. There is some evidence that the Wikipedia model performs worse for countries with a shared language, such as the US and Australia. Nevertheless, the difference is small and the Wikipedia model outperforms the baseline across all countries. Future research could use internet search data where the user's location is known, such as Google Trends, rather than inferred from language.

5.3.4 Robustness

Our main results, depicted in Figure 5.4, set a training window of 12 months. Table 5.2 assesses the robustness of our results to varying training windows. We show that results do not qualitatively change across a range of windows, from 10 months to 14 months. Therefore, our results are robust to varying training windows.

Table 5.2: **Out-of-sample results with different training windows.** This table shows results across a range of training window lengths. The main results use a fixed 12 month window. The first column shows mean absolute error (MAE) from the baseline model, where the error refers to the error in estimating the monthly percentage change in demand. The mean is the mean across drugs. The second column shows MAE from the augmented model that includes Wikipedia data. The Wikipedia model outperforms the baseline across all training windows, with similar performance to the 12 month window that our main results are based on. We therefore conclude our results are robust to varying the training window.

Training window	Baseline MAE	Wikipedia MAE
10 months	0.51	0.30
11 months	0.52	0.30
13 months	0.52	0.31
14 months	0.53	0.29

Table 5.3: **Out-of-sample results at different aggregation frequencies.** This table shows results across a range of aggregation frequencies. Our main results aggregate data to 1 month frequency, but we could analyse in a higher or lower frequency space as our data is daily frequency. Column 1 shows mean absolute error (MAE) from the baseline model, where the error refers to the error in estimating the monthly percentage change in demand. Column 2 shows MAE after inclusion of Wikipedia data. Our results are qualitatively robust to all aggregation frequencies from 2 weeks to 8 weeks. However, the Wikipedia model performance is relatively stronger at lower frequencies. This suggests there may be a trade-off between model speed and accuracy.

Aggregation frequency	Baseline MAE	Wikipedia MAE
2 weeks	0.45	0.30
4 weeks	0.50	0.30
6 weeks	0.55	0.32
8 weeks	0.61	0.29

Table 5.4: **Out-of-sample results using different start dates for the data.** The main text results use a start date of July 2015, which keeps all possible data. However, this may induce coverage errors because of listings removed by vendors (see Section 6.2.1.) The later the start date, the less the darknet scrape's coverage is affected by removed listings. The first column shows mean absolute error (MAE) from the baseline model. The second column shows MAE from the augmented model that includes Wikipedia data. The Wikipedia model outperforms the baseline regardless of the start date, indicating robustness to a variety of start dates.

Start date	Baseline MAE	Wikipedia MAE
April 2016	0.47	0.29
July 2016	0.45	0.27
October 2016	0.50	0.27
January 2017	0.57	0.35

Given the daily frequency of our data, we can vary the time frequency of our nowcasting models. Higher time frequencies may be more useful for policymakers as they would get a faster estimate of drug use. However the sales data is sparser at higher frequency, as shown in Figure 5.2.

Table 5.3 shows results across a range of aggregation frequencies. Our results are qualitatively robust to all aggregation frequencies, but stronger at lower frequencies. This suggests there may be a trade-off between model speed and accuracy. Nevertheless, the monthly frequency of our main results would still be much faster than the current annual survey data.

Finally, our main results use a start date of July 2015. As discussed in Section 6.2.1, this maximises our sample size. However, it is likely to induce coverage errors in the data as some of the earlier listings were removed by vendors. Table 5.4 shows results using different start dates for the data. Our results are qualitatively robust to a variety of start dates.

5.4 Discussion

Table 5.1 shows that past drug sales data alone cannot accurately nowcast current drug sales. However our results consistently show that adding Wikipedia data, which is reliably available in real-time, greatly boosts nowcast performance. We present results using two broad approaches: a pooled approach, where all drugs are modelled jointly, and a second approach where we model each drug and country separately. In all specifications, Wikipedia page views reduce nowcast errors by at least 40% relative to the baseline model.

The average nowcast errors in the pooled model, shown in Figure 5.4, are comparable to the errors when modelling each drug separately, shown in Figure 5.6. This suggests we should use the pooled model, as it may allow nowcasting demand for drugs with little available data. This would be particularly useful when new drugs are entering the market, such as the novel psychoactive substances, which traditional surveys struggle to capture [Smith and Garlich, 2013].

We acknowledge limits to the external validity of extrapolating our darknet results to nowcasting overall drug use. There are known demographic biases with internet usage [Graham et al., 2015], so darknet drug users may not be representative of drug users overall. If so, this may diminish the nowcast value of Wikipedia data for overall drug use. However, previous research found that darknet demand geographically represents overall drug demand well for cannabis, cocaine and heroin [Dittus et al., 2018]. Moreover, the Wikipedia model performs well across a range of drugs,

as shown in Figure 5.6. If demographic bias were affecting our results, we may expect the Wikipedia model to perform better among drugs whose consumers use the internet more, such as DMT, LSD and 2C-B [Kruithof et al., 2016]. We cannot find strong evidence of this, with the Wikipedia model performing well for harder "street" drugs such as heroin and cocaine, whose demographics are less represented among internet users [Kruithof et al., 2016]. Furthermore, nowcasting darknet demand itself may be of interest given its rapid growth over the last decade [Soska and Christin, 2015].

Now that we have shown Wikipedia views can nowcast darknet sales at high time frequency, a next step would be to show that that this relationship generalises to wider drug demand. This is beyond the scope of this chapter, but there is literature supporting this relationship with other internet search data. A recent study showed strong correlations between the volume of Reddit comments in US states and annual opioid use, as measured by CDC surveys [Balsamo et al., 2019]. Another paper showed correlation between Google searches for novel psychoactive substances and changes in their global consumption, measured by UNODC surveys [Perdue et al., 2018]. While out of scope for this paper, future work demonstrating these results hold out-of-sample at high time frequencies would be valuable.

This chapter analysed whether darknet and internet search data can help build a high frequency measure of drug demand. The darknet drug markets are a promising real-time data source, but analysts may not reliably be able to access them due to frequent outages. We show that a model including Wikipedia page views greatly improves nowcast accuracy, particularly when darknet data is unavailable for long periods. These results hold out-of-sample across all drugs and a range of modelling choices. Wikipedia page views most improve nowcast accuracy for less popular drugs, suggesting our model may be particularly useful for detecting newly emerging substances. We acknowledge there are limits on extrapolating results from darknet data to wider drug consumption, due to demographic biases among internet users. Nevertheless, we believe there is strong evidence overall that internet data may greatly improve the speed of official drug statistics, which are currently annual frequency. This may help policymakers respond more quickly to the next drug epidemic.

Chapter 6

Faster indicators of chikungunya incidence using Google searches

6.1 Introduction

Chapter 5 analysed whether online search data could improve nowcasts of illicit drug demand, which is a major public health issue. This chapter analyses a different public health application for nowcasting: monitoring the spread of disease. The analysis focuses on chikungunya, a mosquito-born viral disease or *arbovirosis*, which is a growing global public health challenge. In Brazil, there have been 100,000 to 250,000 reported cases per year since 2016 [Secretaria de Vigilância em Saúde, 2020]. Infections lead to severe health complications in around 25% of cases, such as paralysis and long-term debilitating syndromes [Aguiar et al., 2018; Schilte et al., 2013; de Brito, 2017]. Fatality rates may also be higher than previously recognised, due to challenges in determining the cause of death [de Brito, 2017]. Chikungunya incidence is highly seasonal, with one epidemic per year during the warmer months when mosquitoes are more active. Epidemics in Ceará, Brazil, have caused major disruptions to their healthcare system [Bastos et al., 2018], and the economic costs from treatment and workplace absence are often catastrophic for the low-income households affected by chikungunya [Gopalan and Das, 2009].

Infectious disease surveillance relies on doctors to enter cases into the monitoring system. Disease statistics are prone to delays, as there is often a long lag between a patient seeking treatment and entry of the case [Codeço et al., 2018]. In Rio de Janeiro, chikungunya surveillance delays average around four weeks, with data arriving gradually and inconsistently. Public health policymakers need accurate real-time data on disease incidence in order to respond quickly to epidemics [Olliaro

et al., 2018]. Faster surveillance is helpful in many ways: helping doctors to decide between different diseases with similar symptoms, better targeting of mosquito control activities and increasing awareness in the general population. More broadly, failure to respond at sufficient speed to the spread of other diseases can have catastrophic consequences, and may have raised the death toll from fast-growing epidemics such as the Covid-19 crisis [Sun et al., 2020; Galea et al., 2020].

Online data is a useful source of information for improving the speed of disease surveillance. People experiencing symptoms of a disease may not only consult a medical professional for help, but also search for information on Google. In contrast to official case counts, data on Google searches are reliably available in real-time [Google Trends API, 2020]. Previous studies have shown a relationship between internet search data and case counts for diseases such as the flu [Ginsberg et al., 2009; Preis and Moat, 2014] and dengue [Chan et al., 2011a; Mizzi et al., 2021]. Here, we investigate whether Google search data can help generate faster estimates of chikungunya case counts in Brazil.

As a reminder from previous chapters, estimating the value of statistics before they are officially released is known as "nowcasting". A promising sign that internet search data may help nowcast chikungunya can be found in one study that reports a positive correlation between Google search activity and chikungunya incidence in the Amazon [Naveca et al., 2019]. The question is whether this is a sufficiently strong and consistent relationship to improve the accuracy and precision of chikungunya nowcasts, in comparison to a model that uses historic chikungunya case data alone. To investigate this question, we build on a Bayesian approach specifically designed for nowcasting where case data is subject to long and varied delays [Bastos et al., 2019; Mizzi et al., 2021]. Better chikungunya nowcasts could help public health authorities respond more quickly to future epidemics, therefore mitigating their damage [Coelho and Codeço, 2019].

6.2 Data and methods

6.2.1 Data

We use two main data sources in this chapter. The first is chikungunya case data from Brazil's disease monitoring system, henceforth referred to as SINAN *(Sistema de Informação de Agravos de Notificação)* [SINAN, 2020]. The second is Google search data from the Google Health Trends API.

We obtain chikungunya case data for Rio de Janeiro through the InfoDengue project [Codeço et al., 2018]. Our case data begins in January 2016, shortly after

the start of the first chikungunya epidemic in Rio de Janeiro, and ends in December 2019. The raw data is case-level: for each case we have access to a notification date and an entry date. The notification date is the date at which a doctor first diagnoses a chikungunya case. The entry date is the date that a confirmed chikungunya case is entered into the system. This will usually be based on symptoms alone, as only 10% of cases are confirmed by laboratories. If a laboratory finds that a chikungunya case is falsely diagnosed, it is retroactively removed from the system.

Table 6.1 shows that case entry usually occurs well after notification; only 50% of notified cases are entered into the system within 2 weeks of notification. There are also some very long delays in the data, such that 26% of notified cases are still not entered after 4 weeks, and 13% of notified cases are still not entered after 8 weeks. We verify whether the delays are similar when only considering epidemic periods, as defined by the Moving Epidemic Method (MEM) Vega et al. [2013]. The MEM analyses the frequency of cases across the sample to set a weekly case threshold, above which the week would be defined as an epidemic period. For our data, the epidemic threshold identified by the MEM is 104 cases. The second row of Table 6.1 shows that delays during epidemic periods are similar to delays when considering all of the data. However, Table 6.1 also reveals that 58% of weeks in this dataset fall in an epidemic. For this reason, we further examine the pattern of delays in different years, to help us understand whether delays are impacted by the size of the epidemic. We find little difference between the delays witnessed in individual years and those in the sample as a whole, regardless of the size or presence of an epidemic in each of the years. This suggests that any differences in the performance of the model in these different time periods are unlikely to be due to differences in the structure of the delays.

Table 6.1: **Delays between chikungunya case diagnosis and entry into the disease surveillance system, in days.** There is often a long delay of weeks or months between initial diagnosis of a chikungunya case and entry of that case into the monitoring system. Varied delays make nowcasting on a weekly basis more difficult, as we lack complete data on both the most recent week and the weeks shortly preceding it. The first row describes the distribution of delays across the sample as a whole, in days. Only 26% of cases are reported after 1 week, and 74% after 4 weeks. We further find evidence of a tail of cases with very long delays, with 13% of cases still not entered 8 weeks after notification. The second row describes the distribution of delays during epidemic periods, which are weeks in which the case count exceeds 104 cases. This is very similar to the sample as a whole. Similarly, we find that the distribution of delays in non-epidemic periods (third row) and the distributions of delays by year (fourth to seventh rows) do not differ greatly to the distribution of delays for the sample as a whole. This suggests that any differences in the performance of the model in these different time periods are unlikely to be due to differences in the structure of the delays.

	Number of Weeks	1 Week	2 Weeks	4 Weeks	8 Weeks
All Periods	208	26%	50%	74%	87%
Epidemics	121	26%	51%	76%	90%
Non-Epidemics	87	26%	48%	72%	84%
Only 2016	52	31%	51%	68%	82%
Only 2017	52	27%	49%	76%	88%
Only 2018	52	24%	49%	78%	91%
Only 2019	52	23%	49%	76%	89%

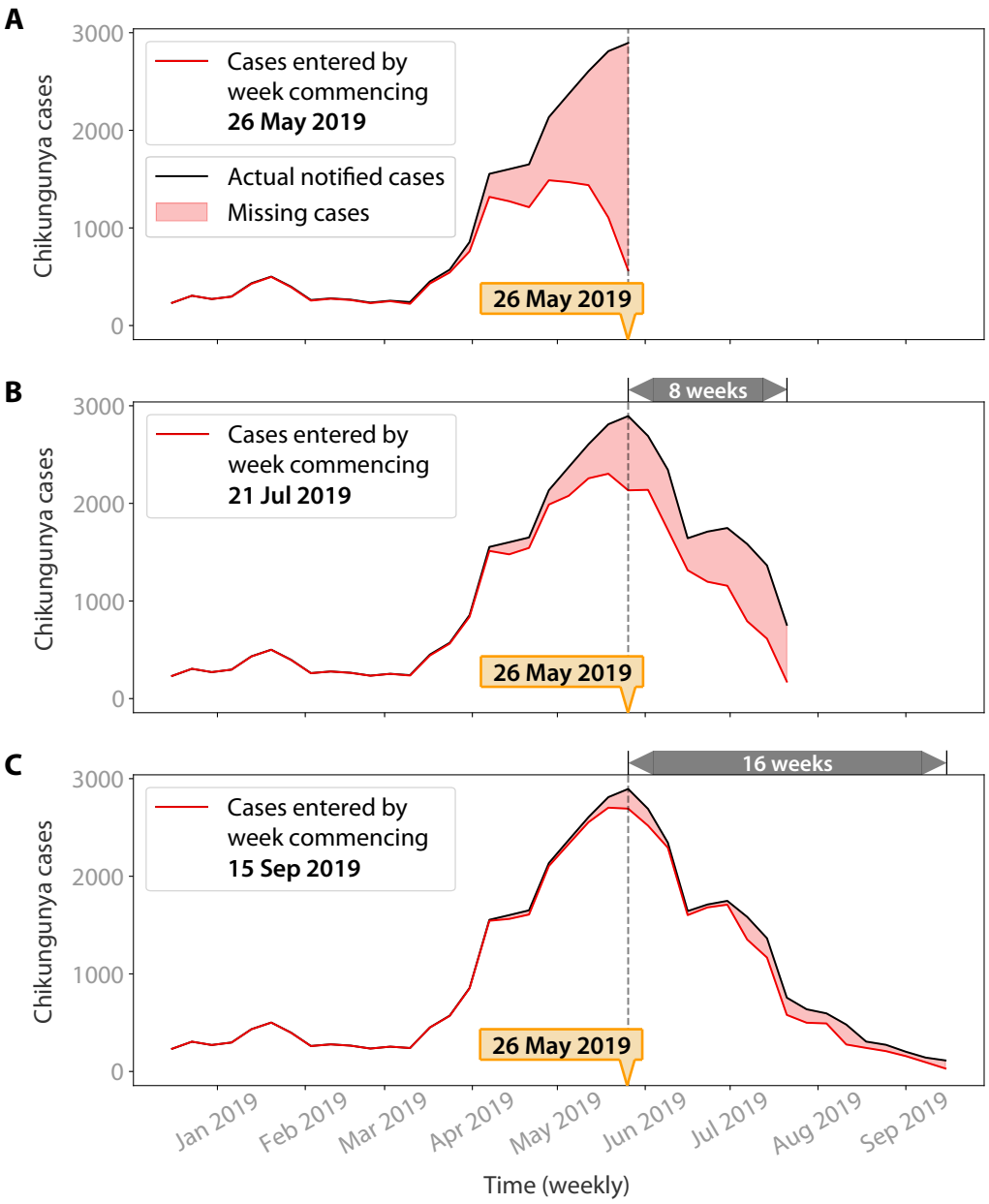

Figure 6.1: **Chikungunya case data availability for a given week.** Data on previous chikungunya cases arrives gradually and inconsistently. We illustrate this problem using the week beginning 26 May 2019 - the peak of the 2019 epidemic in Rio de Janeiro. The black series show the true number of cases in each of the previous weeks. The red series show how many of these cases had been entered into the surveillance system by the end of the week. The upper panel shows that only 20% of the 2895 cases for the week of the 26 May 2019 had been entered by the end of the week. The middle panel shows the data is still very incomplete 8 weeks later, with only 74% of cases from 26 May 2019 being entered. The lower panel shows that the data is still not complete even after 16 weeks, with only 93% of cases being entered.

The length and inconsistency in reporting delays makes nowcasting chikungunya incidence much more difficult. Figure 6.1 shows a snapshot of data availability for the week of 26 May 2019. There were 2895 diagnosed cases during the week, but they were entered into the system only gradually over the following weeks, with around 25% of cases still not entered after two months. At the end of the example week, the data on previous weeks was similarly incomplete, with completeness being worse for more recent weeks. Nowcasting methods traditionally rely on complete data about the past. Therefore, without aggregating this data to a very low temporal resolution, it is difficult to accurately nowcast chikungunya incidence using past case data alone.

Our second source of data is on Google search behaviour, which is available in real-time from the Google Health Trends API. This provides data at weekly frequency from January 2016 to December 2020 on searches related to chikungunya. To retrieve data on searches related to chikungunya, we use Wikidata [Wikidata API] to identify the Freebase topic ID for chikungunya (/m/01__7l), in line with the approach taken in previous work on dengue [Mizzi et al., 2021].

Figure 6.2 shows that spikes in Google searches for chikungunya may provide a rapid indicator of higher case counts. There is visually a strong correlation, with the peaks in Google searches occurring on or before the epidemic peaks in 2016, 2018 and 2019. The magnitude of spikes in Google searches are also visually a good fit for the magnitude of epidemics, with the biggest spikes in searches occurring during the 2016 and 2019 epidemics.

The Google data is reliably available in real-time: at the end of any given week, we have a complete record of search behaviour for that week. This clearly does not apply to the official chikungunya case data, where we lack complete data on both the current week and previous weeks. Therefore, Google data may help us nowcast chikungunya case counts.

We considered including alternative internet data in our nowcasting model, such as Twitter mentions of chikungunya. However, previous research on nowcasting dengue fever showed that adding Twitter data only slightly improved a model that already included Google Trends data [Mizzi et al., 2021]. Therefore, we chose to focus on Google Trends data for this model.

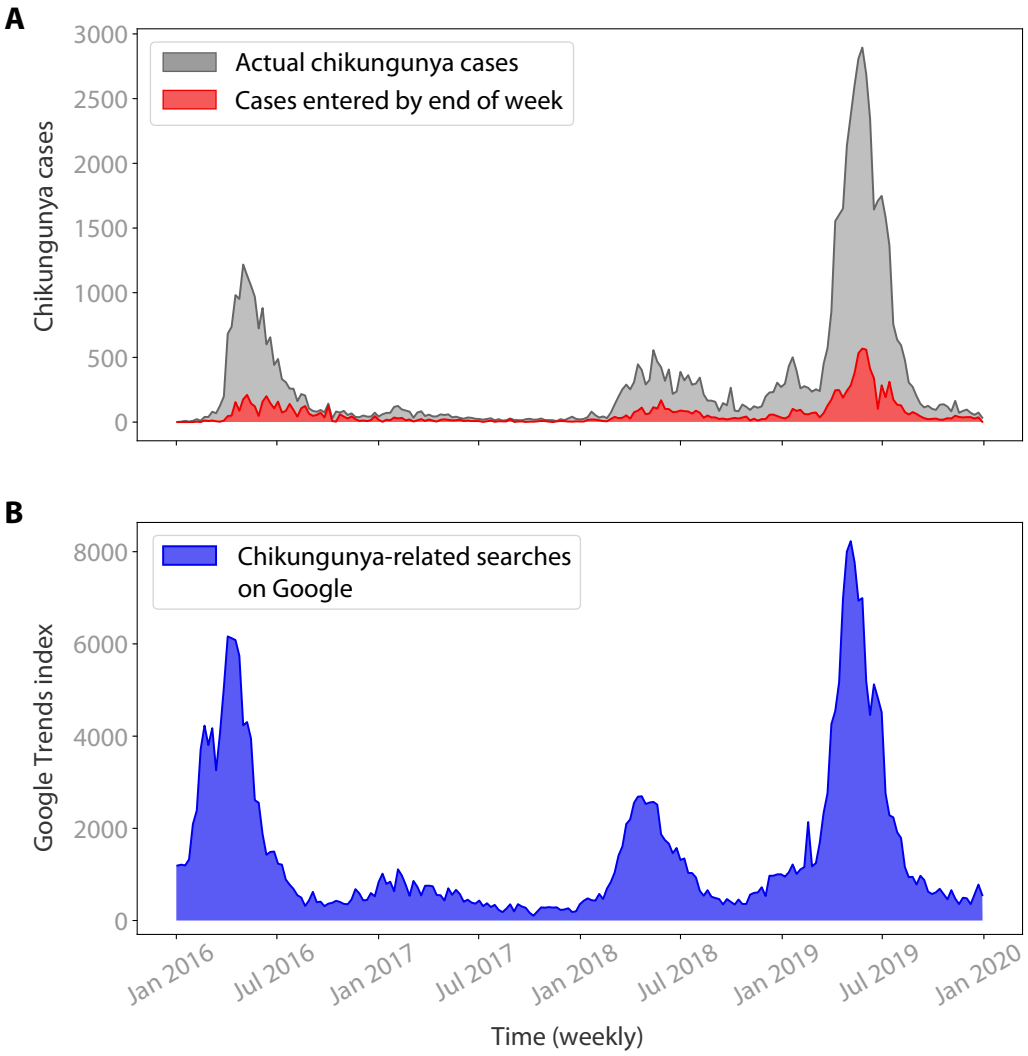

Figure 6.2: **Comparing chikungunya case counts and Google searches over time.** Panel A shows how chikungunya cases have evolved since 2016. The black series shows the number of cases diagnosed in each week. The red series shows the number of cases diagnosed in a given week that have been entered into the disease surveillance system by the end of that week. Entered cases are an inconsistent fraction of notified cases, and this issue is most severe during the large epidemics in 2016 and 2019. This makes estimating current chikungunya case counts from the official data alone particularly challenging. Panel B compares the weekly notified cases against weekly Google searches for chikungunya-related terms, which are available in real-time. Visually there is a strong correlation, with Google searches peaking during the large epidemics in 2016 and 2019. The size of the peaks in Google searches also seems to match the size of the epidemics, with the largest peaks during the 2016 and 2019 epidemics. Google search data may therefore provide a rapid indicator of chikungunya case counts.

6.2.2 Methods

Our objective is to estimate the current weekly case count X_t. The method must be operationally feasible, using only data available by the end of week t.

Our baseline model uses only data on previous chikungunya incidence to perform the estimation. Traditional nowcasting methods assume complete data about previous weekly cases $(X_{t-1}, X_{t-2}, ..., X_0)$ is available at time t. This assumption does not hold for chikungunya data, as shown in Section 6.2.1, where cases are often entered only after a long delay. Following previous work on dengue [Mizzi et al., 2021], we therefore use an alternative baseline nowcasting model developed for case count data that arrives gradually and inconsistently [Bastos et al., 2019]. For each week, this model aims to estimate the number of cases that will be entered into the system with a given number of weeks delay, using data available in week t.

Table 6.2: **Stylised example of chikungunya case count data availability for a given week.** This matrix provides a stylised example of the chikungunya case count data available when nowcasting cases for a given week. In this example, we hold data from week 1 onwards and are currently in week 7. We assume here that the maximum delay in entering a case into the surveillance system is five weeks. Each row represents a previous week (t) of entered cases, and the column represents the entry delay (d) in weeks. For example, we can see that there were initially 15 cases entered into the system in week 2, 8 further cases after a delay of 1 week, 10 cases after a delay of 2 weeks, and so on. Case data is incomplete, not only for week 7 but also weeks 3 through 6. The incompleteness is usually worse the closer we are to the current period, so there is a running "triangle" of unknown case counts and associated delays to be estimated for previous weeks. Estimating each cell in the last row of this triangle yields a nowcast of the total case count for week 7. The method introduced by Bastos et al. [2019] provides an approach for generating these estimates.

		Delay in weeks (d)						
		0	1	2	3	4	5	Total
Week (t)	1	10	12	6	4	1	1	34
	2	15	8	10	2	4	1	40
	3	19	9	13	5	2	?	?
	4	19	9	13	5	?	?	?
	5	17	25	11	?	?	?	?
	6	26	20	?	?	?	?	?
	7	39	?	?	?	?	?	?

Table 6.2 provides a stylised example of this procedure. To estimate total cases in week t, we must therefore estimate how many cases will be entered with a

delay of d weeks ($x_{t,d}$), where

$$X_t = \sum_{d=0}^{D} x_{t,d}$$

Following Bastos et al. [2019], we assume $x_{t,d}$ has a negative binomial distribution:

$$x_{t,d} \sim \mathcal{NB}(\lambda_{t,d}, \phi)$$

We estimate the mean of this distribution, $\lambda_{t,d}$, with the following specification:

$$\log(\lambda_{t,d}) = \alpha + \beta_t + \gamma_d$$

where

- α is a time-invariant constant.
- β_t is a first order random walk (rw1) random effect $\beta_t \sim \mathcal{N}(\beta_{t-1}, \sigma_\beta^2)$ capturing serial correlation in case counts. If we observe larger case counts in the previous week, we estimate a higher case count for the current week.
- γ_d is an rw1 random effect $\gamma_d \sim \mathcal{N}(\gamma_{d-1}, \sigma_\gamma^2)$ capturing serial correlation in the number of cases reported with a given number of weeks delay. If we observe a greater number of cases with $d-1$ weeks delay, we estimate a higher number of cases with d weeks delay too.

We fit the parameters for this specification via Integrated Nested Laplace Approximation (INLA) method [Rue et al., 2009]. We estimate each $x_{t,d}$ via sampling, which yields a posterior distribution of estimates for $X_t = \sum_{d=0}^{D} x_{t,d}$. This distribution provides a natural measure of uncertainty, with wider distributions implying greater uncertainty.

We estimate each week $t >= 21$ following an adaptive nowcasting procedure [Preis and Moat, 2014]. We initially train a model with the first 20 weeks of data. The model then outputs a posterior distribution of estimated case counts for week 21. We record the difference between the mean case count estimate and the true case count as the model's out-of-sample nowcast error.

In each following week $t > 21$, we re-train the model with all available data at week t. Therefore, the model "adapts" over time. Taking into account the conclusions of previous work on dengue in which the benefit of varying training window size was investigated, we always use all available data rather than a fixed training window [Mizzi et al., 2021]. Epidemics are infrequent, so we do not want to risk losing them from the training data with a shorter training window.

We now define our "Google model". This is similar to the baseline model, but also includes Google data G_t:

$$log(\mu_{t,d}) = \alpha + \beta_t + \gamma_d + \delta log(G_t)$$

where δ is a regression coefficient.

Google search data are fully available by the end of the week. We can therefore include G_t directly, rather than having to estimate it in the same way as the chikungunya case data.

Following Yang et al. [2017], our final model is a "heuristic" approach, using the notified cases for week $t-3$ that have been entered by week t. As shown in Panel A of Figure 6.2, this model will not be accurate due to reporting delays. Nevertheless, it provides important context – our baseline already provides a large improvement over a heuristic approach. We are therefore analysing whether Google data can help us further improve the performance of a carefully chosen baseline model that is well-suited to this nowcasting problem.

6.3 Results

Following previous work on nowcasting dengue [Mizzi et al., 2021], we calculate a range of metrics for our models to allow us to investigate both accuracy and precision of our estimates. For all models, we report accuracy in terms of mean absolute error (MAE), where a lower error reflects a more accurate model. We report precision in terms of the mean 95% prediction interval width (MPI), where a smaller interval represents greater model precision. It is important to verify that any reduction in the size of the 95% prediction interval is not simply due to the interval becoming too narrow and no longer reliable. We therefore also report interval reliability, in terms of the percentage of true weekly case counts that fall within the 95% prediction interval for that week. 95% of true weekly cases falling within the 95% prediction interval represents good interval reliability. Finally, we report results both for the full period and when considering epidemic periods alone, as epidemics are likely to be particularly important for policymakers. We define epidemic periods using the 104 weekly case threshold previously calculated by the MEM [Vega et al., 2013].

Table 6.3 compares accuracy across the heuristic, baseline and Google models. The baseline model is much more accurate than the heuristic model, reducing MAE by 34% across the sample. The Google model further improves upon the baseline, reducing MAE by 41% relative to the heuristic model. Panel A of Figure 6.3 shows that the Google model predictions are rarely far from the true case number.

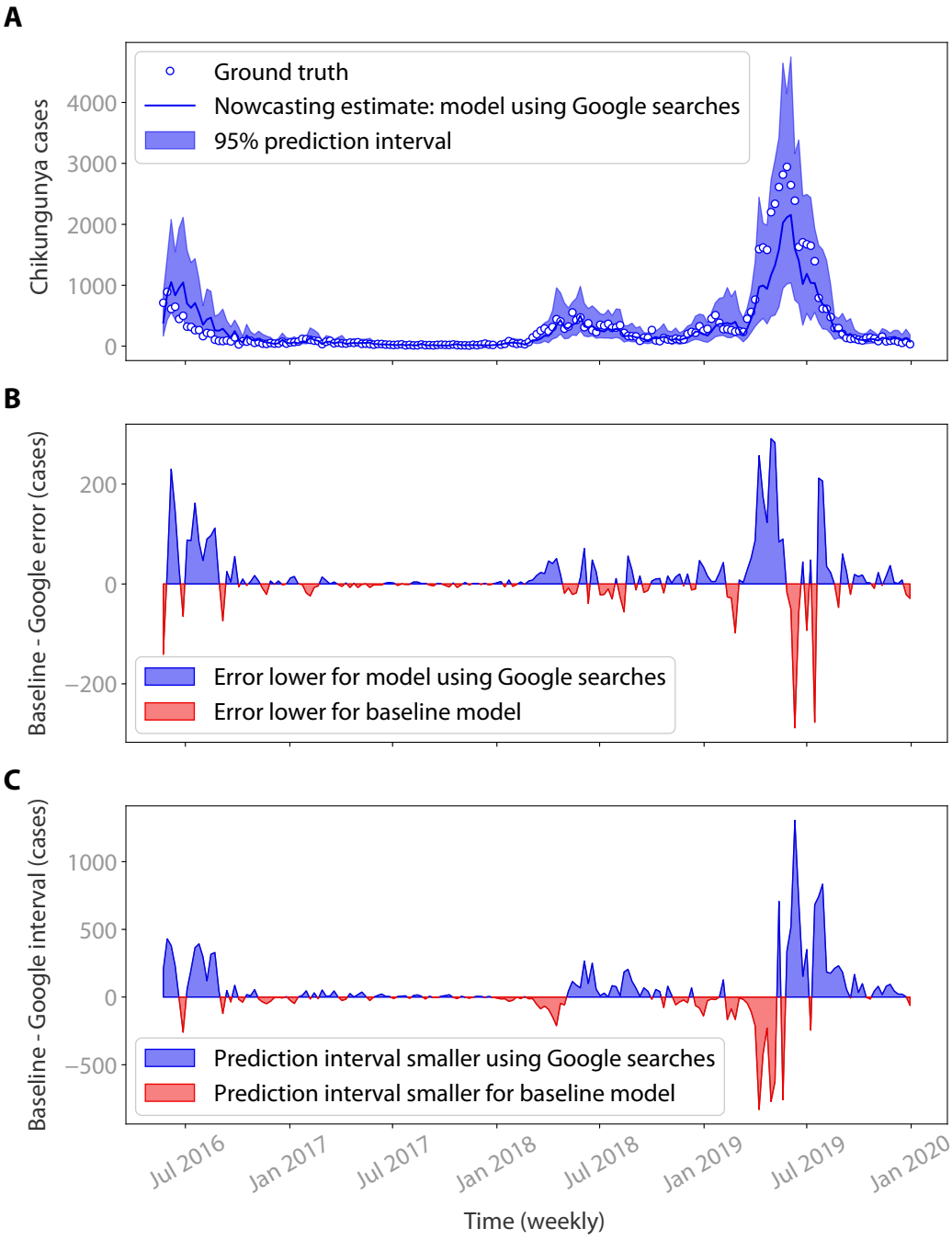

Figure 6.3: **The Google model's performance over time and relative to the baseline model.** Panel A shows nowcast results over time from the Google chikungunya nowcasting model. The Google model's estimates are relatively accurate across the sample. Moreover, the ground truth rarely falls outside the 95% intervals. Panel B compares relative nowcast errors, in case numbers, between the baseline and Google models. Blue indicates the Google model error is lower, and red indicates the baseline model error is lower. Overall the Google model errors are lower, although there is some volatility around epidemic periods. Panel C compares relative prediction interval width, in case numbers, between the baseline and Google models. The Google model interval is generally narrower, except for the period at the start of the 2019 epidemic. However, Table 6.5 shows that the baseline model intervals may be unreliable during an epidemic. Several of the true weekly case counts at the start of the 2019 epidemic fall outside the baseline interval (see 6.4), but within the Google interval.

In Table 6.3, we report the accuracy of the model estimates during epidemics as well as across the full period. The improvement offered by the baseline model over the heuristic model increases further during epidemics, reducing MAE by 35%. In turn, the improvement offered by the Google model over the baseline model increases too, reducing MAE by 43% relative to the heuristic model.

Table 6.3: **Comparison of baseline and Google model chikungunya now-casting accuracy.** This table compares the mean absolute errors (MAEs) for the nowcasts produced by the heuristic, baseline and Google models. The first two columns show results across the sample as a whole, reported in number of cases and relative improvement space respectively. The baseline is far more accurate than the heuristic model, reducing MAE by 34%. The Google model improves further upon the baseline, reducing MAE by 41% relative to the heuristic model. The third and fourth column show results when only considering epidemic periods. Here we see that the accuracy boosts offered by both the baseline and Google models increases further: the baseline model reduces MAE relative to the heuristic model by 35%, and the Google model reduces MAE by 43%.

Model	All periods		Epidemics	
	MAE	Relative MAE	MAE	Relative MAE
Baseline	124.8	0.66	212.8	0.65
Google	110.9	0.59	187.7	0.57
Heuristic	187.9	1.00	327.0	1.00

Panel B of Figure 6.3 shows there are few periods during epidemics where the baseline model outperforms the Google model. This is further corroborated when splitting the errors by year, as shown in Table 6.4. The Google model most outperforms the baseline model during the years with epidemics (2016, 2018 and 2019). There is little difference between the two models in 2017.

Table 6.5 compares the precision of the baseline and Google models in terms of the mean 95% prediction interval width (MPI). The heuristic approach does not allow an interval to be calculated. Panel A of Figure 6.3 shows that ground truth weekly case counts very rarely fall outside the Google model 95% prediction interval. Furthermore, there are no instances where the ground truth falls far outside the prediction interval.

The Google model is more precise than the baseline model, reducing MPI by 7%. While the prediction interval width increases for both models when considering epidemic periods alone, the Google model remains 8% more precise than the baseline. This precision advantage holds for each year in the sample. Table 6.6 shows baseline and Google model intervals for each year individually. The Google model reduces

Table 6.4: **Comparison of model errors by year.** In each year with an epidemic (2016, 2018 and 2019), the Google model outperforms the baseline model. There is little difference between the Google and baseline models in 2017. Note that poor performance in 2016 relative to the heuristic approach is because the models are initiated during an epidemic, therefore lack training data.

Period		Baseline	Google	Heuristic
2016	MAE	199.8	167.8	103.6
	Relative MAE	1.93	1.62	1.0
2017	MAE	9.2	10.0	18.1
	Relative MAE	0.50	0.56	1.0
2018	MAE	54.2	48.0	112.2
	Relative MAE	0.48	0.43	1.0
2019	MAE	264.8	239.8	485.4
	Relative MAE	0.55	0.49	1.0

Table 6.5: **Comparison of baseline and Google model chikungunya nowcasting precision.** This table compares the mean prediction interval widths (MPIs) for nowcasts from the baseline and Google models. The heuristic model is omitted as this approach does not allow a prediction interval to be calculated. The first three columns show results across the sample as a whole. The first column is the MPI reported in number of cases; the second column is the MPI relative to the baseline model; and the third column is the percentage of actual weekly case counts within the prediction interval. The Google model is more precise than the baseline model, reducing MPI by 7%. It is also slightly more reliable, capturing 93% of actual weekly case counts relative to 91% for the baseline model. The final three columns show similar results when considering epidemic periods only. While the intervals are much wider for both models, the Google model again reduces MPI by 8%. The Google intervals are also even more reliable relative to the baseline during epidemics, capturing 93% of weekly case counts relative to 88%.

Model	All periods			Epidemics		
	MPI	Relative MPI	% Correct	MPI	Relative MPI	% Correct
Baseline	524.2	1.00	91.0	873.6	1.00	88.1
Google	485.6	0.93	92.6	805.5	0.92	93.1

the mean prediction interval width by between 4% and 14% each year.

Finally, the Google model is also more reliable than the baseline model. Across the sample as a whole, the Google model interval captures 93% of the ground truth while the baseline model only captures 91%. This difference is larger during epidemics; the Google model captures 92% of the ground truth whereas the baseline

Table 6.6: **Comparison of model intervals by year.** The Google model is more precise than the baseline model in each year. This precision advantage ranges from 4% to 14%. Note that the intervals are overconfident in 2016, likely due to the lack of training data at this point.

Period		Baseline	Google
2016	MPI	650.8	558.3
	Relative MPI	1.0	0.86
	% correct	65.6	62.5
2017	MPI	75.8	71.9
	Relative MPI	1.0	0.95
	% correct	98.1	96.2
2018	MPI	337.4	323.4
	Relative MPI	1.0	0.96
	% correct	96.2	100
2019	MPI	1081.5	1016.8
	Relative MPI	1.0	0.94
	% correct	94.2	100

model captures only 88%. The slight overconfidence of the intervals seems largely driven by the first epidemic, where the model had little data to train on. Table 6.7 shows results excluding the first epidemic. The ground truth falls within the baseline interval 94% of the time, and the Google interval 96% of the time. The intervals are more reliable for both the baseline and Google models This shows that the overconfidence was likely driven by lack of training data for the first epidemic rather than issues with the modelling approach.

Table 6.7: **Prediction intervals excluding the first epidemic.** In Table 6.5, we note that the 95% intervals for both models are slightly overconfident. We suspect that this is due to the first epidemic, where the model must make predictions given little training data. This table shows predictive intervals excluding the first epidemic, as we restrict the sample to dates after 1 September 2016. Both baseline and Google model intervals seem more reliable, with the ground truth falling within the interval 94% and 96% of the time respectively.

	All periods			Epidemics		
Model	MPI	Relative MPI	% correct	MPI	Relative MPI	% correct
Baseline	468.2	1.00	94.2	817.7	1.00	93.1
Google	444.1	0.95	96.0	773.7	0.95	98.9

Panel C of Figure 6.3 shows that some of the periods where the baseline

interval is narrower than the Google model interval occur during epidemics, particularly early in 2019. However, several of the true weekly case counts during the early 2019 epidemic fall outside the baseline interval (see Figure 6.4), but within the Google interval. Therefore, the narrower baseline intervals may not be reliable during an epidemic, which further favours the Google model.

Visual examination of Figure 6.3B suggests that the Google model may be particularly effective relative to the baseline prior to the epidemic peak. We analyse results from 2018 and 2019, as we do not have data from the onset of the 2016 epidemic, and there was no epidemic in 2017. For each year, we consider weeks in which the case count is above the epidemic threshold of 103 cases. We split this data into the period prior to the epidemic peak and the period after the epidemic peak.

Table 6.8 shows that the Google model errors are 16% lower than the baseline in the period prior to the 2018 epidemic peak Similarly, Table 6.9 shows that the Google model errors are 15% lower than the baseline prior to the 2019 peak.

Table 6.8: **Errors during the 2018 epidemic** The Google model errors are 16% lower than the baseline prior to the epidemic peak. Google model errors are 7% lower after the epidemic peak.

	Before peak		Post peak	
Model	MAE	Relative MAE	MAE	Relative MAE
Baseline	99.6	0.44	50.0	0.51
Google	83.5	0.37	46.6	0.48
Heuristic	226.9	1.00	97.1	1.00

Table 6.9: **Errors during the 2019 epidemic** The Google model errors are 15% lower than the baseline prior to the epidemic peak. However, Google errors are 4% higher than the baseline after the epidemic peak.

	Before peak		Post peak	
Model	MAE	Relative MAE	MAE	Relative MAE
Baseline	447.6	0.57	182.7	0.47
Google	380.8	0.49	189.6	0.49
Heuristic	783.7	1.00	385.4	1.00

Similarly, visual comparison of the baseline model's performance in Figure 6.4 to the Google model's performance in Fig. 6.3A suggests that the Google model's prediction intervals may be more reliable in the period prior to the epidemic peak.

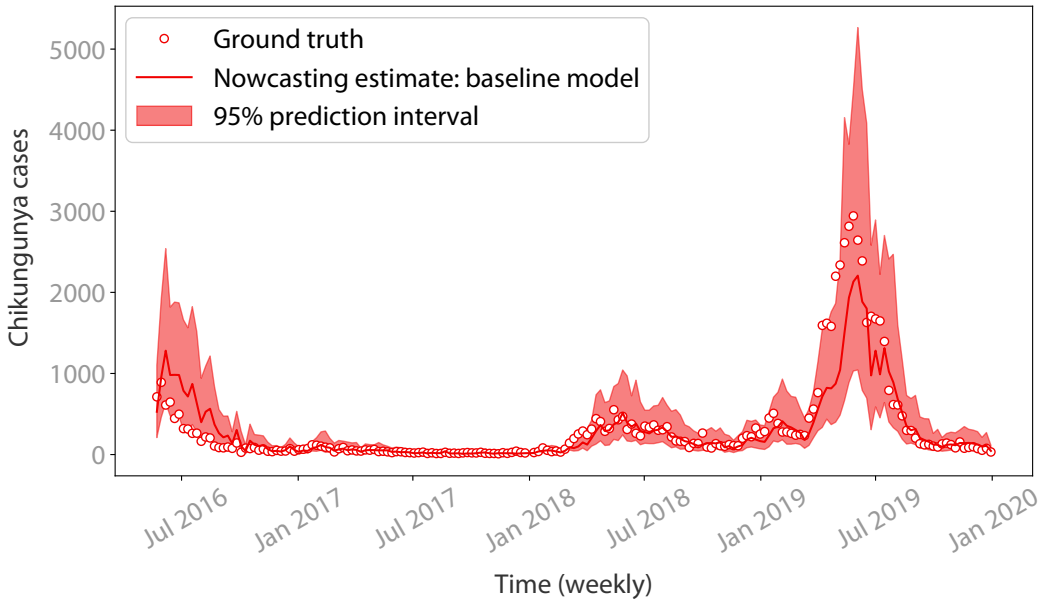

Figure 6.4: **Predictions and intervals from the baseline model.** In Figure 6.3, we note there are some periods where the baseline interval is narrower than the Google interval. However, we note these may be periods where the baseline interval is overconfident. This figure shows baseline model performance over time. At the start of the 2019 epidemic, there are several periods where the ground truth falls outside the baseline model interval. This corresponds to the timespan where the baseline model intervals are narrower than the Google model intervals. Therefore, the narrower baseline intervals may be less reliable in the period before the epidemic peak, which further favours the Google model. We examine this suggestion further in Table 6.10 and Table 6.11.

The prediction intervals produced by the Google model are larger in this period. Table 6.10 shows that the Google model errors are 15% higher than the baseline in the period prior to the 2018 epidemic peak. Similarly, Table 6.11 shows that the Google model errors are 14% higher than the baseline prior to the 2019 peak. However, the actual case counts fall within the baseline prediction intervals only 82% of the time in the period prior to the 2018 epidemic peak and 86% of the time prior to the 2019 peak. By contrast, the actual case counts fall within the Google intervals 100% of the time in the periods prior to both the 2018 and 2019 epidemic peaks. Overall, the Google model appears to be more reliable than the baseline model in the periods before epidemic peaks, displaying lower errors and greater accuracy of the prediction intervals.

There is less of a difference in model performance during the period following the epidemic peak. The baseline model is slightly more accurate during this period.

Table 6.10: **Intervals during the 2018 epidemic** The Google model intervals are 15% higher than the baseline before the epidemic peak. However, the baseline model intervals may be overconfident as the 95% prediction interval is correct only 82% of the time. By contrast, the Google model intervals are correct 100% of the time.

	Before peak			Post peak		
Model	MPI	Relative MPI	% correct	MPI	Relative MPI	% correct
Baseline	383.3	1.00	81.8	406.3	1.00	100.0
Google	439.4	1.15	100.0	360.3	0.89	100.0

Table 6.11: **Intervals during the 2019 epidemic** The Google model intervals are 14% higher than the baseline before the epidemic peak. However, the baseline model intervals may be overconfident as the 95% prediction interval is correct only 86% of the time. By contrast, the Google model intervals are correct 100% of the time.

	Before peak			Post peak		
Model	MPI	Relative MPI	% correct	MPI	Relative MPI	% correct
Baseline	1150.3	1.00	85.7	1369.0	1.00	100.0
Google	1309.2	1.14	100.0	1081.6	0.79	100.0

Table 6.8 shows that, following the 2018 epidemic peak, the Google model errors are 7% lower than the baseline model. Table 6.9 shows that, following the 2019 peak, the Google model errors are 4% higher. These differences are small relative to the differences prior to the epidemic peaks.

There is some evidence the Google model is more precise than the baseline following the epidemic peak. Table 6.10 shows that Google model intervals are 11% narrower than the baseline in 2018. Table 6.11 shows that Google model intervals are 21% narrower than the baseline in 2019. There is no difference in the frequency with which the actual case counts fall within the prediction intervals: for both models this is 100% in both 2018 and 2019.

The strong performance of the Google model during epidemic onset periods may be particularly helpful for providing early warning of epidemics to policymakers. The MEM estimates the threshold for an epidemic to be 103 weekly cases. We investigate which model provides the most timely detection of this threshold being crossed at the beginning of an epidemic. We consider detection of the 2018 epidemic as a case study, as the case count does not sink below 103 cases for more than two continuous weeks before the 2019 epidemic. In 2018, the actual case count crosses

the threshold of 103 cases in epidemiological week 9. In contrast, the heuristic model does not detect threshold crossing until week 15 of the epidemiological year. The baseline model detects the epidemic earlier, estimating that the threshold is crossed in week 12. However the Google model produces the closest estimate, detecting threshold crossing in week 11. The Google model therefore provides four weeks of early warning relative to the heuristic approach, which monitors only entered cases. These four weeks could have been crucial for policymakers seeking to intervene early in order to limit the spread of chikungunya.

6.4 Discussion

This chapter has analysed whether including Google search data improves chikungunya nowcasts in Rio de Janeiro, Brazil. Chikungunya in Rio de Janeiro is seasonal, with attack rates varying from year to year. Early warnings of bad outbreaks are important for delivering timely interventions.

Data on chikungunya cases is usually entered into the surveillance database with a significant delay after diagnosis, making decisions less timely. These delays are also inconsistent, increasing the challenge of estimating current chikungunya case counts from official data alone. Here, we have examined the performance of three approaches to delivering weekly estimates of chikungunya incidence in Rio de Janeiro whilst mitigating against delays and incomplete data. These are a heuristic approach, frequently applied by policymakers in practice, where data on the last few weeks is simply disregarded; a baseline nowcasting model, as currently implemented in the InfoDengue system, where a statistical approach is employed to model the varying delays in the data; and a nowcasting model using Google searches, which augments the baseline nowcasting model with rapidly available Google search data. We evaluate the error of the models' estimates of chikungunya incidence using the data that was available at the end of each week. For the baseline nowcasting model and nowcasting model using Google searches, we also examine the size of the prediction intervals accompanying the estimates, to understand how certain policymakers could be of the estimates delivered.

We find that both the baseline model and the model using Google searches outperform the heuristic approach by some margin. Importantly, while the baseline model performs well, we find that including Google search data reduces both nowcast error and uncertainty relative to the baseline. Our analyses show that including Google search data reduces nowcast errors between May 2016 and December 2019. When considering only epidemic periods, which are particularly important for poli-

cymakers, we find a similar reduction. We further find that including Google search data reduces nowcast uncertainty, reducing prediction intervals by 8% during epidemics and 7% across the sample as a whole. Finally, including Google search data may make prediction intervals more reliable during epidemics. We find that, during epidemics, the prediction interval produced by the model using Google searches captures 93% of weekly case counts compared to 88% for the baseline interval. Our model can be used in practice to generate weekly estimates, despite the significant and varied delays in the entry of chikungunya case count data.

In this analysis, we always train the baseline nowcasting model and nowcasting model using Google searches on all data from earlier weeks, rather than using a fixed training window. Updating the model to include the most recent data is important, as the predictive relationship between search data and disease incidence may change over time [Preis and Moat, 2014; Lazer et al., 2014]. However, previous work on dengue found little evidence that discarding past data in training leads to a reduction in error [Mizzi et al., 2021]. As epidemics are infrequent, using all previous data also helps avoid a situation where the epidemics are lost from the training data due to a shorter training window. However, one important advantage of using a shorter fixed size training window is a reduction in computation time. This advantage accumulates if estimates are being produced for thousands of cities in parallel, as is currently the case for the InfoDengue platform. Future work could further examine the performance of the chikungunya nowcasting approaches outlined here with a fixed size training window, to ensure that this parameter delivers the optimal combination of reduced error and rapid computation.

The analysis we present here focuses on the city of Rio de Janeiro. However, the Google search data is for the state of Rio de Janeiro, rather than the city. Further research could build on our results by testing whether they hold for other cities in the state, whose Google search behaviour may be less well correlated with the state's overall search behaviour. Promising initial indications are provided by previous analyses for dengue that demonstrate that state-level Google data can still help reduce error and uncertainty, even in smaller cities, as shown in Chapter 8 of Mizzi [2019].

Our methods could also be extended to other areas of Brazil, or other arbovirus-prone regions, such as India, which have experienced chikungunya outbreaks [Gopalan and Das, 2009]. Inhabitants of other regions may have a different relationship with the internet. For example, they may use it less frequently to gather information on illness. It would be valuable to analyse whether Google search data is still effective for nowcasting in such scenarios. Future work should therefore con-

sider a wider range of chikungunya-prone regions and states beyond the state of Rio de Janeiro.

A key limitation of the analysis we present here is the relatively short length of the case count time series available for training. This spans four years, and hence approximately 200 weekly data points. Both the baseline model and model using Google searches overestimate the 2016 epidemic. We suggest that this is due to the fact that this epidemic falls at the beginning of the sample, when little model training has been completed. Both models also underestimate the 2019 epidemic. This is likely to be due to that model training through two smaller outbreaks in 2017 and 2018. Nevertheless, the true data points for the 2019 epidemic still fall within the 95% prediction interval of the model using Google data. As further data arrives, it will be possible to continue to monitor the performance of the proposed nowcasting model.

In the analysis described here, we have also only considered one real-time data source. Future nowcasting research could include other data sources, whether measuring other online activity [Mizzi et al., 2021; Marques-Toledo et al., 2017], or properties of the external environment related to arbovirosis incidence, such as the weather [Lowe et al., 2014].

Finally, we note that other arboviruses spread by the *Aedes aegypti* mosquito in Brazil, such as dengue and Zika, exhibit similar symptoms. This can cause challenges for medical practitioners in diagnosis. Similarly, increases in Google searches for one of these diseases could be driven by an increase in cases of another [Mizzi et al., 2021]. Future research could jointly model their incidence, which may be more effective than modelling them independently. If so, policymakers would be able to respond more quickly to epidemics across a range of diseases.

Chapter 7

Discussion

The Covid-19 crisis has emphasised how access to rapid, accurate statistics is crucial across both economics and public health. Policymakers must quickly make decisions on highly impactful policies, such as lockdown, while lacking timely information on both the state of the economy and the spread of disease among the population. Currently, statistics in economics and public health are usually released with a long delay. Statistics authorities are still highly reliant on traditional methods, such as surveys, to produce their statistics. While well-developed, these methods are often costly and take up to several months before producing estimates for crucial statistics. This thesis demonstrates how we can exploit novel data sources, such as aircraft location data, and techniques from machine learning to improve the speed and accuracy of statistics. Some of the statistics we analyse, such as aviation's contribution to GDP, were prone to particularly rapid change during the Covid-19 crisis. Policymakers with more timely data on greatly affected sectors, such as aviation, could in turn have provided support more quickly.

Recent developments in computational social science show great potential for improving nowcasting [Lazer et al., 2009; Moat et al., 2014]. In economics, where the term nowcasting was coined [Giannone et al., 2008], there is now a rich methodological literature on integrating higher frequency data into low-frequency official statistics [Kapetanios and Papailias, 2018]. However, the scope of data sources considered has generally been restricted to existing economic indicators, such as surveys of business sentiment, and online search data [Choi and Varian, 2009b]. There are now far more sources of real-time data available, along with a range of techniques from machine learning with which to transform them into numeric data that is usable in analysis. We introduce two novel sources of real-time data to economic nowcasting: aircraft location data from the ADS-B protocol, and illicit drug transactions

scraped from online "darknet" markets. We analyse concrete policy applications, that use these novel data sources, in nowcasting i) aviation's contribution GDP and ii) illicit drug demand.

Similarly in public health, there is also a well-developed nowcasting literature due the use of nowcasting in epidemiology. Real-time statistics are particularly important for policymakers combatting rapid disease outbreaks [Allard, 1998; Ginsberg et al., 2009], who often need to respond quickly to combat rapidly growing epidemics. However, this literature has often been limited by the short time series of available data on novel data sources for nowcasting disease outbreaks. Studies have therefore been limited to documenting correlations between novel data sources and disease outbreaks, rather than a fuller analysis of out-of-sample nowcast performance. This thesis uses time series of sufficient length to consistently analyse the out-of-sample nowcasting value of novel data sources such as internet search data. A further issue with nowcasting in epidemiology is that ground truth data on infection rates is often delivered with a long and variable delay. Models that rely on having regular, complete information about previous periods in the time series are therefore often unusable in practice. This thesis analyses whether integrating real-time online search data into a modelling approach specifically designed for such situations [Bastos et al., 2019] can help nowcast chikungunya, which has so far received relatively little attention in the nowcasting literature.

7.1 Key contributions

In Chapter 3, we analyse whether novel data sources can help provide faster statistics on airline performance and aviation's contribution to GDP. Aviation is a crucial sector of the global economy, directly contributing at least 3% to GDP in the UK and US. Aviation also provides critical support to other big economic sectors, such as tourism. Current aviation statistics are published only after a two month delay. The Covid-19 crisis has shown how volatile the aviation sector can be, with its contribution to UK GDP shrinking by 97% in April 2020. Rapid estimates of aviation statistics therefore have great value to economic policymakers.

Aircraft now broadcast their location at high frequency in real-time under the ADS-B protocol. In Chapter 3 of this thesis, we have shown that ADS-B data can help nowcast aviation statistics. We first construct a global dataset of flights, covering the period July 2016 to December 2018, from the raw location data. This data is available in real-time, and including it substantially boosts nowcast performance of airline flight volumes, relative to a strong baseline autoregressive model.

These improvements are large: out-of-sample nowcast errors fall by roughly 30% for UK airlines and 20% for US airlines.

We next show that real-time knowledge of flight volumes improves nowcast performance of aviation's contribution to GDP, also relative to a baseline autoregressive model. These improvements are largest during volatile economic times, such as the 2008 recession. Intuitively this may be unsurprising, as autoregressive models perform worst when there are sudden shifts in the series. This is when rapid indicators have the most value to policymakers, as decisions often need to be taken quickly.

Flight volumes are also a key determinant of airport performance. In Chapter 4, we extend the methods from from Chapter 3 to another application of ADS-B data: nowcasting airport performance. We further exploit the spatial dimension of ADS-B data to estimate a takeoff and landing airport for each flight extracted from the historic ADS-B data. We then construct an ADS-B indicator for airport flight volumes, which is available in real-time. Including this indicator alongside a baseline autoregressive model boosts nowcast performance for both UK and US airport flight volumes.

Chapter 4 extends the time series of ADS-B data to April 2020, whereas Chapter 3 ends in December 2018. We am able to build a rich historic dataset of some 117 million flights globally, relative to the 67 million flights identified in Chapter 3. Crucially, the data in Chapter 4 covers the start of the Covid-19 crisis, which marked a sharp decline in air traffic in many countries. The results from Chapter 4 show that models with ADS-B data perform particularly well during this volatile period, which is also when faster statistics are likely to be of greatest value to policymakers.

We build a historic dataset of some 117 million flights in Chapter 4. Both the UK and US customs authorities currently publish granular international trade statistics at monthly frequency. Analysts can further split these statistics by trading partner and port of arrival, so there is a ground-truth measure of how much trade arrives through each UK and US airport from each of their trading partners. Future researchers could construct a matching indicator from ADS-B data: the monthly volume of flights arriving at each airport from each trading partner. Changing flight volumes along a given route could be a leading indicator for trade flows along that route. Economic policymakers would find faster statistics on international trade flows highly useful. It would enable faster responses to supply chain disruption from external shocks, such as the Covid-19 crisis. Policymakers may also be able to more quickly evaluate the impact of major changes in trade policy, such as Brexit,

and respond with appropriately targeted sectoral support.

In Chapter 5, we analyse whether a novel combination of data from online "darknet" markets and Wikipedia page views can be used to nowcast illicit drug demand. Black markets are an important part of the economy, and they have always difficult for economic policymakers to observe. Illicit drugs are one of the largest black markets globally, and the market for illicit drugs is also the source of major public health issues in most developed countries. Currently, policymakers rely on annual surveys to monitor demand for illicit drugs. Arguably, these surveys are too infrequent for responding to rapid changes in drug use, such as the US opioid epidemic. Moreover new "designer" drugs may not appear in these surveys at all.

Owing to recent developments in cryptocurrencies and encrypted web browsing, there are now online markets that enable anonymous trade in illicit goods. On these "darknet" markets, buyers generally must leave feedback in order to enforce a reputation mechanism. We show that scraping this feedback allows for collection of transaction-level data on darknet drug purchases, which is available in real-time. We then show this transaction-level data can build a global dataset of darknet drug demand over time.

However, scraping the darknet is difficult, so this data is not always reliably available. Darknet drug buyers may search online for information about the drugs before making a purchase. These searches may lead them to Wikipedia, which has a unique page for each drug. Wikipedia also has a different page for each major language, which can proxy for country of origin of the searcher. Data on daily Wikipedia page views is reliably available in real-time.

We show that data on Wikipedia page views can help nowcast darknet drug demand across a range of drugs and countries. Augmenting a baseline panel nowcasting model with real-time Wikipedia data substantially boosts accuracy. These improvements are large: out-of-sample nowcast errors are around 40% smaller from the Wikipedia model relative to the baseline. These results hold across a range of time frequencies, from weekly to quarterly. We further show that the Wikipedia-based model has a very direct policy application: spiking page views provided advance warning of the 2016 US fentanyl epidemic, but policymakers failed to declare a state of emergency until 2017.

Nowcasting also has a strong public health policy application through epidemiology. Fast and accurate statistics are crucial to monitoring the spread of disease and responding quickly to epidemics. Chikungunya, a mosquito-born disease (arbovirus), is a growing problem in Brazil, with over 130,000 probable cases recorded in 2019 alone. Infections often lead to severe health complications, with

the resulting workplace absences being economically catastrophic for the low-income households affected. There is a long and variable lag between a patient seeking treatment and entry of the case into the national disease surveillance system. As a result, case count data arrives in a gradual and inconsistent stream. This makes accurate monitoring of chikungunya prevalence particularly difficult, reducing policymakers' ability to respond effectively to fast-growing epidemics.

Data on online searches for chikungunya-related terms, from Google Trends, is reliably available in real-time. In Chapter 6, we analyse whether Google Trends data can improve nowcasts of chikungunya prevalence in Rio de Janeiro. We begin with a Bayesian baseline model [Bastos et al., 2019] designed explicitly for dealing with disease data that arrives gradually and inconsistently. This baseline model far outperforms a "heuristic approach" of estimating based solely on cases already entered into the system. We find that augmenting the baseline with Google Trends data further boosts both model accuracy and precision, which would allow policymakers to be more confident in their estimates of chikungunya prevalence. Moreover, these improvements are largest during epidemics, and particularly at the onset of the epidemic. These are the most crucial periods for policymakers seeking to combat the spread of disease to have fast and accurate statistics.

Importantly, our model in Chapter 6 is operationally usable in situations where data arrives with long and varied delays. This contrasts with traditional nowcasting approaches, which require regular and complete data on previous periods at the time of nowcasting the current period in order to be operationally usable. This is often not the case with epidemiology data, as documented by Mizzi et al. [2021] for dengue fever and further reinforced by the current Covid-19 crisis.

7.2 Key limitations

The main limitation of our results from Chapter 3 is the short time series of the ADS-B data. We found evidence that ADS-B data could nowcast GDP in-sample, but the time series was too short for a more rigorous out-of-sample analysis. Instead, we substituted the ADS-B estimate of aggregate flight volumes in the GDP nowcast for the official estimate of flight volumes, which we showed that ADS-B data could accurately nowcast out-of-sample. Future research will have access to a longer time series of ADS-B data, and should build on our results by rigorously analysing whether ADS-B data can nowcast aviation's contribution to GDP out-of-sample. Moreover, future research will have access to data from the Covid-19 crisis. This will provide an excellent case study of whether ADS-B data is useful during a

crisis that hit aviation particularly hard.

There are also some limitations on the generalisability of our results to sectors other than aviation. Aviation is a significant fraction of GDP in most developed countries, but policymakers would benefit from rapid estimates of other sectors too. Future research could extend the methods in this thesis for analysing aircraft location data to real-time location data from other transport sectors, such as road and shipping.

Lack of geographic coverage limits the generalisability of results from Chapter 4, similarly to Chapter 3. Our results were geographically restricted to the UK and USA. A useful next step would be to check whether these results hold in other regions with good ADS-B coverage, such as mainland Europe. Some countries that are crucial to supply chains and the international trade network, notably China, have very poor ADS-B coverage. It would be particularly useful for future researchers to focus on these regions as the geographic coverage of ADS-B data continues to improve.

Poor ADS-B coverage in one country can negatively impact the accuracy of flight volume estimates for its partners. For example: a researcher attempting to nowcast imports of Chinese goods to UK airports may have good ADS-B coverage around UK airports, but poor coverage around Chinese airports. Therefore they cannot reliably identify that planes arriving at UK airports came from China. Future studies could attempt to infer takeoff and landing locations of flights where one side comes from an area of poor ADS-B coverage.

The primary limit of our results from Chapter 5 is the external validity of extrapolating darknet drug demand to the wider population. Despite growing popularity, darknet drug demand is only a small percentage of overall drug demand. Policymakers are likely to be more interested in overall drug demand, particularly when tracking epidemics. Official statistics on drug use from annual surveys are very low frequency, so there is not a sufficiently long enough time series to directly assess out-of-sample predictive power of data from Wikipedia page views. We instead rely on previous demographic research showing that darknet demand should proxy well for overall drug demand, at least in developed countries. As higher frequency official data on wider drug demand becomes available, future researchers should test directly whether real-time data on Wikipedia page views or darknet transactions can improve nowcast accuracy.

Another limitation comes from proxying the location of the viewer of a Wikipedia page from the page language. This proxy varies in precision. A person viewing a Dutch language Wikipedia page is more likely to be located in the Nether-

lands than elsewhere. However, determining the location of an English language page-viewer is much harder. They could be from several countries in our dataset. There is indeed some evidence that model performance is worse for countries that are less likely to have a precise mapping, such as the UK, US and Australia, than countries with a more precise mapping, such as the Netherlands. Future research could include other sources of online search data, such as Google Trends, which offer more precision regarding the location of the person searching for information.

Data on both darknet drug demand and Wikipedia page views is available at daily frequency, so in principle we could perform the analysis at any frequency up to daily frequency. While including Wikipedia data improved model accuracy at all frequencies up to weekly frequency, the Wikipedia model performed relatively better at lower frequencies. For example, we found Wikipedia data boosted model accuracy more at monthly frequency than weekly. There may therefore be a trade-off between model speed and model accuracy. Faster models are more useful for policymakers, so future research could explore this trade-off in more detail to decide on the optimal frequency of modelling for responding to fast-changing patterns of drug use.

The analysis in Chapter 6 is geographically limited to Rio de Janeiro. It is possible that our results may not generalise as well to other regions. For example, if they have a different relationship to the internet then Google Trends may be a less reliable indicator. Future research should investigate the external validity of our results to other regions, such as other Brazilian cities, as in Chapter 8 of Mizzi [2019], or other chikungunya-prone countries, such as India.

The real-time data we included in Chapter 6 was restricted to Google Trends, which was easy to collect for Rio de Janeiro through the Google Health Trends API. However, the Bayesian modelling framework we used could include more sources of real-time data. Future research could include a much broader set of real-time indicators to further improve nowcast performance. These could be other sources of online data, such as Twitter, or other real-time data already known to affect arbovirus transmission, such as weather data.

In Chapter 6, we analyse chikungunya nowcasting independently of other arboviruses, such as dengue fever. These diseases both share transmission vectors (they are both borne by mosquitoes) and symptoms, which means misdiagnosis is frequently an issue. Through a similar disease monitoring system, data on dengue prevalence is also available in Brazil. Future research may therefore benefit from jointly modelling these diseases when nowcasting their prevalence.

While the novel data sources in this thesis show much promise, we urge

caution in integrating them into the production of official statistics. We emphasise that novel data augments existing data collection, rather than replacing it entirely. We deploy a consistent approach in each chapter of including novel data alongside a baseline model based on existing data collection methods, rather than a model that omits existing data entirely. Indeed, removing existing methods entirely may make nowcast performance worse. By definition, these novel data sources have only been available for a relatively short amount of time compared to traditional data sources. Therefore we generally need a longer time series before we can fully evaluate their efficacy, and determine whether statistics authorities can move on from existing methods.

Another general point of caution is that none of our nowcasting methods represent causal relationships. In each chapter, we claim that novel data sources can help with measurement of official variables, but they cannot provide a policy prescription for affecting the variables that they help to measure. For example, in Chapter 5 we provide evidence that rising numbers of Wikipedia page views for a certain drug may indicate rising demand. If a policymaker were to interpret these results causally, they may conclude that the way to reduce illicit drug demand is to ban Wikipedia pages for those drugs. We are therefore careful to steer clear of any claims of causality in our results, which are limited to helping only with measuring outcomes.

7.3 Future work

There are many avenues for future researchers to build on our results. In doing so, we hope that it will be possible to use nowcasting methodologies and more novel data sources to deliver faster socioeconomic statistics across a range of domains.

Chapter 3 showed that ADS-B data could improve estimates of aviation's contribution to GDP. Future researchers could extend our results by analysing this relationship during the Covid-19 crisis. While we highlight that ADS-B data is most useful during volatile economic times, the data used in Chapter 3 ends before 2020. Similarly, researchers could apply our methodology to real-time data from other methods of transport such as shipping, road and rail. In this way, they could nowcast the contribution to GDP from other sectors. We could then develop similar nowcasting methods for other maroeconomic variables, such as inflation and unemployment, to build a more complete, real-time picture of the economy.

Chapter 4 analysed how ADS-B data could help nowcast airport performance. In doing so, we built a global, real-time dataset of the international flight network.

Future research could use such a dataset to assess whether ADS-B data can help nowcast international trade flows. Authorities collect granular data on the volume of trade going through each airport on a monthly basis. The ADS-B data can match this granularity, and may be a helpful leading indicator for changes in trade dynamics.

Chapter 5 explored the potential for real-time data from darknet drug markets and Wikipedia page views to improve the speed of statistics on drug demand, which are currently released at an annual frequency. To complement this use of data from the online world, future research could also incorporate higher frequency data on drug demand from the physical world, such as drug seizures. In turn, this would strengthen the case of using darknet data to nowcast drug demand in the physical world, rather than just demand on the darknet markets.

Chapter 6 showed how real-time data from Google Trends could improve the accuracy and precision of real-time estimates of chikungunya in Rio de Janeiro, Brazil. Future research could build on these results in several ways. Researchers could incorporate other real-time sources of data that affect the spread of mosquito-borne disease, such as data on humidity. Our methods could also be extended to other regions, although researchers should take care to consider differences in demography if they broaden the geographic scope of analysis. Demographic variables, such as age, race and income, can impact the mechanism underlying the spread of disease. Lastly, as incremental delivery of case data is frequently observed across many disease surveillance scenarios, our methods could be used to nowcast the spread of other diseases such as Zika, or even Covid-19.

7.4 Policy implications

In this thesis, we have demonstrated the value of multiple novel data sources in nowcasting official statistics. These findings cover a wide variety of statistics, ranging from UK GDP to global drug demand and chikungunya incidence in Rio de Janeiro. These findings should help policymakers monitor statistics that are prone to rapid change, thereby responding more quickly to emerging crises such as recessions and pandemics.

Chapter 3 shows that there is great potential for using real-time ADS-B in nowcasting airline performance and aviation's contribution to GDP. These methods could be extended to real-time location data from sectors beyond aviation, such as road and shipping. Nowcasting GDP from an increasing number of sectors may allow for policymakers to have an accurate, real-time estimate of overall GDP in the

future.

Chapter 4 shows that ADS-B data is already a useful indicator for nowcasting airport flight volumes. This indicator is particularly strong during periods of volatility, such as the Covid-19 crisis, when faster statistics are likely to be most valuable to policymakers. Policymakers may therefore benefit greatly in future from ADS-B data as a novel indicator for global measures of economic activity, such as international trade.

Chapter 5 marks the first, to our knowledge, use of darknet or Wikipedia data for nowcasting drug demand. While we only directly analyse demand from online "darknet" markets, we provide evidence that our model could help to nowcast demand from traditional drug markets too. There is a clear policy application in public health for combatting fast-growing drug epidemics, such as the Fentanyl epidemic in the USA. Our results show great potential for this novel data source, and in Section 7.2, we outline several concrete avenues for future researchers using online data to build on our results.

Chapter 6 provides further evidence for the value of online data in nowcasting for epidemiology. It is the first study to nowcast chikungunya in a robust framework where reporting delays are properly accounted for and models are evaluated out-of-sample. As well as suggesting concrete avenues for future research to build upon our results, the methodology deployed could be extended to other diseases with inconsistent reporting, such as Covid-19.

Policymakers are now considering integrating novel data sources into their socieconomic statistics across a range of domains. In their May 2020 Monetary Policy Report, the Bank of England drew on high frequency ADS-B data to illustrate the rapid decline in air traffic at the start of the Covid crisis [Monetary Policy Committee, 2020]. The UK's Office for National Statistics (ONS) recently launched a "faster indicators" initiative, which considers the case for drawing on real-time data from a range of sources to create economic indicators [Nolan, 2019]. In Brazil, the Bastos et al. [2019] methodology that is built upon in Chapter 6 of this thesis is now used to deliver weekly reports of dengue incidence to more than 5,000 local authorities across the entire country.

We conclude that researchers should continue to search for novel, real-time data sources. However, given their novelty, these should be integrated gradually to improve the production of official statistics. Moreover, while helping with measurement, they do not provide an understanding of the causal processes underlying the statistics they help to estimate. These limitations suggested that novel data sources should be considered a complement to existing methods rather than a replacement.

Nevertheless, if they are tested rigorously and integrated cautiously, then novel data sources represent a promising path towards improving the speed and accuracy of official statistics.

References

Knut Are Aastveit and Tørres Trovik. Nowcasting norwegian GDP: The role of asset prices in a small open economy. *Empirical Economics*, 42(1):95–119, 2012.

ADS-B Exchange. Accessible at https://www.adsbexchange.com/data.

Breno S. Aguiar, Camila Lorenz, Flávia Virginio, Lincoln Suesdek, and Francisco Chiaravalloti-Neto. Potential risks of Zika and chikungunya outbreaks in Brazil: A modeling study. *International Journal of Infectious Diseases*, 70:20–29, 2018.

Luca Maria Aiello, Rossano Schifanella, Daniele Quercia, and Francesco Aletta. Chatty maps: Constructing sound maps of urban areas from social media data. *Royal Society Open Science*, 3(3):150690, 2016.

Merve Alanyali. *Quantifying Human Behaviour With Online Images*. PhD thesis, University of Warwick, 2018.

Merve Alanyali, Helen Susannah Moat, and Tobias Preis. Quantifying the relationship between financial news and the stock market. *Scientific Reports*, 3:3578, 2013.

Merve Alanyali, Tobias Preis, and Helen Susannah Moat. Tracking protests using geotagged Flickr photographs. *PLOS ONE*, 11(3):e0150466, 2016.

Robert Allard. Use of time-series analysis in infectious disease surveillance. *Bulletin of the World Health Organization*, 76(4):327–333, 1998.

Richard Alligier and David Gianazza. Learning aircraft operational factors to improve aircraft climb prediction: A large scale multi-airport study. *Transportation Research*, 96:72–95, 2018.

Elena Andreou, Eric Ghysels, and Andros Kourtellos. Should Macroeconomic Forecasters Use Daily Financial Data and How? *Journal of Business & Economic Statistics*, 31(2):240–251, 2013.

Nikoleta Anesti, Simon Hayes, Andre Moreira, and James Tasker. Peering into the present: The Bank's approach to GDP nowcasting. *Bank of England Quarterly Bulletin*, 57(2):122–133, 2017.

Elena Angelini, Gerhard Rünstler, and Marta Bańbura. Estimating and forecasting the Euro area monthly national accounts from a dynamic factor model. Technical Report 953, European Central Bank, 2008.

Sinan Aral and Dylan Walker. Identifying Influential and Susceptible Members of Social Networks. *Science*, 337(6092):337–341, 2012.

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576, Edinburgh, Scotland, UK., 2011. Association for Computational Linguistics.

Manuel Arriaza, Juan Cañas-Ortega, Juan Antonio Cañas-Madueño, and Pablo Ruiz-Aviles. Assessing the visual quality of rural landscapes. *Landscape and Urban Planning*, 69(1):115–125, 2004.

Nikolaos Askitas and Klaus F Zimmermann. Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, 55(2):107–120, 2009.

FRB Atlanta. GDPNow. Accessible at https://www.frbatlanta.org/cqer/research/gdpnow, 2021.

Alan J Auerbach and Yuriy Gorodnichenko. Measuring the Output Responses to Fiscal Policy. *American Economic Journal: Economic Policy*, 4(2):1–27, 2012.

United Kingdom Civil Aviation Authority. UK Airport Data. Accessible at https://www.caa.co.uk/.

Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, 2015.

Duilio Balsamo, Paolo Bajardi, and André Panisson. Firsthand Opiates Abuse on Social Media: Monitoring Geospatial Patterns of Interest Through a Digital Cohort. In *The World Wide Web Conference*, WWW '19, pages 2572–2579, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6674-8.

Daniele Barchiesi, Helen Susannah Moat, Christian Alis, Steven Bishop, and Tobias Preis. Quantifying international travel flows using Flickr. *PLOS ONE*, 10(7): e0128470, 2015a.

Daniele Barchiesi, Tobias Preis, Steven Bishop, and Helen Susannah Moat. Modelling human mobility patterns using photographic data shared online. *Royal Society Open Science*, 2:150046, 2015b.

Monica J. Barratt and Judith Aldridge. Everything you always wanted to know about drug cryptomarkets* (*but were afraid to ask). *International Journal of Drug Policy*, 35:1–6, 2016.

Leonardo Bastos, Theodoros Economou, Marcelo Gomes, Daniel Villela, Trevor Bailey, and Claudia Codeço. Modelling reporting delays for disease surveillance data. *Statistics in Medicine*, 38(22):4363–4377, 2019.

Maria Luiza Almeida Bastos, Francileudo Santos de Abreu, and Geraldo Bezerra da Silva Junior. Inability to work due to Chikungunya virus infection: Impact on public service during the first epidemic in the State of Ceará, northeastern Brazil. *The Brazilian Journal of Infectious Diseases*, 22(3):248–249, 2018.

George E. Battese, Rachel M. Harter, and Wayne A. Fuller. An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, 83(401):28–36, 1988.

Venetia Bell, Lai Wah Co, Sophie Stone, and Gavin‘ Wallis. Nowcasting UK GDP growth. *Bank of England Quarterly Bulletin*, 54(1):58–68, 2014.

Ben S Bernanke and Mark Gertler. Inside the Black Box: The Credit Channel of Monetary Policy Transmission. *Journal of Economic Perspectives*, 9(4):27–48, 1995.

Venkataraman Bhaskar, Robin Linacre, and Stephen Machin. The economic functioning of online drugs markets. *Journal of Economic Behavior & Organization*, 159:426–441, 2019.

Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415), 2012.

Christopher Bonham, Alex Noyvirt, Ioannis Tsalamanis, and Sonia Williams. Analysing port and shipping operations using big data. Technical report, Office for National Statistics, 2018.

Federico Botta, Helen Susannah Moat, and Tobias Preis. Quantifying crowd size with mobile phone and Twitter data. *Royal Society Open Science*, 2:150162, 2015.

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*, 47(9):1690–1703, 2020a.

Federico Botta, Tobias Preis, and Helen Susannah Moat. In search of art: Rapid estimates of gallery and museum visits using Google Trends. *EPJ Data Science*, 9:14, 2020b.

Scott Brave. Another Look at the Correlation Between Google Trends and Initial Unemployment Insurance Claims. Technical report, Chicago Fed Insights, 2020.

Robert Breton, Gareth Clews, Liz Metcalfe, Natasha Milliken, Christopher Payne, Joe Winton, and Ainslie Woods. Research indices using web scraped data. Technical report, Office for National Statistics, 2015.

Robin Burgess, Matthew Hansen, Benjamin A. Olken, Peter Potapov, and Stefanie Sieber. The Political Economy of Deforestation in the Tropics. *The Quarterly Journal of Economics*, 127(4):1707–1754, 2012.

Robin Burgess, Francisco J. M. Costa, and Ben Olken. The Brazilian Amazon's Double Reversal of Fortune. Technical Report 67xg5, Center for Open Science, 2019.

Roberto Calvo-Palomino, Fabio Ricciato, Blaz Repas, Domenico Giustiniano, and Vincent Lenders. Nanosecond-Precision Time-of-Arrival Estimation for Aircraft Signals with Low-Cost SDR Receivers. In *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 272–277, Porto, 2018. IEEE. ISBN 978-1-5386-5298-5.

Yan Carriere-Swallow and Felipe Labbe. Nowcasting with Google Trends in an Emerging Market: Nowcasting with Google Trends in an Emerging Market. *Journal of Forecasting*, 32(4):289–298, 2013.

Alberto Cavallo. Scraped Data and Sticky Prices. *The Review of Economics and Statistics*, 100(1):105–119, 2016.

Alberto Cavallo. Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers. *American Economic Review*, 107(1):283–303, 2017.

Alberto Cavallo and Roberto Rigobon. The Billion Prices Project: Using Online Prices for Measurement and Research. *Journal of Economic Perspectives*, 30(2): 151–178, 2016.

Alberto Cavallo, Brent Neiman, and Roberto Rigobon. Currency Unions, Product Introductions, and the Real Exchange Rate. *The Quarterly Journal of Economics*, 129(2):529–595, 2014.

Alberto Cavallo, W. Erwin Diewert, Robert C Feenstra, Robert Inklaar, and Marcel P Timmer. Using Online Prices for Measuring Real Consumption Across Countries. Working Paper 24292, National Bureau of Economic Research, 2018.

Alberto Cavallo, Gita Gopinath, Brent Neiman, and Jenny Tang. Tariff Passthrough at the Border and at the Store: Evidence from US Trade Policy. Working Paper 26396, National Bureau of Economic Research, 2019.

Jakub Červený and Jan C. van Ours. Cannabis prices on the dark web. *European Economic Review*, 120:103306, 2019.

Graeme Chamberlin. Googling the present. *Economic & Labour Market Review*, 4 (12):59–95, 2010.

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, 2011a.

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, 2011b.

Wesley S. Chan. Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics*, 70(2):223–260, 2003.

Xi Chen and William D. Nordhaus. Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, 108(21):8589–8594, 2011.

Yin-Wong Cheung and Kon S. Lai. Lag Order and Critical Values of the Augmented Dickey Fuller Test. *Journal of Business & Economic Statistics*, 13(3):277–280, 1995.

Hyunyoung Choi and Hal Varian. Predicting initial claims for unemployment benefits. Technical report, Google, 2009a.

Hyunyoung Choi and Hal Varian. Predicting the Present with Google Trends. *Economic Record*, 88:2–9, 2009b.

Alain Yee Loong Chong, Eugene Ch'ng, Martin J. Liu, and Boying Li. Predicting consumer product demands via Big Data: The roles of online promotional marketing and online reviews. *International Journal of Production Research*, 55(17): 5142–5156, 2017.

Nicolas Christin. Traveling the silk road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 213–224, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 978-1-4503-2035-1.

C. Codeço, F. Coelho, O. Cruz, S. Oliveira, T. Castro, and L. Bastos. Infodengue: A nowcasting system for the surveillance of arboviruses in Brazil. *Revue d'Épidémiologie et de Santé Publique*, 66:S386, 2018.

Flavio C. Coelho and Claudia T. Codeço. Precision epidemiology of arboviral diseases. *Journal of Public Health and Emergency*, 3(0), 2019.

European Commission. Commission Implementing Regulation (EU) No 1207/2011, 2011.

Rosaria Conte, Nigel Gilbert, Giulia Bonelli, Claudio Cioffi-Revilla, Guillaume Deffuant, Janos Kertesz, Vittorio Loreto, Suzy Moat, J-P Nadal, Anxo Sanchez, Andrzej Nowak, Andreas Flache, Maxi San Miguel, and Dirk Helbing. Manifesto of computational social science. *The European Physical Journal Special Topics*, 214:325–346, 2012.

Patrick Copeland, Raquel Romano, Tom Zhang, Greg Hecht, Dan Zigmond, and Christian Stefansen. Google Disease Trends: An Update. In *International Society of Neglected Tropical Diseases 2013*, page 3, 2013.

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, 2014.

Chester Curme, Ying Daisy Zhuo, Helen Susannah Moat, and Tobias Preis. Quantifying the diversity of news around stock market moves. *Journal of Network Theory in Finance*, 3(1):1–20, 2017.

Zhi Da, Joseph Engelberg, and Pengjie Gao. In Search of Attention. *The Journal of Finance*, 66(5):1461–1499, 2011.

Francesco D'Amuri. Predicting unemployment in short samples with internet job search query data. MPRA Paper 18403, University Library of Munich, Germany, 2009.

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.

Carlos Alexandre Antunes de Brito. Alert: Severe cases and deaths associated with Chikungunya in Brazil. *Revista da Sociedade Brasileira de Medicina Tropical*, 50: 585–589, 2017.

Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347 (6221):536–539, 2015.

Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

Jakob Demant, Rasmus Munksgaard, David Decary-Hetu, and Judith Aldridge. Going Local on a Global Platform: A Critical Analysis of the Transformative Potential of Cryptomarkets for Organized Illicit Drug Crime. *International Criminal Justice Review*, 28(3):255–274, 2018.

Martin Dittus, Joss Wright, and Mark Graham. Platform Criminalism: The 'Last-Mile' Geography of the Darknet Market Supply Chain. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 277–286, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-5639-8.

Taeyoung Doh and Jaeheung Bae. Tracking U.S. GDP in Real Time. *Federal Reserve Bank Atlanta Economic Review*, Q3:5–19, 2019.

Dave Donaldson and Adam Storeygard. The View from Above: Applications of Satellite Data in Economics. *Journal of Economic Perspectives*, 30(4):171–198, 2016.

J.-P. Eckmann, E. Moses, and D. Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences*, 101 (40):14333–14337, 2004.

Abeer ElBahrawy, Laura Alessandretti, Anne Kandler, Romualdo Pastor-Satorras, and Andrea Baronchelli. Evolutionary dynamics of the cryptocurrency market. *Royal Society Open Science*, 4(11), 2017.

Abeer ElBahrawy, Laura Alessandretti, and Andrea Baronchelli. Wikipedia and Digital Currencies: Interplay Between Collective Attention and Market Performance. *Frontiers in Blockhain*, 2:12, 2019.

EMCDDA. European Drug Report 2019: Trends and Developments. Technical report, European Monitoring Centre for Drugs and Drug Addiction, 2019.

Michael Ettredge, John Gerdes, and Gilbert Karuga. Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, 48(11):87–92, 2005.

Martin D. D. Evans. Where Are We Now? Real-Time Estimates of the Macroeconomy. *International Journal of Central Banking*, 1(2):49, 2005.

FAA. ADS–B Out Performance Requirements To Support Air Traffic Control (ATC) Service; Final Rule, 2010.

FAA. US Airline Data. Accessible at https://www.transtats.bts.gov/, 2020.

Federal Aviation Authority FAA. The Economic Impact of Civil Aviation on the U.S. Economy. Federal Aviation Authority Report, Federal Aviation Authority, 2016.

Neil M. Ferguson, Derek A.T. Cummings, Simon Cauchemez, Christophe Fraser, Steven Riley, Aronrag Meeyai, Sopon Iamsirithaworn, and Donald S. Burke. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 437(7056):209–214, 2005.

Y. Fondeur and F. Karamé. Can Google data help predict French youth unemployment? *Economic Modelling*, 30:117–125, 2013.

Richard G. Frank and Harold A. Pollack. Addressing the Fentanyl Threat to Public Health. *New England Journal of Medicine*, 376(7):605–607, 2017.

John W. Galbraith and Greg Tkacz. Nowcasting with payments system data. *International Journal of Forecasting*, 34(2):366–376, 2018.

Sandro Galea, Raina M. Merchant, and Nicole Lurie. The Mental Health Consequences of COVID-19 and Physical Distancing: The Need for Prevention and Early Intervention. *JAMA Internal Medicine*, 180(6):817–818, 2020.

Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113, 2017.

Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, and Reid Priedhorsky. Global Disease Monitoring and Forecasting with Wikipedia. *PLoS Computational Biology*, 10(11):e1003892, 2014.

Eric Ghysels, Pedro Santa-Clara, and Rossen Valkanov. The MIDAS Touch: Mixed Data Sampling Regression Models. CIRANO Working Paper, CIRANO, 2004.

Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676, 2008.

Domenico Giannone, Lucrezia Reichlin, and Saverio Simonelli. Nowcasting Euro Area Economic Activity in Real-Time: The Role of Confidence Indicators. Technical Report 240, Centre for Studies in Economics and Finance (CSEF), University of Naples, Italy, 2009.

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, 2009.

Edward L Glaeser, Hyunjin Kim, and Michael Luca. Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity. Working Paper 24010, National Bureau of Economic Research, 2017.

Rebecca Tave Gluskin, Michael A. Johansson, Mauricio Santillana, and John S. Brownstein. Evaluation of Internet-Based Dengue Query Data: Google Dengue Trends. *PLOS Neglected Tropical Diseases*, 8(2):e2713, 2014.

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, 2010.

Janaína Gomide, Adriano Veloso, Wagner Meira, 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 1–8, Koblenz, Germany, 2011. Association for Computing Machinery. ISBN 978-1-4503-0855-7.

Google Trends. Accessible at https://trends.google.com/trends/, 2022.

Saji Saraswathy Gopalan and Ashis Das. Household economic impact of an emerging disease in terms of catastrophic out-of-pocket health care expenditure and loss of productivity: Investigation of an outbreak of chikungunya in Orissa, India. *Journal of Vector Borne Diseases*, 46(1):57–64, 2009.

Mark Graham, Stefano De Sabbata, and Matthew A. Zook. Towards a study of information geographies: (im)mutable augmentations and a mapping of the geographies of information. *Geo: Geography and Environment*, 2(1):88–105, 2015.

Robert Griffioen, Jan de Haan, and Leon Willenborg. Collecting clothing data from the Internet. Technical report, Statistics Netherlands, 2014.

Mariaflavia Harari. Cities in Bad Shape: Urban Geometry in India. *American Economic Review*, 110(8):2377–2421, 2020.

J. Vernon Henderson, Adam Storeygard, and David N. Weil. Measuring Economic Growth from Outer Space. *American Economic Review*, 102(2):994–1028, 2012.

Scott Higham, Sari Horwitz, and Katie Zezima. Obama officials failed to focus as fentanyl burned its way across America - Washington Post. https://www.washingtonpost.com/graphics/2019/national/fentanyl-epidemic-obama-administration/, 2019.

Highways England. Highways england open data sources, 2022. URL http://tris.highwaysengland.co.uk/.

Patrick Hochedez, Stephane Jaureguiberry, Monique Debruyne, Philippe Bossi, Pierre Hausfater, Gilles Brucker, Francois Bricaire, and Eric Caumes. Chikungunya Infection in Travelers. *Emerging Infectious Diseases*, 12(10):1565–1567, 2006.

Stefano Maria Iacus, Fabrizio Natale, Carlos Santamaria, Spyridon Spyratos, and Michele Vespe. Estimating and projecting air passenger traffic during the COVID-19 coronavirus outbreak and its socio-economic impact. *Safety Science*, 129: 104791, 2020.

Seema Jayachandran. Air Quality and Early-Life Mortality: Evidence from Indonesia's Wildfires. *The Journal of Human Resources*, 44(4):916–954, 2009.

N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301): 790–794, 2016.

George Kapetanios and Fotis Papailias. Big Data & Macroeconomic Nowcasting: Methodological Review. Technical Report ESCoE DP-2018-12, Economic Statistics Centre of Excellence (ESCoE), 2018.

Mate Kapitány-Fövény and Zsolt Demetrovics. Utility of Web search query data in testing theoretical assumptions about mephedrone. https://pubmed.ncbi.nlm.nih.gov/28657189/, 2017.

Ashish Kapoor, Zachary Horvitz, Spencer Laube, and Eric Horvitz. Airplanes aloft as a sensor network for wind forecasting. In *IPSN-14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*, pages 25–33, Berlin, 2014. IEEE. ISBN 978-1-4799-3146-0 978-1-4799-3147-7.

Gene Kindberg-Hanlon and Andrej Sokol. Gauging the globe: The Bank's approach to nowcasting world GDP. Technical report, Bank of England, 2018.

Michal Kosinski, Sandra C. Matz, Samuel D. Gosling, Vesselin Popov, and David Stillwell. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70 (6):543–556, 2015.

Gueorgi Kossinets and Duncan J. Watts. Empirical Analysis of an Evolving Social Network. *Science*, 311(5757):88–90, 2006.

Ladislav Kristoufek. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, 3(1), 2013.

Ladislav Kristoufek, Helen Susannah Moat, and Tobias Preis. Estimating suicide occurrence statistics using Google Trends. *EPJ Data Science*, 5:32, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

Kristy Kruithof, Judith Aldridge, David Decary Hetu, Megan Sim, Elma Dujso, and Stijn Hoorens. Internet-facilitated drugs trade: An analysis of the size, scope and the role of the Netherlands. Product Page, RAND Corporation, 2016.

Stephen Law, Brooks Paige, and Chris Russell. Take a Look Around: Using Street View and Satellite Images to Estimate House Prices. *ACM Transactions on Intelligent Systems and Technology*, 10(5):54:1–54:19, 2019.

D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Computational Social Science. *Science*, 323(5915):721–723, 2009.

David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6176):1203–1205, 2014.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521 (7553):436–444, 2015.

Allen Yilun Lin, Justin Cranshaw, and Scott Counts. Forecasting U.S. Domestic Migration Using Internet Search Queries. In *The World Wide Web Conference on - WWW '19*, pages 1061–1072, San Francisco, CA, USA, 2019. ACM Press. ISBN 978-1-4503-6674-8.

Ira M. Longini, Azhar Nizam, Shufu Xu, Kumnuan Ungchusak, Wanna Hanshaoworakul, Derek A. T. Cummings, and M. Elizabeth Halloran. Containing Pandemic Influenza at the Source. *Science*, 309(5737):1083–1087, 2005.

Rachel Lowe, Christovam Barcellos, Caio A S Coelho, Trevor C Bailey, Giovanini Evelim Coelho, Richard Graham, Tim Jupp, Walter Massa Ramalho, Marilia Sá Carvalho, David B Stephenson, and Xavier Rodó. Dengue outlook for the World Cup in Brazil: An early warning model framework driven by real-time seasonal climate forecasts. *The Lancet Infectious Diseases*, 14(7):619–626, 2014.

Michael Luca and Georgios Zervas. Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. *Management Science*, 62(12):3412–3427, 2016.

Lawrence C. Madoff, David N. Fisman, and Taha Kass-Hout. A New Approach to Monitoring Dengue Activity. *PLOS Neglected Tropical Diseases*, 5(5):e1215, 2011.

Maimuna S. Majumder, Mauricio Santillana, Sumiko R. Mekaru, Denise P. McGinnis, Kamran Khan, and John S. Brownstein. Utilizing Nontraditional Data Sources for Near Real-Time Estimation of Transmission Dynamics During the 2015-2016 Colombian Zika Virus Disease Outbreak. *JMIR public health and surveillance*, 2(30), 2016.

Christopher Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999. ISBN 978-0-262-30379-8.

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, 2017.

Amaryllis Mavragani. Tracking COVID-19 in Europe: Infodemiology Approach. *JMIR Public Health and Surveillance*, 6(2):e18941, 2020.

Russell McKenna, Jann Michael Weinand, Ismir Mulalic, Stefan Petrović, Kai Mainzer, Tobias Preis, and Helen Susannah Moat. Scenicness assessment of onshore wind sites with geotagged photographs and impacts on approval and cost-efficiency. *Nature Energy*, 6:663–672, 2021.

Nick McLaren and Rachana Shanbhogue. Using Internet Search Data as Economic Indicators. Quarterly Bulletin 134, Bank of England, 2011.

Connor McMahon, Isaac L. Johnson, and Brent J. Hecht. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. In *ICWSM*, volume 11. AAAI Publications, 2017.

Marton Mestyan, Taha Yasseri, and Janos Kertesz. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. *PLoS ONE*, 8(8):e71226, 2013.

Sam Miller, Abeer El-Bahrawy, Martin Dittus, Mark Graham, and Joss Wright. Predicting drug demand with Wikipedia views: Evidence from darknet markets. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen, editors, *Proceedings of The Web Conference 2020*, pages 2669—2675, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450370233.

Sam Miller, Helen Susannah Moat, and Tobias Preis. Using aircraft location data to estimate current economic activity. *Scientific Reports*, 10:7576, 2020b.

Sam Miller, Tobias Preis, Giovanni Mizzi, Leonardo Soares Bastos, Marcelo Ferreira da Costa Gomes, Flávio Codeço Coelho, Claudia Torres Codeço, and Helen Susannah Moat. Faster indicators of chikungunya incidence using Google searches. *PLOS Neglected Tropical Diseases*, 16(6):e0010441, 2022.

Giovanni Mizzi. *Improving Dengue Fever Surveillance with Online Data*. PhD thesis, University of Warwick, 2019.

Giovanni Mizzi, Tobias Preis, Leonardo Soares Bastos, Marcelo Ferreira da Costa Gomes, Claudia Torres Codeço, and Helen Susannah Moat. Faster indicators of dengue fever case counts using Google and Twitter. arXiv [preprint], arXiv:2112.12101, 2021. URL https://arxiv.org/abs/2112.12101.

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, 2013.

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, 2014.

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, 2016.

David Molnar, Serge Egelman, and Nicolas Christin. This is our data on drugs: Lessons computer security can learn from the drug war. In *Proceedings of the 2010 Workshop on New Security Paradigms - NSPW '10*, page 143, Concord, Massachusetts, USA, 2010. ACM Press. ISBN 978-1-4503-0415-3.

Monetary Policy Committee. Monetary Policy Report – May 2020. Technical report, Bank of England, 2020.

Gordon E. Moore. Cramming More Components Onto Integrated Circuits. *Proceedings of the IEEE*, 86(1):82–85, 1998.

National Institute on Drug Abuse. Overdose Death Rates. Accessible at https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates, 2019.

Felipe Gomes Naveca, Ingra Claro, Marta Giovanetti, Jaqueline Goes de Jesus, Joilson Xavier, Felipe Campos de Melo Iani, Valdinete Alves do Nascimento, Victor Costa de Souza, Paola Paz Silveira, José Lourenço, Mauricio Santillana, Moritz U. G. Kraemer, Josh Quick, Sarah C. Hill, Julien Thézé, Rodrigo Dias de Oliveira Carvalho, Vasco Azevedo, Flavia Cristina da Silva Salles, Márcio Roberto Teixeira Nunes, Poliana da Silva Lemos, Darlan da Silva Candido, Glauco de Carvalho

Pereira, Marluce Aparecida Assunção Oliveira, Cátia Alexandra Ribeiro Meneses, Rodrigo Melo Maito, Claudeth Rocha Santa Brígida Cunha, Daniela Palha de Sousa Campos, Marcia da Costa Castilho, Thalita Caroline da Silva Siqueira, Tiza Matos Terra, Carlos F. Campelo de Albuquerque, Laura Nogueira da Cruz, André Luis de Abreu, Divino Valerio Martins, Daniele Silva de Moraes Vanlume Simoes, Renato Santana de Aguiar, Sérgio Luiz Bessa Luz, Nicholas Loman, Oliver G. Pybus, Ester C. Sabino, Osnei Okumoto, Luiz Carlos Junior Alcantara, and Nuno Rodrigues Faria. Genomic, epidemiological and digital surveillance of Chikungunya virus in the Brazilian Amazon. *PLoS Neglected Tropical Diseases*, 13(3), 2019.

Louisa Nolan. Faster indicators of UK economic activity. Accessible at https://datasciencecampus.ons.gov.uk/faster-indicators-of-uk-economic-activity/, 2019.

Office for National Statistics. Air transport index. Accessible at https://www.ons.gov.uk/economy/.

Piero Olliaro, Florence Fouque, Axel Kroeger, Leigh Bowman, Raman Velayudhan, Ana Carolina Santelli, Diego Garcia, Ronald Skewes Ramm, Lokman H. Sulaiman, Gustavo Sanchez Tejeda, Fabiàn Correa Morales, Ernesto Gozzer, César Basso Garrido, Luong Chan Quang, Gamaliel Gutierrez, Zaida E. Yadon, and Silvia Runge-Ranzinger. Improved tools and strategies for the prevention and control of arboviral diseases: A research-to-policy forum. *PLoS Neglected Tropical Diseases*, 12(2), 2018.

Donald R. Olson, Kevin J. Konty, Marc Paladini, Cecile Viboud, and Lone Simonsen. Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. *PLoS Computational Biology*, 9(10):e1003256, 2013.

Oxford Economics. Economic Benefits from Air Transport in the UK. Technical report, 2014.

Michael J Paul and Mark Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. In *The International AAAI Conference on Web and Social Media*, volume 5, pages 265–272, 2011.

Jaroslav Pavlicek and Ladislav Kristoufek. Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries. *PLOS ONE*, 10 (5):e0127084, 2015.

Robert Todd Perdue, James Hawdon, and Kelly M. Thames. Can Big Data Predict the Rise of Novel Drug Abuse? *Journal of Drug Issues*, 48(4):508–518, 2018.

Matija Piškorec, Nino Antulov-Fantulin, Petra Kralj Novak, Igor Mozetič, Miha Grčar, Irena Vodenska, and Tomislav Šmuc. Cohesiveness in Financial News and its Relation to Market Volatility. *Scientific Reports*, 4(1):5038, 2014.

Philip M. Polgreen, Yiling Chen, David M. Pennock, Forrest D. Nelson, and Robert A. Weinstein. Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*, 47(11):1443–1448, 2008.

Tobias Preis and Helen Susannah Moat. Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society Open Science*, 1:140095, 2014.

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 A: Mathematical, Physical and Engineering Sciences*, 368:5707–5719, 2010.

Tobias Preis, Helen Susannah Moat, H. Eugene Stanley, and Steven R. Bishop. Quantifying the advantage of looking forward. *Scientific Reports*, 2:350, 2012.

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, 2013a.

Tobias Preis, Helen Susannah Moat, and H. Eugene Stanley. Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports*, 3:1684, 2013b.

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*, 52:471–477, 2020.

Rob Procter, Farida Vis, and Alex Voss. Reading the riots on Twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, 16(3):197–214, 2013.

Eulogio Real, Constantino Arce, and José Manuel sabucedo. Classification of landscapes using quantiative and categorical data, and prediction of their scenic beauty in North-Western Spain. *Journal of Environmental Psychology*, 20(4): 355–373, 2000.

Edward Rowland. Faster indicators of UK economic activity: Road traffic data for England. Data Science Campus Report, Office for National Statistics, 2019.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2): 319–392, 2009.

Silvia Runge-Ranzinger, Olaf Horstick, Michael Marx, and Axel Kroeger. What does dengue disease surveillance contribute to predicting and detecting outbreaks and describing trends? *Tropical Medicine & International Health*, 13(8):1022–1041, 2008.

Clémentine Schilte, Frederik Staikowsky, Frédérik Staikovsky, Thérèse Couderc, Yoann Madec, Florence Carpentier, Somar Kassab, Matthew L. Albert, Marc Lecuit, and Alain Michault. Chikungunya virus-associated long-term arthralgia: A 36-month prospective longitudinal study. *PLoS Neglected Tropical Diseases*, 7 (3):e2137, 2013.

Secretaria de Vigilância em Saúde. Monitoramento dos casos de arboviroses urbanas transmitidas pelo Aedes (dengue, chikungunya e Zika), Semanas Epidemiológicas 01 a 52. Technical Report 51, Ministério da Saúde, 2020.

Eric W. K. See-To and Eric W. T. Ngai. Customer reviews for demand distribution and sales nowcasting: A big data approach. *Annals of Operations Research*, 270 (1-2):415–431, 2018.

Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. Quantifying the impact of scenic environments on health. *Scientific Reports*, 5:16899, 2015.

Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. Quantifying the link between art and property prices in urban neighbourhoods. *Royal Society Open Science*, 3:160146, 2016.

Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. Using deep learning to quantify the beauty of outdoor places. *Royal Society Open Science*, 4: 170170, 2017.

Chanuki Illushka Seresinhe, Helen Susannah Moat, and Tobias Preis. Quantifying scenic areas using crowdsourced data. *Environment and Planning B: Urban Analytics and City Science*, 45:567–582, 2018.

Chanuki Illushka Seresinhe, Tobias Preis, George MacKerron, and Helen Susannah Moat. Happiness is greater in more scenic locations. *Scientific Reports*, 9:4498, 2019.

SINAN: Sistema de Informação de Agravos de Notificação. Accessible at http://portalsinan.saude.gov.br/.

Aaron Smith and Monica Anderson. Social Media Use 2018: Demographics and Statistics. Accessible at https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/, 2018.

Silas W. Smith and Fiona M. Garlich. Chapter 3 - Availability and Supply of Novel Psychoactive Substances. In Paul I. Dargan and David M. Wood, editors, *Novel Psychoactive Substances*, pages 55–77. Academic Press, Boston, 2013. ISBN 978-0-12-415816-0.

Kyle Soska and Nicolas Christin. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *Proceedings of the 24th USENIX Conference on Security Symposium*, SEC'15, pages 33–48, USA, 2015. USENIX Association. ISBN 978-1-931971-23-2.

Wataru Souma, Irena Vodenska, and Hideaki Aoyama. Enhanced news sentiment analysis using deep learning methods. *Journal of Computational Social Science*, 2(1):33–46, 2019.

Martin Strohmeier, Matthew Smith, Vincent Lenders, and Ivan Martinovic. The Real First Class? Inferring Confidential Corporate Mergers and Government Relations from Air Traffic Communication. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 107–121, London, 2018. IEEE. ISBN 978-1-5386-4228-3.

Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. User Intent, Behaviour, and Perceived Satisfaction in Product Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 547–555, 2018.

Tanya Suhoy. Query Indices and a 2008 Downturn: Israeli Data. Bank of Israel Working Papers 2009.06, Bank of Israel, 2009.

Qin Sun, Haibo Qiu, Mao Huang, and Yi Yang. Lower mortality of COVID-19 by early recognition and intervention: Experience from Jiangsu Province. *Annals of Intensive Care*, 10(1):33, 2020.

Paul C. Sutton and Robert Costanza. Global estimates of market and non-market values derived from nighttime satellite imagery, land cover, and ecosystem service valuation. *Ecological Economics*, 41(3):509–527, 2002.

Paul C Sutton, Christopher D Elvidge, and Tilottama Ghosh. Estimation of Gross Domestic Product at Sub-National Scales using Nighttime Satellite Imagery. *International Journal of Ecological Economics and Statistics*, 8:17, 2007.

Nataliya Tkachenko, Sarunkorn Chotvijit, Neha Gupta, Emma Bradley, Charlotte Gilks, Weisi Guo, Henry Crosby, Eliot Shore, Malkiat Thiarai, Rob Procter, and Stephen Jarvis. Google Trends can improve surveillance of Type 2 diabetes. *Scientific Reports*, 7(1):4993, 2017a.

Nataliya Tkachenko, Stephen Jarvis, and Rob Procter. Predicting floods with Flickr tags. *PLOS ONE*, 12(2):e0172870, 2017b. Publisher: Public Library of Science.

Derek K Tracy, David M Wood, and David Baumeister. Novel psychoactive substances: Identifying and managing acute and chronic harmful use. *British Medical Journal*, 356, 2017.

Margarita Triguero-Mas, Payam Dadvand, Marta Cirach, David Martínez, Antonia Medina, Anna Mompart, Xavier Basagaña, Regina Gražulevičienė, and Mark J. Nieuwenhuijsen. Natural outdoor environments and mental and physical health: Relationships and mechanisms. *Environment International*, 77:35–41, 2015.

Roman Trub, Daniel Moser, Matthias Schafer, Rui Pinheiro, and Vincent Lenders. Monitoring Meteorological Parameters with Crowdsourced Air Traffic Control Data. In *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 25–36, Porto, 2018. IEEE.

David Tsurel, Dan Pelleg, Ido Guy, and Dafna Shahaf. Fun Facts: Automatic Trivia Fact Extraction from Wikipedia. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 345–354, New York, NY, USA, 2017. ACM.

United Nations Office on Drugs and Crime. World Drug Report. UN Technical Report, United Nations, 2019.

United States Bureau of Economic Analysis. GDP by Industry. Accessible at https://www.bea.gov/data/gdp/gdp-industry.

Tomás Vega, Jose Eugenio Lozano, Tamara Meerhoff, René Snacken, Joshua Mott, Raul Ortiz de Lejarazu, and Baltazar Nunes. Influenza surveillance in Europe: Establishing epidemic thresholds by the Moving Epidemic Method. *Influenza and Other Respiratory Viruses*, 7(4):546–558, 2013.

Simeon Vosen and Torsten Schmidt. Forecasting private consumption: Survey-based indicators vs. Google trends. *Journal of Forecasting*, 30(6):565–578, 2011.

M. Mitchell Waldrop. The chips are down for Moore's law. *Nature News*, 530(7589): 144, 2016.

Catharine Ward Thompson, Jenny Roe, Peter Aspinall, Richard Mitchell, Angela Clow, and David Miller. More green space is linked to less stress in deprived communities: Evidence from salivary cortisol patterns. *Landscape and Urban Planning*, 105(3):221–229, 2012.

Alexandra Weilenmann, Thomas Hillman, and Beata Jungselius. Instagram at the museum: Communicating the museum experience through social photo sharing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, page 1843, Paris, France, 2013. ACM Press. ISBN 978-1-4503-1899-0.

Rachel S. Wightman, Jeanmarie Perrone, and Lewis S. Nelson. Comparative Analysis of Opioid Queries on Erowid.org: An Opportunity to Advance Harm Reduction. *Substance Use & Misuse*, 52(10):1315–1319, 2017.

Wikidata API. Accessible at https://www.wikidata.org/wiki/Q243257.

Wikipedia API. Accessible at https://www.mediawiki.org/wiki.

Annelies Wilder-Smith, Duane J. Gubler, Scott C. Weaver, Thomas P. Monath, David L. Heymann, and Thomas W. Scott. Epidemic arboviral diseases: Priorities for research and public health. *The Lancet Infectious Diseases*, 17(3):e101–e106, 2017.

Shihao Yang, Mauricio Santillana, and S. C. Kou. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences of the United States of America*, 112(47):14473–14478, 2015.

Shihao Yang, Samuel C. Kou, Fred Lu, John S. Brownstein, Nicholas Brooke, and Mauricio Santillana. Advances in using Internet searches to track dengue. *PLOS Computational Biology*, 13(7):e1005607, 2017.

Matthew S. Yiu and Kenneth K. Chow. Nowcasting Chinese GDP: Information content of economic and financial data. *China Economic Journal*, 3(3):223–240, 2010.

Mitsuo Yoshida, Yuki Arase, Takaaki Tsunoda, and Mikio Yamamoto. Wikipedia Page Views Reflect Web Search Trends. In *Proceedings of the ACM Web Science Conference - WebSci '15*, pages 1–2, Oxford, United Kingdom, 2015. ACM Press. ISBN 978-1-4503-3672-7.

Jinhui Zhao, Tim Stockwell, and Scott Macdonald. Non-response bias in alcohol and drug population surveys: Non-response bias in surveys. *Drug and Alcohol Review*, 28(6):648–657, 2009.

Andrey Zheluk, Casey Quinn, and Peter Meylakhs. Internet Search and Krokodil in the Russian Federation: An Infoveillance Study. *Journal of Medical Internet Research*, 16(9):e212, 2014.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and Resolution of Rumours in Social Media: A Survey. *ACM Computing Surveys*, 51(2):32:1–32:36, 2018.