Do Google Search Queries Contain Relevant Information on Job Separation?

A Mixed-Frequency Modelling Approach

Master of Research - Department of Economics, Macquarie University

Samir Sultani

December 4, 2015

Statement of Candidature

I certify that the work in this thesis entitled "Do Google Search Queries Contain Relevant Information on Job Separation? A Mixed-Frequency Modelling Approach" has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree to any other university or institution other than Macquarie University.

In addition, I certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged.

I certify that all information sources and literature used are indicated in the thesis.

Samir Sultani

Samir Sultani-42453348

Abstract

There is a growing interest in alternative sources of data within the economic literature. These sources are referred to as 'Big Data'. Internet search queries, are one such data source, providing an avenue to discern the real-time behaviour of users, searching for information online. This thesis explores whether weekly Google search query data contains informationally-relevant signals on the composition of the US labour market. That is, search queries seemingly employed by Internet users in fear of losing their jobs, or planning to quit the labour force altogether, for example *'unemployment insurance'*.

A weekly composite search index is constructed from the search query data, in order to utilise all the data available in the given period. The relationship between Google search and the unemployment rate is modelled using a recently developed technique in the mixed-frequency time series literature to model the weekly Google search data and the corresponding job separations data, specifically, Ghysels et al's (2015) model. To assess the informational content of Internet search query data, a mixed-frequency Granger causality test is conducted.

It is established that there is insufficient evidence to suggest that Internet search query data is useful in predicting future job separation statistics.

Acknowledgements

There are two groups of people who helped me complete my thesis. The first were my supervisors - Dr Chris Heaton and Dr Edwin Franks - who gladly offered their time and expertise, for which I am very thankful.

The second group of people are my parents, who supported me with their love and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent times, the economic discipline has been introduced to a vast range of novel data sources termed 'Big Data'. This is partly due to the growth in information technology, making data promptly accessible, available in larger scales, and formed on unique variables. Internet search query data is one such source, which attempts to discern the real-time behaviour of users, searching for information online. That is, to yield signals of the intent of economic agents on the internet (see Da et al. (2015)). Such novel data sources are sampled more frequently, for example weekly, in comparison to conventional sources. A current venture in the economic literature is to determine whether such novel data sources are informationally relevant in assessing various phenomena.

Most applications of internet search query data have been modelled by temporally-aggregating the data. That is, the high-frequency internet search query data are pre-filtered to align the mixed-frequency time series at one low frequency. It is well known in the existing literature that many economic relationships arise between variables sampled at different frequencies (see for example Ghysels (2015)). Moreover, contemporary econometrics contends that pre-filtering techniques such as temporal aggregation can potentially lead to a loss of information, as one averages over the high-frequency data (see Rossana and Seater (1995), Wohlrabe (2009) and Foroni and Marcellino (2013) among others). As such, directly modelling mixed-frequency data may be of greater use.

This thesis examines the informational content of Google search queries in relation to job separation in the US. We hypothesise that Google search queries are informationally relevant in predicting future labour job separation. The rationale to this hypothesis is that economic agents in fear of losing their jobs, or planning to quit the labour force search for information online on the eligibility of benefits, and other services. Thus, the link between internet search queries and job separation is based on flows into unemployment, and flows out of the labour force. This is in contrast to much of the literature examining the link between internet search query data and the labour market, asserting that Google's search engine is a portal for job search.

A Google search composite measure in line with Smith (2015), is constructed from the relevant search terms associated with the aforementioned flows in the labour market. To model the link between weekly Google search query data and monthly US job separations, a mixed-frequency vector autoregression (henceforth mixed-frequency VAR) developed by Ghysels (2015) is employed. The hypothesis is subsequently tested using a mixed-frequency Granger causality test set out in Ghysels et al. (2015b) on the mixed-frequency variables. To compare the mixed-frequency test to a standard Granger causality test, a low-frequency test is conducted on temporally-aggregated Google search query data.

The outline of this thesis is as follows. The following chapter reviews relevant literature in relation to internet search query data and mixed-frequency time series. Chapter 3 presents the construction of our Google search composite measure, a set up of the employed mixed-frequency VAR model and the subsequent mixed-frequency Granger causality test. Chapter 4 presents the results and discussion of our hypothesis tests. Finally, Chapter 5 concludes.

Chapter 2

Literature Review

2.1 Introduction

A number of decades ago, data on economic activity were relatively scarce. However, growth in modern information technology, and in particular, the Internet, has accelerated the lattice of new data sources available for economic research. New sources of data are available more promptly, voluminous, and on novel variables which were previously unquantifiable. Such sources are commonly referred to as "Big data" (see for example Varian (2014)). The advent of big data guides various avenues for economic research. Equally, big data presents a number of challenges to economists, since the arrangement of such sources are unconventional. In the following chapter, we explore these avenues and challenges in detail.

2.2 Big Data and Economic Analysis

The Internet has made the advent of new data sources largely possible. The behavioural patterns of users are extractable, as every consumption purchase and every search query is captured and stored. Social media posts and messages are equally recorded into databases. Thus, digital footprints are left behind on every instance of mouse click online (see for example Einav and Levin (2013), and Bholat (2015)). In contrast to conventional data sources in many disciplines, most sources are not in a structured form. Hence, utilising data analytic techniques may prove useful, as such data are in need of structure prior to its empirical application (see

for example Varian (2014)).

At present, a number of sources of big data are of interest in economic research. First, social media such as Facebook and Twitter, provide a source of data for which to analyse the behaviour of users online in various market structures. Social media message posts hold information on preferences, and social connections. For example, computational linguistic techniques such as textual sentiment analysis can be utilised to extract potentially meaningful signals from message posts, classified as either positive or negative (see Kearney and Liu (2014) for a survey of textual sentiment methods). This has been demonstrated through the examination of investor sentiment in financial markets using Facebook (see for example Siganos et al. (2014)), and labour market flows using Twitter (see Antenucci et al. (2014)). In particular, Antenucci et al. (2014) utilised tweets based job offers, layoffs and employment to examine labour market inflows and outflows, and whether they coincided with actual data. In general, such data sources are accompanied by application programming interfaces (henceforth API), which are a set of protocols used to extract data from the Internet. This consequently simplifies the process of extracting data.

Second, computer software techniques such as Web-scraping, enable researchers, through a few lines of written code, to "scrape" big data off the Internet. This is principally applicable where certain sources of data do not provide access to their API services. "Scraped" data can hence be used to construct economic indicators[1], and assess the strategic behaviour of economic agents online.[2] For instance, Cavallo (2012) co-developed the Billion Prices Project (BPP), run from the Massachussets Institute of Technology (MIT), aiming to provide an alternative measure of retail price inflation.[3] Specifically, online retail prices are scraped off hundreds of retail store websites and used to construct real-time prices indices to document price patterns across various industries in over 50 countries. In the United States, the BPP index closely tracks the official Consumer Price Index (CPI). Such techniques are

[1] Cavallo (2012)
[2] Bajari and Hortacsu (2003)
[3] http://bpp.mit.edu/

readily available at the disposal of economists in order to extract novel sources of data such as consumer preferences.[4]

Third, Internet search queries are a source of big data, providing a quantifiable measure of the volume of search online. Specifically, Internet search query data derive a measure of search through the keywords entered into search engines such as Google and Yahoo. These sources of data discern the behaviour of economic agents through their search process, potentially revealing their intentions online. According to Ettredge et al. (2005), the applicability of internet search query data hinged on the following key premise - "people reveal useful information about their needs, wants, interests, and concerns through their search behaviour". As such, the utilisation of Internet search query data bridges across multiple disciplines. For example, Chang et al. (2015) examined the association between Internet search, and the incidence of charcoal-burning suicide in Taiwan. They found that keywords associated with charcoal-burning suicide positively related to incidences of charcoal-burning suicide.

The interest into Internet search is that traditional sources of search may, to some extent, become obsolete. For instance, there is evidence to suggest that traditional sources of job-listings are being crowded-out by the Internet (see for example Kroft and Pope (2014)). Moreover, traditional sources of data measuring behaviour lack volume and are relatively costly, such as surveys. In contrast, internet search queries retain key characteristics of interest, including volume, timeliness, and the novelty of variables it can emanate, such as consumer preferences.

[4]Edelman (2012) provides a survey of such applications in the existing literature.

2.3 Internet Search Query Data

Early ventures into internet search query data relied on count data published online (see for example Ettredge et al. (2005)). That is, data on search query volumes were gathered by delving through annual reports and archives to assemble such data. Indeed, Ettredge et al. (2005) utilised WordTracker's 500 Top keywords reports to investigate the applicability of internet search queries to macroeconomic statistics. Fast-forward to 2008, Google introduced *Insights for Search*^{TM}, a public web facility which provided a weekly time series index on the insights into what Internet users search for on Google's search engine.[5] The facility was initially developed "with the advertiser in mind", enabling businesses to examine the extent to which their products were being searched for online.[6]

Subsequently, Google Insights for Search was usurped into Google Trends^{TM} as a single interface, to cater for various other endeavours, including research. Google Trends provides a normalised index of the volume of search for a particular search term, relative to the total volume within a particular week. [7]

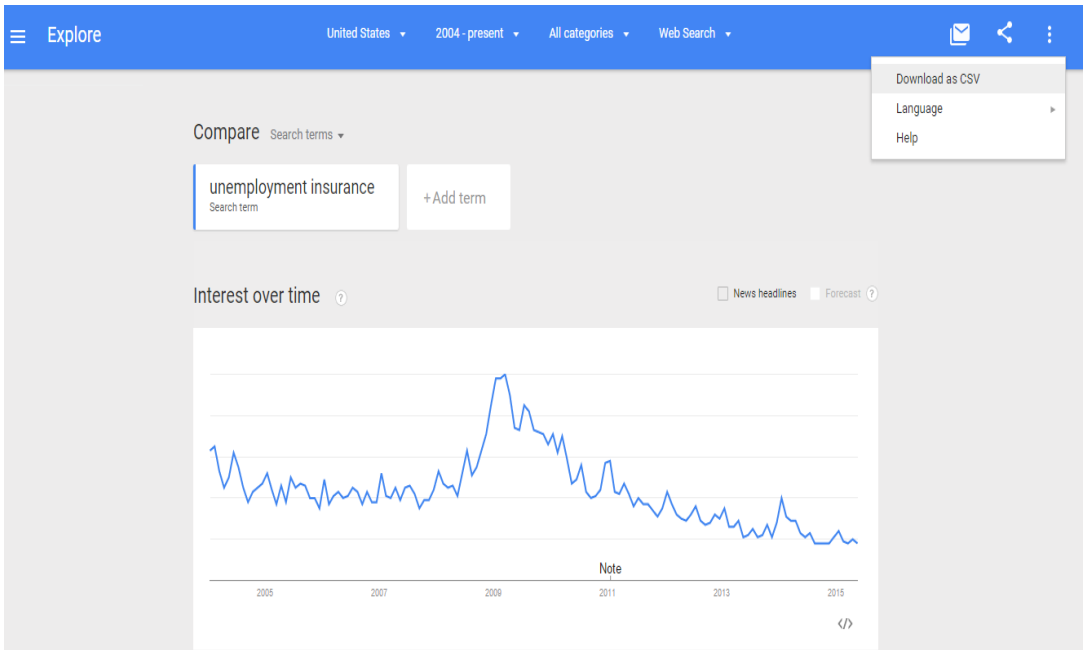

Figure 2.1: Google Trends interface for search query term

[5]Prior to it being shutdown, Yahoo launched a similar tool to Google's called Yahoo Clues.
[6]http://adwords.blogspot.com.au/2008/08/announcing-google-insights-for-search.html
[7]http://insidesearch.blogspot.co.il/2012/09/insights-into-what-world-is-searching.html

Figure 2.1 depicts the search index for the search term 'unemployment insurance'. Note that the index exhibits a peak around the time of the recession. Subsequently, the search query data is conveniently downloadable as a CSV file, filterable through geographical regions and comparable to other search terms.

A simple example motivates the advantages and applicability of internet search queries, for which we discuss in the following section. Choi and Varian (2009a) considered the application of Internet search query data to assess its ability to predict changes in labour market variables. In particular, Google search query data was utilised as a predictor of initial claim for unemployment insurance. The impetus to this investigation is in line with Ettredge et al. (2005). To model the relationship between the aforementioned variables, a baseline autoregressive model (AR) was augmented with Google search queries. It was found that the out-of-sample performance of the augmented model retained significant improvements in mean absolute error (MAE).

2.4 Advantages of Internet Search Query Data

Choi and Varian's (2009a) application, disclose several potential advantages in utilising internet search query data, these include: greater volume, timeliness and the observation of actual behaviour. Below, we consider all three in detail.

2.4.1 Volume

According to Bholat (2015), one of the key characteristics of big data is volume. Volume refers to the scale of the source of data. For Internet search query data (and many others), this is fundamentally attributable to its high sampling frequency. Hence, for a given sample period, such data possess a greater number of observations. For example, as it is noted above, Google Trends publish Google search query data on higher frequencies.[8] This is in contrast to other sources of data which lack granularity, but are aggregate measures. In this regard, it is well known in the

[8]Google Trends publishes mostly weekly data, however various instances result in monthly, and even daily series.

existing literature that voluminous data can be constructive (see for example Hansen and Lunde (2011) and Einav and Levin (2013), among others).

2.4.2 Real-time data source

Complementing the voluminous nature of the data, an additional advantage is that it is timely. That is, Internet search queries are available in real-time. From a policy perspective, the utilisation of real-time data is of primary relevance, as policy-makers need to assess the current state of the economy systematically (see Foroni and Marcellino (2013)). In comparison to conventional sources of data, Internet search query data are frequently updated and published. In particular, Google search query data is updated weekly on the first day of every week (Sunday), and additionally updated every 24 hours.[9] This source of timeliness can prove to be important in prediction, as discussed in Varian (2014).

2.4.3 The observation of actual behaviour

An important distinction between conventional data sources such as surveys, and Internet search query data, is that the former inquire into, whilst the latter, reveals behavioural patterns (see for example Taylor et al. (2014) and Da et al. (2015)). Survey data presuppose that all respondents, respond truthfully to various questions. It is well known that such presuppositions are not necessarily the case, notably with sensitive questions, for example medical/personal (see Singer (2002)). On the other hand, Internet search queries are an objective measure of behaviour, in which such presuppositions are not the case. Thus, an advantage of Internet search query data is that it preserves observations of actual behaviour.

A simple example reiterates this advantage of Internet search query data. Economic theory suggests that consumers search the market, until the marginal cost of additional search outweigh the expected gain from that consumption (see Stigler (1961) and Kogut (1990), among others). From such a perspective, Internet search query data, may potentially disclose the consumption search process of economic agents, in

[9] See Seabold and Coppola (2015) page 6.

their pursuit to compare price of goods and services. Therefore, the aforementioned advantages provide an impetus for the examination of Internet search query data. Subsequently, such advantages cultivate a natural inquisition into the applicability of Internet search queries - that is, what types of questions have been investigated, and what types of questions are worth investigating?

2.5 The Applicability of Internet Search Query Data

The applicability of Internet search query data comprises of a multidisciplinary pursuit in the existing literature. Interesting questions arise which galvanise the endeavour to understand what users are searching for, and whether their discernible search behaviour is meaningful. Examples of such applications include the investigation of early influenza warning systems[10], public attentiveness[11], the measurement of investor and market sentiment[12], the fore-and nowcasting of sales, and other variables.[13] For an example of this multidisciplinary pursuit, we present the following. Ripberger (2011) utilise Google search query as an indicator for public attentiveness to political issues. Specifically, on-going social concerns to the public prompt the search for information by users. Hence, the behaviour of Internet users in a time of social and political unrest is extractable through their search patterns.

Accordingly, various research questions emerge in economics for which Internet search queries may venture into, preserving the aforementioned advantages described above.[14] Below, we provide a number of these research ventures.

First, theories of noise trading propose that 'naive' investors, also known as uninformed traders, are prone to beliefs about future cash flows and risks which are not rationalised by the facts at hand (see Baker and Wurgler (2007)). Internet search query data parallels a measurable proxy for such sentiment, as naive investors are

[10] Ginsberg et al. (2009), Doornik (2009) and Dugas et al. (2012).
[11] Ripberger (2011)
[12] Joseph et al. (2011), Alexander Dietzel et al. (2014), and Da et al. (2015).
[13] Carrière-Swallow and Labbé (2013) and Hand and Judge (2012), among others.
[14] We cover the main research questions which Internet search query data intends to tackle.

likely to gather cost-free information through the Internet. On the other hand, Institutional investors are likely to have access to in-house proprietary sources of information (see for example Vlastakis and Markellos (2012)). Hence, an interesting endeavour within behavioural finance is to examine whether an Internet search proxy for investor sentiment, corresponds to the outcomes various theories of noise trading. This has been investigated by a number of different studies (see for example Joseph et al. (2011) and Da et al. (2015), among others).

Second, as it is noted above, the consumer search process is a key idea in consumer theory. If "search" is subsequently defined as the *"canvassing of various sellers or buyers"*, Stigler (1961, page 213), an ensuing question is whether Internet search query data emulate consumer preferences - Specifically, whether changes in consumption are directly measurable through Internet search behaviour? This question has been undertaken in a number of studies (see for example De los Santos et al. (2012) and, Vosen and Schmidt (2012)). Vosen and Schmidt (2012) for instance, investigate the predictability of private consumption using Google search queries. In comparison to survey-based indicators, it was found that the mean-squared error (MSE) for models augmented with Internet search query data were generally lower, although not significantly different.

A more compelling investigation emerges from the labour market. The Internet has been largely responsible for major structural changes in the labour market (see Autor (2001) and Stevenson (2008)). In particular, Stevenson (2008) notes that the advent of search portals has eased the process of job search. Theoretically, search models of the labour market have exploited various outcomes which emerge from the relationship between job search effort and unemployment insurance, and the incidence of unemployment.[15] Thus, a resulting question may emerge from this exposition - Whether Internet search queries, such as Google search, resonate the job search process?

[15]The literature on job search is relatively extensive. See Lippman and McCall (1976) for an earlier survey into search model of the labour economics, and Rogerson and Shimer (2011) for a recent exposition.

Many studies have attempted to demonstrate that Internet search queries resonate the job search process online (see for example D'Amuri and Marcucci (2010), Fondeur and Karamé (2013), and Baker and Fradkin (2014)). In particular, Baker and Fradkin (2014) examine the validation of Internet search queries from Google as a job search indicator. The search term *"jobs"* is utilised as a proxy, as a means to generalise the term(s) which would be used. It is argued that indirect relationships between job search portals and Google search are uncovered.

Alternatively, it is arguable that Internet search queries in relation to labour market characteristics reflect the inflows to unemployment, and/or out of the labour force. In contrast to the aforementioned exposition above, the rationale to this argument pertains to the fact that there are job search-specific Internet portals such as Monster Jobs.[16] That is, general Internet search query portals such as Google search and Yahoo, may not necessarily reflect signals of job search activity. Studies such as Choi and Varian (2009a) and McLaren and Shanbhogue (2011), seem to suggest a similar argument. For instance, Choi and Varian (2009a) implicitly ask the following question: "What would you search for, if you thought you might lose your job?" Such a question may disclose a direct relationship between Internet search query data and labour market movement, as increases in search for unemployment benefits may resonate lay-offs or quits. Thus, an interesting inquiry is whether changes in the composition of the labour force emulate Internet search queries in relation to labour market outcomes.

2.6 Key Challenges Posed by Internet Search Query Data

The various questions which Internet search query data may tackle as those which we discuss above, reiterate the advocacy of utilising sources of big data for research, as it can provide potentially meaningful signals of behaviour. However, there are two key on-going challenges posed by big data, and in particular Internet search query data. These include data snooping/term selection and mixed-frequencies. Below,

[16] http://www.monster.com/

we provide a description of each challenge, and ways in which the existing literature has handled each of them.

2.6.1 Data snooping/Term selection

Term selection refers to the process of selecting terms which best reflect intuition on a phenomenon of interest. That is, keywords are selected which to some degree, resonate with what Internet users would enter into a search engine on a particular subject. For example, an individual who wants to find out when and where the screening of the "Everest" is taking place, he/she may enter the terms *"Everest movie", "Everest movie time", "What time is the Everest movie screening"*.[17] These particular search terms might indicate that the individual may (or may not) attend that screening. Hence, such search terms may provide an indication of a users behaviour.

In this regard, several issues arise. Firstly, term selection is not a trivial task, as there are millions of search terms which are employable by users. Many terms may, either directly or indirect associate with the phenomenon of interest. Thus, it is not clear which terms are most appropriate for a particular subject. Some level of theory may provide a basis for the pursuit of selecting particular search terms. Secondly, several terms may yield excessively noisy signals. Given the immoderate use of Internet search, and the possibility of having several meanings for one term, extracting a meaningful signal may prove difficult (see for example Smith (2015) and Seabold and Coppola (2015)). For example, the use of the term *"Jobs"* could refer to either a search for employment, or information on Steve Jobs. Finally, term selection induces a level of data-snooping bias. As asserted by Lazer et al. (2014), this is a caveat of Internet search query data, for which researchers must keep in mind.

A question consequently arises: "How do you select the appropriate search terms whilst avoiding the aforementioned issues as much as possible?" There are various

[17]Note that the capitalisation of letters and the use of stop words are irrelevant, when entering a term into a search engine.

methods readily utilised throughout the existing literature. These methods generally include, economic and financial dictionaries, and texts[18], keyword tools such as Google Correlate[19], theory[20], and Google Trends' automated category system[21].

Each method comprises of its own particular technique. Google Trends' automated category algorithm is a system in which a base term is fed into the web facility, and the term can be filtered to a particular subject matter. This method is notably useful when there are multiple meanings for a particular search term. Economic and financial dictionaries, and texts for example, have been commonly used in the analysis of financial markets (see for example Da et al. (2011)). Similar to Trends' category system, a base term is chosen from which it is cross-referenced with other textual sources based on their count frequency - that is, terms which occur most often. As a simple example, Perlin et al. (2014) used finance textbooks and the online dictionary 'Investopedia', to select a pool of terms in relation to stock market returns. Unlike dictionaries/textbooks, word tools such as Google Correlate perform correlations between different terms to identify additional search words on the basis of a time series.[22]

2.6.2 Mixed-frequencies

A key consideration in utilising Internet search query data is mixed-frequencies. Mixed-frequencies refer to the mismatch in the sampling frequency between variables. Ordinary time series regressions are often estimated with same-frequency variables. However, it is well known in the existing literature that most economic relationships emerge between variables sampled at different frequencies (see Ghysels et al. (2015b) and Silvestrini and Veredas (2008)). For instance, GDP is sampled quarterly, whilst associated macroeconomic indicators such as the unemployment rate, are monthly. The datasets are subsequently unbalanced, due to the misalignment of the sampling frequencies. How does one consequently take into account

[18]Perlin et al. (2014) and Da et al. (2015).
[19]Baker and Fradkin (2014).
[20]Wu and Brynjolfsson (2014) and D'Amuri and Marcucci (2010).
[21]Vosen and Schmidt (2012).
[22]See Mohebbi et al. (2011) for a discussion of the methodology behind the algorithm in Google Correlate.

the mismatch in sampling frequency between variables? Thus, in embracing new sources of data such as Internet search queries, such a challenge is deliberated. In utilising Internet search query data, this challenge is in need of consideration.[23]

There are a number of methods available to take this challenge into account. One of the most common solutions is temporal aggregation. Temporal aggregation, as the name suggests, is a process in which variables are aggregated and averaged across time. Specifically, temporal aggregation involves the aggregation of high-frequency variables to a common low-frequency.[24] This ensures that the mixed-frequency variables are aligned at the same frequency. Known as a 'pre-filtering method', it is often considered to be the simplest approach to deal with mixed-frequency sampling.

Notation

In order to avoid confusion further into this thesis, we set out our necessary notation.[25] Let τ denote the basic time unit where $\tau = 1, ..., T$. τ_L denotes the time unit of a lower frequency variable. The number of times a higher-frequency observations appears between two low-frequency periods is denoted m, where $j = 1, ..., m.$, which is called the ratio of sampling frequencies. Hence, the time unit attached to the higher-frequency variable is denoted τ_H. Subsequently, we will denote x_H and $x(\tau_L)$ as the high-frequency variable, and the low-frequency variable respectively. $x_H(\tau_L, j)$ will denote *each* high-frequency observation for $j = 1, ..., m$. For a quarterly-monthly relationship, we would consequently denote $x_H(\tau_L, 1)$ as the first monthly observation of x_H in quarter τ_L, $x_H(\tau_L, 2)$ to represent the second, and $x_H(\tau_L, 3)$ to represent the last.

Temporal Aggregation

For temporal aggregation, two aggregation schemes exist in which their applications depend on the nature of the variable. That is, whether it is a stock or flow variable. Stock variables are the result of systematic sampling, in which high-frequency vari-

[23] A notable exception is Choi and Varian (2009a).

[24] Interpolation is another, yet uncommon technique which is not covered in this thesis. See Wohlrabe (2009) for an in-depth survey of this approach.

[25] We adopt notation throughout the thesis from Ghysels et al. (2015a) and Ghysels et al. (2015b).

ables are sampled at every low-frequency observation. For instance, it is possible to take the latest available value of a variable to form the aggregated variable. On the other hand, flow variables are more dedicate in their aggregation procedure, as it is necessary to consider the whole time period in question. With regards to weekly data, flow variables require the full month of observations as oppose to the 'stock-end' value of the month (see Wohlrabe (2009)).[26]

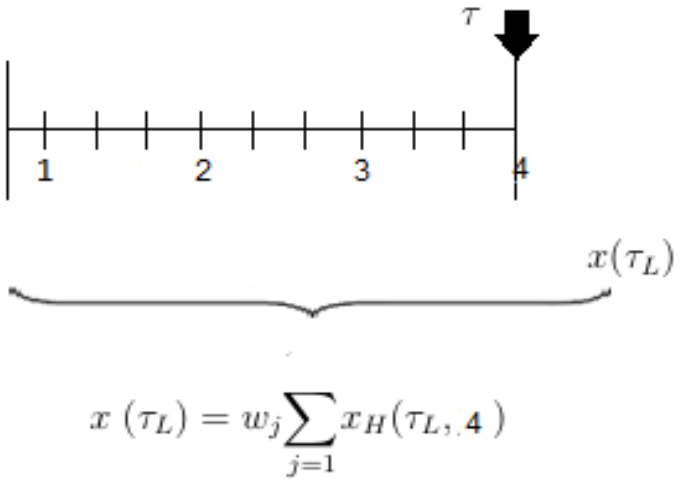

Figure 2.2: Temporal aggregation mechanism for $m = 4$.
(Adapted version of Figure 1 in Silvestrini and Veredas (2008), pg. 462)

Figure 2.2 depicts a visual representation of the temporal aggregation mechanism for a weekly variable and a corresponding monthly variable, and the formula for which to calculate the aggregate value. As illustrated, each high-frequency observation $x_H(\tau_L, j)$ is observed $m = 4$ times within each monthly observation $x(\tau_L)$. Temporally-aggregating a flow variable entails the summation of the m-weekly observations over its monthly counterpart. On the other hand, temporally-aggregating a stock variable entails the sampling at each m observation. Thus, in the example above, each temporally-aggregated stock observation would appear on every 4th period.

With regards to Internet search query data, temporal aggregation has been implemented to unburden the problem of mixed-frequencies. This has been demonstrated

[26]In addition to simple time averaging, different weighting systems are possible. For example step-weighting, is considered when the high-frequency values are not equidistant.

by D'Amuri and Marcucci (2010), Carrière-Swallow and Labbé (2013) and Vicente et al. (2015). Most notably, Carrière-Swallow and Labbé (2013) aggregated Google search query data related to automobiles, to fore-and nowcast automotive sales in Chile. A linear regression is fit from which an aggregated monthly measure of Google search queries is constructed. This is in line with various others, including Seabold and Coppola (2015). Therefore, temporal aggregation is a particular solution to the challenge of mixed-frequencies with Internet search query data.

Mixed-frequency modelling

The simplicity of temporal aggregation has been readily identified as a reason for its persistent application (see Wohlrabe (2009)). However, such a solution is scarcely considered satisfactory. Pre-filtering may have serious consequences for the sampling information variables possess. That is, the structure of the time series could change, potentially altering any subsequent inferences made about the series in question (see for example Rajaguru (2004)). The econometric literature on temporal aggregation has examined the time series properties of temporally aggregated variables, and its conclusions propose a level of caution.

For example, various econometric methodologies are effected by temporal aggregation. For instance, the dynamic relationship between variables is often negatively impacted on due to a loss of information (see for example Sims (1971) and William and Wei (1990), among others). Specifically, by temporally aggregating over a given number of high-frequency observations, a large number of low-frequency observations are created. With reference to Figure 2.2, the four observations within the span of the low-frequency observation are lost. Moreover, the loss of potentially useful information greater with a higher ratio of sampling frequency. This was demonstrated by Wei and Mehta (1980) through the examination of parameter estimates and subsequent Monte Carlo simulations. The subsequent conclusions were that efficiency of aggregated parameter estimates decreased with an increase in the ratio.

In addition, time series properties attached to particular variables may not nec-

essarily be invariant to temporal aggregation (see for example Marcellino (1999)). That is, if time series properties such as causality, hold for a disaggregated variable, the property may not potentially hold for its aggregated counterpart.[27] According to the existing literature, causal inferences formed on temporally-aggregated variables could potentially be spurious. This was demonstrated by Wei (1982), in which it was shown to convert one-sided causal relationships to pseudo two-sided feedback systems. Analogous results were illustrated empirically by Marcellino (1999), in which causality was tested between 10-year government bond yields and 90-day deposit rates, and more recently illustrated by Rajaguru and Abeysinghe (2012).[28]

In order to avoid the issues of temporal aggregation, a recent strand of the econometric literature has developed models which avoid pre-filtering methods. It is argued that a more desired outcome, is to model the mixed-frequency variables directly, in order to avoid any loss of information (see for example Wohlrabe (2009)). Bridge equations, were an early attempt to resolve the misalignment of sampling frequencies (see for example Baffigi et al. (2004)). Bridge equations involved the estimation of high-frequency variables, from which the estimates were aggregated and fed through a Bridge equation. The Bridge equation was subsequently used to estimate the low-frequency variable. Thereafter, mixed-frequency models were developed, including mixed-frequency vector autoregressions (MF-VAR) (see Mariano and Murasawa (2010) Ghysels et al. (2015b)), as oppose to same-frequency VARs, and Mixed Data Sampling models (MIDAS) (see Ghysels et al. (2004) and Ghysels et al. (2007) and Andreou et al. (2010) among others), which rely on distributed lag polynomials.[29]

Therefore, the mixed-frequency modelling literature is ubiquitous, however it is not the most common of solutions. With regards to Internet search query data, this is indeed the case. Only a small number of studies have explicitly modelled the mixed-frequency of Internet search query data directly. In particular, Smith (2015) for instance, utilised a mixed-frequency time series model to nowcast UK unemploy-

[27] Note that by causality, we mean in Granger's sense.
[28] Marcellino (1999) empirically investigated many other time series properties for invariance.
[29] For a survey of the different modelling approaches, see Foroni and Marcellino (2013).

ment data using weekly Google search queries. Moreover, Bangwayo-Skeete and Skeete (2015) employed a mixed-frequency modelling approach to forecast tourist arrivals using Google search query data. Both studies employ a MIDAS model in which the timing of the information is preserved - Distributed lag polynomials are used to ensure parsimonious specifications (see for example Armesto et al. (2010)).

2.7 Contribution to the Literature

The following thesis aims to contribute to the increasing literature on Big Data in economic analysis. In particular, Internet search query data, as a key source of understanding user behaviour. In line with studies such as Choi and Varian (2009a), McLaren and Shanbhogue (2011) and Smith (2015), we pursue an investigation into whether Internet search queries contain relevant information on changes in the composition of the labour force. Explicitly, our main hypothesis is that Google search queries related to unemployment insurance benefits, government services and legal information are useful in predicting future US job separation statistics.

Given the mixed-frequencies among economic variables and Internet search query data, adopting a mixed-frequency modelling approach, as oppose to temporal aggregation is desirable. Hence, we pursue an auxiliary hypothesis to complement to use of Internet search query data. This auxiliary hypothesis is to test whether mixed-frequency modelling yields distinct results to standard low-frequency modelling. This hypothesis is pursued in line with Smith (2015) and Bangwayo-Skeete and Skeete (2015).

The rationale behind the main hypothesis is that there are explicit search portals such as Monster.com for job search. Hence, utilising Google search as a job search proxy may not necessarily yield meaningful signals on the inflows and outflows of the labour market. On the other hand, Google search queries may therefore be related to flows into unemployment, and out of the labour force, rather than flows into employment (see Smith (2015)). The rationale behind the complementary hypothesis is as mentioned, the desirability to avoid temporal aggregation. We test both

hypotheses through Granger causality. Specifically, we apply a mixed-frequency Granger causality test recently developed by Ghysels et al. (2015b), and a general low-frequency Granger causality test for comparisons. In the following chapter, we set out our methodology.

Chapter 3

Methodology

3.1 Data

Two data sources were employed for this thesis. Seasonally-adjusted monthly US Total Separations: Total Non-farm data was retrieved from FRED®, Federal Reserve Bank of St. Louis.[1] Total Separations, commonly referred to as turnover, represents the total number of employee quits, lay-offs and discharges, and other separations. Other separations includes retirement, deaths and separations due to disability. For the purposes of this thesis, we subtracted 'Other separations' as it was not directly related to our underlying hypothesis. As such, our separations data only consisted of quits and, lay-offs and discharges. Separations data is available from 1st December 2000. However, for the purposes of this research, we specified a sample period from the 1st January 2005, to 1st May 2015. This was to align the job separations data to the Google Trends data, which is only available from 2004 onwards.

Internet search data was retrieved from the Google Trends™ tool provided by Google Inc.[2] Google Trends provides weekly time series data on the relative volume of searches for a particular search term. Weekly values are accumulated over the number of searches every day for that particular week. The search volume is not in absolute terms, hence in a specific period of time, each series returns an index

[1] https://research.stlouisfed.org/fred2/series/JTSTSL
https://research.stlouisfed.org/fred2/series/JTSOSL

[2] https://www.google.com.au/trends/

with values between 0 and 100. A volume of 0 represents the lack of relative search, whilst a volume of 100 represents the most in relative terms. It was not possible to find out how many people actually searched a specific keyword. Thus, internet penetration rates did not necessarily matter. We now discuss the process of linking Google search terms to our measure of US job separations.

3.1.1 Linking Google search terms with US Job Separations

As it is discussed in detail in Chapter 2, Section 2.6, terms should reflect the behaviour of individuals searching for particular information. Thus, the choice of search terms is an important consideration. In selecting a set of relevant terms, we assumed each reflected the majority of searches conducted within a labour market context. This assumption was critical to our methodology, as it required a selection of terms associated with our main hypothesis, prior to estimation.

Initially, we conceived that an appropriate choice related job separations was job search, as many users search for employment on the internet (see Baker and Fradkin (2014)). However, further inspection disclosed four concerns. Firstly, the possibility that most volumes of search for vacancies were initiated by employed persons, which could potentially mask the search activity of actual joblessness. Secondly, job search is generally considered pro-cyclical (see for example Burda and Wyplosz (1994) and Shimer (2005) among others). Hence, "on-the-job" search is likely to increase with an increase in the business cycle, potentially counteracting other labour market searches - a similar concern to the first. Thirdly, job search cites such as Monster.com, and Indeed.com are specific portals for job search, which raised doubt on the validity of using Google search as a direct job-search tool. Lastly, a large proportion of separations may be quits, yet there is an increasing trend number of persons moving out of the labour force.[3] It was subsequently conceived that job-search terms may not be suitable. We consequently resolved such concerns by selecting search terms ubiquitous to *flows into* unemployment, and flows out of the labour force (see for example Smith (2015)).

[3]Data on the number of people not in labour force is retrievable from the Bureau of Labor Statistics; code no. LNS15000000.

Analogous to Choi and Varian (2009a), the term selection process began by considering what an individual would search for, if he/she (a) thought they may lose their job, or (b) intended to quit the labour force. A "root" term was chosen which best reflected the flow into unemployment and out of the labour force. According to Koop and Onorante (2013), a "root" search term refers to a keyword directly related to a target variable. For example, the root term in an investigation into the relationship between internet search and stock market volatility, may be based on a particular stock ticker. Consequently, the root term chosen was 'unemployment insurance'.[4] To ensure the relationship between internet search and US job separations remained stable, direct and indirect terms related to the root term were downloaded. See the Appendix A.1 for the full list of search terms. Below, we discuss the challenges faced in using Google Trends data.

3.1.2 Challenges in Using Google Trends Search data

Utilising Google Trends data presented a number of practical challenges:

1. Google Trends data updates regularly on any particular day (see for example Carrière-Swallow and Labbé (2013) and Seabold and Coppola (2015)). That is, for a given week, Google Trends' weekly values may be different on two separate days.

2. According to Google, certain search terms which do not accumulate enough volume during the week, are published in a lower frequency (either monthly, biweekly or daily). As such, it is not known a priori, what the frequency of a search term's series will be. Note that Google does not specify what the minimum volume for a search term should be.

3. In relation to 2, search terms accumulating insufficient volume may in addition, return zero values[5]. Indeed, this was the case for our data, in which certain

[4]For example, McLaren and Shanbhogue (2011) chose the term "JSA", which stands for job seekers allowance, a term reflecting the UK system of unemployment benefits.

[5]https://support.google.com/trends/answer/4355213?hl= $en\&ref\_topic = 4365599\&vid = 1\text{-}635774781534888206\text{-}420084290$

terms returned a string of zeros for a number of weeks. A caveat to this is that it is not known at which point the data become censored.

4. A property of the calendar itself is that there are an inconsistent number of weeks per calendar month. That is, most months in a calendar year do not contain exactly 4 weeks. To directly map weekly data onto our monthly job separations data, a consistent number of high-frequency observations is required per-low frequency observation.

For the first challenge, we initially favoured retrieving search volume for the same terms across a number of days, and averaging the data (see for example Seabold and Coppola (2015)). However, data transformations in addition to the construction of our search composite index, remained an undesirable outcome. The second and third challenge led to a reluctance in adopting each search term as its own variable, and therefore its own potential regression. This was undesirable as the processing of each variable would have been difficult to handle for a larger number of search terms. A simple and efficient method was to construct a composite index of the search terms. This method is in line with Carrière-Swallow and Labbé (2013) and Smith (2015), who overcame challenges two and three in this manner. The fourth challenge required much more effort to overcome. Specifically, the data required a transformation into a balanced weekly series with a consistent number of weeks per month. We now discuss our solution and reasoning in detail.

Weekly Google Trends data

To map the weekly Google Trends series onto our monthly US job separation series, we required a balanced dataset of weekly observations. The following strategy was adopted from Smith (2015).[6] Specifically, we defined Week 1 as the days spanning day 1 to 7 for a particular month, Week 2 as days 8 to day 14, Week 3 as days 15 to day 21, and Week 4 which generally ran from day 22 to the end of the month. Week 4 varied depending on the month and year the observation was cited. In other words, for months which had 29 (for leap years), 30 or 31 days, the number of days in week 4 were generally longer.

[6]According to Smith (2015), this method is an adaptation from Hamilton and Wu (2014).

We converted our weekly series into a daily series, recovering the missing observations through simple interpolation. By averaging across the daily data, we attained a transformed weekly series, structured by our defined weeks above (see Appendix A.2 for code).[7]

The purpose of this strategy was two-fold. Firstly, by creating balanced weeks, we recovered a consistent number of high-frequency observations per low-frequency observation - In this particular case, a consistent number of weeks per month. Secondly, Google publishes its Google Trends data every Sunday, defining a week spanning from Sunday to Saturday. Every weekly observation subsequently occurred a Sunday. However, the first day of every calendar month did not, which generated an overlap of Google weeks across multiple months. To the best of our knowledge, this strategy has only been documented by Smith (2015), for the purposes of employing internet search query data. Alternatively, studies in the literature either take the first weekly value for each month as its representative value, or do not necessarily comment/document on this account (see for example Choi and Varian (2009b) and Hand and Judge (2012) among others). Given the first concern we alluded to earlier, we perceived this process as necessary. In the following subsection, we set out the process of constructing a Google search composite measure.

3.1.3 Constructing a Google Search Composite measure

With a desire to retain parsimony whilst accounting for excessively low volumes, a Google search composite measure was constructed. Formerly, studies such as Carrière-Swallow and Labbé (2013) employed a linear index approach, where a linear regression was fit to construct the composite index. For the purposes of exposition we present Carrié-Swallow and Labbé's method. Let X be the matrix of Google search terms and y_t, the year-on-year percentage change in the variable of interest. In each sample period, a weight $\hat{\beta}$ is estimated using observations up to time $t-1$,

[7]This was accomplished using the statistical package R, R Core Team (2015) and various associated packages. Full details in Appendix A.2.

from which a linear model

$$y_t = \alpha + \beta X_t + \epsilon_t$$

was fit. The index I_t for period t is then computed as the fitted values of the estimated linear model

$$I_t = \hat{E}_t[\beta | y_{t-1}, X_{t-1}] \cdot X_t$$

Despite the simplicity of Carriére-Swallow et al.'s method, the index relies on temporal aggregation. In other words, applying this approach is conditional on having both dependent and independent variables in the same frequency. In this thesis, we required an approach which did not hinge on temporal aggregation.

Consequently, we employed the composite index approach of Smith (2015). Smith's (2015) approach relies on the construction of a dynamic weighting system which adapts weekly, with the addition of a weekly observation. Therefore, this method comprised of summing the readings for q Google search terms (GST) for a particular week t, and constructing weights for each Google search term, based on their individual contribution for that week (see below)

$$W_{i,t} = \frac{GST_{i,t}}{\sum_{i=0}^{q} GST_{i,t}} \tag{3.1}$$

where $i = 1, ..., q$ represents the number of search terms. Equation 3.1 would therefore ensure that searches which achieved relatively greater increases in volume attracted greater weight. The subsequent Google Search Composite Index was ultimately defined as the sum of the weighted search term volumes for each week (see Appendix A.2 for code).

Before presenting preliminary sample statistics, we supplement our existing notation from Chapter 2, Section 2.6.2 with additional notation to define the weekly Google Search Composite Index (GSCI) we constructed above, and the subsequent monthly aggregated variable. Let $GSCI_j(\tau_L)$ denote the $j-th$ week of month τ_L. Hence, $GSCI_1(\tau_L)$ denotes the first week of month τ_L, $GSCI_2(\tau_L)$ is the second week, and so on. The aggregated Google search composite index is defined $GSCI^M$, where M

denotes the monthly frequency of the variable. This notation will assist in understanding what is set out in the next subsection, as well as the section which follows in setting out our model.

3.1.4 Preliminary Sample Statistics

Figure 1 plots the 100× log difference of our measure for US job separations (primary y-axis) and Google Search Composite Index (secondary y-axis) respectively.

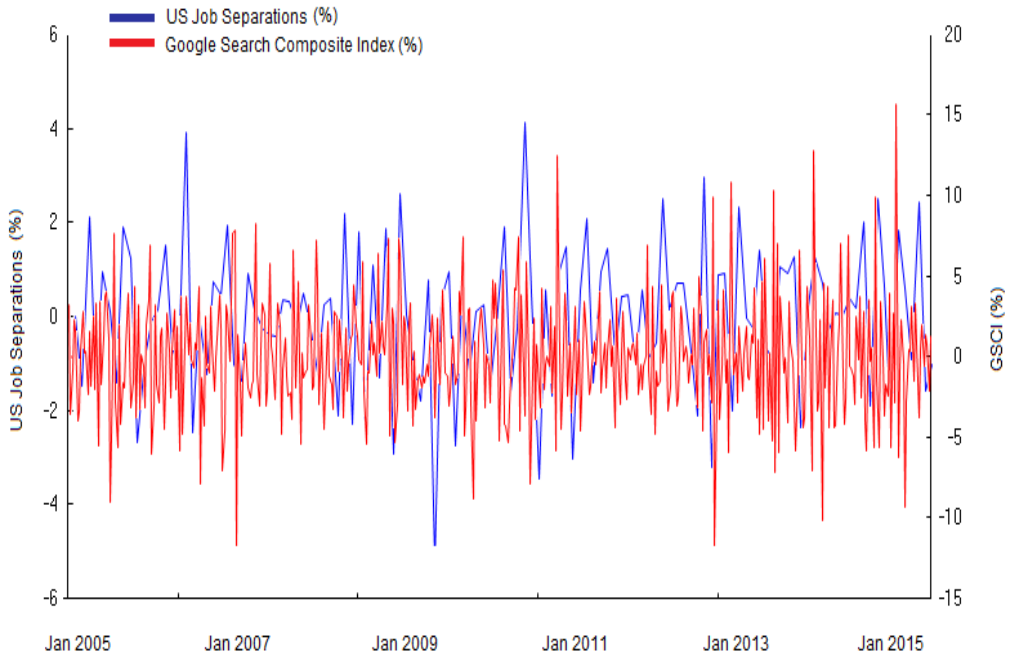

Figure 3.1: Percentage growth rates in US job separations and GSCI

Illustrated in Figure 3.1, the misalignment of sampling frequencies between the percentage growth rate in US job separations and the percentage growth rate in Google search composite measure is exemplified. As is typically the case, the weekly search composite index observations fluctuate between the monthly job separation observations. The growth rate in $GSCI$ is highly volatile, as the percentage change from previous weeks seem to fluctuate heavily. As such, it is not clear from Figure 3.1 whether each series is related to the other. We reserve any testing for Section 3.3.

For each series, we took the log-difference to account for any non-stationarity issues and seasonality. We subsequently multiplied each log-differenced series by 100 for ease of interpretation. Both series were therefore interpretable as approximate

percentage growth rates. We subsequently denoted $GSCI$ as the percentage growth rate of our Google search composite measure, and SEP was denoted as the percentage growth rate in US job separation. These definitions are kept throughout the rest of the thesis.

Table 3.1 provides a list of the univariate sample statistics and p-values for the Kolmogorov-Smirnov test of normality for our weekly Google search composite index $\{GSCI_1, GSCI_2, GSCI_3, GSCI_4\}$, monthly aggregated Google search measure $GSCI^M$, and US job separations SEP in 2005M1 - 2015M5 (500 weeks and 125 months). Specifically, the table reports the minimum, median, maximum, mean, standard deviation, skewness, kurtosis and subsequent p-values for the aforementioned test.

Table 3.1: Univariate Statistics

	SEP	$GSCI_1$	$GSCI_2$	$GSCI_3$	$GSCI_4$	$GSCI^M$
Minimum	-5.27	-10.13	-12.77	-11.78	-7.92	-4.18
Median	0.13	-0.49	-0.44	0.06	0.37	0.04
Maximum	4.11	10.34	9.90	7.64	15.71	3.40
Mean	-0.03	-0.49	-0.42	-0.12	0.92	-0.03
Std. Dev.	1.58	3.62	3.26	3.30	3.99	1.27
Skewness	-0.21	0.15	-0.14	-0.40	0.95	-0.08
Kurtosis	3.36	3.48	4.14	3.49	4.43	3.25
KS-pvalue	0.04**	0.00***	0.00***	0.00***	0.00***	0.19

For the p-values of the KS, we put *** if the null hypothesis of normality is rejected at the 1% level (strong rejection), ** if rejected at 5% but not at 1%, and * if rejected at 10% but not 5%.

All four weekly $GSCI$ variables ($GSCI_{1,2,3,4}$) have heterogeneous features. Firstly, the mean of the first three $GSCI$ are negative whilst of the three, only the third has a positive median, 0.06%. That is, the third variable $GSCI_3$, is skewed to the left. Indeed, this is equally the case for the aggregated variable, $GSCI^M$. Secondly, amongst the weekly variables, the skewness of $GSCI_2$ is the lowest. Thus, it indicates the weakest asymmetry amongst the weekly variables. The minimum and maximum of the weekly variables range from as low as -12.77% for the $GSCI_2$ to as high as 15.71% for $GSCI_4$. The range of values for $GSCI_4$ is reiterated through its standard deviation. For our dependent variable $SEP(\tau_L)$, the sample statistics are as follows. For example, its minimum and maximum values are -5.27 and 4.11 re-

spectively. Moreover, the mean and the median of the variable have opposing signs. As it is the case with the first three weekly variables $\{GSCI_1, GSCI_2, GSCI_3\}$ and $GSCI^M$. This consequently implies left-skewness.

For the Kolmogorov-Smirnov test of normality, the null hypothesis for normality was strongly rejected for all the weekly variables $\{GSCI_1, GSCI_2, GSCI_3, GSCI_4\}$. On the other hand, the null hypothesis for normality was not rejected at all significance levels for the aggregated variable $GSCI^M$. Moreover, for our measure of US job separation, the null hypothesis was rejected at 5%, but not at 1%. It should be noted, Ghysels et al. (2015b) demonstrate that the asymptotic theory of mixed-frequency VARs do not require the normality assumption (Section 2 in Ghysels et al. (2015b)).

In addition, correlation coefficients between each pair of variables were computed. Both contemporaneous and lagged correlation coefficients up to 1 lags are reported in the Table B.1-B.2 in Appendix B. First, the correlation coefficient between $GSCI^M(\tau_L)$ and its high-frequency counterparts $\{GSCI_1, GSCI_2, GSCI_3, GSCI_4\}$, range between 0.215 and 0.485, with the the fourth weekly variable $GSCI_4$, the highest. This is expected as the monthly aggregated variable is a simple average of the sum of the original disaggregated weekly variables. Similar correlation coefficients range for a number of the weekly variables, including $GSCI_3$ and $GSCI_4$. Second, correlation coefficients between $SEP(\tau_L)$ and the $GSCI^M(\tau_L - k)$ seem to suggest an increase around 2 months. There does not seem to be any persistence among the variables across time. This consequently suggests that log-differencing took care of any autocorrelation.

3.2 Mixed-frequency VAR

Since the work of Sims (1980), VAR models have become a common tool for analysing the co-movements of economic time series. This generally involved regressing a real economic variable (for example, unemployment rates) on some other series (for example, initial claims for unemployment insurance). As such, economic relationships would typically emerge between variables sampled at different frequencies. As it was noted in Chapter 2, researchers readily deal with mixed-frequency time series data through temporal aggregation and interpolation. However, Wohlrabe (2009) notes, if the purpose were to analyse time series dynamics between variables, pre-filtering methods may have adverse effects on parameter estimates, and impulse responses alike (see Chapter 2, Section 2.6.2).

Common mixed-frequency models have relied on latent processes and shocks, where the model is cast into a state-space representation (see for example Zadrozny (1988) and Mariano and Murasawa (2010) among others). The state-space representation of the model match the latent process with corresponding mixed-frequency data, where low-frequency variables are treated as high-frequency variables with missing observations. Zadrozny's (1988) method, estimates a VARMA model with different frequencies, whilst Mariano and Murasawa (2010) set a mixed-frequency VAR model for partially latent time series. Computational devices such as Kalman filters, are subsequently employed to extract the missing observations, and relate the dynamics between low, and high-frequency variables. As demonstrated by Bai et al. (2013), if the model is specified correctly, and the parameters are known, the Kalman filter would yield an optimal outcome in population. That is, under ideal conditions, the Kalman filter would attain relatively low forecasting errors.

Latent processes and state-space representations however, entail computational complexities (see for example Wohlrabe (2009), Ghysels et al. (2015b) and Bai et al. (2013) among others). Firstly, the complexity of estimation increases with the number of variables involved in the model. Namely, measurement equations, latent low-frequency and high-frequency variables result in many parameter estimations. In-

variably, this elicits the estimation of only small-scale models (see Wohlrabe (2009)). Secondly, as noted above, the success of state-space models hinge on correct specification. Correct specifications are increasingly difficult with missing observations in the dependent variable, which appends to the difficulty of estimating many parameters (see for example Foroni and Marcellino (2013)).

To circumvent these issues, we applied the mixed-frequency vector autoregression model of Ghysels et al. (2015b). A multivariate extension of the MIDAS regression model, Ghysels' (2015) model employs a stacking vector system in which the high-frequency observations within each low-frequency observation are stacked on-top of each other. For example, in a monthly-weekly relationship, in which there are approximately four weeks per month, we would observe four weekly observations per month, hence each weekly observation over the data sample is considered a variable in itself (see for example Ghysels et al. (2015b) page 4). Any potential deviation from this ratio requires data transformation, as per subsection 3.1.2 above.

Mixed-frequency VAR models were independently introduced by Mariano and Murasawa (2010), Ghysels et al. (2015b), and McCracken et al. (2013). In comparison to other mixed-frequency VAR models, Ghysels' (2015) model does not rely on latent processes. Rather, the model is observation-driven - the model is formulated exclusively from observable data. This approach directly relates to standard VAR models, for which common tools used are directly exploitable. Consequently, Ghysels' (2015) model is easier to estimate. Moreover, contrary to classical VAR models, high-frequency observations are allowed to have heterogeneous effects on a low-frequency variable (see Sadahiro and Motegi (2014)). Sadahiro and Motegi (2014) note that standard same-frequency VAR models implicitly require homogeneous impacts on low-frequency variables. In other words, for the mixed-frequency VAR, the lagged information of high-frequency observations may take on different values, and possible seasonal effects (see Equation 3.5 below). To compare a standard bivariate VAR to our mixed-frequency VAR, we set out both specifications below.

The mixed-frequency VAR model, requires a reconsideration of notation from Chapter 2, Section 2.6.2 through the following key assumption (see Ghysels et al. (2015b) page 3):

1. *A K-dimensional process is considered where* $K = K_L + m \times K_H$. *The first* $K_L < K$ *elements are collected in the low-frequency process* $x_L(\tau_L)$, *whilst the remaining* $K = K - K_L$ *elements* (K_H) *are collected in the high-frequency process* $x_H(\tau_L, j)$, *where* $j = 1, ..., m$;

$x_L(\tau_L)$ and $x_H(\tau, j)$ subsequently referred to the low-frequency process and high-frequency process respectively.

3.2.1 Bivariate Monthly VAR

To estimate a standard low-frequency VAR, our Google search composite measure $GSCI$ was aggregated into a monthly index $GSCI^M$. $SEP(\tau_L)$ denotes our variable of interest, the percentage growth rate in US job separations. We first fit the following bivariate standard monthly VAR of order 4.

$$\begin{bmatrix} GSCI^M(\tau_L) \\ SEP(\tau_L) \end{bmatrix} = \sum_{k=1}^{4} \begin{bmatrix} a_{11,k} & a_{12,k} \\ a_{21,k} & a_{22,k} \end{bmatrix} \begin{bmatrix} GSCI^M(\tau_L - k) \\ SEP(\tau_L - k) \end{bmatrix} + \begin{bmatrix} \epsilon_1(\tau_L) \\ \epsilon_2(\tau_L) \end{bmatrix} \tag{3.2}$$

where $GSCI^M(\tau_L)$ is calculated as the monthly average of the high-frequency weekly observations. Subsequently, using our notation from above, $GSCI^M(\tau_L)$ is defined as

$$GSCI^M(\tau_L) = \frac{1}{4}\sum_{j=1}^{4} GSCI_j(\tau_L) \tag{3.3}$$

3.2.2 Bivariate Mixed-Frequency VAR

We fit the following bivariate mixed-frequency VAR of order 4, which consisted of our *weekly* Google Search Composite Index, $\{GSCI_1, GSCI_2, GSCI_3, GSCI_4\}$ and

our dependent variable from above, $SEP(\tau_L)$

$$\underbrace{\begin{bmatrix} GSCI_1(\tau_L) \\ GSCI_2(\tau_L) \\ GSCI_3(\tau_L) \\ GSCI_4(\tau_L) \\ SEP(\tau_L) \end{bmatrix}}_{\equiv X(\tau_L)} = \sum_{k=1}^{4} \underbrace{\begin{bmatrix} a_{11,k} & a_{12,k} & a_{13,k} & a_{14,k} & a_{15,k} \\ a_{21,k} & a_{22,k} & a_{23,k} & a_{24,k} & a_{25,k} \\ a_{31,k} & a_{32,k} & a_{33,k} & a_{34,k} & a_{35,k} \\ a_{41,k} & a_{42,k} & a_{43,k} & a_{44,k} & a_{45,k} \\ a_{51,k} & a_{52,k} & a_{53,k} & a_{54,k} & a_{55,k} \end{bmatrix}}_{\equiv A_k} \underbrace{\begin{bmatrix} GSCI_1(\tau_L - k) \\ GSCI_2(\tau_L - k) \\ GSCI_3(\tau_L - k) \\ GSCI_4(\tau_L - k) \\ SEP(\tau_L - k) \end{bmatrix}}_{\equiv X(\tau_L - k)} + \underline{\epsilon}(\tau_L) \tag{3.4}$$

where $\epsilon(\tau_L) \overset{mds}{\sim} (0, \Sigma)$[8]. Note that $K_L = K_H = 1$, which reduces the model to an $m+1$-dimensional VAR. Moreover, when $m = 1$, the model collapses to a single-frequency VAR. $X(\tau_L)$ is the mixed-frequency vector, A_k is the matrix of coefficients, and $X(\tau_L - k)$ is the vector of lagged values.

Note in this particular case, $m = 4$, we specify a 5-dimensional mixed-frequency VAR model. Ghysels' (2015) model is primarily designed to handle small sampling frequency ratio, for example 3 or 4. Models handling a larger ratio of sampling frequency (for example, weekly-to-quarterly) are an on-going agenda within the literature (see for example Götz and Hecq (2014)). We did not include a constant term, since both series were demeaned prior to estimation. GSCI weekly observations $GSCI_j(\tau_L)$ are stacked together in the mixed-frequency vector $X(\tau_L)$, corresponding to each monthly low-frequency observation (see below).

	$\tau_L = M_1$	$\tau_L = M_2$	...	$\tau_L = M_{11}$	$\tau_L = M_{12}$
$GSCI_1$	Week 1	Week 1	...	Week 1	Week 1
$GSCI_2$	Week 2	Week 2	...	Week 2	Week 2
$GSCI_3$	Week 3	Week 3	...	Week 3	Week 3
$GSCI_4$	Week 4	Week 4	...	Week 4	Week 4

Table 3.2: Structure of the stacked vector

Table 3.1 above illustrates the structure of the stacked vectors. The monthly low-frequency observations are displayed across the top panel. On the left, the four weekly observations spanning each monthly observation are displayed. Note that for each monthly low-frequency observation τ_L, which spans 12 months, there exists

[8]Note that the mds is a weaker restriction on the error terms.

4 weekly high-frequency observations $GSCI_1$, $GSCI_2$, $GSCI_3$ and $GSCI_4$. In our particular case, the weekly data were set via the data transformations we conducted in subsection 3.1.2.

As maintained by Sadahiro and Motegi (2014), the stacking of the high-frequency variable allows for the existence of heterogeneous effects on the low-frequency variable. To see this, we extract the last row of the mixed-frequency VAR(4) model in (3.4)

$$SEP(\tau_L) = \sum_{k=1}^{4}\left[\sum_{j=1}^{4} a_{5j,k} GSCI_j(\tau_L - k) + a_{55,k} SEP(\tau_L - k)\right] + \epsilon_5(\tau_L) \tag{3.5}$$

Note that $a_{51,k}, a_{52,k}, a_{53,k}$ and $a_{54,k} \equiv a_{5j,k} GSCI_j$ may take on different values from each other. This ultimately implies that the lagged values of GSCI, that is, $GSCI_1(\tau_L - k)$, $GSCI_2(\tau_L - k)$, $GSCI_3(\tau_L - k)$ and $GSCI_4(\tau_L - k)$, have heterogeneous impacts on $SEP(\tau_L)$ (see Sadahiro and Motegi (2014)). On the other hand, the standard monthly VAR model implicitly assumes that the lagged values have a homogeneous impact on $SEP(\tau_L)$. Recalling (3.2)

$$\begin{aligned} SEP(\tau_L) &= \sum_{j=1}^{4}\Big[a_{21,k} GSCI^M(\tau_L - k) + a_{22,k} SEP(\tau_L - k)\Big] + \epsilon_2(\tau_L) \\ &= \sum_{j=1}^{4}\left[a_{21,k}\left\{\frac{1}{4}\sum_{j=1}^{4} GSCI_j(\tau_L - k)\right\} + a_{22,k} SEP(\tau_L - k)\right] + \epsilon_2(\tau_L), \end{aligned} \tag{3.6}$$

Note that the last equality is held due to Equation (3.3). The lagged values of each weekly observation have a homogeneous impact of $a_{21,k}/4$ on $SEP(\tau_L)$. Hence, the mixed-frequency VAR model in (3.4) is more general than the standard monthly VAR in (3.2).

The mixed-frequency VAR(4) is mechanically identical to the single-frequency VAR. That is, standard conditions on single-frequency VARs carry over to Ghysels' (2015) mixed-frequency model. For instance, it is assumed that all the polynomial roots lie outside the unit circle and that $\epsilon(\tau)$ is a strictly stationary martingale difference se-

quence with finite variance, among others (see Ghysels et al. (2014) for a theoretical exposition). Consequently, equation (3.4) was estimated using standard Ordinary Least Squares (OLS) methods. Model estimation was conducted using GNU Octave, and the code to conduct the estimation was retrieved from a closed-source.[9]

3.3 Mixed-frequency Granger Causality Test

Granger's (1969) concept of causality is a key tool utilised to investigate the dynamic relationship between economic variables. Causality, is defined in terms of the predictability of future time periods. Specifically, that "the past and present may cause the future, but the future cannot cause the past" (Granger 1980, page 330). It is the imposition of a temporal ordering which subsequently constructs a causal connection between variables (see Kuersteiner (2008)). Thus, a variable x is said to cause another variable y, if at time t the x helps to predict the variable y_{t+1}.

To test the hypothesis that internet search query data (that is, $GSCI$), contains relevant information beyond that which is reflected in previous job separation values, a test for Granger causality is conducted. However, recall from Chapter 2, that a number of time series properties are not invariant to temporal aggregation, including Granger causality. Temporal aggregation is said to confound parameter estimates across variables, from which aggregated variable may yield different causal patterns. This was demonstrated by Tiao and Wei (1976), and Wei (1982), who have shown that a given one-sided causal pattern between disaggregated variables, could potentially return a feedback causal pattern between temporally-aggregated counterparts. Furthermore, existing literature has demonstrated that temporally-aggregating variables may yield a loss of useful information (see Wohlrabe (2009)). Since temporal aggregation averages over intra-period (high-frequency) observations, the resulting parameter estimate may lose information, which could otherwise be useful. Thus, it is generally considered that variables remain un-filtered.

Recent work by Ghysels et al. (2015b) develop a new class of Granger causality tests

[9]Kaiiji Motegi, the author of the code, provides the code on their website. *http* : *//www.aoni.waseda.jp/motegi/Matlab_Codes.html*

which explicitly account for mixed-frequency time series. Indeed, this is in contrast to standard Granger causality tests in which the variables are all of the same frequency. To circumvent the aforementioned issues related to temporal aggregation, the mixed-frequency class of Granger causality tests supplement the mixed-frequency VAR model we specified in subsection 3.3.2.

To set out Ghysels et al.'s (2015b) mixed-frequency Granger causality test, preliminary notation is initially defined. As the investigation is within a bivariate case, definition/notation is employed from Ghysels et al. (2014)[10]. Let the information set be defined over a Hilbert space $\mathbf{X}(-\infty, \tau_L]$ spanning $\{\mathbf{X}(\tau)|\tau \leq \tau_L\}$, denoted Ω_τ. Ω_τ^H and Ω_τ^L are analogously defined as all the information in the universe except x_L and x_H respectively. Furthermore, $P[x(\tau_L + 1)|I(\tau)]$ is the best linear projection of $x(\tau_L + 1)$ conditional on $I(\tau)$. Consequently, the mixed-frequency information set is denoted $\mathcal{I} = \{\Omega_\tau|\tau_L \in \mathbb{Z}\}$. Ghysels et al. (2014) define Granger non-causality in the following manner:

Definition 1. *The high-frequency variable x_H does not Granger cause the low-frequency variable x_L given the mixed-frequency information set $\mathcal{I}$ if*

$$P[x_L(\tau_L + 1)|\Omega_\tau^L] = P[x_L(\tau_L + 1)|\Omega_\tau]$$

Similarly, x_L does not Granger cause x_H given $\mathcal{I}$ if

$$P[x_H(\tau_L + 1)|\Omega_\tau^H] = P[x_H(\tau_L + 1)|\Omega_\tau]$$

In other words, x_H is said to not Granger-cause x_L if past information on the high-frequency variable, do not help predict future values of the low-frequency variable. Correspondingly, x_L is said to not Granger-cause x_H if past information on the low-frequency variable, do not help predict future values of the low-frequency variable.

From the mixed-frequency VAR specification in Subsection 3.3.2 the causality test is expressed as a test of zero restrictions in VAR (see Dufour and Renault (1998)).

[10]In the short version of Ghysels et al. (2015b), only a bivariate case is considered. An arbitrary number of variables are dealt within the full paper, in which case, testing multi-horizon causal chains are possible i.e Ghysels et al. (2015b).

Taking a closer look at the coefficient matrix in Equation (3.4)

$$\mathbf{A}_k = \begin{bmatrix} a_{11,k} & \dots & a_{14,k} & a_{15,k} \\ \vdots & \ddots & \vdots & \vdots \\ a_{41,k} & \dots & a_{44,k} & a_{45,k} \\ a_{51,k} & \dots & a_{54,k} & a_{55,k} \end{bmatrix} \tag{3.7}$$

we have elements $a_{5j,k}$ from the bottom left of the matrix representing the causality from $GSCI_{1,2,3,4}$ to our dependent variable $SEP(\tau_L)$, and elements $a_{j5,k}$ from the top right representing causality from $SEP(\tau_L)$ to $GSCI_{1,2,3,4}$. It then follows that our internet search query data measure, $GSCI_{1,2,3,4}$ *does not Granger-cause* $SEP(\tau_L)$ if and only if elements $a_{5j,k} = 0_{m\times 1}$ for all $k = 1, ..., p$. Similarly, the US job separations, $SEP(\tau_L)$ *does not Granger-cause* $GSCI_{1,2,3,4}$ if and only if elements $a_{j5,k} = 0_{m\times 1}$ for all $k = 1, ..., p$. Zero restrictions are testable via a linear Wald test, and the asymptotic distribution of the test statistic W_{T_L} under the null hypothesis of Granger non-causality is χ^2_q, where q is the number of restrictions (see Ghysels et al. (2015b)).

In regards to the testing procedure, we note that an i.i.d assumption is relatively strict. A weaker restriction is to assume the error terms follow mds, as noted in Section 3.3.2. To allow this restriction, and the possibility of conditional heteroskedasticity of the unknown form, the Gonçalves and Kilian (2004) (GK) wild bootstrap is employed. The GK bootstrap additionally assists in controlling for potential size distortions due to the small sample period of our Google search query data $\tau_L = 125$).[11] As per the mixed-frequency VAR estimation, the mixed-frequency Granger causality test was conducted in GNU Octave. The test's code was retrieved from the same closed-source as the mixed-frequency VAR code. To compare the mixed-frequency procedure to a single-frequency procedure, a standard low-frequency Granger-causality was applied to the aggregated Google search composite measure, $GSCI^M$. That is, we test the null hypothesis that $GSCI^M \not\rightarrow SEP(\tau_L)$, and correspondingly, $SEP(\tau_L) \not\rightarrow GSCI^M$, where $\not\rightarrow$ represents non-Granger causal-

[11] As we only dealt with a single horizon $h = 1$, Newey and West's (1987) HAC estimator was not used. Ghysels et al. (2015b) note that for $h > 1$, the estimator of the long-run variance may not be positive semi-definite. See page 11.

ity. In the next chapter, we present the Granger causality tests results, and discuss the results in detail.

Chapter 4

Results and Discussion

4.1 Model specification

The mixed-frequency VAR model and its low-frequency counterpart are specified with a lag order of 4, as noted in the previous chapter. However, Akaike's (1974) Information Criterion (henceforth AIC) is minimised at a lag order of 1. It is well known in the econometric literature that existing test for joint whiteness of residuals do not perform well with large dimensional models, such as ours (see for example Lütkepohl (2005) and Sadahiro and Motegi (2014)). Specifically, residual Lagrange Multiplier (LM) tests and multivariate Ljung-Box Q tests among others, do not perform well in small samples with many parameters. Yet we do not simply assume no serial correlation, rather, we conduct univariate residual LM tests for both mixed-frequency and low-frequency models respectively.[1]

The null hypothesis for the residual LM test is that there is no serial correlation of any order up to lag p. In our case, we test up to a maximum lag length of 10. Under the null, the test statistic nR^2 is $\sim \chi^2_p$. Rejecting the null hypothesis would imply that the model specified has serially-correlated residuals. Table 4.1 presents the results of test for the 4th lag order. Each column exhibits the variables for each model, the maximum number of lags regressed, the LM test statistic, and the p-value for each respective lag. P-values marked by asterisk/s indicate the rejection of the null at each lag.

[1]Note that Sadahiro and Motegi (2014) assume no serial correlation in the errors.

Table 4.1: Residual LM test (Lag order of 4)

	Lags	1	2	3	4	5	6	7	8	9	10
MFVAR(4)											
$SEP(\tau_L)$	χ^2	0.000	0.001	0.599	1.495	1.348	4.067	11.206	12.352	14.116	14.497
	p-value	0.986	0.999	0.897	0.827	0.930	0.668	0.130	0.136	0.118	0.151
$GSCI_1$	χ^2	0.066	0.320	0.176	0.600	0.863	1.075	1.712	3.643	4.692	5.750
	p-value	0.798	0.852	0.981	0.963	0.973	0.983	0.974	0.888	0.860	0.836
$GSCI_2$	χ^2	0.134	0.297	0.503	0.354	1.715	5.595	6.566	7.099	8.046	8.377
	p-value	0.714	0.862	0.918	0.986	0.887	0.470	0.475	0.526	0.530	0.592
$GSCI_3$	χ^2	0.008	0.336	0.313	0.216	1.062	2.387	3.761	4.393	6.148	6.803
	p-value	0.927	0.845	0.958	0.995	0.957	0.881	0.807	0.820	0.725	0.744
$GSCI_4$	χ^2	0.010	0.026	0.083	0.228	0.252	2.833	3.112	5.612	6.244	7.051
	p-value	0.921	0.987	0.994	0.994	0.998	0.830	0.875	0.691	0.715	0.721
VAR(4)											
$SEP(\tau_L)$	χ^2	0.004	0.004	0.391	0.977	0.898	3.642	9.616	10.034	12.477	12.822
	p-value	0.952	0.998	0.942	0.913	0.970	0.725	0.211	0.263	0.188	0.234
$GSCI^M$	χ^2	0.0723	0.154	0.732	0.765	2.137	2.561	2.830	3.289	2.865	10.790
	p-value	0.787	0.926	0.866	0.943	0.830	0.862	0.900	0.915	0.969	0.374

For the p-values of the test, we put *** if the null hypothesis of no serial correlation up to p lags is rejected at the 1% level (strong rejection), ** if rejected at 5% but not at 1%, and * if rejected at 10% but not 5% (weak rejection).

As it can be seen, at the lag order of 4, we do not reject the null hypothesis for no serial correlation at every p. Hence, there was insufficient evidence to suggest that at the lag order of 4, there is serially correlated errors in each model. We should note that tests were sequentially conducted, beginning with the AIC-specified lag order of 1. The LM tests for orders 1, 2 and 3 all rejected the null hypothesis. Therefore, in accordance with the VAR specification in the previous chapter, we pursued a (mixed frequency-) VAR model of lag length 4.

4.2 Granger Causality Test

The mixed-frequency Granger causality test set out by Ghysels et al. (2015b), tests the null hypothesis that the percentage growth rate of all the high-frequency Google search composite variables $\{GSCI_1, GSCI_2\ , GSCI_3, GSCI_4\}$ do not Granger-cause the percentage growth rate of the low-frequency US job separation measure $SEP(\tau_L)$. Correspondingly, the null hypothesis that $SEP(\tau_L)$ does not Granger-cause $\{GSCI_1$, $GSCI_2, GSCI_3, GSCI_4\}$ is tested. Such causal patterns correspond to Case I and II in Ghysels et al. (2015b) (page 8) respectively. Case I and Case II treat bivariate causal patterns such that causal chains do not exist.[2] Accordingly, Case I treats

[2]Causal chains can emerge from trivariate relationships.

Granger non-causality from all high-frequency variables to all low-frequency variables, whilst Case II treats Granger non-causality from all low-frequency variables to all high-frequency variables (see Ghysels et al. (2015b) for more details).

In the event that the Wald statistic W_{T_L} for each causal pattern is significant, a rejection of the null for Case I would imply that the lagged values of the high-frequency variables $GSCI_{1,2,3,4}$ do help predict future values of $SEP(\tau_L)$. Equally, a rejection of the null for Case II would imply that the lagged values of the low-frequency variable $SEP(\tau_L)$ do help predict future values of $GSCI_{1,2,3,4}$. We present the results below for this test below.

Table 4.2: Mixed-frequency Granger Causality Wald Test

Null Hypothesis	χ^2 (Test Statistic)	GK (2004) p-value	Conclusion
$GSCI_{1,2,3,4} \not\rightarrow SEP(\tau_L)$	33.473	0.581	Do not reject
$SEP(\tau_L) \not\rightarrow GSCI_{1,2,3,4}$	51.349	0.390	Do not reject

For the p-values of the test, we put *** if the null hypothesis of non-Granger causality is rejected at the 1% level (strong rejection), ** if rejected at 5% but not at 1%, and * if rejected at 10% but not 5% (weak rejection). GK stands for the Gonçalves and Kilian (2004) bootstrapped p-values.

Table 4.2 presents the results of the mixed-frequency Granger causality test. The null hypothesis for each causal pattern, the chi-squared distributed Wald statistic, the Gonçalves and Kilian (2004) bootstrapped p-values and the subsequent conclusion for each test are summarised respectively. As it can be seen, the null hypothesis that $GSCI_{1,2,3,4}$ does not Granger-cause $SEP(\tau_L)$ is not rejected at all significant levels. Equally, the null hypothesis that $SEP(\tau_L)$ does not Granger-cause $GSCI_{1,2,3,4}$ is not rejected at all significant levels. The result in the first row implies that there is insufficient evidence to suggest that the percentage growth in the weekly Google composite measures ($GSCI_{1,2,3,4}$) are useful in predicting future percentage growth in job separation ($SEP(\tau_L)$). The result in the second row implies that there is insufficient evidence to suggest that the percentage growth in $SEP(\tau_L)$ is useful in predicting future percentage growth in $GSCI_{1,2,3,4}$.

To compare the mixed-frequency Granger causality results to the low-frequency counterpart, the test results for the standard low-frequency VAR(4) are presented.

Table 4.3: Low-frequency Granger Causality Wald Test

Null Hypothesis	χ^2 (Test Statistic)	GK (2004) p-value	Conclusion
$GSCI^M \not\to SEP(\tau_L)$	6.554	0.240	Do not reject
$SEP(\tau_L) \not\to GSCI^M$	1.965	0.833	Do not reject

For the p-values of the test, we put *** if the null hypothesis of non-Granger causality is rejected at the 1% level (strong rejection), ** if rejected at 5% but not at 1%, and * if rejected at 10% but not 5% (weak rejection). GK stands for the Gonçalves and Kilian (2004) bootstrapped p-values.

Summarised in Table 4.3, the null hypothesis that $GSCI^M$ does not Granger-cause $SEP(\tau_L)$ is not rejected at all significant levels. Equally, the null hypothesis that $SEP(\tau_L)$ does not Granger-cause $GSCI^M$ is not rejected at all significant levels. The result in the first row implies that there is insufficient evidence to suggest that percentage growth in the *monthly aggregated* Google Search Composite Index ($GSCI^M$) is useful to predict future percentage growth in US job separations ($SEP(\tau_L)$). Equally, the result in the second row implies that there is insufficient evidence to suggest that $SEP(\tau_L)$ is useful to predict future percentage growth in $GSCI^M$.

The results reflected in Table 4.2 are not directly interpretable through the empirical literature since no other study, to the best of our knowledge, has applied Ghysels et al's (2015b) mixed-frequency Granger causality test.[3] However, we may discuss the results in relation to a number of possible explanations drawn from the existing literature. Accordingly, we provide such a comparison in the next section.

The results presented in Table 4.3 are generally in contrast to the existing literature which has examined whether Google search queries contain relevant information in predicting economic variables (see for example Askitas and Zimmermann (2009) and Alexander Dietzel et al. (2014) among others). For instance, Askitas and Zimmermann (2009) find a Granger-causal relationship between aggregated Google search data related to various labour market characteristics and the German unemployment rate. Their conclusion is drawn from a pool of search terms which seek to reflect

[3]Note that although Ghysels et al. (2015b) provide a simple empirical application, it is not based on Internet search query data.

flows into unemployment and flows into employment, similar to the terms we adopt. As we discuss in the succeeding section, the distinction between the aforementioned results and the existing literature may emerge from the selection of search terms.

4.3 Discussion of Results

In Section 2.4, we set out two hypotheses - a main hypothesis, and an auxiliary hypothesis. The main hypothesis was that Google search query data related to unemployment insurance, government benefits and services are useful in predicting future US job separation statistics. The auxiliary hypothesis was that we yield distinct results to temporal aggregation by modelling the relationship between Google search query data and US job separation. As it was presented above, the results reflect the insufficiency of evidence to reject the null hypothesis of Granger non-causality for both mixed-frequency modelling and low-frequency modelling. In particular, evidence in favour of our main and auxiliary hypotheses was not found. Consequently, an important question emanates in this regard - what are the possible explanations for the aforementioned results?

Main hypothesis

In relation to the main hypothesis, two possible accounts provide a basis for the results. Firstly, one of the challenges in utilising Google Trends search query data is that Google updates its weekly Trends data on a daily basis. That is, for a given calendar week, Google Trends can yield different weekly values on two separate days. However, queries sent to Google on the same day, yield identical values (for a given search term). This sampling procedure induces a source of measurement error, which is likely to weaken the informational content of the signal in internet search query data (see for example Carrière-Swallow and Labbé (2013) and Seabold and Coppola (2015)). Therefore, the daily variation in the weekly data could mask the true relation between the two variables we employed. Despite the construction of the composite index, the results in Table 4.2 and 4.3 may consequently reflect the outcome of excessive noise induced by the Google search composite measure $GSCI$.

Secondly, term selection is an additional explanation. The selection of terms should reflect what an internet user may enter into a search engine.[4] In part of the researcher utilising internet search query data, an element of subjective judgement is

[4]Section 2.6.

involved. Despite adapting a similar process of term selection to the existing literature (see for example Askitas and Zimmermann (2009), Koop and Onorante (2013) and Smith (2015)), the results do not necessarily reflect this process. This is due primarily to the complexity in perceiving which keywords users enter into online search portals.[5] As noted in Chapter 2, this complexity can induce a data-snooping bias, which is an on-going challenge in utilising internet search query data.

Auxiliary hypothesis

In relation to our auxiliary hypothesis, a number of explanations may clarify the results which we obtain. Firstly, in implementing the mixed-frequency Granger causality test above, the mixed-frequency VAR model employed comprised on $K = 5$ dimensions, where $K = K_H + m \times K_L$. That is, for the given sampling frequency ratio between weekly Google search query data and the monthly job separation data, $m = 4$, $K = m + 1 = 5$. For the conclusive lag length we set in testing for serial independence in the residuals ($k = 4$), there are as many as $\{pK \times K\} = 100$ parameters in the model. This can have an adverse impact on the power of the mixed-frequency Granger causality test, as demonstrated by Ghysels et al. (2015b) (Section 6.2), although it is less of an issue for the low-frequency test.[6] The results presented above could potentially reflect the significant loss of power in the test due to the large dimensionality present in a observation-driven mixed-frequency VAR model.

Secondly, the stacked mixed-frequency vector $X(\tau_L)$, of the mixed-frequency VAR is assumed to follow a specific ordering. The ordering of the observations in the mixed-frequency vector characterise the timing of intra-τ_L period releases (see for example Ghysels (2015)). As such, the order of the information releases in the vector determine the impact and timing of subsequent shocks. In our particular case, the timing, and subsequent ordering of the observations in the stacked mixed-frequency

[5]The examination of user search strategies in the context of search engine portals and databases is explored in the information sciences. See for example Teevan et al. (2004).

[6]See Section 7 of Ghysels et al. (2015b).

vector are set in Equation 4.1.

$$\mathbf{X}(\tau_L) = [GSCI(\tau_L, 1)', GSCI(\tau_L, 2)', GSCI(\tau_L, 3)', GSCI(\tau_L, 4)', SEP(\tau_L)']' \tag{4.1}$$

Observe that the low-frequency variable $SEP(\tau_L)$ is the last block in the mixed-frequency vector. The rationale behind this ordering is due to the observation of the monthly $SEP(\tau_L)$ series at the end of a respective month. The first observation of the high-frequency variable is subsequently made public from the 1st of the month, followed by the additional weekly observations. This is in line with Ghysels et al. (2015b) who note that it is the conventional ordering (which we adopt) is applicable to most cases. Alternative orderings may correspondingly yield different outcomes, yet it is distinctive from application to application.

Finally, an alternative explanation is related to the sampling frequency ratio between the mixed-frequency series. It is well known in the existing literature that temporal aggregation can potentially lead of a loss of information (see for example Wei and Mehta (1980), Marcellino (1999) and Foroni and Marcellino (2013) among others). This may likely occur since temporal aggregation involves aggregating over the high intra-τ_L observations. However, Wei and Mehta (1980) demonstrate that the loss of information is generally more pronounced for larger m values. Recall Equation 3.3 in Chapter 3

$$GSCI^M(\tau_L) = \frac{1}{4}\sum_{j=1}^{4} GSCI_j(\tau_L)$$

The sampling frequency ratio is 4 such that the aggregated variable is defined as an equally-weighted average of the weekly observations. The aforementioned ratio is relatively small, in comparison to alternatives (weekly-quarterly and monthly-yearly), where the ratio is much larger. Therefore, in lieu of information loss, the parallel results between the mixed-frequency and low-frequency models may potentially arise from a similar account to our main hypothesis - namely, measurement error due to noise. The following chapter concludes this thesis and provides further considerations for future research.

Chapter 5

Conclusion

The main purpose of this thesis was to contribute to the expanding literature on the applicability of big data to economic phenomena - specifically, Internet search query data. In this regard, we sought to assess whether weekly Google search queries were informationally relevant in predicting future monthly US job separations. The rationale behind this assessment was that Internet users who intend to quit the labour force, or in fear of losing their jobs, search the Internet for information on unemployment eligibility, social services and welfare benefits. Search terms were pooled together to construct a dynamically-weighted Google search composite measure which seemingly reflected the information search behaviour of users online. With respect to this assessment, we found no evidence in support of this affirmation.

To directly model our measure of Internet search queries and monthly labour turnover, whilst avoiding temporal aggregation, Ghysels et al's (2015) mixed-frequency vector autoregression model was employed. According to the existing literature, temporal aggregation is known to potentially yield spurious outcomes on various time series properties, including causality. In this regard, the informational content of the weekly Google search query data was assessed via a mixed-frequency Granger causality test, as set out by Ghysels et al. (2015b). To compare the mixed-frequency outcome to a temporally-aggregated outcome, a standard low-frequency vector autoregression, and subsequent Granger causality test were applied. We found that both models yielded the same conclusion.

With regards to the results we attained, there are several avenues for future research. Firstly, as we noted in the previous chapter, the mixed-frequency VAR model we employed, held as many as 100 parameters in the model. Ghysels (2015) proposes various non-linear parsimonious specifications from the MIDAS literature, which could be applied to account for the dimensionality in the model. Secondly, as is contended within the existing literature, term selection is an important element in applying Internet search query data. However, the difficulty lies in nominating appropriate search terms which best reflect the search behaviour of Internet users. Thus, future consideration may adopt an algorithmic method, such as Bayesian variable selection, to select search terms. This could prove to be a more robust and objective process for term selection.

Appendix A

Google Search Composite Index material

A.1 List of search terms

Search Keywords
unemployment insurance
unemployment benefits
eligibility for unemployment
requirements for unemployment
wrongful termination
wrongful termination lawyers
state unemployment insurance
unemployment insurance benefits
state minimum wage
laid off

Table A.1: List of Google search terms used to construct the Google Search Composite Index.

A.2 Code

Below, we present the code of the function written in R to download Google Trends data, and subsequently construct the Google Search Composite Index (GSCI) from the pool of terms presented in Appendix A.1. Each line of code corresponds to a number, and comments on the code are coloured in red.

```
1  get_terms <- function(x) {
2    # Create character vector "terms.str" of search terms from file.
3    terms.str <- as.character(read.csv(file.choose(),
4                                       header=FALSE)[,1], stringsAsFactor=FALSE)
5    # Detach each search term from the vector as single elements.
6    terms <- strsplit(terms.str, split=0, fixed=FALSE, perl=FALSE, useBytes=FALSE)
7    # load googletrend package (Chris Okugami).
8    message("retrieving multiple search volume data from Google Trends...")
9    require(googletrend)
10   # Retrieving Google Trends data via API which,
11   # is fed with elements of the .csv file in line 3.
12   terms.list <- gettrend(keyword=(terms), geo='US', simple = TRUE)
13   # load xts package.
14   require(xts)
15   # Create object 'xts.list' to hold two xts objects -
16   # One for weekly data and the other for monthly data.
17   xts.list <- vector("list", 2)
18   weekly = NULL # create weekly xts object
19   monthly = NULL # create monhtly xts object
20   # From terms.list (line 10), run a loop to identify which search term,
21   # is returned as weekly and monthly:
22   for (i in 1:length(terms.list)) {
23   # creates list object of column names from the downloaded search terms.
24     dimNames <- list(NULL, names(terms.list)[i])
25   # Calculate the difference in days between
26   # the first and second elements of each xts object in list.
27     diff.days <- difftime(as.Date(terms.list[[i]][2,1]),
28                           as.Date(terms.list[[i]][1,1]))
29     if (diff.days < 10) {
30     freq <- 52 # if difference is < 10, its frequency is weekly.
31   # column bind each weekly-identified terms into an object called 'weekly'
```

```
32      weekly = cbind(weekly, xts(terms.list[[i]][,2], order.by=terms.list[[i]][[1]],
33                     frequency=freq, dimnames = dimNames))
34      } else {
35      freq <- 12 # if difference is otherwise greater, identify as monthly.
36    # column bind each monthly-identified terms into an object called 'monthly'
37    monthly = cbind(monthly, xts(terms.list[[i]][,2], order.by=terms.list[[i]][[1]],
38                                         frequency=freq, dimnames = dimNames))
39      }
40    # subset Google Trends data to required time period
41    weekly <- weekly["2004-01-01/2015-05-31"]
42    monthly <- monthly["2004-01-01/2015-05-31"]
43    # combined weekly & monthly objects into a list.
44    xts.list <- list(weekly, monthly)
45    # name each element of list respectively.
46      names(xts.list) <- c("weekly", "monthly")
47    }
48    # Create balanced weeks in line with Smith (2015):
49      message("disaggregating data into daily series")
50    # Define new daily sequence of dates.
51      new1 <- seq(from=as.Date("2004-01-01"), to=as.Date("2015-05-31"), by = "day")
52    # weekly observations for each search term are merged into a daily series:
53      daily_series = NULL
54      daily_series <- merge(xts.list$weekly, xts(, new1))
55    # We assume that the weekly values are unchanged,
56    # from weekly object throughout a particular week.
57    # Then, the weekly values are interpolated into daily observations.
58    #
59    daily_series <- na.locf(daily_series, fromLast = TRUE)
60    message("creating a weekly average from the daily series")
61    # Daily observations are split into months according to calendar.
62    split_list <- split(daily_series, f = "months", drop = FALSE, k = 1)
63    # Create large list of elements which are lists of
64    # 4 weekly elements corresponding to weeks defined (page 34 of thesis).
65    splitlist = NULL
66    for (i in 1:length(split_list)) {
67      intervals <- cut(.indexmday(split_list[[1]]),
68                       c(0, 7, 14, 21, 31), paste0("W", 1:4))
69      splitlist[[i]] <- split(split_list[[i]], intervals)
```

```
70      splitlist
71    }
72    # Loop through each monthly element i and weekly element j,
73    # and combined rows into weekly observations.
74    dat1 = NULL
75    for (i in 1:length(splitlist)){
76      for (j in 1:4){
77      dat1 = rbind(dat1, sapply(splitlist[[i]][[paste0("W", j)]], FUN = mean))
78      dat1
79      }
80    }
81    # create sequence of structured dates as defined on page 34 of this thesis :
82    require(lubridate)
83    v1 <- seq(as.Date("2004-01-01"), as.Date("2015-05-31"), by = "week")
84    # split vector into months and years
85    lst <- split(v1, list(month(v1), year(v1)), drop=TRUE)
86    # substring extracts first 4 observations from the month.
87    days <- substr(v1[1:4],9,10)
88    v2 <- unlist(lapply(lst, function(y) {
89      sprintf('%s%s', substr(y[1:4], 1,8), days)}), use.names=FALSE)
90
91    # convert structured weekly object to dataframe,
92    # then to xts object with sequence above:
93    weekly_df <- data.frame(week=as.Date(v2), dat1)
94    weekly_series <- xts(weekly_df[-1], order.by = as.Date(v2))
95
96    # Calculating weekly Composite Index as per Smith (2015)
97    # (subsection 3.2.3 page 35-36 of thesis):
98    W = NULL # weights defined by equation 3.1 page 36 of thesis.
99    for (t in 1:nrow(weekly_series)) {
100     W = rbind(W, weekly_series[t][,]/sum(weekly_series[t][,]))
101     W
102   }
103   # Define Google Search Composite Index (GSCI) object:
104   # Smith (2015) page 8. Definition of GSCI is,
105   # the sum of weighted individual searches for each search term:
106   GSCI = NULL
107   dat <- W*weekly_series
```

```
108     GSCI = xts(rowSums(dat), order.by = index(dat))
109     GSCI <- GSCI["2004-01-01/2015-05-01"]
110     GSCI # return the Google composite measure
```

Description of code

Two key parts of the code are explained in greater detail for clarity. Firstly, the code explicitly takes into account the fact that, in downloading Google search query data, the sampling frequency of the data may either be monthly, weekly or daily (rarely). This occurs if a particular search term, does not accumulate enough volume. We noted this issue in Chapter 3 Subsection 3.3.2. This issue is taken into account through an identification process in lines 21-38. Specifically, the frequency of each search term's time series is identified by calculating the difference in the number of days between the second element and first element. That is, the difference between the second data point, and the first data point. If the difference < 10, the data are identified as weekly, otherwise it is identified as monthly.

Secondly, as per Smith (2015), lines 50-94 construct the balanced calendar weeks due to monthly overlapping values from the Google Trends data, and subsequently creates a structured weekly dataset (See Chapter 3 page 34 for an explanation). The process involved placing the weekly and monthly identified data into a list, from which we split each element by months. Hence, each split element contained daily data depending on which month it fell in. Each monthly split elements was subsequently split again into 4 weeks based on the definition we set on page 34 of this thesis. Finally, each weekly series of daily data were aggregated and averaged into a weekly value, corresponding to the sequence of dates created (lines 81-89).

A number of packages were utilised within the function. For example, in line 12 we call the 'gettrend' function from the 'Google Trend' package to import Google Trends data from its API. [1] Throughout the function, the 'xts' package[2] was employed for the time series objects that were created. Finally, we utilise the 'lubridate'

[1] Okugami (2015).
[2] Ryan and Ulrich (2014).

package to deal with different time intervals.[3]

[3]Grolemund and Wickham (2011).

Appendix B

Statistics

B.1 Correlation coefficients

Table B.1: Contemporaneous Correlation Coeffcients

	$GSCI_1(\tau_L)$	$GSCI_2(\tau_L)$	$GSCI_3(\tau_L)$	$GSCI_4(\tau_L)$	$GSCI^M(\tau_L)$	$TURN(\tau_L)$
$GSCI_1(\tau_L)$	1					
$GSCI_2(\tau_L)$	-0.323	1				
$GSCI_3(\tau_L)$	-0.044	-0.178	1			
$GSCI_4(\tau_L)$	-0.060	-0.026	-0.369	1		
$GSCI^M(\tau_L)$	0.429	0.275	0.215	0.485	1	
$TURN(\tau_L)$	0.033	0.086	-0.028	0.035	0.087	1

Table B.1 reports contemporaneous correlation coefficients between each variable. For example, the correlation between $TURN(\tau_L)$ and $GSCI_1(\tau_L)$ is 0.033.

Table B.2: Correlation Coefficients (Lag 1)

	$GSCI_1(\tau_L-1)$	$GSCI_2(\tau_L-1)$	$GSCI_3(\tau_L-1)$	$GSCI_4(\tau_L-1)$	$GSCI^M(\tau_L-1)$	$TURN(\tau_L-1)$
$GSCI_1(\tau_L)$	0.097	0.162	-0.113	-0.021	0.083	0.024
$GSCI_2(\tau_L)$	-0.085	-0.054	0.052	-0.052	-0.102	-0.082
$GSCI_3(\tau_L)$	-0.125	0.036	0.047	-0.208	-0.198	-0.104
$GSCI_4(\tau_L)$	0.011	-0.306	-0.079	0.105	-0.157	0.164
$GSCI^M(\tau_L)$	-0.058	-0.136	-0.079	-0.101	-0.258	0.027
$TURN(\tau_L)$	0.038	-0.055	0.084	-0.078	-0.014	-0.427

Table B.2 reports correlation coefficients with 1 time lag.

Bibliography

Alexander Dietzel, M., Braun, N., and Schäfers, W. (2014). Sentiment-based Commercial Real Estate Forecasting with Google search volume data. *Journal of Property Investment & Finance*, 32(6):540–569.

Andreou, E., Ghysels, E., and Kourtellos, A. (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics*, 158(2):246–261.

Antenucci, D., Cafarella, M., Levenstein, M., Ré, C., and Shapiro, M. D. (2014). Using Social media to measure Labor market flows. Technical report, National Bureau of Economic Research.

Armesto, M. T., Engemann, K. M., Owyang, M. T., et al. (2010). Forecasting with mixed frequencies. *Federal Reserve Bank of St. Louis Review*, 92(6):521–36.

Askitas, N. and Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *German Council for Social and Economic Data (RatSWD) Research Notes*, (41).

Autor, D. H. (2001). Wiring the labor market. *Journal of Economic Perspectives*, pages 25–40.

Baffigi, A., Golinelli, R., and Parigi, G. (2004). Bridge models to forecast the Euro Area GDP. *International Journal of forecasting*, 20(3):447–460.

Bai, J., Ghysels, E., and Wright, J. H. (2013). State space models and Midas regressions. *Econometric Reviews*, 32(7):779–813.

Bajari, P. and Hortacsu, A. (2003). The winner's curse, reserve prices, and endogenous entry: Empirical insights from ebay auctions. *RAND Journal of Economics*, pages 329–355.

Baker, M. and Wurgler, J. (2007). Investor sentiment in the Stock Market. *The Journal of Economic Perspectives*, pages 129–151.

Baker, S. R. and Fradkin, A. (2014). The Impact of Unemployment Insurance on job search: Evidence from Google search data.

Bangwayo-Skeete, P. F. and Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? mixed-data sampling approach. *Tourism Management*, 46:454–464.

Bholat, D. (2015). Big data and central banks. *Big Data & Society*, 2(1).

Burda, M. and Wyplosz, C. (1994). Gross worker and job flows in europe. *European economic review*, 38(6):1287–1315.

Carrière-Swallow, Y. and Labbé, F. (2013). Nowcasting with Google Trends in an emerging market. *Journal of Forecasting*, 32(4):289–298.

Cavallo, A. (2012). Scraped data and sticky prices.

Chang, S.-S., Kwok, S. S. M., Cheng, Q., Yip, P. S., and Chen, Y.-Y. (2015). The association of trends in charcoal-burning suicide with google search and newspaper reporting in taiwan: a time series analysis. *Social psychiatry and psychiatric epidemiology*, pages 1–11.

Choi, H. and Varian, H. (2009a). Predicting initial claims for unemployment benefits. *Google Inc.*

Choi, H. and Varian, H. R. (2009b). Predicting the present with Google trends. Technical report.

Da, Z., Engelberg, J., and Gao, P. (2011). In search of attention. *The Journal of Finance*, 66(5):1461–1499.

Da, Z., Engelberg, J., and Gao, P. (2015). The sum of all fears investor sentiment and asset prices. *Review of Financial Studies*, 28(1):1–32.

De los Santos, B., Hortaçsu, A., and Wildenbeest, M. R. (2012). Testing models of consumer search using data on web browsing and purchasing behavior. *The American Economic Review*, 102(6):2955–2980.

Doornik, J. A. (2009). Improving the timeliness of data on influenza-like illnesses using google search data.

Dufour, J.-M. and Renault, E. (1998). Short run and long run causality in time series: theory. *Econometrica*, pages 1099–1125.

Dugas, A. F., Hsieh, Y.-H., Levin, S. R., Pines, J. M., Mareiniss, D. P., Mohareb, A., Gaydos, C. A., Perl, T. M., and Rothman, R. E. (2012). Google flu trends: correlation with emergency department influenza rates and crowding metrics. *Clinical infectious diseases*, 54(4):463–469.

D'Amuri, F. and Marcucci, J. (2010). 'Google it!'forecasting the us unemployment rate with a google job search index.

Edelman, B. (2012). Using internet data for economic research. *The Journal of Economic Perspectives*, pages 189–206.

Einav, L. and Levin, J. D. (2013). The Data revolution and Economic Analysis. Technical report, National Bureau of Economic Research.

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

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

Foroni, C. and Marcellino, M. G. (2013). A Survey of Econometric methods for mixed-frequency data.

Ghysels, E. (2015). Macroeconomics and the reality of mixed frequency data. *Journal of Econometrics (forthcoming)*.

Ghysels, E., Hill, J. B., and Motegi, K. (2014). Testing for Granger Causality with Mixed frequency data (short version).

Ghysels, E., Hill, J. B., and Motegi, K. (2015a). Simple Granger Causality tests for Mixed frequency data.

Ghysels, E., Hill, J. B., and Motegi, K. (2015b). Testing for Granger Causality with Mixed Frequency data. *Journal of Econometrics (forthcoming)*.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The MIDAS touch: Mixed data sampling regression models. *Finance*.

Ghysels, E., Sinko, A., and Valkanov, R. (2007). Midas regressions: Further results and new directions. *Econometric Reviews*, 26(1):53–90.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

Gonçalves, S. and Kilian, L. (2004). Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. *Journal of Econometrics*, 123(1):89–120.

Götz, T. B. and Hecq, A. W. (2014). Testing for Granger causality in large mixed-frequency vars.

Grolemund, G. and Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25.

Hamilton, J. D. and Wu, J. C. (2014). Risk premia in crude oil futures prices. *Journal of International Money and Finance*, 42:9–37.

Hand, C. and Judge, G. (2012). Searching for the picture: forecasting uk cinema admissions using google trends data. *Applied Economics Letters*, 19(11):1051–1055.

Hansen, P. and Lunde, A. (2011). Forecasting volatility using high frequency data. *The Oxford Handbook of Economic Forecasting', Oxford: Blackwell*, pages 525–556.

Joseph, K., Wintoki, M. B., and Zhang, Z. (2011). Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. *International Journal of Forecasting*, 27(4):1116–1127.

Kearney, C. and Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185.

Kogut, C. A. (1990). Consumer search behavior and sunk costs. *Journal of Economic Behavior & Organization*, 14(3):381–392.

Koop, G. and Onorante, L. (2013). Macroeconomic nowcasting using Google probabilities.

Kroft, K. and Pope, D. G. (2014). Does online search crowd out traditional search and improve matching efficiency? evidence from craigslist. *Journal of Labor Economics*, 32(2):259–303.

Kuersteiner, G. (2008). Granger-sims causality. *The New Palgrave Dictionary of Economics, 2nd Edition, Palgrave Macmillan. Doi*, 10(9780230226203.0665).

Lazer, D. M., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google flu: traps in big data analysis.

Lippman, S. A. and McCall, J. (1976). The economics of job search: A survey. *Economic inquiry*, 14(2):155–189.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Marcellino, M. (1999). Some consequences of temporal aggregation in empirical analysis. *Journal of Business & Economic Statistics*, 17(1):129–136.

Mariano, R. S. and Murasawa, Y. (2010). A coincident index, common factors, and monthly real gdp*. *Oxford Bulletin of Economics and Statistics*, 72(1):27–46.

McCracken, M., Owyang, M., and Sekhposyan, T. (2013). Real-time forecasting with a large bayesian block model. Technical report, Discussion Paper, Federal Reserve Bank of St. Louis and Bank of Canada.

McLaren, N. and Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, (2011):Q2.

Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H., and Kumar, S. (2011). Google correlate whitepaper. Technical report, Google Inc.

Okugami, C. (2015). *Googletrend: R Google trend (automated download Google trend data to R)*. R package version 1.1.

Perlin, M., Caldeira, J., Santos, A. A., and Pontuschka, M. (2014). Can we predict the financial markets based on internet search queries?

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rajaguru, G. (2004). *Effects of Temporal Aggregation and Systematic sampling on Model Dynamics and Causal inference*. PhD thesis, PhD Thesis, Department of Economics, National University of Singapore.

Rajaguru, G. and Abeysinghe, T. (2012). The distortionary effects of temporal aggregation on Granger causality. In *Some Recent Developments in Statistical Theory and Applications: Selected Proceedings of the Interenational Conference on Recent Developments in Statistics, Econometrics and Forecasting, University of Allahabad, India, December 27-28, 2010*, page 38. Universal-Publishers.

Ripberger, J. T. (2011). Capturing curiosity: Using internet search trends to measure public attentiveness. *Policy Studies Journal*, 39(2):239–259.

Rogerson, R. and Shimer, R. (2011). Search in macroeconomic models of the labor market. *Handbook of Labor Economics*, 4:619–700.

Rossana, R. J. and Seater, J. J. (1995). Temporal aggregation and economic time series. *Journal of Business & Economic Statistics*, 13(4):441–451.

Ryan, J. A. and Ulrich, J. M. (2014). *xts: eXtensible Time Series*. R package version 0.9-7.

Sadahiro, A. and Motegi, K. (2014). Sluggish Private investment in Japan's Lost Decade: Mixed frequency vector autoregression approach.

Seabold, S. and Coppola, A. (2015). Nowcasting prices using Google Trends.

Shimer, R. (2005). The Cyclicality of Hires, Separations, and job-to-job transitions. *REVIEW-FEDERAL RESERVE BANK OF SAINT LOUIS*, 87(4):493.

Siganos, A., Vagenas-Nanos, E., and Verwijmeren, P. (2014). Facebook's daily sentiment and international stock markets. *Journal of Economic Behavior & Organization*, 107:730–743.

Silvestrini, A. and Veredas, D. (2008). Temporal aggregation of univariate and multivariate time series models: a survey. *Journal of Economic Surveys*, 22(3):458–497.

Sims, C. A. (1971). Discrete approximations to continuous time distributed lags in econometrics. *Econometrica: Journal of the Econometric Society*, pages 545–563.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pages 1–48.

Singer, E. (2002). The use of incentives to reduce nonresponse in household surveys. *Survey nonresponse*, 51:163–177.

Smith, P. (2015). Predicting UK Unemployment with Internet Search and Survey data.

Stevenson, B. (2008). The internet and job search. Technical report, National Bureau of Economic Research.

Stigler, G. J. (1961). The economics of information. *The journal of political economy*, pages 213–225.

Taylor, L., Schroeder, R., and Meyer, E. (2014). Emerging practices and perspectives on big data analysis in economics: Bigger and better or more of the same? *Big Data & Society*, 1(2):2053951714536877.

Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R. (2004). The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 415–422. ACM.

Tiao, G. C. and Wei, W. S. (1976). Effect of temporal aggregation on the dynamic relationship of two time series variables. *Biometrika*, 63(3):513–523.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, pages 3–27.

Vicente, M. R., López-Menéndez, A. J., and Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? *Technological Forecasting and Social Change*, 92:132–139.

Vlastakis, N. and Markellos, R. N. (2012). Information demand and stock market volatility. *Journal of Banking & Finance*, 36(6):1808–1821.

Vosen, S. and Schmidt, T. (2012). A monthly consumption indicator for Germany based on internet search query data. *Applied Economics Letters*, 19(7):683–687.

Wei, W. W. (1982). Comment: The effects of systematic sampling and temporal aggregation on causality—a cautionary note. *Journal of the American Statistical Association*, 77(378):316–319.

Wei, W. W. and Mehta, J. (1980). Temporal aggregation and information loss in a distributed lag model. *Analyzing Time Series*, pages 271–281.

William, W. and Wei, S. (1990). Time series analysis: univariate and multivariate methods.

Wohlrabe, K. (2009). *Forecasting with Mixed-frequency time series models*. PhD thesis, lmu.

Wu, L. and Brynjolfsson, E. (2014). The Future of Prediction: How Google Searches foreshadow Housing Prices and Sales. In *Economics of Digitization*. University of Chicago Press.

Zadrozny, P. (1988). Gaussian likelihood of continuous-time ARMAX models when data are stocks and flows at different frequencies. *Econometric Theory*, 4(01):108–124.