FORECASTING THE CANADIAN UNEMPLOYMENT RATE USING INTERNET SEARCHES

by Devon Mitchell

(3439728)

Major Paper presented to the
Department of Economics of the University of Ottawa
in partial fulfillment of the requirements of the M.A. Degree
Supervisor: David Gray

ECO 6999

Ottawa, Ontario
August 2015

I. Introduction

The expansion and ubiquity of the internet has had a broad and profound impact on many aspects of society.[1] Nie and Erbring (2000) encapsulate this influence: "IT [Information Technology] innovations are revolutionizing information and entertainment delivery, affecting their production and consumption, transforming our social life and behavior, even our political institutions and the role of citizens within them" (p.1). Accordingly, internet use in Canada has become tremendously prevalent and is rapidly growing. The 2012 *Canadian Internet Use Survey* shows that 83% of households had access to the internet in 2012, up from 79% in 2010 (Statistics Canada 2013).

The internet's proliferation is significantly changing the way people search and consume information as it is becoming the primary source of information in economic decision-making; for instance in purchasing goods, investing, or searching for employment (Chamberlin 2010). A Statistics Canada report states further, "with 97% of [internet] users conducting online searches and 93% using the Internet for communication in 2007, uptake is approaching a maximum, with almost all Internet users now engaging in these activities" (Middleton et al. 2010, p.10). As internet use booms, search engines have become the principal mainstay of online browsing.[2] A February 2012 Pew survey found search engines are used by 91% of adult internet users to find information and this percentage has been increasing in the past decade (Purcell et al. 2012). Consequently, it seems natural to ask if any usable information can be acquired from this plethora of internet searches, and if it is possible to discern signals about socio-economic variables. Furthermore, one may ponder if

[1] Middleton and Ellison (2008) further discuss the social impact of the internet.

[2] Askitas and Zimmermann (2009) mention that search engines are also used as a directory where people search for familiar websites as opposed to typing in the URL address that is already known.

data on searches can be used as an accurate measure of these phenomena and perhaps even be used in forecasting. Since individuals search for information in order to make future decisions, these searches may reveal their preferences before an economic action occurs, such as the purchase of non-durable goods in which there is an evident period between acquiring information and making the purchase. Thus, it appears plausible that aggregate search data could potentially be exploited for economic analyses.

The formative study by Varian and Choi (2009) initially sparked interest in relating internet search data to economic data.[3] Subsequent analyses broadened the scope and began to probe the robustness of internet search-derived forecasting. Studies have examined the relationship between internet searches and a myriad of socio-economic phenomena. Examples include: consumer confidence (Della Penna and Huang 2009; Goel et al. 2010; Vosen and Schmidt 2011), detecting and tracking diseases (Brownstein et al. 2009; Ginsberg et al. 2009; Polgreen et al. 2008), real estate (Chauvet et al. 2013; Wu and Brynjolfsson 2014), equity markets (Bank et al. 2011; Smith 2012; Vlastakis and Markellos 2012), and labour market forecasting (Choi and Varian 2009b; D'amuri and Marcucci 2010; D'Amuri 2009; Ettredge et al. 2005; Fondeur and Karamé 2013; McLaren and Shanbhogue 2011; Suhoy 2009).

Because employment is a major economic indicator and a focal gauge for the overall well-being of the economy, it appears to be a vital variable of study. For instance, it is used as a coincident index in economic analyses by the Federal Reserve Bank of Philadelphia.[4] Thus, useful economic information may perhaps be garnered from individuals searching for employment information online. An exciting notion is the potential practicality of

[3] Hal Varian has been Chief Economist at Google Inc. since 2007.
[4] See http://www.philadelphiafed.org/research-and-data/regional-economy/indexes/coincident/

internet searches on economic 'nowcasting' or possibly forecasting.[5] In their own work on internet search data and the unemployment rate, D'Amuri and Marcucci (2010) state that "it is easy to guess that the use of internet-based data will become widespread in economic research in the near future" (p.20). This study will evaluate this claim by assessing the association of Google searches with the unemployment rates published by Statistics Canada. Previous analyses have been conducted in various countries but none as of yet in Canada.

Hence, it is proposed that Google Trends search indices may be of benefit for the purposes of forecasting various Canadian unemployment rates. To assess this proposition, the error correction model (ECM) is employed to statistically analyze the short- and long-run relationships between unemployment rates and Google internet searches. A comparative analysis will be conducted to determine which model/data pair produces the best unemployment rate prediction model.

This paper is organized as follows: Section II gives a theoretical background on the relationship between the internet and job searching, then there is a review of the pertinent literature in Section III. Section IV follows with a description of the data, and then there is an overview of the methodology in Section V. Finally, Section VI contains a summary of the results, and Section VII presents the conclusion.

II. Theoretical Background

The premise of this study is that individuals are increasingly using the internet to acquire information. Due to the incredible vastness of material available on the internet, the primary tool that is employed to get any specific knowledge is a search engine. Thus, if

[5] Choi and Varian (2009a) introduce 'nowcasting' to describe producing data contemporaneously with internet searches before official data are released.

internet usage is indicative of actual economic activity, then practical information can possibly be extracted from aggregated search data. The main objective is to determine the validity of the association between aggregate internet searches with socio-economic indicators, specifically the unemployment rate. More broadly, it is to determine whether peoples' individual demands for information (via internet searches) when aggregated have some meaningful association with socio-economic phenomena, and in addition, whether that search information has predictive qualities. Specifically, the purpose is to investigate the efficacy of internet searches as predictors of unemployment rates and consequently of economic behaviour.

Because internet search data (specifically *Google Trends*) are available freely and continually, they may provide information about the labour market and economy at a greater frequency than is the case for traditional government or survey sources. Also, traditional sources often release their data with a lag of at least a couple of weeks and are subjected to revisions, whereas internet search data can be obtained basically contemporaneously (Fondeur and Karamé 2013). Also, *Google Trends*, Google's online search data application, consists of a direct and automatic data entry routine, and thus should be less prone to errors associated with human data entry survey methods. The next step is to decide where attention should be placed in terms of internet search topics. Presumably, people use the internet to search for a myriad of topics, some of which are more economically important than others.

The labour market appears to be a suitable focus for tackling this question, as it is an essential aspect and foremost gauge of the economy. Also, it is a fundamental influence in most people's economic life, and use of the internet for employment information has

become noteworthy. The United Kingdom's June 2009 Labour Force Survey (LFS) reveals that four in five job seekers used the internet in their job search process (Green et al. 2011). A British study of employment and internet use states that, "unemployed people tended to use more job-search methods than either the economically inactive or the employed, suggesting that job-search intensity is greater amongst the unemployed" (Green et al. 2011, p.2). The authors also go on to write that multivariate analysis of this LFS data shows no substantial gap in internet job searching between men and women, but that age has a significant negative relationship with respect to online job searches. Thus, internet job searches are skewed towards younger individuals.

Accordingly, it appears that investigating the relationship between internet job searches and the unemployment rate seems warranted through the channel of inflows into the labour market. Internet searching is becoming pervasive, and as Askitas and Zimmermann (2009) state in their examination of these indicators, the importance of the unemployment rate is paramount in recent times: "given the severe economic crisis and the sudden strong decline in economic activity, the unemployment variable is currently of particular interest to the general public and for scientists" (p. 8). Thus, the advantages of internet search data being easily accessible at a high frequency shows promise as a suitable instrument for modeling and forecasting unemployment rates.

III. Literature Review

Numerous studies have been conducted on connecting a multitude of socio-economic variables and internet searches. Here, the literature focus is narrowed to studies involving the unemployment rate. For instance, Askitas and Zimmermann (2009) use

internet searches from Google as well as German unemployment data to produce a new predictor for the unemployment rate. The authors justify using *Google Trends* by the fact that data are available on a regular and repeated basis.[6] The authors assert that "expecting that search engine keyword searching contains information which correlates with people's lives is a natural and, we believe, commonly accepted expectation" (2009, p. 7).

Their analysis uses the monthly unemployment rate while combining the weekly search indices into two-week periods.[7] The German unemployment rate is announced at the end of the month, so for the weekly Google search indices, Askitas and Zimmermann combine the last two weeks of the preceding relevant month into one time interval (denoted W34M-1), and likewise they use the first two weeks of the particular month for the other interval (denoted W12M). They model both intervals to determine which has superior predictive power. For the search index variables they use four different query terms: unemployment office or agency, unemployment rate, Personnel Consultant, and various names for German employment websites (e.g. Monster, etc.).[8] The authors' goal is to examine the degree to which each of their constructed Google search indices can predict the monthly unemployment rate.

Askitas and Zimmermann use a 'time-series causality approach' by applying the error correction model (ECM) for forecasting. Their endogenous variable is the seasonally-unadjusted monthly German unemployment rate, and their study period runs from January 2004 to April 2009. Because they use a lag-length of 12, they do not directly adjust for seasonality in their data. They estimate a number of models, which differ by the weekly-

[6] Google Trends was originally named Google Insights. A detailed description of Google Trends is given below.
[7] As this technique was used for this study, a description of these indices can be found in the data section.
[8] These are the English translations provided by Askitas and Zimmermann (2009). The original German terms can be found in their paper.

average search variables, to determine if a legitimate predictor can be found. After estimating their models, they evaluate various information criteria (R-squared, log-likelihood values, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC)), where the BIC was the litmus test for a model's capability. The BIC was used to compare models based on their fit with the data, in which smaller values of the statistic are desirable. However, the statistic penalizes overfit models with unnecessary explanatory variables. Askitas and Zimmermann summarize: "The correct choice of model has to be seen in the context of parsimony, prediction success, usefulness, and sound economic basis" (2009, p. 9).

Their results reveal that the coefficients for the lagged-level unemployment rates are negative and statistically significant, suggesting that there are stable long-run solutions to the models. Additionally, the conclusions from the BIC assessments indicate that the earlier Google search intervals (i.e. W34M-1) are statistically superior. Thus, using an earlier time period ensures that the search index will be increasingly effective at predicting the unemployment rate, since the data will be accessible beforehand.

Likewise, Choi and Varian (2009) conduct a study to follow up their initial Google Trends research in which they use Google searches to try to predict U.S. Initial Jobless Claims. The reasoning is that the number of Initial Claims filed is recognized as a leading indicator of the labour market in the United States. To predict Initial Jobless Claims, Choi and Varian use the Google Trends search categories 'Jobs' and 'Welfare & Unemployment'.[9] They then implemented an ARIMA model and selected an AR(1) specification as their baseline. Subsequently, they supplemented the baseline model with the Google search

[9] The 'Welfare & Unemployment' category is no longer available on Google Trends.

category index to test whether the search information improved forecasting.

The authors also point out that a structural break is apparent in the Initial Claims after 2008, a consequence of the major global recession. Thus, Choi and Varian assert that it may be astute to use short-run models (before and after 2008) for their predictions. However, they found that their model produced a better fit using the longer time series. Their conclusion is that their baseline model is improved with the inclusion of the Google Trends data as measured by an appreciable decrease (15.74%) in the out-of-sample estimated mean-absolute error (MAE).

By using a different estimation approach, D'Amuri and Marcucci (2010) likewise attempt to predict U.S. unemployment rates by using Google search and Initial Jobless Claims data. They further try to establish that their Google Index constructed from job-search related queries is the best leading indicator to predict the U.S. unemployment rate by comparing it to the forecast quality of the Initial Jobless Claims variable.

The authors use seasonally-adjusted data for both the endogenous unemployment rate as well as the exogenous Initial Claims and Google search variables where the latter was adjusted manually. They elect to use the single Google search term 'jobs' for two reasons: First, they wanted to use the most popular search term because Google Trends releases their data as relative indices (representing the relative popularity of a search term in terms of all searches done at the time period and geographic area in question), and secondly, they believe that this term covers a large cross-section of job seekers. D'Amuri and Marcucci conjectured that narrower search terms would characterize only a subset of internet-using employment seekers and thus introduce bias into their forecasts.[10] Weekly

[10] The present study ascertains that 'jobs' is too broad of a term to produce a high resolution forecast.

time-series data for Initial Claims and Google searches were combined in order to search for an association with the monthly unemployment rate. For each month, Google search index variables were constructed so as to include the week that contains the 12th day, and then the three preceding weeks were added in order to produce a monthly index.

To assess the validity of their Google search-based index as a predictor of the U.S. unemployment rate, D'Amuri and Marcucci use a 'long horizon out-of-sample comparison' of more than five hundred forecasting models. Three different groups of models were estimated which differed in terms of their exogenous variables and time frames. Models excluding the Google index consisted of a large sample starting in 1967, while those including Google data required a shorter sample that begins in 2004.[11] Specifically, the researchers use ARMA and ARMAX models supplemented with Initial Claims and likewise for the Google index for comparison. Both linear and non-linear models were used to properly account for short-term movements and long-term cyclicality (i.e. the business cycle). This approach is in contrast to the one adopted in this paper and in Askitas and Zimmermann (2009), both of which resolve the short- and long-run dynamics by utilizing the error correction model. Modeling both short-run and long-run dynamics is difficult in a single model specification. However, the ECM is advantageous because both the unemployment rates and internet searches display a long-run cointegrating relationship (explained below) which is incorporated in the model, while short-term coefficients for the lags of both the unemployment rates and internet searches are also included. Thus, the ECM is an efficient way to integrate both the short-run and long-run relationships.

[11] This is the first availability of Google Trends and most studies begin at this point. However, D'Amuri and Marcucci (2010) produce an Initial Claims estimator from data dating from 1967 and compare it to the Google predictor estimated from 2004.

Their results show that the models augmented with the Google index 'jobs' variable vastly outperformed other models that included traditional indicators (e.g., Initial Claims), as measured by a lower mean-squared error (MSE). D'Amuri and Marcucci conclude that "for all forecast horizons the best model (i.e. the model with the lowest MSE out-of-sample) always includes the GI [Google Index] as the exogenous variable" (2010, p.10). This was the case over the longer time horizon even though the equation including the Google index is estimated over a much shorter period than the estimate with the more traditional Initial Claims indicator (i.e., 1976 vs. 2004). They conclude that the Google index is the best leading indicator to forecast the U.S. unemployment rate.

To further strengthen their findings, they also estimated augmented models for all 50 U.S. states plus the District of Columbia. Again, models including the Google Index were superior, as measured by a lower MSE, than all other models in 70% of the states. They also found Google Searches to be a more robust predictor than the projections from the Survey of Professional Forecasters produced by the Federal Reserve Bank of Philadelphia. Their results were also found to be robust to various changes in the functional form, i.e. linear and non-linear model transformations (logarithm, logit, first differences and log-linear de-trended data).

Another study conducted by Fondeur and Karamé (2013) uses Google search data to attempt to forecast youth unemployment in France, for which they constrain the unemployment data to 15-24 year olds. They claim that this should reduce the selection bias towards younger individuals inherent in internet job-related search data. For their Google search term, they selected 'emploi' (French for job and/or employment) due to its simplicity and breadth.

Their study period runs from January 2004 to July 2011, in which seasonally unadjusted data are used and due to an apparent structural break in their data in 2009, a Fourier decomposition is executed to model seasonality. Furthermore, the authors also acknowledge a limitation due to the dissonant frequencies of the weekly Google and monthly unemployment data. Simply choosing a one- or two-week interval or averaging weeks over a particular month creates, what Fondeur and Karamé term, an 'impoverished' dataset. To overcome this they engage a 'robust non-parametric' strategy. Due to the unobserved components in their time-series and because employment-related search data could potentially include non-labour market-related information, their Google index will encompass a fair amount of noise. Given these challenges, Fondeur and Karamé utilize an "unobserved-variables decomposition-based model" (2013, p.119). Additionally, this model is estimated with a Kalman filter and is estimated by maximum likelihood to determine the unknown parameters. The model they construct appears to be excessively complex, and consequently, simple models such as the error correction model should be an improvement in its ease and efficiency.

Their forecast method was to estimate a bivariate model which includes the Google data and to compare it to a baseline counterpart (i.e., an AR model with only lagged endogenous variables). Estimation of the model that includes the Google data reveals "that the assumed relation between the Google index and the DEFM [unemployment] in the bivariate representation is quite natural and pertinent" (2013, p. 122). They produce three distinctive unemployment rate forecasts driven by the Google search index that differ in their timing (2 week lead, 1 week lead, and contemporaneous).

Similar to the other studies conducted, Fondeur and Karamé found that the inclusion of Google search data in the equation comparatively improves their models' predictive power. They determined that including the Google index in their forecast models improved both precision and accuracy. They also did separate forecasts for the 25-49 and 50 and over age groups and found that the inclusion of the Google index improved RMSE (root-mean-square error) predictions by 17.5% and 9.7% respectively. The authors conclude, "this observation most likely illustrates the selection bias noted by D'Amuri (2009) in favor of the young regarding the Internet as a job search tool" (2013, p. 124).

Thus, numerous earlier studies show that forecasting with the inclusion of Google search data is a worthwhile endeavour. Wu and Brynjolfsson (2014) proclaim:

> We believe that we are only at the beginning of the data revolution. Newer and more fine-grained data are becoming available every day from various search, social media and micro-blogging platforms. These data are made available instantaneously, allowing consumers, business managers, researchers, and policy makers to tap into the pulse of economic activities as they are happening (p. 9).

IV. Data

To determine whether aggregate internet search variables have information and forecast potential with respect to socio-economic indicators, the unemployment rate will be the variable to be explained. Canadian national data are used and collected from the CANSIM Multidimensional @ CHASS database. These rates are derived from Statistics Canada's Labour Force survey (LFS), and are published monthly. Statistics Canada typically performs the survey on the third week of the month and then releases the results at the

beginning of the following month.[12] This study will attempt to forecast several measures of the unemployment rate. The variants are the official rate for individuals 15 years of age and older (designated UR4), the official rate for youths aged 15 to 24 (URyo), and Statistics Canada's more broadly defined unemployment rate (UR8) that supplements the official rate with the inclusion of discouraged job searchers and involuntary part-time workers.[13] All series are composed of both males and females and are seasonally unadjusted.

The explanatory variables are derived from *Google Trends*, Google's freely accessible search data tool. Google is the foremost internet search engine used in Canada with a 71.56% market share, while Microsoft's Bing is in second place at 8.89% as of May 2015, according to Experian Marketing Services.[14] The data are released weekly, commence in January 2004, and are available almost immediately, unlike data from other traditional sources (e.g. government agencies). D'Amuri and Marcucci (2010) affirm that, "notwithstanding its limited time availability … we believe that the GI [Google Index] should routinely be included in time series models to predict unemployment dynamics" (p. 20).

Google Trends publishes data as a weekly index that represents the relative popularity of search terms.[15] Individual search terms are computed as a proportion of the total searches done in a particular geographical area for a particular week. The index is computed as: (Number of searches for term)/(Total searches conducted). This proportion is then normalized on a scale of 1 to 100, with 100 corresponding to the most popular search term during the relevant time period and geographical area. Additionally, particular

[12] More information on the LFS is available at http://www.statcan.gc.ca/eng/survey/household/3701.
[13] The three series are V2440389, V2440413, and V2064894 and were obtained on May 26, 2015.
[14] https://www.experian.com/marketing-services/online-trends-canada.html.
[15] Google Trends data information is available at: https://support.google.com/trends.

search terms are aggregated into categories and these are given as a negative or positive percentage representing their relative popularity. These were subsequently converted to decimal numbers for this study, e.g. 21.4% was changed to 0.214. This corrects for the upward trend in searches over time from the evident increased internet use and allows for robust analysis (McLaren and Shanbhogue 2011). Moreover, it also allows for direct comparison between areas with different populations as data are available along a geographical dimension, including national and sub-national entities (e.g. for Canada and the U.S.). In their paper, *Using Internet Search Data as Economic Indicators*, McLaren and Shanbhogue state that "the ability to track the popularity of such a wide range of search terms makes this the most suitable data source for this type of study [economic analysis]" (2011, p. 135).

However, D'Amuri and Marcucci (2010) allude to several limitations in using searches to acquire information about the labour market. For instance, selection bias is apparent as individuals that use the internet for job-searching activities are not likely to be randomly selected. Also, a constructed search index will contain more than merely information about unemployed job-seekers, but will also comprise information about all searches related to employment. However, the benefits of internet search data availability and frequency appear to outweigh any detrimental effects that may occur from selection bias and noise in this and many previous studies.

While some studies use individual search terms (Askitas and Zimmermann 2009, D'Amuri and Marcucci 2009), others use aggregated search categories (Choi and Varian 2009, Fondeur and Karamé 2013). In this paper, the *Google Trends* category 'Job Listings' (denoted 'JobList'), which is a sub-category of the larger category 'Jobs', is used in an

attempt to forecast the unemployment rate. The reasoning is that many different search terms can be used for job searches. Firstly, Canada is not a unilingual country and searches in other languages (most notably French) will complicate the observation of single search terms. Also, a simple search could encompass an unnecessary amount of noise. For example, the search term 'jobs' could potentially be convoluted by the popularity of Steve Jobs and his death (2011), and biographical film (2015), or any number of spurious associations. Moreover, it seems completely reasonable that people doing searches that fall into the 'Job Listing' sub-category would represent the bulk of people who are actively searching for work. Thus, it should be possible to find a relationship with respect to the unemployment rate. In addition to JobList, to broaden the study, the individual search terms 'employment insurance' (EI) and 'employment' (EM) are also used. Employment insurance was selected as it is conjectured that people that have lost their jobs will search for information about their eligibility for EI and how they can attain it. Likewise, 'employment' appears to be a suitable simple search term to round out the analysis.

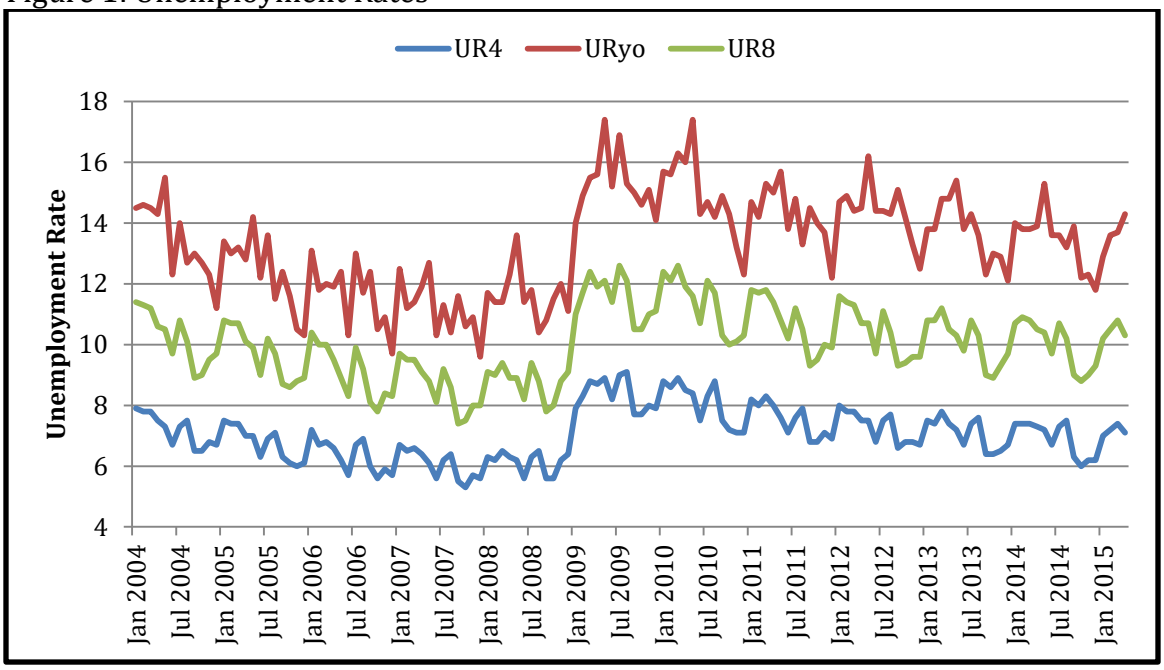
Due to the different frequencies of the unemployment rates and Google search indices, it was required to take four- and two-week averages of the weekly searches for each month. The study period is from January 2004 (the commencement of Google Trends) to April 2015 (latest available period). None of the data are seasonally adjusted because the frequency of the time-series variables is monthly and up to 13 lags are used in the error correction model.

The explanatory variables are labeled generally as *Searchterm_xxxx* where *xxxx* indicates the weeks that have been averaged for each month. For each month t that the unemployment rate is released there is a period of weeks in which the search terms are

averaged. The selections are: *1234* for the four weeks of month *t*; *4123* for the 4th week of month *t-1* and the subsequent three weeks of month *t*; *3412* for the 3rd and 4th week of month *t-1* and the first two weeks of month *t*; *12* for the first two weeks of month t; and finally *34* for the 3rd and 4th week of month *t-1*.

The unemployment rates are plotted in Figure 1, while a subset of Google search indices (ones averaged over weeks *3412*, as these should be representative of the other average intervals) are in Figure 2.

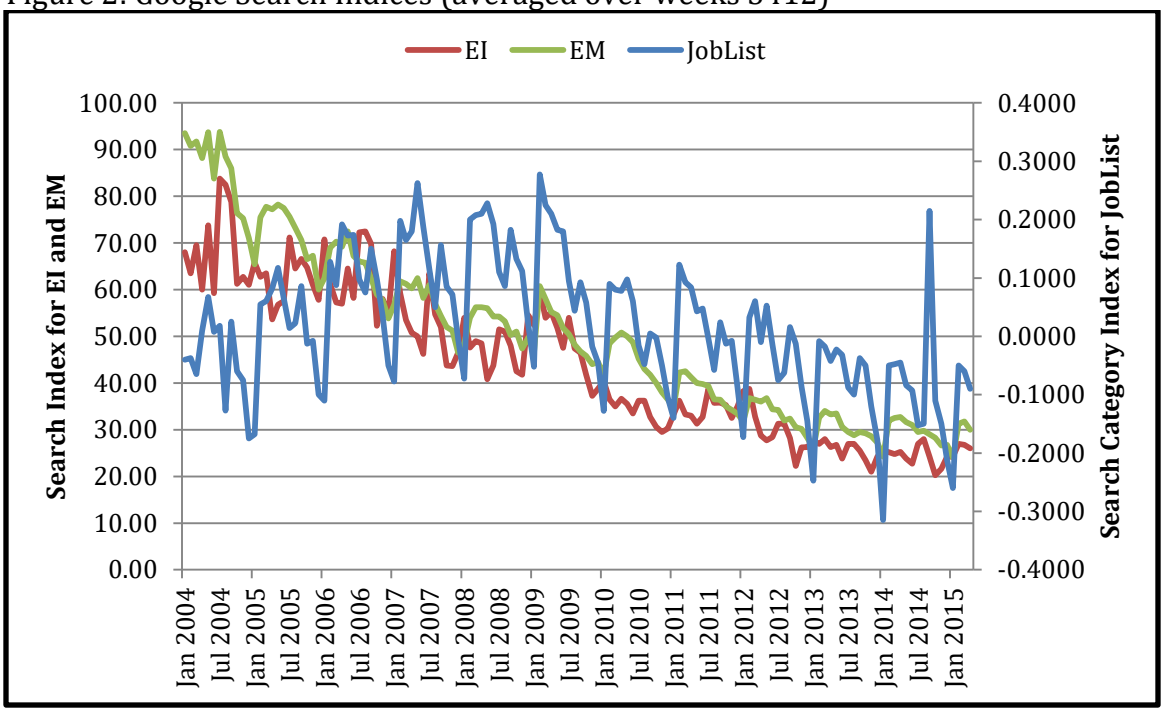
Figure 1: Unemployment Rates



An examination of the data clearly shows that all the time-series are highly cyclical and demonstrate a high degree of seasonality. It is also quite apparent that they are all non-stationary and may contain trend and drift components. Moreover, there appears to be a structural break for the unemployment rate that shows two distinct periods before and after January 2009. Similar to previous studies, this structural break signifies a major event

in the labour market. This appears to be the manifestation of the major recession in Canada brought on by the global financial crisis, as Canada was officially in a recession from the fourth quarter of 2008 to the second quarter of 2009 (Trichur 2009).

Figure 2: Google Search Indices (averaged over weeks 3412)



The search indices show similar seasonal variability but not the same long-term trend and structural break. The aggregated search category JobList follows the unemployment rate in the long-run, but the search terms EI and EM are decreasing in relative popularity over the study period. However, there appears to be a shorter-run association which will be important for prediction. There also appears to be an outlier data point for JobList, but because time-series models require that there be no gaps in the data, it was not removed.

V. Methodology

Following Askitas and Zimmermann (2009), the error correction model (ECM) is the primary specification used to determine the relationship between several measures of the unemployment rate and various Google search indices. This method is used to model cointegrated relationships among variables that individually have unit roots, and it can also be used to test for forecasting potential. The principal conceptual foundations of the ECM are non-stationarity, order of integration, and cointegration. The following description of my methodology is based on Greene (2008), the *Stata Time-series Reference Manual 13* (StataCorp 2013), and online tutorials by Ben Lambert (2013).[16]

The ECM can be particularly useful under particular circumstances. If two seemingly related time-series are non-stationary, then a simple practical exercise is to take their first differences, and if that transformation results in stationarity, then ordinary least squares can be used. If such a relationship exists, however, it will only be characterized by a short-run association. Yet, if the two variables are cointegrated, then a more complete relationship can be discerned, and both the short- and long-run relationships can be determined through implementation of the ECM (Lambert 2013).

A time-series is said to be integrated of order d (denoted I(d)) if it is stationary after being differenced d times (Best 2008). Moreover, two series are cointegrated if they are both integrated of degree one, but a linear combination of both time-series is stationary, (i.e., I(0)). Cointegration implies that the two series have a long-run equilibrium relationship. The ECM incorporates the first-difference OLS (short-term) with the cointegrated relationship (long-term) (Lambert 2013).

---

[16] Lambert is an Oxford University Ph.D candidate who offers free econometrics tutorials online.

To determine if the ECM can be implemented, first it must be shown that the data are I(1) (Greene 2008). The Augmented Dickey-Fuller (ADF) test for unit roots was performed to determine if the null hypothesis of a unit root can be rejected.[17] This test requires that the number of autoregressive lags be specified correctly. Ng and Perron (1995) suggest starting with a large number of lags, testing the significance of the coefficient of the highest lag, and then reducing by one lag until there is statistical significance. If non-stationarity is not rejected in variable levels ADF tests are done on the first differences to determine if variables are stationary and thus I(1).

Once the variables are shown to be I(1), Johansen's test of cointegration is used to determine if two series have a long-run association. The test statistic is derived from a log likelihood test in order to determine the number (if any) of cointegrating equations that are present in the model (Greene 2008). However, this test is particularly sensitive to the specification of the lag-length. The proper lag-order can be determined by computing the underlying vector autoregressive model (VAR) with differing numbers of lags up to a particular maximum. For each lag p, a likelihood ratio (LR) test is performed that compares it to the previous lag (i.e., p–1), where the null hypothesis is that all the coefficients of lag p of the VAR are zero. Starting from the maximum lag, then progressively decreasing, the first lag for which the null hypothesis is rejected is the appropriate lag-order. In addition, for each VAR computed the final prediction error (FPE), Akaike's information criterion (AIC), Schwarz's Bayesian information criterion (SBIC), and the Hannan and Quinn information criterion (HQIC) are reported (StataCorp 2013, p. 701). Together, these statistical values assess the optimal number of lags necessary for Johansen's cointegration test.

[17] D'Amuri and Marcucci (2009) use the ADF with Generalized Least Squares de-trending because it has a high power with small samples. This is less important here due to the longer sample time-frame.

The approach here is to pair each version of the unemployment rate with every search index averaged over a particular interval. There are three endogenous unemployment rates coupled with three search indices, each having five different weekly-average schemes. Then, each individual unemployment rate-SeachTerm_xxxx pair has five different trends (explained below). There are 225 specifications that are tested for cointegration.

The trace statistic method of Johansen's test is used to determine if there is a cointegrating relationship between the unemployment rates and Google search indices. The test is based off of Johansen's maximum likelihood (ML) estimator where the null hypothesis of the trace statistic is that there is no cointegrating equation (see Statacorp 2013, p. 770 for the derivation of the trace statistic). Only one cointegrating relationship is tested since there are only two variables, however in general the test can be for multiple cointegrating equations. If the null hypothesis is rejected, then the ECM cannot be used and the alternative is to run a simple autoregressive model to establish a statistical relationship; however for the purpose of this study, these models are simply discarded since a long-run association is being sought.

After determining if a cointegrating equation exists, then determining the lag-length, and after choosing a trend specification, the ECM can be correctly specified for estimation. It should be noted that the maximum lag in the ECM is one less than that established from the VAR process done above due to the fact that first-differences are used in the ECM.

The general error correction model can be expressed as:

$$\Delta Y_t = \gamma + \Sigma_{i=1}^{p} \eth_i \Delta X_{t-i} + \Sigma_{j=1}^{p} \mu_j \Delta Y_{t-j} - \lambda(Y_{t-1} - \alpha - \rho t - \beta \Delta X_{t-1}) + \tau t + \varepsilon_t \quad (1)$$

where the terms at time *t* are:

Y is the dependent variable
X is the explanatory variable
ð and μ are the short-run coefficients for the lagged explanatory and dependent variables respectively
p is the number of lags
γ is the constant intercept term
λ is the adjustment coefficient or error-correction term
β is the parameter of the cointegrating equation[18]
α, τ, ρ are trend parameters
ε is the error term

Referring to the equation above, the main components of the ECM are the short-run coefficients, the bracketed cointegrating equation with the adjustment error-correction coefficient, and the trend parameters.

The ECM can contain five different trend specifications which are based on the following assumptions (StataCorp 2013, p. 745):

i) An unrestricted trend (denoted trend) does not constrain any of the trend parameters and indicates that the variable levels have quadratic trends and that the cointegration equation is trend stationary.
ii) A restricted trend (rtrend) sets $\tau = 0$ and indicates that the variable level trends are no longer quadratic but are linear and the cointegrating equation remains trend stationary.
iii) An unrestricted constant (constant) sets $\tau = 0$ and $\rho = 0$, still allows linear trends in the variables, but the cointegrating equation is now only stationary around a constant mean.
iv) A restricted constant (rconstant) sets $\tau = 0$, $\rho = 0$, and $\gamma = 0$ which implies that there are no linear trends in the data; however, the cointegrating equation is still stationary around a constant mean.

[18] There is another β parameter for Y_{t-1} in the cointegration expression, but if the cointegration is exactly specified, as is the case here, the coefficient is simply unity so it is omitted.

v) No trend (none) sets $\tau = 0$, $\rho = 0$, $\gamma = 0$, and $\alpha = 0$; this is the most restrictive specification and assumes that the variables and cointegrating equation are stationary around a mean of zero.

The summation expressions represent the short-run component of the model, whereas the term in the parentheses is the long-run or cointegrating characteristic. The adjustment coefficient (λ) indicates the error-correction and should be negative and between 0 and -1 if there is a long-run equilibrium between the unemployment rate and internet searches. It measures the rate at which a deviation in the preceding period is corrected towards the long-run equilibrium (Best 2008).

For the purposes of this study, it is assumed that the unemployment rate and Google search indices are the dependent variable and explanatory variable, respectively. If the value of Google searches are above (below) its long-run equilibrium with respect to the unemployment rate for time $t-1$, then the cointegrating equation is positive (negative), and the adjustment parameter (if the term containing λ is correctly specified) will incrementally adjust the cointegrating equation towards its long-run equilibrium at time t. The magnitude of λ determines the speed at which the correction occurs. Likewise, the cointegrating parameter (β) should be positive, as an increase in employment searches should be associated with a higher unemployment rate. In this model, the change in the unemployment rate is determined by the short-run past values of the unemployment rate and Google searches, as well as the long-run cointegrating relationship.

To validate the predictive qualities of the Google search indices, the conventional practice of comparing the model in question to the simple autoregressive model containing only lags of the dependent variable is done. This is to establish if the explanatory variables

contribute more to forecasting than simply a model that relies on past values of itself (Clements and Hendry 2005). This will be done by a simple RSME comparison.

After the ECM has been estimated, all specifications that have a negative and statistically significant estimate for the adjustment parameter are retained since these models exhibit a long-run statistical relationship in-line with the assumed exogeneity of the Google search variables (Alogoskoufis and Smith 1990). The superior models (judged by a low RMSE) are then subjected to an in-sample forecasting procedure. Afterwards, an out-of-sample forecast is performed with approximately three-quarters of the sample being used for estimation with the last quarter being forecasted and compared to the actual data. Whereas an in-sample examination is commonly used to fit a model, an out-of-sample investigation measures the predictive power of a model (Duy and Thoma 1998). As Carruth et al. (1998) judge, "in a study of US unemployment: If a dynamic modeling approach is to be convincing, it needs to say something about the behavior of unemployment out of sample" (p. 626). An out-of-sample examination is executed due to the fact that models can have an exceedingly good in-sample fit while not necessarily comprising a suitable out-of-sample prediction quality.

VI. Results

The results of the ADF unit root tests with trend specifications and lag-lengths are presented in Tables 4 and 5 in Appendix A. The proper trend specification was determined by inspection of the time-series plots. Figure 1 shows that all three of the unemployment rates display a downward trend over the study period, except for the interval of the structural break, i.e. the major recession of early 2009. Meanwhile, Figure 2 shows that

both the Google search indices EI and EM contain a downward trend, however, JobList does not appear to contain a trend over the sample period. Taking the first differences of all the unemployment rates and search indices variables eliminates any trends, and accordingly the second set of ADF tests were carried out with constant trend specifications. Again, the method of Ng and Perron (1995) discussed above is used to determine the appropriate number of lags for each ADF test. The results in Tables 4 and 5 show that all of the unemployment rates and all the Google search indices are level non-stationary and first-difference stationary, and are thus I(1).

Next, a list of 225 models were then compiled in which the selections varied by the combination of unemployment rate, weekly-averaged search indices, and cointegration trend specification. Then, the appropriate lag-length was found to be 13 for all models, and using this information, Johansen's trace test for cointegration was performed on all model combinations. Models that have a positive result for cointegration (and thus contain a cointegrating rank of one) with a level of significance of 5% were estimated using the ECM, and the results are presented in Tables 6 to 13 of appendix A. It is interesting to note that URyo and EI were found to be not cointegrated for any trend specification. This seems reasonable since young people are probably less likely to apply for employment insurance if not employed and they may not have worked enough to entitle them to EI. A total of 105 models were cointegrated and thus selected for ECM estimation.

The implementation of the ECM produces estimates for the adjustment parameters (λ), the cointegrating coefficients (β), and also the RMSE, AIC, and SBIC. Additionally, the test statistics from the test of joint significance of the short-run coefficients are presented in Table 1. Because the estimates of the ECM are in levels and not logs, the coefficients of an

ECM regression are not directly interpretable (Clements and Hendry 2005). The litmus test is a statistically significant negative estimate for λ, and this infers a long-run causal relationship running from internet searches to the unemployment rate. The RMSE, AIC, and SBIC were then used to compare the fit and forecasting capability of models with the same endogenous variable.

The elimination procedure of the models is continues wherein the best 32 models (presented in Table 1) are subjected to diagnostic tests to ensure their robustness. A Lagrange multiplier test of the residuals of the ECMs for autocorrelation was performed as its presence is possibly problematic for forecasting efficiency. Another important diagnostic issue is to investigate whether the disturbances are normally distributed and the Jarque-Bera test was done to this effect. The likelihood method that is used for estimating the ECM assumes that the errors are independently, identically, and normally distributed with zero mean and finite variance (StataCorp 2013, p. 766). If the disturbances are not normally distributed, then the parameter estimates are consistent but not efficient in small samples.

Table 1: ECM results with long-run equilibriums (negative estimated λ)

Endo-genous Variable	Exogenous Variable	Trend specification	RMSE	AIC	SBIC	Lags w/ Auto-correlation	Jarque-Bera (p-value)*
UR4	**JobList_12**	**constant**	**.2203**	**-3.224**	**-2.012**	**1, 8, 12**	**0.339****
UR4	EI_4123	trend	.2344	5.523	6.781	***	***
UR4	EI_3412	trend	.2303	5.461	6.718	***	***
UR4	**EI_12**	**rconstant**	**.2319**	**6.128**	**7.317**	**1, 12**	**0.0446**
UR4	EI_12	trend	.2324	6.052	7.310	***	***
UR4	**EM_1234**	**none**	**.2221**	**4.321**	**5.487**	**3, 5, 12**	**0.989****
UR4	EM_4123	none	.2305	4.221	5.388	3, 12	0.017
UR4	EM_3412	none	.2171	3.897	5.063	12	0.012
UR4	EM_12	rconstant	.2293	4.794	5.983	3, 12	0.625**
UR4	EM_12	none	.2241	4.805	5.971	1, 3, 12	0.857**
UR4	**EM_34**	**none**	**.2098**	**4.576**	**5.742**	**1, 3, 5, 12**	**0.906****
URyo	JobList_1234	constant	.6341	-1.030	-0.182	13	0.000

URyo	JobList_1234	rconstant	.6313	-1.044	0.145	13	0.000
URyo	JobList_4123	constant	.6298	-1.558	-0.346	13	0.147**
URyo	JobList_3412	constant	.6135	-1.105	0.1063	13	0.000
URyo	**JobList_3412**	**rconstant**	**.6108**	**-1.120**	**0.0689**	**13**	**0.000**
URyo	JobList_12	constant	.6936	-1.074	0.1379	***	***
URyo	JobList_12	rconstant	.6908	-1.088	0.1013	8, 12	0.138**
URyo	JobList_34	rconstant	.6158	-0.270	0.9184	8, 12	0.328**
URyo	**EM_1234**	**none**	**.6457**	**6.588**	**7.754**	**1, 4**	**0.359****
URyo	EM_4123	none	.6846	6.421	7.587	4, 12	0.002
URyo	EM_3412	none	.6759	6.339	7.505	None	0.000
URyo	EM_12	none	.6946	7.027	8.193	12	0.385**
UR8	**JobList_1234**	**rtrend**	**.2719**	**-3.000**	**-1.766**	**12**	**0.000**
UR8	EI_4123	trend	.2880	5.962	7.220	***	***
UR8	EI_3412	trend	.2817	5.887	7.144	***	***
UR8	**EI_12**	**rconstant**	**.2854**	**6.541**	**7.730**	**12**	**0.006**
UR8	EI_12	trend	.2865	6.480	7.737	***	***
UR8	**EM_1234**	**none**	**.2752**	**4.720**	**5.886**	**3, 5, 12**	**0.871****
UR8	EM_3412	none	.2819	4.428	5.594	12	0.004
UR8	EM_12	none	.2896	5.256	6.423	1, 7, 12	0.808**
UR8	**EM_34**	**none**	**.2611**	**5.028**	**6.194**	**3, 5, 12**	**0.623****

* This is the probability of rejecting the null hypothesis that the residuals are normally distributed.
** Indicated normally distributed residuals at 5% level of confidence.
*** Indicates that there was an error computing temporary VAR estimates in STATA and thus diagnostic tests on the residuals could not be carried out. No solution was found.
Boldface indicates retained models

Afterwards, the ten superior models (presented in Table 2) were selected for forecasting. However, to ensure that Google searches do in fact help predict the unemployment rate, the implied causality of the model must be determined.

Table 2: "Best" ECM results with diagnostic tests (subset of Table 1)

Endo-genous Variable	Exogenous Variable	Trend specification	RMSE	AIC	SBIC	Lags with Auto-correlation	Jarque-Berra (p-value)*
UR4	**JobList_12**	**constant**	**.2203**	**-3.224**	**-2.012**	**1, 8, 12**	**0.339****
UR4	EI_12	rconstant	.2319	6.128	7.317	1, 12	0.0446
UR4	EM_1234	none	.2221	4.321	5.487	3, 5, 12	0.989**
UR4	EM_34	none	.2098	4.576	5.742	1, 3, 5, 12	0.906**
URyo	**JobList_3412**	**rconstant**	**.6108**	**-1.120**	**0.0689**	**13**	**0.000**
URyo	EM_1234	none	.6457	6.588	7.754	1, 4	0.359**
UR8	**JobList_1234**	**rtrend**	**.2719**	**-3.000**	**-1.766**	**12**	**0.000**
UR8	EI_12	rconstant	.2854	6.541	7.730	12	0.006
UR8	EM_1234	none	.2752	4.720	5.886	3, 5, 12	0.871**
UR8	EM_34	none	.2611	5.028	6.194	3, 5, 12	0.623**

* This is the probability of rejecting the null hypothesis that the residuals are normally distributed
** Indicated normally distributed residuals
Boldface indicates models that were retained for forecasting analysis

To affirm that causality is running in the expected direction of internet searches to the unemployment rates, a causality test is done. However, the conventional Granger causality test is a Wald test that requires that the test statistic have an asymptotic χ^2 distribution. This assumption is not true for non-stationary data, making the Granger causality test invalid (Lin 2008). Nevertheless, Toda and Yamamoto (1995) showed that causality can be tested when time-series variables are characterized by unit-roots. The method requires the estimation of the VAR model where the total number of lags is the sum of the lags determined by the likelihood ratio test (discussed above), denoted by *p*, plus the maximum order of integration of the variables that are to be tested, denoted by *d*. In this case, the VAR was estimated with 13 + 1 = 14 lags. Toda and Yamamoto (1995) showed that the Wald test of the null hypothesis that the coefficients of the first *p* lags of explanatory variables are jointly zero is asymptotically valid. Thus, a rejection of the null hypothesis suggests that the Google search indices 'Granger-causes' the unemployment rates. The results for the estimation of the VARs and the causality tests are presented in Table 3.

Table 3: VAR and Causality Test Results

Endogenous Variable	Exogenous Variable*	RMSE	AIC	SBIC	Causality test (p-value)*
UR4	**JobList_12**	**.2214**	**-3.207**	**-1.973**	**0.001****
UR4	EI_12	.2336	6.088	7.323	0.069
UR4	EM_1234	.2213	4.263	5.498	0.001**
UR4	EM_34	.2074	4.533	5.767	0.000**
URyo	**JobList_3412**	**.6150**	**-1.096**	**.1382**	**0.000****
URyo	EM_1234	.6468	6.540	7.775	0.000**
UR8	**JobList_1234**	**.2749**	**-2.837**	**-1.602**	**0.002****
UR8	EI_12	.2876	6.507	7.741	0.103
UR8	EM_1234	.2741	4.669	5.904	0.001**
UR8	EM_34	.2580	4.991	6.225	0.000**

* This is the p-value that the chi-squared distributed test statistical is larger than the critical value
** Indicates that the exogenous variable 'Granger-causes' the endogenous variable at the 5% significance level
Boldface indicates models that were retained for forecasting analysis

The RMSEs for the VAR models with the exogenous variables are all higher than for the ECMs suggesting that the latter is the superior model specification. The RMSEs for the VAR models with only the lagged values of the endogenous variables are 0.2354 for UR4, 0.7194 for URyo, and 0.2892 for UR8. These values are all larger than their ECM counterparts and this is to be expected from the results of the causality tests.

Finally, the conclusion of the investigation is to evaluate the forecasting proficiency graphically (presented in Appendix B). The aggregate category 'JobList' was the best forecaster for all three unemployment rate variables, although no consensus regarding the optimal average-period could be established. However, the averaged Google searches from an earlier period appear to be superior for model fit. The predictive superiority of 'JobList' lends support to the hypothesis that because the 'Job Listings' *Google Trends* category includes searches done in multiple languages (at least for both English and French), it will encompass more employment related search information than just the English-language search terms 'Employment Insurance' and 'Employment'. The plots of the forecasts of UR4 using JobList_12, URyo using JobList_3412, and UR8 using JobList_1234 are located in Appendix B (Figures 3 to 14).

Four forecasts were generated for each unemployment rate: three in-sample forecasts consisting of the whole sample period, and one of each before and after January 2009. Also, an out-of-sample forecast was generated from June 2012 to April 2014 that was estimated using only data from before May 2012. This interval for the out-of-sample forecast was chosen due to there being no structural breaks (i.e. major recessions).

It is evident that internet searches do not accurately predict the unemployment rate during the major recession that occurred from late 2008 to early 2009. Forecasts diverge

from actual rates during this period in both the long- and short-term forecast graphs. This is clearly indicative of a structural break in the data, and as Clements and Hendry (2005) point out in their paper *Evaluating a Model by Forecast Performance*, "explanations for forecast failure might point to the occurrence of extraneous or atypical events in the forecast period" (p.2).

UR4 and UR8 appear to be more amenable to forecasting than URyo, which disagrees with the assertion that more precise information can be garnered from younger individuals due to their increased likelihood of internet use. It is difficult to determine why the youth unemployment rate was not forecasted as accurately as the other rates. However, the cause may be due to the higher volatility in the URyo series. The coefficients of variation of UR4, URyo and UR8 are 0.122, 0.125, and 0.118, respectively. This suggests that the youth unemployment rate could be more difficult to predict.

Nevertheless, the Google category 'Job Listings' forecasted the official and supplemental unemployment rates well, and especially on smaller time intervals. 'JobList' was also, very encouragingly, able to forecast the unemployment rate out-of-sample over a two-year period. Inevitably, this supports the objective that internet searches can give accurate information on the unemployment rate.

VII. Conclusion

The idea of discovering a data source to obtain accurate and rapidly updated information about the economy is important, and internet activity provides a promising innovative solution. Individuals are likely to express their economic and decision-making information through their internet use. Thus, internet search data ought to be able to

produce valuable information on people's preferences as well as forecast key economic indicators, such as labour market activity. To assess this assertion, Google internet search data was used to predict three versions of the unemployment rate in Canada.

The methodology involved developing a large number of models that were constructed by combining the unemployment rate with a *Google Trends* search index averaged over various 4- and 2-week intervals. Variables tested included the Google search category 'Job Listings' as well as the individual search terms 'employment insurance' and 'employment'. Models found to have a cointegrating, or long-run, association with the official unemployment rate, the unemployment rate of young individuals (15-24 years old), as well as Statistics Canada's supplementary unemployment rate were estimated using the error correction model with various trend specifications.

Through a process of eliminating steps, models with the lowest RMSE that were well behaved with respect to diagnostic tests were retained. Evidently, the Google aggregate search category 'Job Listings' was the search index that provided the best model fit and predictive power. A long sample period was used that included a significant structural break that would be difficult for any model to predict, no matter how sophisticated. Consequently, none of these models were able to correctly forecast the severity of the increase in the unemployment rate in January 2009. However, over a shorter time period, Google Searches do forecast the unemployment rate remarkably well. As the primary objective of this study was to uncover an indicator that is easily accessible to forecast the unemployment rate

precisely at a high frequency, indices created from *Google Trends* internet searches performed commendably.

Thus, indices constructed from Google search activity are evidently capable forecasters of the unemployment rate in Canada. This supposes that internet search activity can provide accurate information of the labour market, and lends credence to the idea that this approach can easily be extended to other areas of society, and can be specifically beneficial to swiftly fluctuating areas of the economy. It seems exceedingly likely that internet activity will become an increasingly common source of information used to forecast economic activity.

References

Alogoskoufis, George, and Ron Smith. 1990. “On error correction models: specification, interpretation, estimation.” *University of London, Birkbeck College*.

Askitas, Nikos and Klaus F. Zimmermann. 2009. "Google Econometrics and Unemployment Forecasting." *German Council for Social and Economic Data (RatSWD) Research Notes* (41).

Bank, Matthias, Martin Larch, and Georg Peter. 2011. "Google Search Volume and its Influence on Liquidity and Returns of German Stocks." *Financial Markets and Portfolio Management* 25 (3): 239-264.

Best, Robin. 2008. "An Introduction to Error Correction Models." Oxford Spring School for Quantitative Methods in Social Research.

Brownstein, John S., Clark C. Freifeld, and Lawrence C. Madoff. 2009. "Digital Disease detection—harnessing the Web for Public Health Surveillance." *New England Journal of Medicine* 360 (21): 2153-2157.

Carruth, Alan A., Mark A. Hooker, and Andrew J. Oswald. 1998. "Unemployment equilibria and input prices: Theory and evidence from the United States." *Review of economics and Statistics* 80, no. 4: 621-628.

Chamberlin, Graeme. 2010. "Googling the Present." *Economic & Labour Market Review* 4 (12): 56.

Chauvet, Marcelle, Stuart A. Gabriel, and Chandler Lutz. 2013. "Fear and Loathing in the Housing Market: Evidence from Search Query Data." *Available at SSRN 2148769*.

Choi, Hyunyoung and Hal Varian. 2009. "Predicting Initial Claims for Unemployment Benefits." *Google Inc*. 1-5.

Clements, Michael P. and David F. Hendry. 2005. "Evaluating a Model by Forecast Performance*." *Oxford Bulletin of Economics and Statistics* 67 (s1): 931-956.

D’Amuri, Francesco and Juri Marcucci. 2010. "'Google it!' Forecasting the US Unemployment Rate with a Google Job Search Index.

D'Amuri, Francesco. 2009. *Predicting Unemployment in Short Samples with Internet Job Search Query Data*. University Library of Munich, Germany.

Della Penna, Nicolás and Haifang Huang. 2009. "Constructing Consumer Sentiment Index for US using Internet Search Patterns." *Department of Economics, University of Alberta, WP* 26.

Duy, Timothy A. and Mark A. Thoma. 1998. "Modeling and Forecasting Cointegrated Variables: Some Practical Experience." *Journal of Economics and Business* 50 (3): 291-307.

Ettredge, Michael, John Gerdes, and Gilbert Karuga. 2005. "Using Web-Based Search Data to Predict Macroeconomic Statistics." *Communications of the ACM* 48 (11): 87-92.

Experian Marketing Services. “Hitwise Canada online trends,” Accessed June 4, 2015. https://www.experian.com/marketing-services/online-trends-canada.html.

Federal Reserve Bank of Philadelphia. “State Coincident Indexes,” Accessed on May 27, 2015. http://www.philadelphiafed.org/research-and-data/regional-economy/indexes/coincident

Fondeur, Y. and F. Karamé. 2013. "Can Google Data Help Predict French Youth Unemployment?" *Economic Modelling* 30: 117-125.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. "Detecting Influenza Epidemics using Search Engine Query Data." *Nature* 457 (7232): 1012-1014.

Goel, S., J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts. 2010. "Predicting Consumer Behavior with Web Search." *Proceedings of the National Academy of Sciences of the United States of America* 107 (41): 17486-17490.

Google Inc. "Google Trends Help Center." Accessed on May 23, 2015. https://support.google.com/trends/?hl=en#topic=6248052.

Green, Anne E., Maria De Hoyos, Yuxin Li, and David Owen. 2011. *"*Job Search Study: Literature Review and Anlaysis of the Labour Force Survey*" Department for Work and Pensions, UK*.

Greene, William H. 2008. *Econometric Analysis.* Granite Hill Publishers.

Lambert, Ben. 2013. "An undergrad course in econometrics," Accessed on May 13, 2015, http://oxbridge-tutor.co.uk/undergraduate-econometrics-course/.

Lin, J. 2008. "Notes on testing causality." *Institute of Economics, Academia Sinica,* Department of Economics, National Chengchi University.

McLaren, Nick and Rachana Shanbhogue. 2011. "Using Internet Search Data as Economic Indicators." *Bank of England Quarterly Bulletin* (2011): Q2.

Middleton, Catherine, and Jonathan Ellison. 2008. "Understanding Internet Usage Among Broadband Households: A Study of Household Internet Use Survey Data." *Statistics Canada*.

Middleton, Catherine Ann, Ben Veenhof, and Jordan Leith. 2010. "Intensity of internet use in Canada: Understanding different types of users." *Statistics Canada. Business Special Surveys and Technology Statistics Division*.

Ng, Serena and Pierre Perron. 1995. "Unit Root Tests in ARMA Models with Data-Dependent Methods for the Selection of the Truncation Lag." *Journal of the American Statistical Association* 90 (429): 268-281.

Nie, Norman H. and Lutz Erbring. 2000. "Internet and Society." *Stanford Institute for the Quantitative Study of Society*.

Polgreen, P. M., Y. Chen, D. M. Pennock, and F. D. Nelson. 2008. "Using Internet Searches for Influenza Surveillance." *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America* 47 (11): 1443-1448.

Purcell, Kristin, Joanna Brenner, and Lee Rainie. 2012. "Search Engine use 2012." *Pew Internet & American Life Project.*

Smith, Geoffrey Peter. 2012. "Google Internet Search Activity and Volatility Prediction in the Market for Foreign Currency." *Finance Research Letters* 9 (2): 103-110.

Statacorp. 2013. "Stata Time-series: Reference Manual." *StataCorp LP*.

Statistics Canada. 2013. "Canadian Internet Use Survey, 2012." *The Daily*, no. 11-001-X, November 26, 2013. http://www.statcan.gc.ca/daily-quotidien/131126/dq131126d-eng.pdf.

Statistics Canada. "Labour Force Survey," Accessed on May 21, 2015. http://www.statcan.gc.ca/eng/survey/household/3701.

Suhoy, Tanya. 2009. “Query Indices and a 2008 Downturn: Israeli Data.” *Research Department, Bank of Israel.*

Toda, Hiro Y., and Taku Yamamoto. 1995. "Statistical inference in vector autoregressions with possibly integrated processes." *Journal of econometrics* 66, no. 1: 225-250.

Trichur, Rita. “It's official: Canada's in a recession,” *Toronto Star*, June 2, 2009, accessed May 29, 2015, https://www.thestar.com/business/2009/06/02/its_official_canadas_in_a_ recession.html.

Varian, Hal R., and Hyunyoung Choi. 2009. "Predicting the present with Google Trends." *Google Research Blog http://googleresearch. blogspot. com/2009/04/predicting-present-with-google-trends.html.*

Vlastakis, Nikolaos and Raphael N. Markellos. 2012. "Information Demand and Stock Market Volatility." *Journal of Banking & Finance* 36 (6): 1808-1821.

Vosen, Simeon and Torsten Schmidt. 2011. "Forecasting Private Consumption: Survey-based Indicators Vs. Google Trends." *Journal of Forecasting* 30 (6): 565-578.

Wu, Lynn and Erik Brynjolfsson. 2014. "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales." In *Economics of Digitization*: University of Chicago Press.

Appendix A: Results Tables

Table 4: Results of ADF of levels

Variable	Number of Lags	Trend Specification	p-value*	Stationarity
UR4	12	trend	0.0573	Non-stationary
URyo	12	trend	0.6006	Non-stationary
UR8	12	trend	0.0576	Non-stationary
JobList_1234	13	constant	0.8712	Non-stationary
JobList_4123	13	constant	0.8606	Non-stationary
JobList_3412	13	constant	0.8921	Non-stationary
JobList_12	13	constant	0.8968	Non-stationary
JobList_34	13	constant	0.8925	Non-stationary
EI_1234	12	trend	0.4109	Non-stationary
EI_4123	12	trend	0.4495	Non-stationary
EI_3412	12	trend	0.3317	Non-stationary
EI_12	11	trend	0.5130	Non-stationary
EI_34	12	trend	0.3293	Non-stationary
EM_1234	13	trend	0.8820	Non-stationary
EM_4123	13	trend	0.9404	Non-stationary
EM_3412	13	trend	0.9392	Non-stationary
EM_12	13	trend	0.6637	Non-stationary
EM_34	13	trend	0.9074	Non-stationary

* This is the probability of rejecting the null hypothesis of a unit root with the specified number of lags and trends.
All of the variable levels are non-stationary at the 5% significance level.

Table 5: Results of ADF of first differences (denoted by d)

Variable	Number of Lags	Trend Specification	p-value*	Stationary
d.UR4	10	constant	0.0000**	Stationary
d.URyo	11	constant	0.0262**	Stationary
d.UR8	10	constant	0.0000**	Stationary
d.JobList_1234	12	constant	0.0229**	Stationary
d.JobList_4123	12	constant	0.0126**	Stationary
d.JobList_3412	12	constant	0.0271**	Stationary
d.JobList_12	12	constant	0.0314**	Stationary
d.JobList_34	12	constant	0.0002**	Stationary
d.EI_1234	12	constant	0.0071**	Stationary
d.EI_4123	12	constant	0.0139**	Stationary
d.EI_3412	12	constant	0.0106**	Stationary
d.EI_12	12	constant	0.0003**	Stationary
d.EI_34	12	constant	0.0089**	Stationary
d.EM_1234	11	constant	0.0060**	Stationary
d.EM_4123	12	constant	0.0195**	Stationary
d.EM_3412	12	constant	0.0044**	Stationary
d.EM_12	12	constant	0.0213**	Stationary
d.EM_34	12	constant	0.0297**	Stationary

* This is the probability of rejecting the null hypothesis of a unit root with the specified number of lags and trends.
** All of the variable first differences are stationary at the 5% significance level.

Thus, the order of integration for each variable pair is 1 (i.e. I(1)).

Johansen's trace test of cointegration was done at the 5% level of significance. All ECM models were estimated with a lag-length of 13 and all have 123 observations with one cointegrating equation. Thus, the estimation period is from February 2005 to April 2015 due to lagging.

Table 6: ECM Results for UR4 and JobList

Exogenous Variable	**Trend specification**	**Adjustment Coefficient (λ)**	**Cointegrating Parameter (β)**	**Short-Run Causality***	**RMSE**	**AIC**	**SBIC**
JobList_1234	trend	.0254	14.83**	30.11**	.2245	-3.451	-2.193
JobList_4123	trend	.0313	16.48**	35.23**	.2218	-3.894	-2.637
JobList_4123	rtrend	.0352	17.21**	36.56**	.2206	-3.910	-2.676
JobList_3412	trend	.0252	13.55**	38.81**	.2177	-3.495	-2.237
JobList_3412	rtrend	.0370	15.35**	41.20**	.2163	-3.508	-2.273
JobList_12	constant	-0.07550**	-1.686	29.13**	.2203	-3.224	-2.012
JobList_12	trend	-0.02942	6.009**	25.64**	.2241	-3.351	-2.093
JobList_12	rtrend	.01320	8.462**	30.00**	.2244	-3.355	-2.121
JobList_34	trend	.02239	20.95**	33.49**	.2223	-2.520	-1.262
JobList_34	rtrend	.02465	21.96**	34.22**	.2210	-2.536	-1.301

* This is the χ^2 distributed F test statistic for joint significance of the short-run coefficients
** Indicates statistical significance at the 5% level

Table 7: ECM Results for UR4 and EI

Exogenous Variable	**Trend specification**	**Adjustment Coefficient (λ)**	**Cointegrating Parameter (β)**	**Short-Run Causality***	**RMSE**	**AIC**	**SBIC**
EI_4123	trend	-.09286**	.2271**	23.20**	.2344	5.523	6.781
EI_3412	trend	-.09441**	.2377**	26.85**	.2303	5.461	6.718
EI_12	rconstant	-.08293**	.0071	17.18	.2319	6.128	7.317
EI_12	trend	-.08726**	.2512**	23.69**	.2324	6.052	7.310

* This is the χ^2 distributed F test statistic for joint significance of the short-run coefficients
** Indicates statistical significance at the 5% level

Table 8: ECM Results for UR4 and EM

Exogenous Variable	**Trend specification**	**Adjustment Coefficient (λ)**	**Cointegrating Parameter (β)**	**Short-Run Causality***	**RMSE**	**AIC**	**SBIC**
EM_1234	constant	.01728	.1741**	29.19**	.2273	4.315	5.527
EM_1234	rconstant	.02000	.1921**	30.50**	.2261	4.301	5.490
EM_1234	rtrend	.01970	.0961	29.63**	.2270	4.327	5.562
EM_1234	none	-0.00833**	-0.4664**	35.78**	.2221	4.321	5.487
EM_4123	constant	.007148	.1482**	20.84	.2351	4.212	5.424
EM_4123	rconstant	.01290	.1663**	22.10**	.2339	4.199	5.387
EM_4123	rtrend	.00935	.0532	21.25**	.2349	4.220	5.455
EM_4123	none	-0.00520**	-0.5392**	26.18**	.2305	4.221	5.388
EM_3412	constant	.01033	.1908**	37.12**	.2211	3.904	5.116
EM_3412	rconstant	.01213	.2111**	38.00**	.2198	3.889	5.078
EM_3412	rtrend	.01246	.0603	37.25**	.2209	3.913	5.147
EM_3412	none	-0.00494**	-0.5548**	41.94**	.2171	3.897	5.063

EM_12	constant	.02128	.1498**	25.11**	.2310	4.804	6.016
EM_12	rconstant	.02433**	.1887**	26.95**	.2293	4.794	5.983
EM_12	rtrend	.01866	.2202	24.22**	.2313	4.818	6.052
EM_12	none	-0.01028**	-0.4266**	33.33**	.2241	4.805	5.971
EM_34	constant	.00894	.2950**	47.19**	.2132	4.586	5.798
EM_34	rconstant	.00929	.3150**	48.04**	.2121	4.570	5.759
EM_34	rtrend	.01206	-0.0081	48.75**	.2123	4.578	5.812
EM_34	none	-0.00608**	-0.4914**	51.97**	.2098	4.576	5.742

* This is the χ^2 distributed F test statistic for joint significance of the short-run coefficients
** Indicates statistical significance at the 5% level

Table 9: ECM Results for URyo and JobList

Exogenous Variable	**Trend specification**	**Adjustment Coefficient (λ)**	**Cointegrating Parameter (β)**	**Short-Run Causality***	**RMSE**	**AIC**	**SBIC**
JobList_1234	constant	-0.13948**	1.6196	43.81**	.6341	-1.030	-0.182
JobList_1234	rconstant	-0.13641**	1.6679	43.86**	.6313	-1.044	0.145
JobList_1234	trend	-0.01846	18.225**	39.03**	.6534	-1.107	0.1503
JobList_1234	rtrend	-0.01313	18.505**	39.09**	.6501	-1.123	0.1113
JobList_4123	constant	-0.12121**	0.54645	45.22**	.6298	-1.558	-0.3458
JobList_4123	trend	0.04132	23.398**	43.03**	.6439	-1.691	-0.4334
JobList_4123	rtrend	0.02363	22.104**	42.77**	.6414	-1.706	-0.4715
JobList_3412	constant	-0.12237**	1.4147	52.29**	.6135	-1.105	0.1063
JobList_3412	rconstant	-0.11963**	1.4540	52.38**	.6108	-1.120	0.0689
JobList_3412	trend	0.01282	20.801**	45.23**	.6295	-1.240	0.0176
JobList_3412	rtrend	0.00690	20.436**	46.80**	.6263	-1.256	-0.0214
JobList_12	constant	-0.16168**	-0.83636	20.20	.6936	-1.074	0.1379
JobList_12	rconstant	-0.15839**	-0.80161	20.22	.6908	-1.088	0.1013
JobList_12	trend	-0.14096	5.9891	17.61	.7068	-1.110	0.1479
JobList_12	rtrend	-0.06568	8.8107**	16.70	.7107	-1.116	0.1190
JobList_34	constant	-0.13302	4.6608	52.74**	.6187	-0.255	0.9565
JobList_34	rconstant	-0.13129**	4.7043	52.76**	.6158	-0.270	0.9184
JobList_34	trend	0.01300	26.334**	44.47**	.6368	-0.328	0.9298
JobList_34	rtrend	-0.00598	24.688**	44.67**	.6339	-0.343	0.8918

* This is the χ^2 distributed F test statistic for joint significance of the short-run coefficients
** Indicates statistical significance at the 5% level

Note: URyo and EI are not cointegrated for any trend specification.

Table 10: ECM Results for URyo and EM

Exogenous Variable	**Trend specification**	**Adjustment Coefficient (λ)**	**Cointegrating Parameter (β)**	**Short-Run Causality***	**RMSE**	**AIC**	**SBIC**
EM_1234	constant	0.00840	0.2334**	37.28**	.6545	6.558	7.770
EM_1234	rconstant	0.02808	0.2577**	36.71**	.6542	6.553	7.742
EM_1234	rtrend	0.01521	0.0849	36.95**	.6541	6.566	7.801
EM_1234	none	-0.00670**	-1.1262**	40.60**	.6457	6.588	7.754
EM_4123	constant	0.0222	0.2862**	22.32**	.6946	6.415	7.627
EM_4123	rconstant	0.0323	.3177**	22.51**	.6920	6.404	7.593
EM_4123	rtrend	0.0291	0.0485	22.30**	.6940	6.418	7.653
EM_4123	none	-0.00752**	-1.1798**	25.33**	.6846	6.421	7.587
EM_3412	constant	0.01914	0.3167**	25.87**	.6845	6.346	7.558

EM_3412	rconstant	0.02583	0.3533**	26.28**	.6814	6.333	7.522
EM_3412	none	-0.00606**	-1.2800**	28.52**	.6759	6.339	7.505
EM_12	constant	0.00671	0.2007**	18.86	.7047	6.992	8.203
EM_12	rconstant	0.03196	0.2216**	18.25	.7047	6.988	8.177
EM_12	rtrend	0.00618	0.2437	18.74	.7047	7.007	8.242
EM_12	none	-0.00757**	-1.0047**	21.75**	.6946	7.027	8.193
EM_34	constant	-0.01011	0.2788**	46.09**	.6339	7.042	8.254
EM_34	rconstant	0.00522	0.3120**	45.63**	.6335	7.034	8.223
EM_34	rtrend	-0.00210	-0.0149	45.56**	.6345	7.034	8.268
EM_34	none	-0.00284	-1.1779**	47.37**	.6306	7.059	8.225

* This is the χ^2 distributed F test statistic for joint significance of the short-run coefficients
** Indicates statistical significance at the 5% level

Table 11: ECM Results for UR8 and JobList

Exogenous Variable	**Trend specification**	**Adjustment Coefficient (λ)**	**Cointegrating Parameter (β)**	**Short-Run Causality***	**RMSE**	**AIC**	**SBIC**
JobList_1234	trend	0.03229	21.924**	35.73**	.2736	-2.985	-1.727
JobList_1234	rtrend	0.03668**	23.577**	36.41**	.2719	-3.000	-1.766
JobList_4123	trend	0.02757	24.569**	30.59**	.2792	-3.476	-2.219
JobList_4123	rtrend	0.03087	26.216**	31.19**	.2775	-3.491	-2.257
JobList_3412	trend	0.01305	17.250**	28.41**	.2797	-2.969	-1.711
JobList_3412	rtrend	0.03197	21.146**	30.09**	.2781	-2.980	-1.745
JobList_12	trend	-0.03976	7.983**	23.93**	.2786	-2.972	-1.714
JobList_12	rtrend	0.00986	12.302**	27.34**	.2795	-2.975	-1.741
JobList_34	trend	0.02770	31.678**	34.34**	.2758	-2.020	-0.763

* This is the χ^2 distributed F test statistic for joint significance of the short-run coefficients
** Indicates statistical significance at the 5% level

Table 12: ECM Results for UR8 and EI

Exogenous Variable	**Trend specification**	**Adjustment Coefficient (λ)**	**Cointegrating Parameter (β)**	**Short-Run Causality***	**RMSE**	**AIC**	**SBIC**
EI_4123	trend	-0.08757**	0.2878**	22.91**	.2880	5.962	7.220
EI_3412	trend	-0.08709**	0.3124**	28.34**	.2817	5.887	7.144
EI_12	rconstant	-0.08435**	0.0139	15.99	.2854	6.541	7.730
EI_12	trend	-0.08476**	0.3107**	23.71**	.2865	6.480	7.737

* This is the χ^2 distributed F test statistic for joint significance of the short-run coefficients
** Indicates statistical significance at the 5% level

Table 13: ECM Results for UR8 and EM

Exogenous Variable	**Trend specification**	**Adjustment Coefficient (λ)**	**Cointegrating Parameter (β)**	**Short-Run Causality***	**RMSE**	**AIC**	**SBIC**
EM_1234	constant	0.01630	.2293**	27.50**	.2814	4.721	5.932
EM_1234	rconstant	0.01856	0.2555**	28.51**	.2798	4.707	5.896
EM_1234	rtrend	0.01904	0.1126	27.64**	.2811	4.732	5.967
EM_1234	none	-0.00704**	-0.6674**	33.20**	.2752	4.720	5.886
EM_4123	constant	0.00490	0.1892**	12.47	.2997	4.668	5.880
EM_4123	rconstant	0.01153	0.2164**	13.46	.2982	4.655	5.844
EM_4123	rtrend	0.00673	0.0548	12.70	.2996	4.675	5.909
EM_4123	none	-0.00431	-0.7787**	16.78	.2942	4.672	5.838

EM_3412	constant	0.00806	0.2090**	22.80**	.2869	4.434	5.646
EM_3412	rconstant	0.01203	0.2391**	23.71**	.2853	4.419	5.608
EM_3412	rtrend	0.00998	0.0682	22.81**	.2868	4.442	5.678
EM_3412	none	-0.00396**	-0.8229**	27.04**	.2819	4.428	5.594
EM_12	constant	0.01623	0.1862**	14.32	.2972	5.257	6.468
EM_12	rconstant	0.02118	0.2286**	15.59	.2954	5.246	6.435
EM_12	rtrend	0.01420	0.2756	13.96	.2973	5.270	6.505
EM_12	none	-0.00788**	-0.6095	20.52	.2896	5.256	6.423
EM_34	constant	0.00828	0.4227**	43.94**	.2652	5.046	6.257
EM_34	rconstant	0.00838	0.4580**	44.38**	.2638	5.030	6.219
EM_34	rtrend	0.01239	-0.0274	44.90**	.2641	5.037	6.272
EM_34	none	-0.00533**	-0.7192**	47.77**	.2611	5.028	6.194

* This is the χ^2 distributed F test statistic for joint significance of the short-run coefficients
** Indicates statistical significance at the 5% level

Appendix B: Graphs of Forecasts

Figure 3: In-sample forecast of UR4 using JobList_12. January 2007 to April 2015

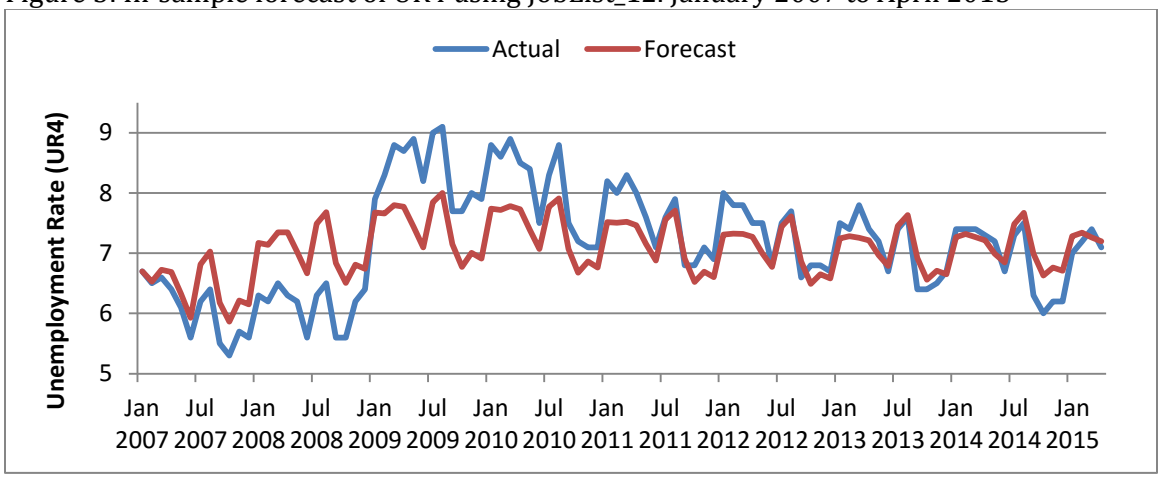


Figure 4: In-sample forecast of UR4 using JobList_12. January 2005 to December 2008

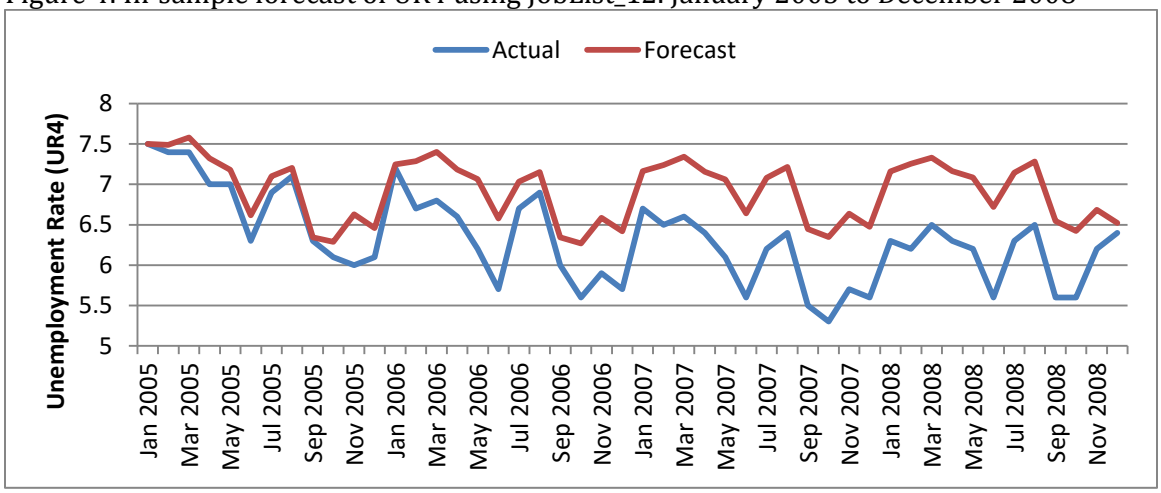


Figure 5: In-sample forecast of UR4 using JobList_12. January 2009 to April 2015

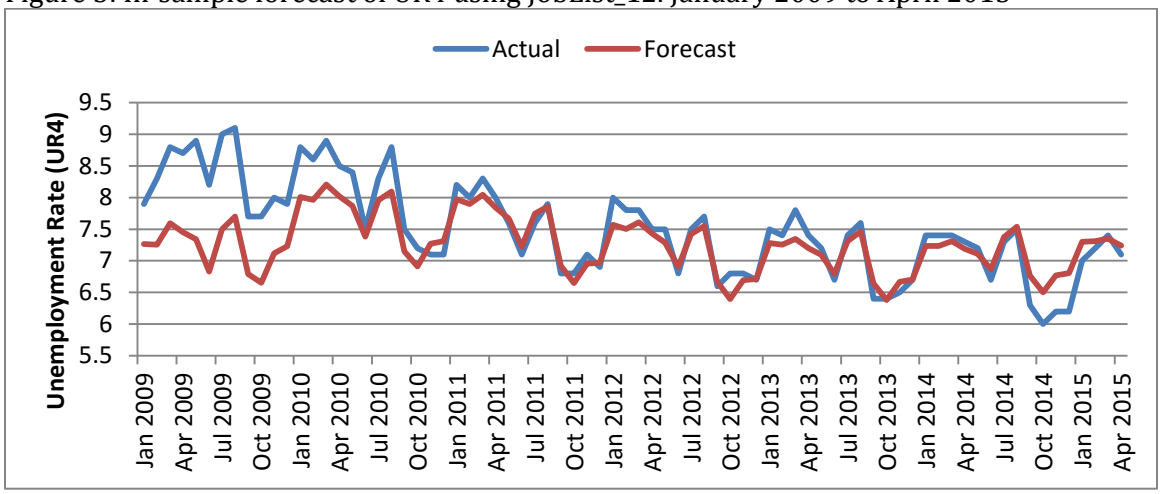


Figure 6: Out-of-sample forecast of UR4 using JobList_12. June 2012 to April 2014

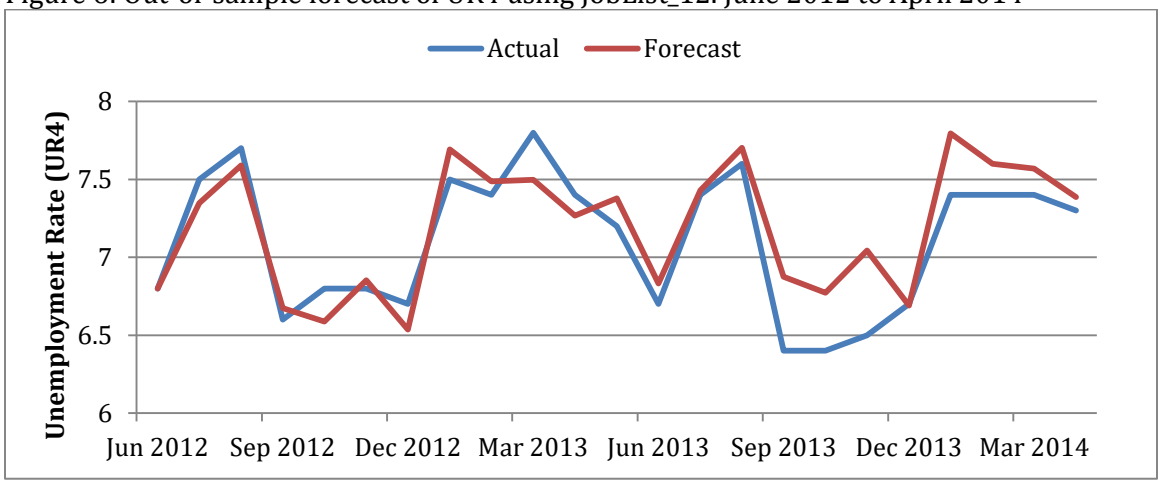


Figure 7: In-sample forecast of URyo using JobList_3412. January 2007 to April 2015

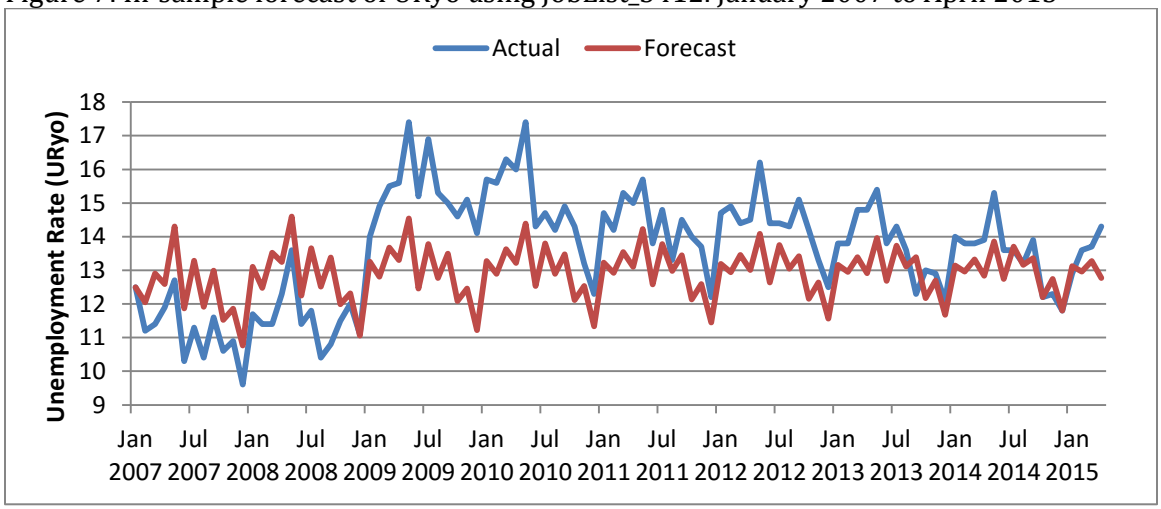


Figure 8: In-sample forecast of URyo using JobList_3412. January 2005 to December 2008

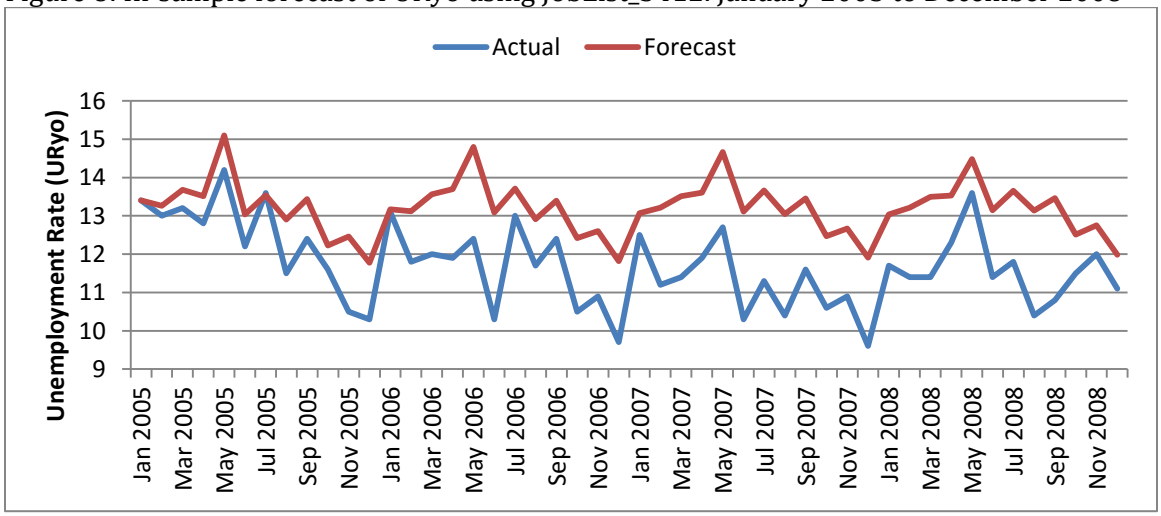


Figure 9: In-sample forecast of URyo using JobList_3412. January 2009 to April 2015

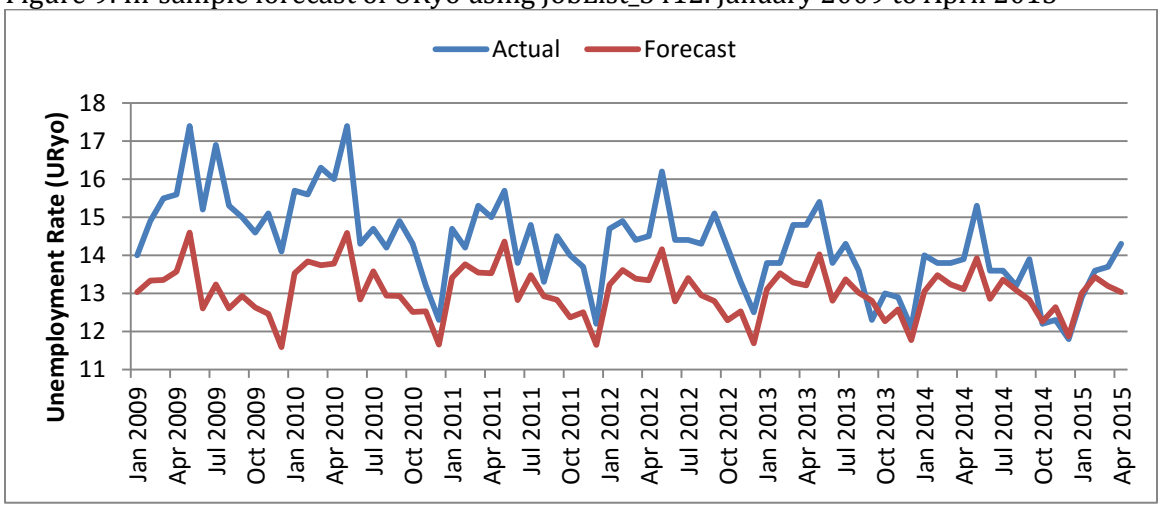


Figure 10: Out-of-sample forecast of URyo using JobList_3412. June 2012 to April 2014

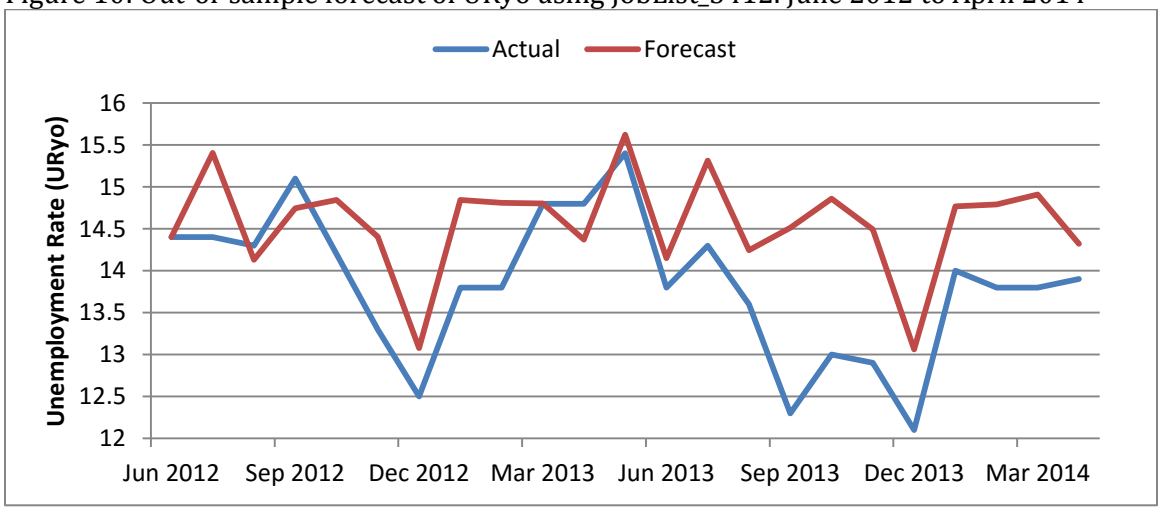


Figure 11: In-sample forecast of UR8 using JobList_1234. January 2007 to April 2015

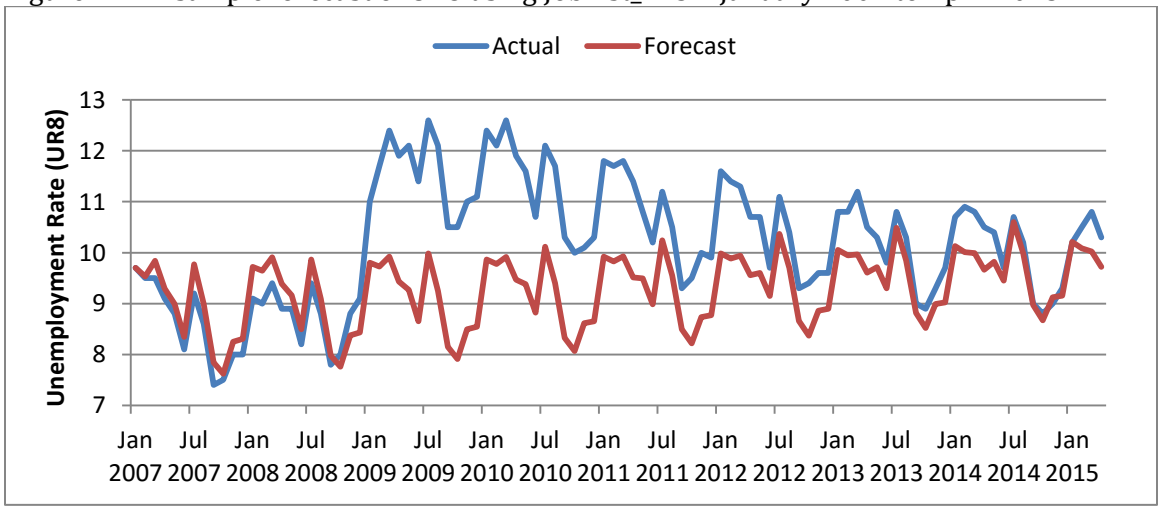


Figure 12: In-sample forecast of UR8 using JobList_1234. January 2005 to December 2008

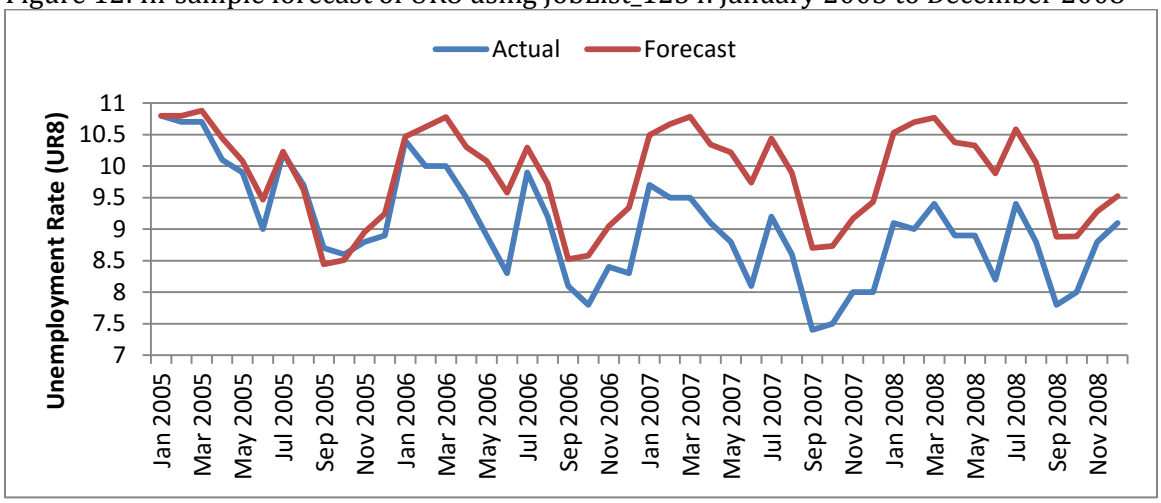


Figure 13: In-sample forecast of UR8 using JobList_1234. January 2009 to April 2015

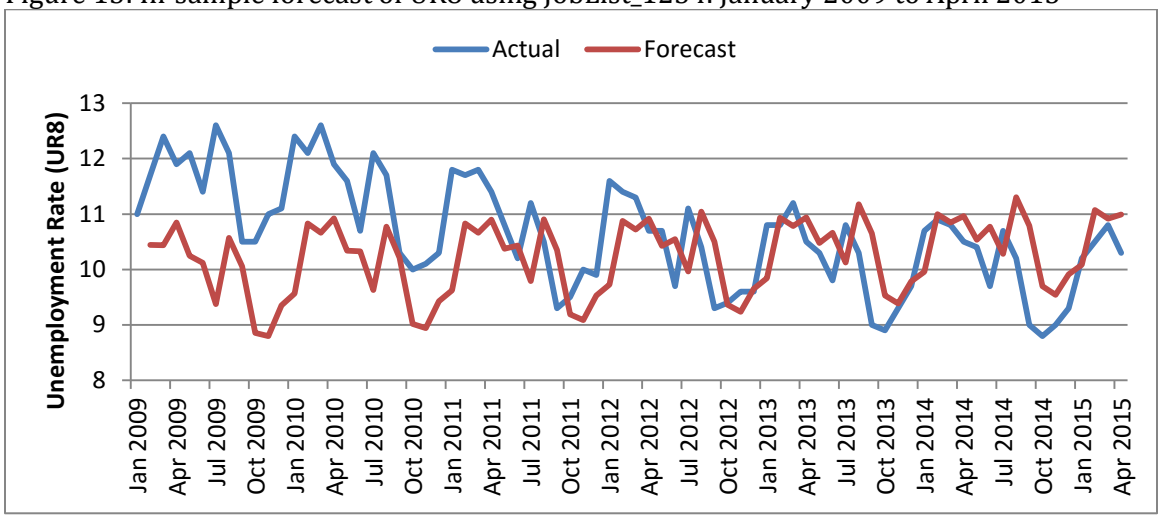


Figure 14: Out-of-sample forecast of UR8 using JobList_1234. June 2012 to April 2014