**Chapter 2**

# Literature overview

## 2.1 Google Trends data

Recently, researchers across various fields like finance, marketing, and economics, have been exploring Google Trends data as a valuable source of real-time information. This can provide unique insights and potentially improve our ability to predict important economic variables. In this literature review, we’ll examine the pros and cons of Google Trends data and their general limitations. We’ll take a close look at key studies that have used this data, especially those focusing on predicting inflation or other large-scale economic indicators. We’ll discuss the methods used in these studies, the challenges they faced, and what their findings might mean for future research in this area.

### 2.1.1 Advantages

The growing interest in the use of GT can be attributed to several factors:

* Real time data: The data is available in real-time. Thus, it can function as a source source of up-to-date insights into consumer behavior, market sentiment, and other factors that influence economic indicators. The real-time nature of the data can be particularly helpful and valuable in dynamic and rapidly changing economic environments or during time periods when traditional sources of data may be lagging behind or outdated.
* High frequency: GT provides high-frequency data, therefore short-term fluctuations and trends can be captured more accurately. This granularity can might help to improve accuracy and responsiveness of forecasting models.
* Broad coverage: GT covers any topic for which search-related terms are searched with non-negligible intensity. As a result, variables for which traditional data are non-existent or hard to collect and quantify, GT can serve as an accesible alternative.
* Behavioral insights: GT offers a unique perspective on consumer behaviour, reflecting collective concerns and interest of internet users. Although internet users cannot be considered as accurate and representative sample of whole population (see chapter 3), GT might provide insight into sentiment, preferences and expectations, which are influenced and possibly can influence economic outcomes among others.
* Cost effectiveness: GT is publicly available and free of charge. That makes it accessible and cost-effective source of information for researchers, policy-makers businesses and overall anyone with access to internet connection.
* Easy integration: GT can complement already existing traditional economic indicators and other data sources in order to create more comprehensive and robust forecasting models. This complementation can help diminish limitations of each dataset in order to enhance predictive power of the models.
* Leading indicator potential: For some economic variables, GT can serve as a leading indicator, which allows for earlier detection of changes in trends or market conditions, since it can provide a representation of consumer expectations which consequently affect the real outcome. This might allow researchers and policy-makers to react more quickly, flexibly and effectively to emerging economic developments.

### 2.1.2 Disadvantages and limitations

Despite the promising findings of prior research of these studies to **??**, there are several notable challenges and limitations associated with using Google Trends data in terms of representativeness of general population. We will review them in general in this chapter, while elaborating furtherly on this topic in specific perspective of our results in to **??**.

Primarily, access to the internet is an obligatory prerequisite. Between 2004 and 2022, the percentage of the Czech population utilizing the internet has increased significantly from 36% to 90% (Kemp (2022), **?**). Subsequently, the specific search engine utilized plays a substantial role. From 2009 to 2023, Google’s market share in the Czech Republic has generally exceeded 75%, with it being particularly prominent (exceeding 80%) during the period from 2013 to 2023 (a20 (2023)). Notwithstanding, Google’s local competitor, Seznam, may attract a unique subset of users due to its distinct user experience. This implies that there may be unobserved characteristics correlated with the use of Seznam, such as age or preference for the Czech language, which could influence the representativeness of Google Trends data.

The third consideration is the inherent bias of the users. Individuals with an interest in financial markets or economics are considerably more likely to search for "inflation" than an average user. For certain professionals, such online searches could even be a fundamental part of their daily work routine. To accurately assess this potential bias, comprehensive data collection from the general populace would be required, a task which is beyond the scope of this thesis.

The fourth constraint is related to privacy measures adopted by some users. Use of privacy-focused search engines or browsing in incognito mode can exclude these users from our data set.

Lastly, Google Trends data may be susceptible to spurious correlations.

In light of these observations and assumptions, it can be inferred that the demographic sample of Google Trends data likely changes over time, impacting its representativeness for the period from 2004 to 2022. As a result, judicious selection of methodological approach, combined with critical evaluation of these constraints, is crucial to ensure that Google Trends data is leveraged effectively and unbiasedly for forecasting purposes.

## 2.2 Preceeding research on Google Trends

Ettredge *et al.* (2005) emerged as early pioneers in the successful application of web search data for economic forecasting. Their research focused on the prediction of unemployment rates utilizing web-based search data. The findings from their study indicate that web search data possess the potential for effective application in forecasting other pertinent macroeconomic indicators.

In addition, various other scientific disciplines have capitalized on the use of web search data. Polgreen *et al.* (2008) investigated connection between search queries related to influenza and actual instances of the illness. Their models demonstrated success in predicting increases in positive influenza cultures and rises in pneumonia and influenza-related mortality several weeks in advance. Subsequently, Ginsberg *et al.* (2009) further substantiated the efficacy of web searches as a tool for early detection in the field of epidemiology. This line of inquiry has been supported and expanded upon by several researchers, including but not limited to Brownstein *et al.* (2009), Pelat *et al.* (2009), and Wilson & Brownstein (2009).

Lets return back to economic-related topics. D’Amuri & Marcucci (2010) examined the relation of web search to evolution of unemployment rate in Germany and USA respectively. Baker & Fradkin (2011) estimated inversely proportional relationship between unemployment benefits and intensity of job search.

Choi & Varian (2009) used engine search data to forecast values of economic indicators such as consumer confidence, unemployment claims, travel destination planning and automobile sales. Authors employ various time-series models including autoregressive models and state-space models.

Askitas & Zimmermann (2009) investigate the use of GT for predicting unemployment rates in Germany, France, and the United States. The authors employ Granger causality tests and vector autoregressive (VAR) models to analyze the relationship between search volume data and unemployment rates. This study serves as an example of how to employ advanced econometric techniques, such as VAR models, when using Google Trends data for forecasting purposes.

Guzmán (2011) was the first one to delve into utilization of GT in order to forecast specifically inflation. The results suggest a statistically significant and positive relationship between search volume data for inflation related terms and inflation expectations. This implies that search data can complement traditional data sources, offering real-time insights and improved predictive accuracy.

Vosen & Schmidt (2011) show that a nowcasting model using Google Trends data outperformed traditional time-series models in predicting consumer price inflation.

Preis *et al.* (2013) utilize GT in such way that they find patterns which can be interpreted as "early warning signs" for otherwise sudden changes in markets. Results of their research align with the idea that significant declines in the financial market are preceded by periods of investor anxiety, as manifested in GT. We believe that this idea can be applied to inflation, as expectations about rising inflation can be conveyed in GT and may precede individual related actions purposely lessening short-term impacts inflation for the individual, while they may spiral inflation upwards even further.

Krištoufek (2013) studies relationship between bitcoin price and GT data, as GT can represent quantified driving factors behind bitcoin price, again providing further evidence for the use of VAR models. Next, he delves deeper into identifying the driving forces of bitcoin price, using GT and Wavelet Coherence analysis. This study provides further evidence for the use of GT when modelling time series data (Krištoufek (2015)).

Lazer *et al.* (2014) reflects on "Big data hubris" problem of GT. Reliability and representativeness of GT can be attributed to many factors such as changes in search algorithms, data sampling method and user behavior to **??** and to **??**, therefore they are prone to inconsistency of quality of their results.

Li *et al.* (2015) set a MIDAS modelling framework to forecast inflation in China using Google search data. They find statistical evidence supporting further use of GT for inflation rate prediction, emphasizing potential of GT as variable that can play pivotal role in business decisions, as it captures relevant effects from user generated content.

Baker & Fradkin (2017) examine the impacts of job search behavior in the order to capture job search intensity. They employ difference-in-differences estimation, utilizing Google Search Data and variations in unemployment benefits across states and time in the United States of America. Authors provide ground on usage of GT data in a quasi-experimental setting, along with methodological approach accounting for confounding factors.

Bulut (2017) employs Vector Autoregression (VAR) model with GT as external regressors as a predictor of Exchange rates. He finds that incorporating GT data in the model improves the model significantly, which provides even more evidence that there exists potential in GT data for financial forecasting.

Kundu & Singhania (2020) aim to obtain best and accurate forecasts of US unemployment using GT data. They apply Neural Networks model and VAR models. VAR models were outperformed by a significant margin, implying new possibilities for modelling of GT.

Mulero & García-Hiernaux (2021) forecast spanish unemployment with GT. They employ ARIMA models and use Principal Component Analysis as a dimension reduction technique to explore hidden driving factors behind GT data.

The authors conclude that to some extent, GT can be used to predict unemployment. They remark on two questions that arise when conducting the analysis:

1. How to choose the best search queries?
2. How to work with the amount of information available in GT?

These questions will arise again in direct specification of search queries in to **??**.

Haile & Strømmen (2022) apply various selection of ARIMA models and find mixed results of usage of GT to predict inflation in Norway. They employ ARIMA models with simple AR(1) model as a benchmark. While they observe increase in prediction quality, they attribute the increase to settings of the ARIMA model than GT being used as external regressor, thus providing no clear consensus on effectivity of GT.

Adu *et al.* (2023) forecast unemployment rates in Ghana using ARIMA, ARIMAX and VAR models). They conclude that not only are GT able to improve unemployment rates, but in addition highlight its potential to assist the entire government system. They propose further work to be based on machine learning and AI models.

**Chapter 3**

# Data and Methodology

## 3.1 Inflation data

As inflation data, we use monthly data for the Czech republic in the time period from January 2004 up to February 2023. In the choice of which attitude to adopt, we decided to use month-on-month inflation data, as they are more likely to be able to capture the month-on-month trend for which it makes most sense to look for in the Google Trends data. Next, the data are gotten rid of seasonality.

## 3.2 Google Trends data

Google Trends data serve as our source of external regressors. We use GT specifically in czech language and only from the geopolitical area of the Czech republic. We begin by specifying our inflation-related terms. We decided to use for now to use just several; 3 specific inflation related terms, word "price" and 8 most searched combinations of word "price" with something else (see Appendix A). Next, Principal Component Analysis (PCA) is introduced. Out of 13 PCA components, only 5 turn out to be valuable with standard deviations being higher than 1. Thus, these are added to our dataset of external regressors for ARIMA models. Last, all GT data are deseasonalized.

## 3.3 ARIMA

According to Autocorroleation and Partial Autocorrelation function, the best attitude would be to try different Arimas with different settings for its parameters. The autoregressive and moving average settings of the model were both ordered to domain of [1,2,3] and the degree of differencing to [0,1]. As input, we use inflation data and external regressor and three ways: aligned, lagged or lagged but to opposite direction. We discard the models with external regressor having p-value higher than 0.1. For every model that comes out with external regressor being statistically significant, we build benchmark model without the external regressor in order to observe and compare information criteria. We save all these models.

We repeat this process for three different time intervals due to recent years of instability. We consider three breaking points, to which it makes sense to break the set to three intervals; Before Covid-19 pandemic, before War on Ukraine and whole available time interval 2004-2023. We do this as otherwise models might not provide good predictions in contrast with varying inconsistent conditions. We exclude 2007-2008 financial crisis as breaking point as we are able to obtain data at 2004 at latest. Thus, including it as a breaking point would leave us with not enough observations to set up an ARIMA model in reliable manner.

## 3.4 ARIMA - Results

For every on of the three time periods, we list for each search term 3 best models with statistically significant coefficient for the external regressor. For each of the models a benchmark model is present, thus we can make comparison of information criteria. To sum it up, for most of the models following holds: AIC and AICc tend to have lower values than the benchmark model, while value of BIC tends to be higher. Lower values indicate an increase in quality of the model, for which AIC and AICc tend to provide evidence. If we take into account propensity of BIC to be over-estimated when adding an external regressor to the model, we mostly observe the models with external regressor to perform better than the benchmark models.

### 3.4.1 2004-2019

For this time period, only one inflation related term achieved statistical significance. "Real estate prices" with opposite lag. The value of real estate prices from the following month is used to estimate the present month’s value of inflation. This obviously has no useful value in terms of predictive power, as we are not aware of future values of external regressor, however it bears possible socio-economic implications about real estate market in the Czech republic.

Considering negative coefficient for the term, one theory might be that increase in inflation causes increase in cost of living as nominal wage stagnates and purchasing power parity of money decreases. This decrease in the standard of living inherently causes lowering of demand for purchasing real estates. It makes sense for this relationship to work vice versa, as lower inflation implicates lower decrease, if not an increase in the standard of living thus implicating opportunity to invest.

### 3.4.2 2004-2022

For this time interval, we have 4 search terms which make up for models with external regressor being statistically significant. These are following:

* a) First principal component, e.i. the component explaining the highest portion of variability in the data from all principal components. If we focus on the degree of correlation between the component and other search queries (see table below), we can see that there is no strong correlation present. However, there is moderate positive correlation present for 4 search queries: "Price", "Price of gas", "Diesel price" and "The price of oil"[[1]](#footnote-1). Remaining search queries are only weakly correlated. Thus, we consider this component as representation of fuel prices in general.

Out of all models using the component as external regressor, only the three with best scores in terms of information criteria are saved (see PC1 - models).

For first and second best model, we can see observe decrease both in AIC and AICc. However, at the same time we observe increase in BIC. We consider it to be less relevant than AIC and AICc, as adding external regressor might cause over-estimation of BIC. This, along with third model, provides sufficient evidence that both no-lag value and one period lagged value of external regressor contain potential to increase predictive power of inflation models.

* Diesel price: three best models all come to have external regressor lagged in opposite direction. As per Real estate prices in previously inspected Table 3.1: Degree of correlation between First principal component and other search queries

Degree of correlation with the first principal component

|  |  |
| --- | --- |
| inflace | -0.01 |
| inflation | -0.10 |
| prices | 0.13 |
| cena.elektriny | -0.14 |
| cena.plynu | 0.14 |
| cena.zlata | 0.23 |
| cena.nafty | 0.50 |
| cena.benzinu | 0.47 |
| cena.nemovitosti | 0.17 |
| cena.bydleni | 0.02 |
| cena | 0.40 |
| cena.ropy | 0.46 |
| nakup.zlata | -0.08 |

Table 3.2: PC1 - models

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ~~AR~~ | ~~I~~ | ~~MA~~ | ~~p-value~~ | ~~lag~~ | ~~AIC~~ | ~~AICc~~ | ~~BIC~~ | ~~sAIC~~ | ~~sAICc~~ | ~~sBIC~~ |
| 1 | 0 | 1 | 0.0829 | delay | 348.90 | 349.19 | 365.80 | 349.92 | 350.11 | 363.44 |
| 1 | 0 | 1 | 0.083 | regresor | 349.50 | 349.78 | 366.42 | 350.51 | 350.70 | 364.05 |
| 2 | 0 | 1 | 0.0613 | delay | 349.38 | 349.78 | 369.66 | 351.82 | 352.10 | 368.72 |

time period, this would rather imply that value of inflation is related to search intensity of diesel prices following month. Again, information criteria in comparison with information criteria of benchmark models imply usefulness of first lag and first opposite lag of the search term.

* The price of oil: indicator of positive relationship between inflation and price of oil with no lag and one lag both. For both models, decrease in AIC and AICc in comparison with the benchmark model can be observed, with BIC again slightly higher for models with external regressor. Nevertheless, we are aware of the tendency BIC has for over-estimation when external regressor is added.
* Real estate prices search term works very same as per 2004-2019, implying negative relationship between inflation and furtherly search intensity of Real estate prices.
* Inflation: all three best models capture negative relationship between

1-period lag and 0 -period lag of search intensity of inflation with real inflation data.

### 3.4.3 2004-2023

For time period 2004 to 2023 we observe only one inflation-related term to be able to make up models with external regressor to be statistically significant - "inflation"[[2]](#footnote-2). From the models created, we see that potential for predictive power lies likely in zero lag and first lag of inflation. We mostly observe same information criteria comparison as for other models.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | row\_names | promenna | p-value | AIC | AICc | BIC | AR | I | MA | coef |
| ~~1~~ | ~~2019/2~~ | ~~cena.nemovitosti~~ | ~~0.0806~~ | ~~291.08~~ | ~~291.43~~ | ~~307.08~~ | ~~1~~ | ~~0~~ | ~~1~~ | ~~-0.0017~~1 |
| 2 | 2022/2 | cena.nafty | 0.0884 | 347.24 | 347.52 | 364.14 | 1 | 0 | 1 | 0.02448 |
| 3 | 2022/2 | cena.nemovitosti | 0.0965 | 347.37 | 347.66 | 364.27 | 1 | 0 | 1 | -0.00113 |
| 4 | 2023/2 | inflace | 0.037 | 423.05 | 423.32 | 440.19 | 1 | 1 | 2 | 0.00054 |

If we look for in every one of the time periods for model with best criteria (both external regressor model and benchmark model), we end up with following models:

Table 3.3: Alfa verze, finalni verze potrebuje jeste doplnit data

As we can see, one and only model for time period 2004-2023 has potential for predictive power, as from search intensity of inflation we derive future value of real inflation. Remaining models rather bear possible socio-economic implications, as external regressors for the models are opposite-lagged.

## 3.5 ARIMA - Forecasts

Therefore we will focus on this specific model[[3]](#footnote-3). We construct rolling window forecasts and expanding window forecasts both for external regressor model and benchmark model for the time period January 2004 - February 2023. The prediction is always made only one step ahead, as nowcasting is likely be more useful than long-ahead predictions due to the nature of GT data.

In the dilemma of choosing the right size of the window for the forecasts, we do this for all window settings from 72 to 200. Minimal length of the window is set this low due to final number of variables to estimate is only 4[[4]](#footnote-4).

We calculate Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Predictive power of the models with given settings can be seen in figures below.

### 3.5.1 Expanding window forecasts

It can be observed that model with the external regressor always performs slightly better than its benchmark counterpart. We run Paired T-tests and find that all measures are lower for external regressor model than for its benchmark counterpart with statistical significance on alpha level 0.01.

Overall, we find the three best settings for the window length to be 87, 88 and 90, given error measures are minimized for them.

**Expanding window forecast**

80

100

120

140

160

180

200

0.3

0.4

0.5

0.6

0.7

0.8

0.9

Starting window length

Value

mae

mse

rmse

mae − benchmark

mse − benchmark

rmse − benchmark

### 3.5.2 Rolling window forecasts

In most cases (ratio) can be better prediction power of the external regressor model than benchmark model observed. However, we run Paired T-test and find that all measures are lower for external regressor model than for its benchmark counterpart with statistical significance on alpha level 0.01. Again, we find the three best window settings in ascending order to be 86,88,89, as error measures are minimized for them.

In conclusion, ARIMA models lead us to conclusion that GT data can in fact be used as valuable external regressor used in order to predict inflation.

**Rolling window forecast**

80

100

120

140

160

180

200

0.3

0.4

0.5

0.6

0.7

0.8

0.9

Window length

Value

mae

mse

rmse

mae − benchmark

mse − benchmark

rmse − benchmark

Firstly, this provides evidence for potential for further use of GT and their possible advantages (see Chapter Literature overview) in context of improving quality of inflation predictions. Secondly, this finding bears possible socioeconomic implications.

## 3.6 Arima - Results: Socio-economic implications

Let’s consider the possibility of causal relationship between GT data and inflation. How can we interpret GT data? In case of our data, we can consider it to be rate of relative popularity of a search term on the geopolitical area known as the Czech republic throughout time. If that be so, how much does GT data source sample differ from general population? We review reliability and representativeness of GT (see chapter Literature overview) in the context of used economic-related variables.

* Internet access: While 90% of czech population have access to Internet now, it has not been the case for long. Thus, several of our variables might be product of varying data sample. In addition, prices of fuel, electricity and especially search term "prices" themselves might be variables for which even those with no internet access might be concerned.
* The use of Seznam is likely to correlate with czech preference, which is likely to be correlated for example with age or political preference. Then, our overall GT sample might represent general population even less.
* Thirdly, the bias of users themselves (see chapter Literature overview). While we have no way to effectivelly measure it, it makes sense to presume that bias is present.
* Fourth, we are not sure how large part of overall bias should we attribute to privacy measures. As motivation for the use of privacy measures we consider a need for anonymity. However, we do not see any clear connection between our general, broadly used terms and any possible need for anonymity.
* Fifth and last, as we consider all of our variables, all of them might be subjects to spurious correlations. It is very much possible that there exists a factor, or more likely many factors influencing both inflation and search terms each with different intensities.

What biases will we presume in our data in contrast with general population? It depends on each of the variables.

Nevertheless, we presume second and fourth point to hold more or less for all variables the same, as users tend to be rather rigid in their search engine use and that we see no reason to use privacy measures in this context. Therefore, we believe that the amount of bias stemming from these two points to be rather lower than higher.

Magnitude of individual bias for each variable thus stems from first and third point. Beyond those mentioned in the first point, all of the search terms can be associated to an existing job where one of the tasks is to google the given search term. In addition, preferences of individuals apart from there occupations may play significant role. Therefore we believe the bias stemming from point one and three to be relatively higher to bias stemming from points two and four.

Last but not least, fifth point reminds us to take into account possibility of spurious correlation when inspecting specific perspectives.

Therefore, we believe that GT sample is likely to be biased in contrast to general population and as a result, in cases of all of our variables, we cannot treat it as an unbiased sample.

However, we defined rate of relative popularity of a search term on the geopolitical area known as the Czech republic throughout time. There are several possible ways to interpret increase in the relative popularity:

* Increased Public Interest or Awareness: When awareness or public interest is increased rapidly, mass of not casual Google users might join data pool, especially for search query "inflation". This can occur as a reaction to an experience, such as news about future expectations about inflation, government announcements, rising prices etc.
* Economic uncertainty: Increased searches for an economic term can be manifestation of uncertain expectations about future economic conditions. For example, if people expect an increase in inflation, they might conduct more searches in order to better understand current situation and possible future scenarios and their implications.
* Reflection of actual economic conditions: While an increase in the relative popularity might imply expectations about future, it can be a reflection of actual state of the economy. Thus an increase in searches for term "inflation" can be just a reflection of currently rising inflation rates. This idea has been developed on in various studies, where GT are used as a real-time indicator (Preis *et al.* (2013), Choi & Varian (2012)). This would provide us with otherwise complex quantification of price sensitivity of users. Nevertheless, as we explained earlier, it is rather problematic to transfer implications of GT to general population.
* Information-seeking behaviour: People might search for economic terms to get a better understanding of them. Apart from previous points, we can consider information-seeking behaviour as self-education. In terms of self-educating adults, we have no way to estimate the impact among overall search intensity. However, we see no reason to assume strong variation throughout time. On the subject of education of adolescents, czech educational system is one of the most strongly de-centralized in Europe. Ministry of education publishes "Rámcový vzdělávací program" (from now on refferred to by acronym RVP) which is binding general document which provides ground for schools themselves to create their specific educational plan. Nonetheless, the RVP states only content outline and expected output, not school year or month in which for instance economic topics are taught. We would need micro-level data for every school possible in order to estimate decomposed ratio of the impact throughout time. Following estimation of magnitude of the impact is faced with similar problems as overall representativeness of GT in contrast with general population. Therefore we neglect the impact, as its estimation is beyond the scope of this thesis.
* Noise: Not all searches reflect meaningful information. People might search economic-related terms for various reasons apart from those stated above, such as curiosity, boredom or coverage of the topic in media. Distinguishing this ’noise’ from economically meaningful signals is a significant challenge in interpreting Google Trends data Choi & Varian (2012).

**Bibliography**

(2023): “Search engine market share czech republic.”

Adu, W. K., P. Appiahene, & S. Afrifa (2023): “Var, arimax and arima models for nowcasting unemployment rate in ghana using google trends.”

*Journal of Electrical Systems and Information Technology* **10**.

Askitas, N. & K. F. Zimmermann (2009): “Google econometrics and unemployment forecasting.”

Baker, S. & A. Fradkin (2011): “What drives job search? evidence from google search data.” *core.ac.uk* .

Baker, S. R. & A. Fradkin (2017): “The impact of unemployment insurance on job search: Evidence from google search data.” *The Review of Economics and Statistics* **99**: pp. 756–768.

Brownstein, J. S., C. C. Freifeld, & L. C. Madoff (2009): “Digital disease detection â€” harnessing the web for public health surveillance.” *New England Journal of Medicine* **360**: pp. 2153–2157.

Bulut, L. (2017): “Google trends and the forecasting performance of exchange rate models.” *Journal of Forecasting* **37**: pp. 303–315.

Choi, H. & H. Varian (2009): “Predicting initial claims for unemployment benefits.”

Choi, H. & H. Varian (2012): “Predicting the present with google trends.”

*Economic Record* **88**: pp. 2–9.

D’Amuri, F. & J. Marcucci (2010): “’google it!’ forecasting the us unemployment rate with a google job search index.” *SSRN Electronic Journal*

**31.2010**.

Bibliography 23

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

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

Guzmán, G. (2011): “Internet search behavior as an economic forecasting tool: The case of inflation expectations.” *Journal of Economic and Social Measurement* **36**: pp. 119–167.

Haile, S. Y. & L. B. Strømmen (2022): *Searching for Inflation*. Master’s thesis.

Kemp, S. (2022): “Digital 2022: Czechia.”

Krištoufek, L. (2013): “Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era.” *Scientific*

*Reports* **3**.

Krištoufek, L. (2015): “What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis.” *PLOS ONE* **10**: p. e0123923.

Kundu, S. & R. Singhania (2020): “Forecasting the united states unemployment rate by using recurrent neural networks with google trends data.”

*International Journal of Trade, Economics and Finance* **11**: pp. 135–140.

Lazer, D., R. Kennedy, G. King, & A. Vespignani (2014): “The parable of google flu: Traps in big data analysis.” *Science* **343**: pp. 1203–1205.

Li, X., W. Shang, S. Wang, & J. Ma (2015): “A midas modelling framework for chinese inflation index forecast incorporating google search data.”

*Electronic Commerce Research and Applications* **14**: pp. 112–125.

Mulero, R. & A. García-Hiernaux (2021): “Forecasting spanish unemployment with google trends and dimension reduction techniques.” *SERIEs* **12**: pp. 329–349.

Pelat, C., C. Turbelin, A. Bar-Hen, A. Flahault, & A.-J. Valleron (2009): “More diseases tracked by using google trends.” *Emerging Infectious Diseases* **15**: pp. 1327–1328.

Bibliography 24

Polgreen, P., Y. Chen, D. Pennock, & F. Nelson (2008): “Using internet searches for influenza surveillance.” *Clinical Infectious Diseases* **47**: pp. 1443– 1448.

Preis, T., H. S. Moat, & H. E. Stanley (2013): “Quantifying trading behavior in financial markets using google trends.” *Scientific Reports* **3**.

Vosen, S. & T. Schmidt (2011): “Forecasting private consumption: surveybased indicators vs. google trends.” *Journal of Forecasting* **30**: pp. 565–578.

Wilson, K. & J. S. Brownstein (2009): “Early detection of disease outbreaks using the internet.” *Canadian Medical Association Journal* **180**: pp. 829–831.

**Appendix A**

**Title of Appendix A**

**Appendix B**

**Appendix B**

1. for czech translations used in data, see Appendix A [↑](#footnote-ref-1)
2. "inflace" in czech. [↑](#footnote-ref-2)
3. settingy modelu (1,1,2). [↑](#footnote-ref-3)
4. external regressor, 1 autoregressive part and 2 moving average parts [↑](#footnote-ref-4)