In recent years, researchers have been exploring the potential of using Google Trends data as a source of real-time information, as they might provide insight and improve predictive power of various variables in diverse fields such as finance, marketing and economics among others. This literature overview revolves around key studies whose main focus was the use of Google Trends data as a tool to predict inflation or other macro-level indicator. We review the employed methodologies, challenges faced, and the implications of these findings for future research.

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

1 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rapidly changing economic environments or during time periods when traditional sources of data may be lagging behind or outdated.

2 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.

3 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.

4 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, GT might provide insight into sentiment, preferences and expectations, which are influenced and possibly can influence economic outcomes among others.

5 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 overal anyone with access to internet connection.

6 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 overal predictive power.

7 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.

Challenges and Limitations:

Despite the promising results of these studies, there are several challenges and limitations associated with using Google Trends data for inflation prediction. The low representativeness of Google Trends data in comparison with general population can be attributed to various factors:

First of all, user needs to have internet access. However, we consider this to be rather lower issue as for 2022, 90\% of people in the Czech republic have internet access.

https://datareportal.com/reports/digital-2022-czechia

Secondly, given user needs to use given search engine, Google in our case. For the time period of 2009 to 2023, markets shares of Google held mostly higher than 75% in the Czech republic, especially more than 80% in the 2013-2023 period. However, with Seznam being Google‘s only relerevant rival market shareholder, the Google sample is more likely to be different, as seznam may feel attractive to particular type of users as user experience is different. There can be attributes related to Seznam use, such as age, czech languaage preference or other, but for now, we are unable to find any paper

https://gs.statcounter.com/search-engine-market-share/all/czech-republic/#monthly-200901-202304

Third bias is the bias of users themselves. Someone interested in financial markets, economy or etc. Is much more likely to be google “inflation“ than casual user. For some, doing this may even be a part of their job, giving them much more weight in opposition to casual users. In order to compare the populations, we would need to perform wide data colletion in general population. That is beyond the scope of this thesis.

Fifth, privacy measures of some users eliminate them from the our data pool. These can be privacy-focused search engines or incognito mode.

Sixth and last, GT can be susceptible to spurious correlations.

According to our observations and assumptions, population sample of GT data likely varies throughout time period 2004 – 2022, along with its representativeness.

Thus, careful consideration of conditions along with choice of appropriate methodological approach is necessary in order to ensure GT is applied effectively and without bias in terms of forecasting purposes.

Despite the promising findings of prior research, several notable challenges and constraints arise when utilizing Google Trends data for the purpose of predicting inflation. The under-representativeness of Google Trends data when compared to the general populace can be attributed to a number of critical factors:

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% (\cite{kemp\_2022\_digital}, \cite{dataworldbank}).

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 (\citet{a2023\_search}). 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.

can be by changes in search algorithms, data sampling methods, and user behavior (Lazer et al., 2014).

Lazer 2014

<https://sci-hub.se/10.1126/science.1248506>

Pejpry:

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 possesses 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 the association 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 (2009).

Polgreen et al. 2008

<https://academic.oup.com/cid/article/47/11/1443/282247>

Ginsberg et al. 2009

<https://www.nature.com/articles/nature07634>

Corley

<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=bab035bbc719ecdb3ffcb2a14b97f1390939a744>

Brownstein

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2917042/>

Pelat

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2815981/>

Wilson

<https://www.cmaj.ca/content/180/8/829.short>

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

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

Askitas and 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.

Choi and Varian (2012) 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.

Choi & Varian

<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=4d91786f9f88e0ec8dd5a25ca7c08f4d8e693b53>

Aksitas and zimmerman

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1465341>

askitas zimmerman 2009

<https://docs.iza.org/dp4201.pdf>

d amuri and marcucci

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1594132>

Guzmán

<https://sci-hub.se/10.3233/JEM-2011-0342>

Baker and fradkin

<https://core.ac.uk/reader/6400010>

Baker Fradkin 2017

<https://sci-hub.se/https://doi.org/10.1162/REST_a_00674>

Vosen and Schmidt

<https://onlinelibrary.wiley.com/doi/full/10.1002/for.1213?casa_token=wDqoyslfLtQAAAAA%3A_Br_ThERGB4wNuDXBZebA-tpq9GOIqJic4_9PjvQNswbugKz7aOH8DopCiYqOFklEmlXdGJUY9eedYmsPA>

Preis et al. (2013) utilize GT in such way that they find patterns which can be interpreted as „early warning signs“. 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.

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 (see Chapter disadvantages of GT and Results), 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.

https://www.sciencedirect.com/science/article/abs/pii/S1567422315000022

Baker and 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.

<https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2500>

Kundu and 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.

<http://www.ijtef.org/index.php?m=content&c=index&a=show&catid=110&id=1047>

García-Hiernaux and Mulero (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: How to choose the best queries and how to work with the amount of information available in GT.

<https://link.springer.com/article/10.1007/s13209-021-00231-x>

Haile and 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.

<https://openaccess.nhh.no/nhh-xmlui/handle/11250/3014071>

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 unemploynt rates, but also highlight its potential to assist the entire government system. They propose further work to be based on machine learning and AI models.

<https://jesit.springeropen.com/articles/10.1186/s43067-023-00078-1>

1. Askitas, N., & Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. Applied Economics Quarterly, 55(2), 107-120.
2. Baker, S. R., & Fradkin, A. (2017). The Impact of Unemployment Insurance on Job Search: Evidence from Google Search Data. The Review of Economics and Statistics, 99(5), 756-768.

Overview: Baker and Fradkin (2017) examine the impact of unemployment insurance on job search behavior, utilizing Google search data to capture job search intensity. The authors employ a difference-in-differences estimation strategy, exploiting variations in unemployment insurance benefit durations across states and time in the United States. The study provides an example of how to use Google Trends data in a quasi-experimental setting, along with a methodological approach that can account for potential confounding factors.

These three papers offer a range of methodological approaches for incorporating Google Trends data into econometric models for forecasting and causal analysis. By reviewing these studies, you can gain insights into different techniques, such as time-series models, vector autoregressive models, and difference-in-differences estimation, which can be adapted to your own research on using Google Trends data for predicting inflation.