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 much rather able to capture the monht-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.

GT

Google Trends data serve as our source of external regressors. We begin by specifying our inflation-related terms. We decided to use for now to use just several; 3 specific inflation related terms, word „“cena“ and 8 most searched combinations of word “cena“ with something else. 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.

Něco o granger causalite

Then, Arima models are run. 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 that might not provide good information background in contrast of recently calmer decades before locally in the Czech Republic. These time intervals start at January 2014 and end in December 2019, February 2022 and February 2023 respectively.

In table tabulka\_arima\_modelu Appendix A we store all models with p-values for external regressor being lower than 0.1.

For every period, 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. If we taka into account propensity of BIC to be over-estimated with external regressor, we mostly observe the models with external regressor to bet he better performing ones.

1. 2004 – 2019

For this time period, only one inflation related term achieved statistical significance. “Real estate prices“ with opposite lag. Thus next month’s value of Real estate prices is used to estimate present month 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 implications.

Self fullfilling prophecy idea

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 lower 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.

1. 2004 – 2022

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

1. 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 tabulka\_pc\_1), 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“ (footnote o tom, jak byly původně česky, případně odkaz na můj slovník). Remaining search queries are only weakly correlated. Thus, we consider this component as representation of price of fuel in general.

Out of all models using the component as external regressor, only with three best scores in terms of information criteria are saved (see tabulka\_pc1\_modely). 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.

1. Diesel price – three best models all come to have external regressor lagged in opposite direction. As per Real estate prices in previously inspected 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.
2. The price of oil – indicator of positive relationship between inflation and price of oil with no lag and one lag bothFor both models, decrease in of AIC and AICc in comparison with benchmark model can be observed, with BIC again slightly higher for models with external regressor. Nevertheless, we are aware of its tendency to over-estimate.
3. 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.
4. 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.
5. 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.

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.

1. Asd

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:

Tabulka(min\_rows)

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.

Therefore we will focus on this specific model \footnote{settingy modelu (1,1,2).}. We construct rolling window forecasts and expanding window forecasts both for external regressor model and benchmark model. For our time period January 2004 – February 2023. The prediction is always made only one step ahead, as information are likely

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 is set this low due to final number of variables to estimate is only 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 below (FigureXY, FigureYX).

Expanding window forecasts

It can be observed that model with the regressor always performs slightly better than its benchmark counterpart. We ran 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 in descending order to be 87,88 and 90, given error measures are minimized for them.

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 descending order to be 86,88,89, as error measure are minimized for them.

Hodit druhej graf

This leads us to conclusion that GT data can in fact be used as valuable external regressor used in order to predict inflation. Firstly, this provides evidence of potential for further use of GT adn their possible advantages (see Chapter Literature overview). Secondly, this finding bears possible socio-economic implications.

Lets consider the possibility of causal relationship between GT data and inflation. How can we interpret GT data? We can consider it to be as rate of interest in given time period about an information? If that be so, how much does GT data source sample differ from general population? We need to consider disadvantages of GT (see chapter Literature overview).

1. 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 “prices“ themselves might be variables from which even those with no internet access might be concerned.
2. 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.
3. Thirdly, the bias of users themselves. See Literature overview. While we have no way to effectivelly measure it, it makes sense to presume that bias is present.
4. 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, as they have no potential even for commercial ads.
5. 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 rigid in their search engine use and that we se 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 infividual bias for each variable thus stems from first and third point. Beyond those mentioned in thie 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 také 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 such.