Charles University

Faculty of Social Sciences
Institute of Economic Studies

BACHELORS'S THESIS

Utilizing Online Data in Modelling Unemployment Rates in the Czech Republic

Author: **Kristýna Křížová**

Supervisor: **doc. PhDr. Ladislav Krištoufek, Ph.D.**

Academic Year: **2018/2019**

Declaration of Authorship

The author hereby declares that she complied this bachelor thesis on her own under the leadership of her supervisor and that the references include all resources and literature she has used.

The author grants to the Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, May 10, 2019

Signature

Acknowledgments

I would like to express my gratitude to my supervisor doc. PhDr. Ladislav Krištoufek, Ph.D. for his willingness to find time and help me anytime I asked for it, and his useful comments that improved the thesis, especially the part focusing on knowledge of econometrics. Special thanks belong to Tomáš Dombrovský of LMC (www.jobs.cz) for his helpfulness and provision of valuable data that I used in the thesis.

I would like to specially thank my family and the closest friends for their permanent support during the process of writing and mainly during the whole studies. I cannot forget my patient boyfriend, who provides the control of this thesis and gives me helpful comments how to improve it.

Bibliographic note

KŘÍŽOVÁ Kristýna. *Utilizing Online Data in Modelling Unemployment Rates in the Czech Republic*. 73. Bachelor thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies. Supervisor doc. PhDr. Ladislav Krištoufek, Ph.D.

Abstract

Unemployment rate is a crucial macroeconomic aspect for each state, which aim to have it as low as possible. However, if it is too low, many problems could arise due to a large number of job vacancies and a small number of people needed for market. As the Internet is very useful nowadays, the main aim of the thesis is to investigate the relationship between the Czech unemployment rate and job search on the Internet by users who are interested in changing jobs or are unemployed and need to find some work. Thanks to the relationship, we can conclude whether online data could improve unemployment prediction, which is needed to make effective government decisions. This thesis should also provide easier and better prediction of movements in the unemployment rate, which is inaccurate as most data sources used in economics are commonly available only after a substantial lag. The study applies data freely available on the website of Integrated Portal of the Ministry of Labour and Social Affairs, which provides statistics of unemployment rates, as well as data from portal Jobs.cz, where are information about job vacancies on the portal and response of candidates to occupied positions. The thesis uses a simple autoregressive model of the unemployment in the Czech Republic and extends it with extra variables containing data from the portal Jobs.cz. In addition to the augmented autoregressive model of the Czech Republic, the study estimates the same models for 14 regions of the Czech Republic separately. The results indicate that data from the job search portal Jobs.cz improve nowcasts of the Czech unemployment rate as well as base models with relationship between the unemployment rate and data on number of job vacancies and responses to them. Nevertheless, our findings show that the job-related data do not improve forecasts of the unemployment rate.

JEL Classification	C51, C53, E24, E27
Keywords	unemployment rates, unemployment prediction, nowcasting, Czech Republic, online data, job vacancies, job search, regions

Author's e-mail	krizovi4@seznam.cz
Supervisor's e-mail	ladislav.kristoufek@fsv.cuni.cz

Abstrakt

Míra nezaměstnanosti je klíčovým makroekonomickým atributem každého státu, jehož cílem je, aby byla co nejnižší. Pokud je však příliš nízká, mohlo by dojít k mnoha problémům v důsledku velkého počtu volných pracovních míst a malého počtu lidí potřebných pro trh práce. Vzhledem k tomu, že Internet je velmi užitečný v dnešní době, hlavním cílem této práce je zkoumání vztahu mezi českou mírou nezaměstnanosti a hledáním práce na Internetu uživateli, kteří mají zájem o změnu zaměstnání, či jsou bez práce a potřebují si nějakou najít. Díky tomuto vztahu můžeme dospět k závěru, zda by online data mohla zlepšit predikci nezaměstnanosti, která je potřebná pro efektivní rozhodování vlády. Práce má také zajistit snazší a lepší odhadnutelnost pohybů míry nezaměstnanosti, která je nepřesná, protože většina dat používaná v ekonomice je běžně dostupná pouze se značným zpožděním. Studie používá data volně přístupná na stránce Integrovaného portálu MPSV (Ministerstvo práce a sociálních věcí), která poskytuje statistiky míry nezaměstnanosti, a současně také data poskytnutá z internetového portálu Jobs.cz, kde jsou informace o počtu pozic na Jobs.cz a reakce kandidátů na obsazované pozice. Práce používá jednoduchý autoregresní model nezaměstnanosti v České republice a rozšiřuje ho o další proměnné obsahující data z portálu Jobs.cz. Kromě rozšířeného autoregresního modelu České republiky studie zkoumá stejné modely pro jednotlivých 14 krajů České republiky. Výsledky dokazují, že data z internetového portálu Jobs.cz zlepšují nowcasty (krátkodobé předpovědi) české míry nezaměstnanosti a také zlepšují základní modely se vztahem nezaměstnanosti a údajů o počtu volných pracovních míst a odpovědích na ně. Nicméně naše zjištění ukazují, že údaje týkající se pracovních míst nezlepšují předpovídání míry nezaměstnanosti.

Klasifikace JEL	C51, C53, E24, E27
Klíčová slova	míra nezaměstnanosti, predikce nezaměstnanosti, nowcasting, Česká republika, online data, volná pracovní místa, hledání práce, kraje

E-mail autora	`krizovi4@seznam.cz`
E-mail vedoucího práce	`ladislav.kristoufek@fsv.cuni.cz`

Contents

List of Tables

List of Figures

Acronyms

ADF Augmented Dickey-Fuller

AFT Accelerated Failure Time

AR Autoregressive

ARIMA Autoregressive Integrated Moving Average

ARIMAX ARIMA with additional explanatory variable(s)

ARMA Autoregressive Moving Average

ARMAX ARMA with additional explanatory variable(s)

ARPA Advanced Research Project Agency

ARPANET ARPA Net

BIC Bayesian Information Criterion

BLS Bureau of Labor Statistics

BP Breusch-Pagan

CESNET Czech Education and Scientific NETwork

CPS Current Population Statistics

CTU Czech Technical University in Prague

CZSO Czech Statistical Office

DM Diebold-Mariano

EU European Union

GDP Gross Domestic Product

GLS Generalised Least Squared

GTAI Google Trends Automotive Index

HAC Heteroskedasticity and Autocorrelation Consistent

HPI House Price Index

ILO International Labour Organization

KPSS Kwiatkowski-Phillips-Schmidt-Shin

LFSS Labour Force Sample Survey

MA Moving Average

MAE Mean Absolute Error

MSE Mean Squared Error

NSFNET National Science Foundation NETwork

NUTS Nomenclature of Units for Territorial Statistics

OLS Ordinary Least Squares

RMSE Root Mean Squared Error

SUR Seemingly Unrelated Regression

SW Shapiro-Wilk

U.S. United States

WWW World Wide Web

2SLS Two-Stage Least Squares

Bachelor's Thesis Proposal

Author	Kristýna Křížová
Supervisor	doc. PhDr. Ladislav Krištoufek, Ph.D.
Proposed topic	Utilizing Online Data in Modelling Unemployment Rates in the Czech Republic

Research question and motivation Unemployment is a big issue in many countries. Some countries have problems with high unemployment rate. In Czech Republic we have the opposite problem. Czech unemployment rate got so low in the past three years, that there are not enough people for job's positions.

The goal of this paper is to study unemployment across 14 regions of the Czech Republic. Also, an employee working at jobs.cz, who is dealing with the labour market and human resources, will cooperate with the production of this work by offering their data. So, statistics of the job search on the web page (jobs.cz) will be mentioned. The statistics are about the number of ads on the website and the number of responses to vacant positions.

The unemployment has been a popular subject of study. For instance, the paper: Modelling of unemployment duration in the Czech Republic (Čabla, Malá, 2017). The subject of the paper was examining the duration of unemployment in the Czech Republic in the three selected years (2008, 2010 and 2014).

Inspiration for this paper was the thesis from school year 2013/2014 with name: Unemployment in the Czech Republic and Job Search on the Internet, which is about analysing historical data about search queries from Seznam.cz related to the job search. In my thesis, I am investigating progression in unemployment rate over time and focusing on following research questions:

- How does unemployment develop over the years?
- What are the differences of unemployment according to regions?
- Are the statistics useful for estimating rate of unemployment?

Contribution The thesis should bring new results and information considered the unemployment in regions of the Czech Republic. It should contribute better understanding of statistical information about our unemployment. Moreover, the thesis should bring information about researching jobs on the websites. In further analysis I may discover usability of the statistics for estimating rate of unemployment.

Methodology I will use data from the Czech Statistical Office, the Ministry of Labour and Social Affairs for analysing the unemployment. For researching jobs, I will use selected data from Jobs.cz (the Czech website). The data will be used subsequently in OLS model.

Outline

1. Introduction
2. Unemployment in the Czech Republic
   - 2.1 Comparison of Czech and European unemployment
   - 2.2 Reasons and development of unemployment
3. OLS model
   - 3.1 Data collection
   - 3.2 Representation of data
   - 3.3 Results
4. Conclusion

Bibliography:

ČABLA, Adam, MALÁ, Ivana, Modelling of unemployment duration in the Czech Republic. Prague Economic Papers, 2017, 26(4): 438-449. Available at: https://doi.org/10.18267/j.pep.620

GITTER, J. Robert, Scheuer, Markus, Unemployment in the Czech Republic. Monthly Labor Review, August 1998, 31. Available at: https://heinonline.org/HOL/LandingPage?handle=hein.journals/month121div=85id=page

KRIŠTOUFEK, Ladislav, PAVLÍČEK, Jaroslav, Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries. PLoS ONE, 2015, 10(5): e0127084.doi:10.1371/journal. pone.0127084

D'AMURI, Francesco, MARCUCCI, Juri, The predictive power of Google searches in forecasting unemployment. Temi di discussion (Economic working papers), 2012, 891, Bank of Italy, Economic Research and International Relations Area

Unemployment statistics. The Czech Statistical Office [online]. Available at: https://www.czso.cz/

Unemployment statistics. Integrated portal MPSV (English: The Ministry of Labour and Social Affairs) [online]. Available at: http://portal.mpsv.cz/sz/stat/nz

Chapter 1

Introduction

Unemployment prediction is a substantial issue in many countries as it is one of the main parts of the macroeconomic problems. Most data sources and statistics used in economics are commonly available only after a substantial lag. It hampers the effectiveness of real-time predictions. Therefore, the thesis investigates whether online data provide a higher accurate way how to predict future movements of Czech unemployment rate.

Unemployment rate has different value pursuant on economics of the state concerned. Development countries, such as Germany, France, the United States; and developing countries, e.g. Brazil, South Africa, Pakistan, have a very different level of unemployment. However, that is not a surprise as poverty or economic crisis affect the economy of a whole state significantly.

The situation in the Czech Republic is a little bit different than in most countries. Nowadays, respectively in the last 3 years, the unemployment rate has been falling, i.e. there are more job vacancies than there are people available for job positions. In December 2018, the unemployment rate in the Czech Republic was 3.1%, which is 2.1% less than it was in December 2016. And this change is in this case quite considerable. The number of people looking for job declined in 2018, nevertheless, the number of job vacancies more than doubled.[1]

The unemployment rate often changes every month, mostly it fluctuates. However, as we mentioned in the first paragraph, it is often published with time delay and macroeconomic indicators (e.g. GDP, unemployment rate, etc.) should be known with enough time ahead. The serious delay causes the impossibility of timely and accurate future estimates of these indicators. Punctual

[1] According to the *Integrated Portal of the Ministry of Labour and Social Affairs.*

unemployment predictions are needed to make efficient government decisions as it is an important aspect for performance of economy. Fluctuating of unemployment has impact on industrial production, movement of wages, and inflation, therefore timely predictions make better estimation of economic situation for government.

In order to figure out the issue about time delay, economic researchers started to use nowcasting as standard measures of prediction. It has been used in meteorology for a long time and recently it has become popular in economy for preventing delayed published data. The term nowcasting is defined as "the prediction of the present, the very near future and the very recent past" (Banbura *et al.* (2010)). Process of nowcasting is functional when it is working with monthly information, e.g. GPD or unemployment rate. Monthly data are a key part of the whole process in order to make the information with small delay or even no delay. This method is often applied in economics and finance, e.g. to nowcast the annual growth rate of Gross Domestic Product (GDP). It is shown in an article about nowcasting Chinese GDP (Yiu & Chow (2011)), where the authors used a large data set that contained economic and financial data with several categories to determine the number of common factors in a factor model. Similarly, articles concerning U.S. and Norwegian GDP nowcasts apply a system of three commonly used model classes (Aastveit *et al.* (2011)) and a dynamic factor model (Aastveit & Trovik (2012)), respectively. Modugno (2011) proposed in his article a methodology using data even with a daily sampling frequency to nowcast inflation. Nowcasting frequently works with online data, mostly with search activity on queries entered into Google. Choi & Varian (2009a;b; 2012) dealt with a usage of search engine data to predict simultaneous values of macroeconomic indicators. Wu & Brynjolfsson (2009) described a usage of data from Google's search to predict housing market trends.

Most data sources used in economics are commonly available only after a substantial lag, which makes predictions about macroeconomic factors inaccurate. Online data have been proven to be a useful source for solving forecasting problems, so we also use online data in this thesis to improve unemployment prediction in the Czech Republic. Unemployment forecasting has been frequently applied since 2008, when Stevenson wrote a published work about the role of the Internet in job search activity in U.S. After that, there were other papers that examine whether online search activity enhances accuracy of predicting a real economy indicator such as the unemployment rate: Askitas &

Zimmermann (2009), Suhoy (2009), Simionescu (2015), or Tuhkuri (2016).

The main aim of the thesis is to provide easier and better prediction of movements in the unemployment rate in the context of the Czech economy. In addition, the study investigates the relationship between the Czech unemployment rate and job search on the Internet by users interested in changing/improving their current job positions or finding job vacancies in order to become employed. Specifically, we work with freely available data from the website of Integrated Portal of the Ministry of Labour and Social Affairs where everyone can find detailed information about the unemployment processed into many tables. Moreover, we use data from the portal Jobs.cz, one of the most popular Czech job search portals, which provides us with job vacancies and responses to them. We are interested in data where job seekers are aged between 15 and 64.

We used obtained data to create a simple autoregressive model of the unemployment rate. To evaluate the prediction of the Czech unemployment, we consequently add extra variables containing data from the job search portal Jobs.cz to the model and examine whether the extra variables improve predictive ability of the model. In other words, we determine their usefulness for unemployment forecasting in the Czech Republic. In addition to the augmented autoregressive model of the Czech Republic, we estimate the same models for 14 regions of the Czech Republic separately.

Our results indicate that data from the job search portal Jobs.cz enhance nowcasts of the Czech unemployment rate as well as fundamental models with basic relationship between the unemployment rate and data on number of job vacancies and responses to them. However, we find that the job-related data do not improve predictive ability of the unemployment rate.

The thesis is structured as follows. Chapter 1 presents brief introduction of the thesis. Chapter 2 discusses theoretical background divided into two main topics of the thesis: Unemployment rate and Utilizing online data, which are highly correlated. Chapter 3 covers the related literature and published studies (papers) related to the topics. In Chapter 4, we describe the utilized and possible data and mention their resources. Chapter 5 provides methodology applied to analyse the dataset used in the thesis. In Chapter 6, we elaborate on empirical results of our study and Chapter 7 summarises the whole thesis and suggests potential continuation of further research.

Chapter 2

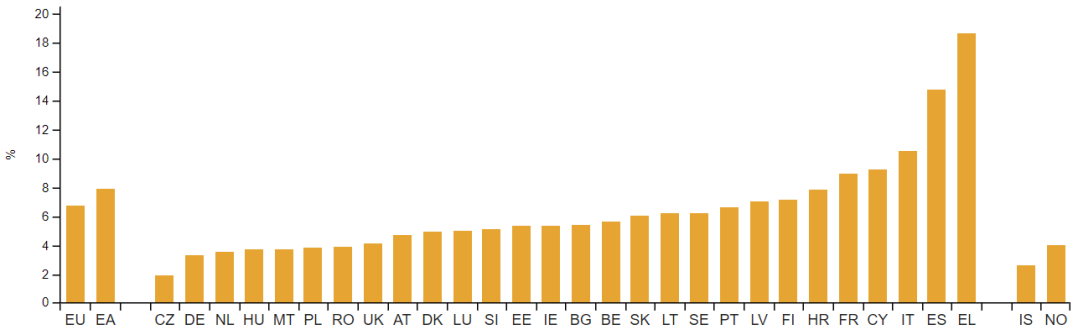
Theoretical Background

2.1 Unemployment in the Czech Republic

The Czech Republic is an industrially advanced country and belongs to the most developed economies in the world. The basis of this stable and very prosperous economy is primarily industry, services, manufacturing and innovation. According to Deloitte economists, it is expected that the economy in 2019 will still grow at 2.2% and inflation will remain above 2%.[1]

Therefore, it is not a surprise that the Czech unemployment rate is very low, even one of the lowest in Europe. In November 2018, unemployment in the Czech Republic was 1.9%, which was the totally lowest in the EU.[2]

Figure 2.1: Unemployment rates, seasonally adjusted, November 2018 (%)



Source: Eurostat Statistics Explained: Unemployment statistics

[1]Deloitte: Czech Economy in 2019: Further Slowdown and a Shortage of Workforce [online]. Available from: https://www2.deloitte.com/cz/en/pages/press/articles/cze-tz-ceska-ekonomika-v-roce-2019-zpomalovani-rustu-a-nedostatek-zamestnancu.html

[2]Eurostat Statistics Explained: Unemployment statistics [online]. Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Unemployment_statistics

Nowadays, the number of people in the Czech Republic without job is the lowest since 1997. Looking at October 2018, the share of unemployed people was 2.8% and jobs were searched by the least number of people since June 1997. There were 215 622 people without job and the number of job vacancies was 316 900.[3]. As reported by the Labour Offices' representatives, it is more attainable to find work for people who have lost motivation and are not interested in working. In October 2018, the number of people looking for job was about 8700 less than in September and nearly about 56,000 less than a year ago. There is also a decrease in the share of people who are unemployed for more than 12 months. The unemployment rate in the Czech Republic is already so low that it cannot decline anymore, which means that in 2019 the rate will probably stay slightly below 3%.[4]

As mentioned above, the Czech Republic has one of the lowest average values of an internationally comparable unemployment rate in the European Union. However, the situation varies among regions. Differences can also be found in gender classification, especially in socio-economic classes. In the Czech Republic there are 14 greater territorial self-governing units, 13 of these are regions (Central Bohemian Region, South Bohemian Region, Plzeň Region, Karlovy Vary region, Ústí Region, Liberec Region, Hradec Králové Region, Pardubice Region, Vysočina Region, South Moravian Region, Olomouc Region, Zlín Region, Moravian-Silesian Region) and one is the City of Prague, which is the capital city of the Czech Republic.

From a territorial point of view, unemployment is unevenly distributed. The individual regions of the Czech Republic have different levels of unemployment. The differences are mainly caused by lower production of heavy industry (in Northern Bohemia and Northern Moravia) or lower production of agriculture or textile industry.

In contrast to the situation of large European cities, the Czech and Moravian cities (Prague, Brno, České Budějovice, Plzeň, Hradec Králové, Zlín) are characterized by quite low unemployment. For instance, the unemployment

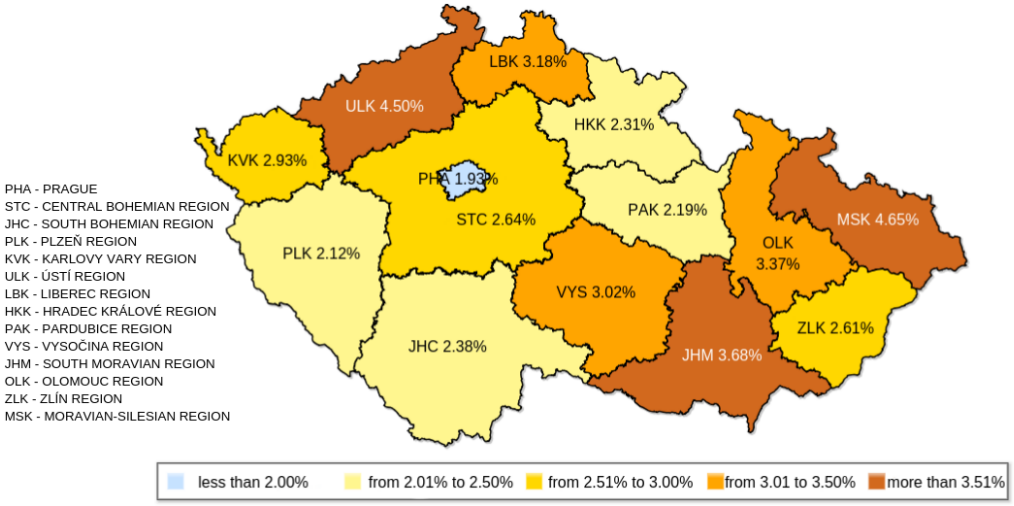
[3] According to the *Integrated Portal of the Ministry of Labour and Social Affairs.*

[4] *Tisková zpráva: Pokles nezaměstnanosti pokračoval i v říjnu* [online]. 08.11.2018, 8. Available from:
https://portal.mpsv.cz/upcr/media/tz/2018/11/$2018_1 1_0 8_t z_n ezamestnanost_r ijen_2$018.*pdf*

rate of Berlin was 7.6%[5] in November 2018, and Madrid had 11.5%[6] unemployment rate as of 31.12.2018.

Figures 2.2 and 2.3 are sufficient for a better understanding of the distribution of unemployment in the Czech Republic. Figure 2.2 shows the unemployment rates of individual regions as of 31.12.2018 in the map of the Czech Republic. The regions are divided exactly how they are geographically situated. As we can see, the lowest unemployment rate is in the City of Prague, then in Plzeň Region and Pardubice Region.

Figure 2.2: The share of unemployed people aged between 15 and 64 in individual regions of the Czech Republic as of 31.12.2018



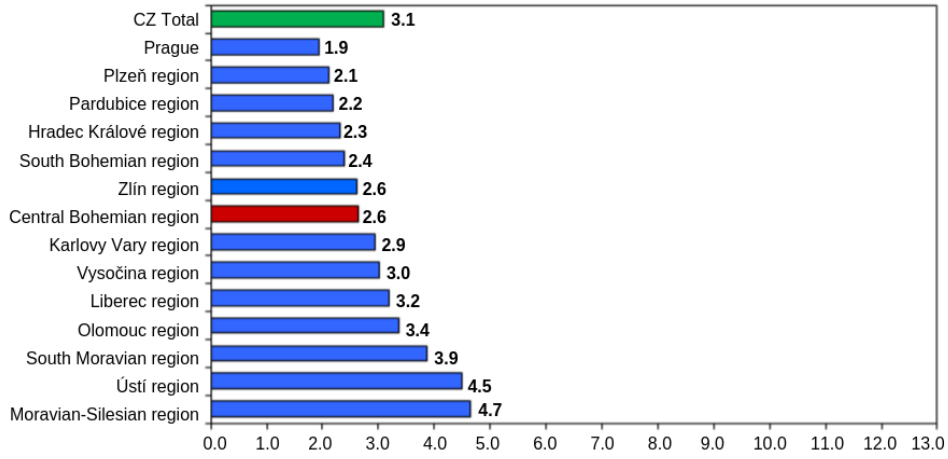
Source: The Czech Statistical Office

Figure 2.3 presents also the unemployment rates of individual regions of the Czech Republic as of 31.12.2018, however, now it is shown in the bar chart. We can notice the same values as in the map, but it is organized from the region with the lowest degree of unemployment, to the region with the highest degree of unemployment rate.

[5] *Arbeitslosenquote in Deutschland nach Bundesländern. Statista: Das Statistik-Portal (Statistiken und Studien aus über 22.500 Quellen)* [online]. Available from: https://de.statista.com/statistik/daten/studie/36651/umfrage/arbeitslosenquote-in-deutschland-nach-bundeslaendern/

[6] Autonomous Communities of Spain Unemployment rate: Madrid - Unemployment rate. Countryeconomy.com [online]. Available from: https://countryeconomy.com/labour-force-survey/spain-autonomous-communities/madrid

Figure 2.3: The share of unemployed people in individual regions of the Czech Republic as at 31.12.2018 in (%)



Source: The Integrated portal of the Ministry of Labour and Social Affairs

Table 2.1: The unemployed aged 15+ years and their structure by region (2010-2017)

Thousand person

Territory	2010	2011	2012	2013	2014	2015	2016	2017
	Total	Total	Total	Total	Total	Total	Total	Total
Region								
Prague	25.6	23.7	20.9	21	16.5	18.8	15.2	12
Central Bohemian	33.3	32.6	30.3	34.4	34.3	23.2	20.9	14.5
South Bohemian	16.8	17.5	17.8	16.2	18.7	12.6	8.8	7.1
Plzeň	17.1	15	14.1	15.4	15	11.3	10.2	5.8
Karlovy Vary	17.5	13.2	16.2	16	14	10.4	8.3	5.1
Ústí	45.4	39.6	42.7	37.8	34.1	30	20.7	13.9
Liberec	15.1	15.4	20	17.7	14	11.9	9.6	8
Hradec Králové	18.7	19.3	19.3	22.5	16.8	15.4	11.2	6.2
Pardubice	18.4	14	19.9	22.1	16.8	12	9.7	7.2
Vysočina	17.7	16.1	15.8	17.1	14.1	11.7	8	6.9
South Moravian	44.4	43.3	47.4	40.6	36.2	29.8	23.2	19.9
Olomouc	27.7	23.1	24.1	28.2	23.5	18.2	11.5	9.8
Zlín	24.5	22.2	21.3	20.2	17.8	13.9	11.7	10.5
Moravian-Silesian	61.4	55.5	57.1	59.7	51.9	48.8	42.3	28.8

Source: The Czech Statistical Office

In Tables 2.1 and 2.2, we can see the number of unemployed aged 15+. Table 2.1 shows the distribution of the unemployed according to the regions in the Czech Republic. We can notice that the Moravian-Silesian Region has unambiguously the most unemployed people in all 7 mentioned years.

Table 2.2: The unemployed and their structure by educational attainment and age group

Thousand person

Indicator	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
	Total	**Total**	**Total**	**Total**	**Total**	**Total**	**Total**	**Total**
The unemployed	383.7	350.6	366.9	368.9	323.6	268	211.4	155.5
Educational attainment								
Primary education	79.5	70.4	83.4	71.5	58.1	58.7	54	33.4
Secondary education								
- *without A-level examination*	174	155.9	157.3	163.3	138.2	108.5	80.4	61.9
- *with A-level examination*	104.5	96.3	96	102.9	94	72.1	53.3	40.6
Higher education	25.6	27.6	30.2	31.3	33.3	28.7	23.6	19.4
Age group (years)								
15–19	16.1	15.5	16.2	12.7	12.6	9.7	6.8	6.5
20–24	57.3	51.3	56.5	56.1	43.9	34	27.8	18.6
25–29	56.6	45.5	50.7	46.8	42.2	37.5	30.6	19.2
30–34	45.9	43	43.4	47.4	41.5	35.7	26.4	17.5
35–39	43	43.2	46.8	52	45.4	36.3	24.3	22.3
40–44	36.6	34.3	35.2	37.2	33.5	29	24.4	18.7
45–49	37.8	35.8	34.7	34.7	33.2	230	17	14.3
50–54	40.9	36.9	35.8	34.4	30.4	26.3	21.6	17
55–59	40.5	37	37.2	36.4	30.8	28	22.7	14.7
60–64	7.6	7	8	9.9	9	7.3	8.9	5.6
65+	1.1	1.1	2.4	1.4	1.1	1.2	1	1.2

Source: The Czech Statistical Office

Table 2.2 does not show the distribution in regions, but it devides the unemployed according to educational attainment and age group. With educational attainment, it is not surprising that we find the lowest number in higher education, because it is naturally the most desirable.

2.2 Utilizing online data in the Czech Republic

The history of the Internet begins in the early 1990s. The first Internet predecessor was created by the Advanced Research Project Agency (ARPA) under the patronage of the U.S. Department of Defense in 1969. The network was called ARPANET. The big problem was communication on many different platforms. That was the reason why an intensive research took place in this area and its result (in 1983) was the TCP/IP protocol, which is used today. In

1986, the National Science Foundation Network (NSFNET) was created and it replaced the ARPANET network in 1990.

The origin of today's Internet is dated back to 1989 when Englishman Tim Berners Lee invented a new way of exchanging information, known as the World Wide Web (WWW), which was designed specifically for the NeXT computer at CERN.

While the history of the global Internet is dated to the 1960s, the history of the Internet in the Czech Republic dates to the early 1990s. In October 1990, the Internet was connected to the ERAN (European Academic and Research Network) in the Czech Republic. The date of the first connection of the Czech Republic to the Internet is February 13, 1992, which was in CTU (Czech Technical University in Prague). In 1991, the first Czech nationwide network has been created and named CESNET (Czech Education and Scientific NETwork).[7]

The emergence of the first commercial providers of Internet access was in 1995. While the official Internet connection in the Czech Republic was dated to 1992, it was possible to connect only through the CESNET (Czech Education and Scientific NETwork, the first Czech nationwide network) for the next 3 years. The main reason was the fact that Eurotel had a state monopoly on the provision of data services. Until the abolition of this monopoly, the first commercial providers of Internet access started to emerge, and thus it came to the great development of the Czech Internet.[8]

Initially, the Internet access was limited only to people owning desktop computer. However, subsequent technological and commercial developments led to a broader range of devices that people could use to go online. These developments meant that people did not have to be fixed on a desktop computer at home or at work, but they could have Internet access on the move. This trend became popular, which resulted in an expansion of Internet use.

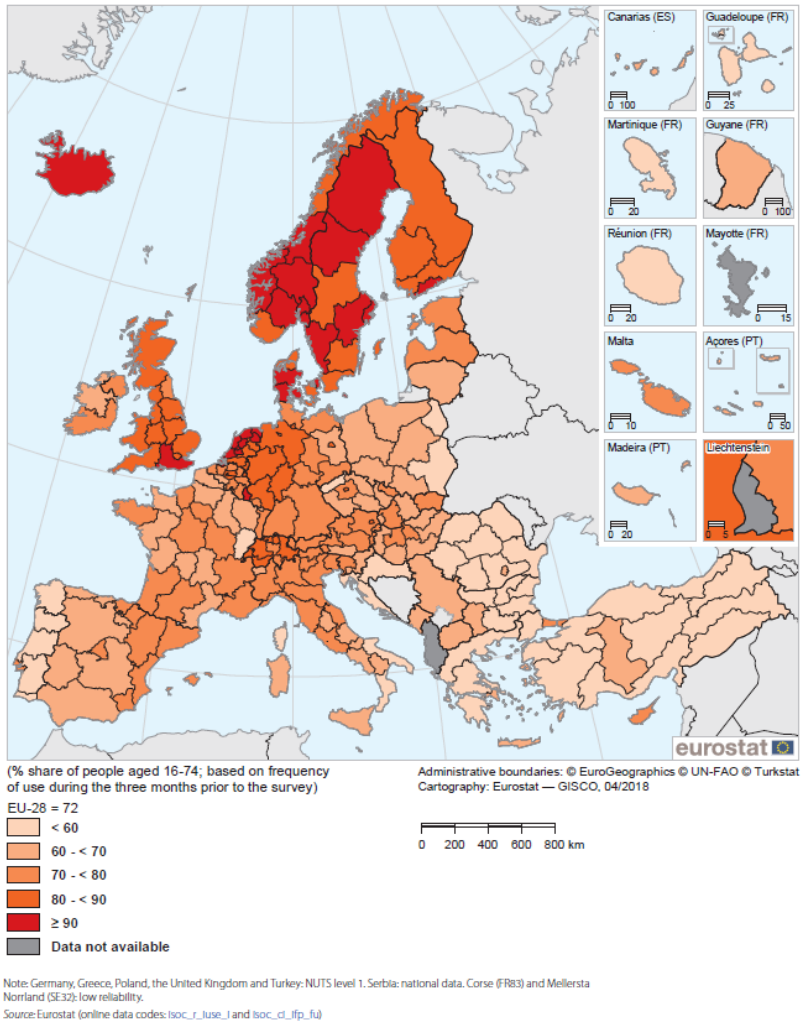
Figures 2.4 and 2.5 show the proportion of Internet use and broadband access in member states of the EU (% share of people aged 16-74; based on frequency of use during the three months prior to the survey). Both Figures analyse the results by NUTS level 2 regions, except Germany, Greece, Poland and the United Kingdom, where the data are related to NUTS level 1 regions

[7]Historie Internetu v datech. *SCIENCEmag.cz* [online]. 7. 2. 2017. Available from: https://sciencemag.cz/historie-internetu-v-datech/

[8]CHLAD, Radim. *Historie Internetu v České republice* [online]. 2000. Available from: https://www.fi.muni.cz/usr/jkucera/pv109/2000/xchlad.htm. Masarykova Univerzita.

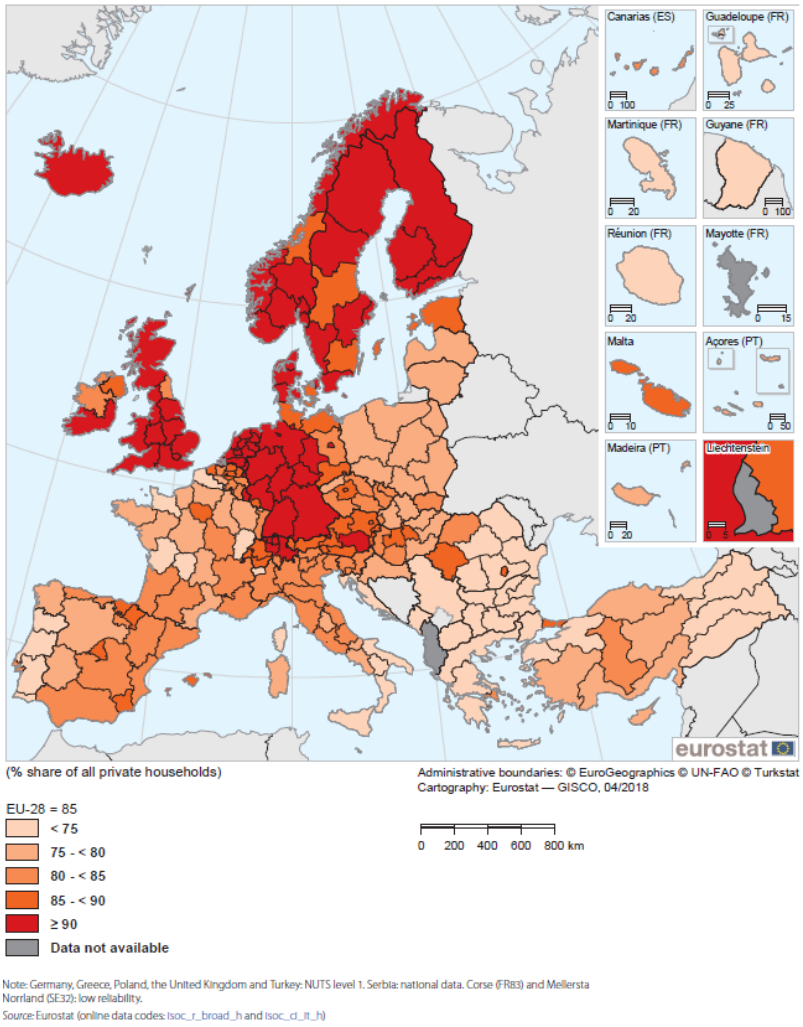
(NUTS is the Classification of Territorial Units for Statistics). From Figure 2.4, we can see that the highest shares of people using the Internet every day were in regions across the Netherlands and the Nordic Member States. Just less than three quarters (72%) of the EU-28 population (aged 16-74 years) used the Internet daily in 2017. In Figure 2.5, we can notice that more than half of all households had broadband access at home. The highest shares (89%) of households with broadband access in the eastern regions of the EU were recorded in the capital city regions in Hungary and in the Czech Republic.

Figure 2.4: Proportion of daily internet users, by NUTS 2 regions, 2017



Source: The Czech Statistical Office

Figure 2.5: Proportion of households with broadband access at home, by NUTS 2 regions, 2017



Source: The Czech Statistical Office

Utilizing online data in the Czech Republic have mainly started in 1996, when Seznam.cz was founded.[9] This foundation made online searches easier, because it allowed not only a general search for information, but also a news website, free e-mail service and a map server. Lots of Czech Internet users chose Seznam.cz as their homepage due to these advantages. Seznam.cz was and still is very important for the Czech Internet. However, as time went on, its popularity dropped when Google's search has started to get to the fore. It offered a much more complex search with a large quantity of information available from various websites. Similarly to Seznam.cz, Google had its own free

[9] O Seznamu. *Seznam.cz* [online]. Available from: https://o.seznam.cz/

e-mail service and other benefits, such as a shared calendar, shared documents, possibility of data storage, etc.

To search information about the unemployment rate in the Czech Republic nowadays, we use special portals focusing on the issue. In the case of unemployment rate and subsequent job search, the main online sources are Integrated Portal of the Ministry of Labour and Social Affairs of the Czech Republic and portals Jobs.cz and Prace.cz. The job search portal Jobs.cz offers a wide range of job vacancies and closely cooperates with the portal Prace.cz – it means that Prace.cz recommends job vacancies, which are less paid and do not have high requirements, and conversely better job vacancies with higher salary, but more stringent conditions, occur on Jobs.cz.

As the unemployment rate is accessible for the public, it can be noticed that it varies over time in accordance to lots of circumstances (e.g. financial crisis, significant decline/rise in population, etc.); and number of available vacancies is changing along with the unemployment. The ideal relationship between unemployment rate and job vacancy rate is described by Beveridge (1944). The Beveridge curve describes the negative relationship, i.e. with high unemployment, the number of vacancies is low (e.g. during recessions), and conversely with low unemployment, the number of vacancies is high (e.g. during expansions). The curve is from the macroeconomic point of view significant as the position of the economy on the curve that gives an idea about a situation on the labour market.

Chapter 3

Literature Review

Unemployment is defined as a situation where a person does not have a job, is actively looking for some job vacancy but cannot find any. It is important to mention that the person is currently able to work, meaning he/she does have no disabilities. Mothers on maternity leave are not classified as unemployed as they are not actively looking for a job. The special group is voluntary unemployment, where are people who decide to leave their job or refuse a job offer because of poor working conditions, low salary, etc. [1]

As a big macroeconomic concept, there are many studies having the unemployment as a main topic. The issue of unemployment is viewed from various points of view. Some studies dealt with a development of unemployment over several decades, which is helpful for economists to derive the future unemployment rate, and possibly to determine a situation on labour market and the whole economy. Other studies investigate causes and consequences of unemployment and look for a solution about avoiding and minimizing these as much as possible. Unemployment has some relationship with other factories, such as inflation or job vacancies. The negative relationship between unemployment and inflation is reflected in the Phillips curve (Phillips (1958)). The Beveridge curve, mentioned in the previous chapter, describes the inverse relationship between unemployment and job vacancies.[2]

Generalisation of unemployment is not suitable, as it differs across countries depending on a size of the country, a quality of the economy, a situation of the labour market, etc. Therefore, we have many studies comparing unemployment in different countries within a single continent or across continents. European

[1] According to the *Czech Statistical Office.*

[2] MANKIW, N. Gregory. Macroeconomics. Eighth Edition. New York: Worth Publishers, 2013. ISBN 978-1-4292-4002-4.

Union countries' unemployment rate is compared with the unemployment of the European Union as a whole, from which various tables and graphs are created and subsequently used as a source for further studies.

3.1 Use of online data in Economics

Thanks to Google's growing popularity, it created a website Google Trends that analyses the latest news and statistics, the popularity of top queries in Google's search and adds current trends from all corners of the world. The website is primally based on graphs, which help to compare the search volume of different queries over time. On 5th August 2008, Google has introduced Google Insights for Search that displays search trends data in more elaborated and advanced way. It studies the trend of searches in different fields using keyword search, which is used as the main variable, to create and estimate econometric models.

One specific field is Google Econometrics. It uses data from Google Trends and subsequently creates econometric models. This technique has become very popular and has been used for numerous issues, for example: portfolio risk (Krištoufek (2013): *Can Google Trends Search Queries Contribute to Risk Diversification?*) or stock market moves (Preis *et al.* (2013): *Quantifying Trading Behavior in Financial Markets Using Google Trends*).

Wu & Brynjolfsson (2009) described a usage of data from Google's search to predict housing market trends. The authors used Google Trends in their study of a relationship between the search index and the housing market indicators – the volume of housing sales and the house price index (HPI). A simple seasonal autoregressive (AR) model is applied to estimate the relationship. In addition, in terms of control for the influence of any time invariant properties (e.g. demographics of a state, any state-wide policies affecting real estate purchase decisions), the authors used a fixed-effect method to the models. Wu & Brynjolfsson (2009) found out that the housing search index had strong ability to predict the future housing market sales and prices.

Kholodilin *et al.* (2010) explored whether Google search query data can help in nowcasting of US monthly private consumption. The study compared Google-based forecasts with forecasts based on an autoregressive benchmark (AR(1)) model and with models including survey-based indicators and financial variables using real-time data set. Moreover, as multiple search queries are investigated, the authors utilized principal components analysis to extract a

reduced number of the data. The results showed that Google searches significantly improved the nowcasts of US private consumption.

Yiu & Chow (2011) used a large data set that contained economic and financial data with 189 indicator series of several categories (e.g. prices, industrial production, fixed asset investment, external sector, money market and financial market) for nowcasting of the annual growth rate of Chinese quarterly GDP. The study applied a factor model proposed by Giannone *et al.* (2005) on this large Chinese data panel from January 1998 to June 2009. The authors also applied Bai & Ng (2002) to determine the number of common factors in the factor model. The empirical results pointed out that the lower dimensional factor model was (e.g. with only two common factors), the more likely it was to produce useful nowcasts for GDP growth rate in China. In addition, the authors found that interest rate data used in the factor model were the most important block in estimating China's GDP. However, the interest rate data were not the only one important block. The other blocks were consumer and retail prices data and fixed asset investment indicators.

Modugno (2011) proposed in his article a methodology using data even with a daily sampling frequency to nowcast and forecast inflation. In order to produce an accurate estimate of inflation for the current and following months, the author used data that contained the World Market Price of Raw Materials and energy prices for the euro area and the United States. The data set ranged from April 1996 to December 2009. The article applied a trading day frequency factor model on the data set. The author concluded that the data improved forecast accuracy over models that used data available only at monthly frequency for both euro area and the United States.

Carrière-Swallow & Labbé (2013) examined usefulness of Google search queries for nowcasting of automobile sales in Chile's emerging market. The study was based on a simple nowcasting model that included a Google Trends Automotive Index (GTAI). The author came out with a novel procedure of downloading the Google Trends series and computing the average for each series. This procedure reduced the bias incurred by Google's sampling method. Due to the absence of search query categories for Chile, the authors fit the Google Trends series to a linear model and allow the weights to take on any value on the real line. Even though the population in Chile used the Internet in a quite low rates, the authors concluded that models including the GTAI improved fit and efficiency of automobile sale's nowcast.

3.2 Use of online data and nowcasting in unemployment

There is a close link between unemployment and job search, where the latter currently uses Google's search as the main source. An issue job search and predicting unemployment was firstly published by Ettredge *et al.* (2005), where the authors investigated the potential of using web job search data for predicting macroeconomic statistics - unemployment. In this case, the data were obtained from WorldTracker's Top 500 Keyword Report, not from Google Trends. The study is working with six most common job-search terms, taking values from their daily search volumes and calculating short-term and long-term usage rate. The data about unemployed workers in the United States (U.S.) were extracted from the website of the Bureau of Labor Statistics (BLS). Because of limited data, the authors did not use time-series models, but they estimated single-variable regressions to investigate whether unemployment level is associated with search-term usage. The results of the regressions showed a positive, significant relationship between job-search variables and the number of unemployed in the US. For the future research, the authors suggested to obtain data from large job portals in order to gain less limited data and use search terms for other important macroeconomic predictions.

Unemployment forecasting has been frequently applied since 2008, when Stevenson wrote a published work about the role of the Internet in job search activity in the United States. Stevenson (2008) examined how the Internet impacted job search behaviour using data from Current Population Statistics (CPS) Computer and Internet Use Supplements. The author pointed out that using the Internet had increased from effectively zero to 70% of the population in the past ten years. It caused that the diversity of job search methods applied by the unemployed had risen and job search behaviour had become more extensive. Moreover, the Internet had led to reallocation of various job search activities (e.g. looking at ads, contacting an employer directly, etc.). Stevenson (2008) figured out that the bigger amount of information is available about a given job for unemployed, the better they target for job search activities. While the role of the Internet in job search activity had risen, there is no clear evidence that the Internet could reduce unemployment duration. The author concluded that currently employed are those who use the Internet for job seeking purposes the most, as they are more likely to make an employment-to-employment tran-

sition. As a result of the conclusion, Stevenson (2008) suggested that future research could investigate whether the Internet is affecting wage compression.

Another study examining the role of the Internet in job search activity is the article written by D'Amuri (2009). The article examined statistical significance of a new indicator based on job search related web queries in forecasting unemployment in Italy. The author used data from various sources. The Italian Labor Force Survey was the main source of official data on unemployment, which is available only on a quarterly base. The new indicator, and also explanatory variable, was Google Index that is the impact of job-related queries ("job offer") over total queries. To unify data, Google Index was averaged quarterly. D'Amuri (2009) concluded that Google Index is a significant indicator that could predict unemployment rate. Furthermore, the author found out that models estimated on small samples with Google Index perform better than large samples without Google Index.

Askitas & Zimmermann (2009) investigated usefulness of the Google's search query data for predicting German monthly unemployment rate from January 2004 to April 2009. They measure Google activity along the division of 4 groups of keywords in German language: "unemployment office or agency", "unemployment rate", "personal consultant or consultancy" and "most popular job search engines in Germany", using Google Insights for Search as their source. The study is working with data about Google activity divided into weeks 1 and 2, and weeks 3 and 4 of each month. The authors created several econometric models where they used combinations of keyword groups and mentioned time periods. Their finding based on Bayesian Information Criterion (BIC) showed that Google's search query data of weeks 3 and 4 from the previous month are statistically acceptable and that it could be used for prediction purposes of German unemployment rate in the current month.

The study *Predicting the Present with Bayesian Structural Time Series* written by Scott & Varian (2013) did not exactly investigate using job search data for creating and estimating econometric models about unemployment, however, it described robust and automatic system for selecting predictors in a nowcasting model, which could be used for nowcasting unemployment rates. The automatic system utilized time-series models to capture the trend, seasonality, and similar components of the target series. The authors combined three Bayesian techniques: Kalman filtering, spike-and-slab regression, and model averaging. These techniques were used for modelling initial claims for unemployment benefits and for retail sales.

Simionescu (2015) with her paper named *The improvement of unemployment rate predictions accuracy* examined during years 2001-2014 whether three anonymous forecasters (F1, F2 and F3) could provide sufficient predictions of Romanian unemployment rate. The author figured out that F3 provided the most accurate forecasts, followed by F1 and F2, respectively, according to multi-criteria ranking and assessment of five accuracy indicators (U1 and U2 Theil's Statistics, mean errors, mean squared error, root mean squared error). Simionescu (2015) concluded that the combined forecasts of forecasters' predictions based on Hodrick-Prescott filter, or more precisely Holt-Winters technique, were the best strategy to improve the accuracy of unemployment predictions in Romania during 2001-2014 sales.

Tuhkuri (2016) examined forecasting unemployment rate in the United States with Google searches. The primary data sources were the Google Trends database by Google Inc. The data about unemployment in U.S. were extracted from the website of the Bureau of Labor Statistics (BLS). In order to investigate prediction of the unemployment rate, the author used (pseudo) out-of-sample comparison. It showed that models with Google variables provides on average more accurate forecasts than models without the variables of Google data, i.e. the author found out that Google searches predict unemployment rate in U.S. However, three findings arose: the predictive power of Google searches was limited to short-term predictions; the value of Google data for forecasting purposes was episodic; and the improvements in forecasting accuracy were modest. Overall, the results depicted both the potentials and limitations of using big data (such as Google data) to predict macroeconomic indicators such as unemployment rate.

Studies mentioned above this subchapter focused on unemployment predictions generally. In the following part, we present some researches specialized on unemployment nowcasting in the Czech Republic. First of all, Platil (2014) examined the applicability of Google Econometrics in the Czech Republic. The author used Google search query data for years 2003-2014 to test unemployment, consumer confidence, and overall economic situation. The thesis analysed the contribution of Google data in three related areas: using an autoregressive (AR) model for unemployment, Granger causality test for consumer confidence, and vector auto-regression and logit models for Gross Domestic Product (GDP) and household consumption. Platil (2014) compared out-of-sample nowcasting performance and in-sample fit with control variables in the areas in order to test the contribution of Google data. For the unemployment model, search queries

about job searching process (e.g. "job", "job offer", "Labour Office") was used and added into the AR model independently. The author analysed improvement of 10% in nowcast of the baseline models measured by Mean Squared Error (MSE). The improvement was statistically significant, and even though Google does not dominate the Czech search engine market, the best models in this thesis contained Google data.

The other study about unemployment nowcasting in Visegrad Group (the Czech Republic, Hungary, Poland and Slovakia) was written by Pavlíček & Krištoufek (2015). The authors worked with data from the Eurostat database and Google Trends webpage between January 2004 and December 2013. Their methodology was based on a set of three autoregressive models with different lag structures and used for each country in the same way. The models were augmented with Google Trends data for a query "job" translated into the appropriate languages of the countries. The study showed that the models for the Czech Republic enhanced significantly with every additional number of lags. Overall, the authors concluded that Google searches improved nowcasting models for unemployment rates for the Czech Republic and Hungary. The results for Poland and Slovakia are mixed.

Zacha (2015) examined a relationship between unemployment rate and job search activity by Internet users in the Czech Republic. The thesis used data from the Czech Statistical Office (CZSO), two most popular Czech search engine – Google and Seznam, and from job search portal Jobs.cz. The methodology was based on a simple autoregressive model augmented with search query data and data on numbers of job vacancies and reactions to them. The study tested how useful is online data related to job search for unemployment prediction in the Czech Republic in terms of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The results showed that short-term unemployment nowcasting was slightly improved with data on numbers of job vacancies. However, search query data from Google and Seznam failed to improve prediction of Czech unemployment rate.

One of the novel studies is written by Čabla & Malá (2017), where the authors focused on examining the duration of unemployment in the Czech Republic in the three selected years (2008, 2010 and 2014). The thesis applied interval censored data from the Labour Force Sample Survey (LFSS) conducted quarterly by the Czech Statistical Office (CZSO) to create the Accelerated Failure Time (AFT) regression model using log-normal probability distribution. The model consisted of the following explanatory variables: years (2008, 2010,

2014), gender, education, five-year age groups and municipality size. To overcome the problems with the selection of appropriate probability distributions, the authors used Turnbull's nonparametric estimator of the survival function in terms of evaluation for subsamples defined by the year, gender and education. Subsequently, it was compared with the parametric AFT model. The results of the study showed that the effects of education and gender were quantified and evaluated, where the effect of education was strong positive and more significant than the effect of gender.

Chapter 4

Data

The data used in this thesis come out from different sources. Unemployment rates of the Czech Republic as a whole and of the 14 regions separately have been obtained from the website of Integrated Portal of the Ministry of Labour and Social Affairs. The data on job search activity have been obtained from the Czech job search portal Jobs.cz. In this chapter, all used data sources are described.

Moreover, two possible data sources are defined in this section: Czech Statistical Office, the Czech job search portal Prace.cz. The Czech Statistical Office provides data about the unemployment rates in the Czech Republic as well as the Ministry of Labour and Social Affairs. The portal Prace.cz cooperates with the portal Jobs.cz and also provides posts of job vacancies. The difference between these two Czech job search portals is described as well.

Czech Statistical Office

The Czech Statistical Office (CZSO) is a publicly available source that offers statistical information to state authorities, local government authorities, the public and abroad. It ensures the comparability of statistical information on a national and an international scale. In case of the Czech labour force, it provides several indicators, such as general unemployment rate, employment rate, shared of unemployed, absolute number of employed and unemployed people, and economic activity rate. Moreover, the statistics are broken down by gender and applied to people aged between 15 and 64. All the time series are seasonally adjusted in order to remove seasonal calendar effects.

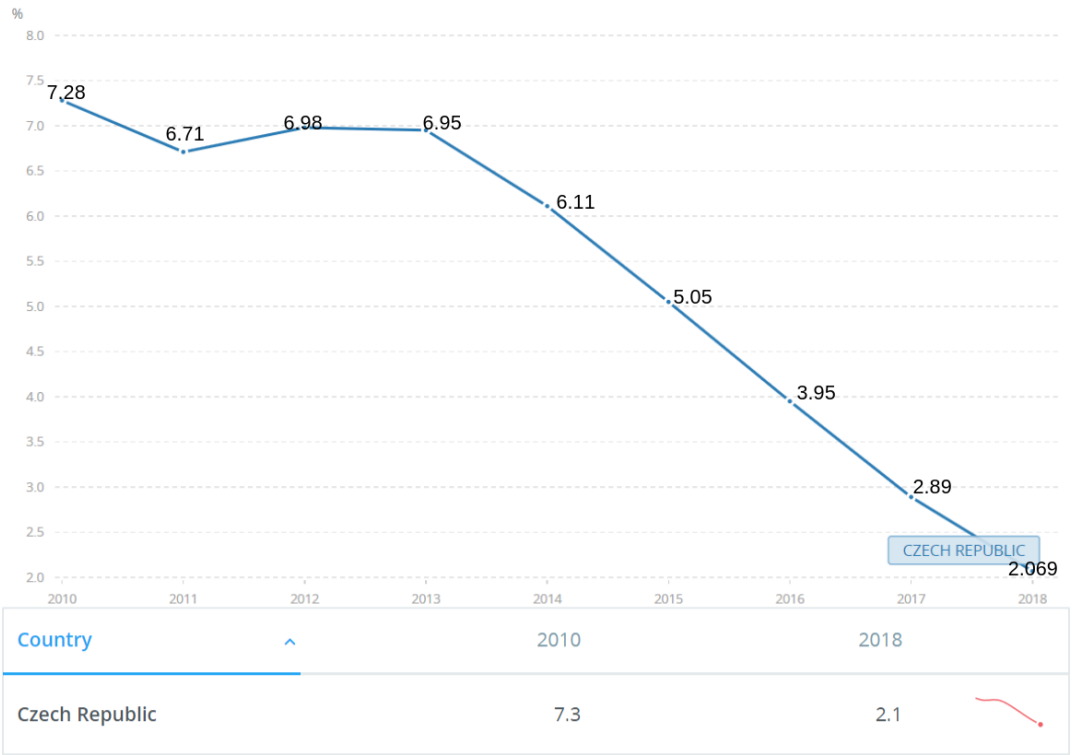
The Czech Statistical Office publishes data on employment and unemploy-

ment mostly from results of the Labour Force Sample Survey (LFSS), which mainly focuses on obtaining regular labour market information. Data result from the Population and Housing Census, final demographic data, and short-term projections. The last census was made in 2011. Data are freely available from January 1993, as CZSO provides equal access to statistical information and identical extent of the published information for all users.

The general unemployment rate for people of age 15-64 in the Czech Republic refers to the percentage of unemployed under the International Labour Organization (ILO) standard within all economically active people. Figure 4.1 depicts the unemployment rate in the Czech Republic between years 2010 and 2018. Specifically, our data range from January 2010 to April 2018.

Although some indicators are known on time, or even earlier, the general seasonally adjusted unemployment rate is published with approximately a month delay. It means that the CZSO publishes this indicator at the end of the following month.

Figure 4.1: The curve of the unemployment rate in the Czech Republic, 2010 - 2018



Source: The World Bank

Ministry of Labour and Social Affairs

The main aim of this thesis is to provide easier and better prediction of movements in the unemployment rate in the context of the Czech economy. The Ministry of Labour and Social Affairs is another publicly available source that offers statistical information about unemployment in the Czech Republic. It cooperates with the Czech Statistical Office, as on its website we can find a direct link to the website of the Ministry of Labour and Social Affairs, which offers statistical yearbooks, basic indicators of labour and social protection in the Czech Republic, information on paid benefits, etc.

The Ministry of Labour and Social Affairs publishes its own statistics of unemployment based on own sources of data. They publish an indicator named "share of unemployed persons in the population", which is a share of people aged between 15-64 registered at the Labour Office to the total population of that age. This series differs from the official unemployment rate of CZSO as the numerator of the fraction contains only people registered as unemployed, which does not follow the general definition of unemployment. Another difference is that the denominator of the fraction contains whole population of the Czech Republic, not only economically active people, as the CZSO counts it.

As it mentioned above, the general unemployment rate is published with approximately a month delay by the Czech Statistical Office. The Ministry of Labour and Social Affairs publishes the share of unemployed persons with an approximately 10-14 days delay. It means that it is published around the middle of the following month.

For our study, we chose monthly share of unemployed people of age 15-64 for the Czech Republic, as well as for the 14 regions separately. For each month, there is a package of data including unemployment rates of individual months, the number of unemployed people, the number of positions registered at the end of the month at the Labour Office. In addition, all data are also available for each individual region. The range of our data is from January 2010 to April 2018.

Jobs.cz

In addition to the main aim of this thesis, the study investigates the relationship between the Czech unemployment rate and job search on the Internet by users interested in changing/improving their current job positions or finding

job vacancies in order to become employed. Data for the explanatory variable come out from portal Jobs.cz, one of the most popular and largest Czech job search portals. Jobs.cz offers mostly highly specialized jobs and requires job seekers to have tertiary education. Job seekers use it to change their job rather than to find a completely new job. Nevertheless, even this subset of the job market is important and has valuable information about the rest of the market.

The data we used in this thesis are not publicly available. It has been obtained from the analytical department of the job search portal Jobs.cz. We contacted Tomáš Dombrovský, the head of the department, via e-mail and arranged a personal meeting with him. He provided us the data set and useful information about data collection and their utilization within the company.

Two types of data are applied. Firstly, the thesis uses data of total numbers of job vacancies published on the portal Jobs.cz in a given month. Job postings include both full-time and part-time jobs, but not temporary jobs. In addition, job postings may stay published on the portal more than one month. Secondly, we work with monthly data of total numbers of reactions to the job postings published in a given month.

It is important to note that, since March 2016, the monthly metric of the total number of job postings has been adjusted. Data after this change vary by a few percent (downward, by older metric it would be higher) than the originally used metric. All data range from January 2010 to April 2018.

Prace.cz

The portal Jobs.cz offers a wide range of job vacancies and closely cooperates with the other Czech job search portal Prace.cz. This portal offers rather jobs that are less paid and does not have such high requirements (e.g. tertiary education) as Jobs.cz offers have. Prace.cz is suitable for people who are unemployed and need to find a completely new job.

Comparable data for all job positions occupied through Prace.cz are not available, as job vacancies from Labour Office are imported on a daily basis, which greatly mix results (roughly triple the total numbers of job vacancies). Thus, the data of reactions to the job postings on the portal Prace.cz are not available as well.

Chapter 5

Methodology

In this thesis, methodology works with a standard framework commonly used for estimating time-series models of macroeconomic indicators, especially the unemployment rate. In this chapter, all used processes needed before estimation of the models and methods using for the estimation are described.

5.1 ARMA, ARIMA, ARMAX

Autoregressive Moving Average (ARMA) model is a time-series model that provides a description of a weakly stationary stochastic process. It consists of two parts: autoregressive (AR) and moving average (MA). A general ARMA model is of order (p, q), where p is the order of the AR part and q is the order of the MA part. Both, AR(p) and AM(q), can be defined as special cases of the ARMA(p, q) model.

A general autoregressive (AR) model of order p is defined as:

$$y_t = c + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \cdots + \rho_p y_{t-p} + \epsilon_t, \tag{5.1}$$

where ρ_i are coefficients, y_{t-i} are lagged values of the dependent variable y_t and ϵ_t are independent and identically random variables distributed as $N(0, \sigma_\epsilon^2)$.

A general moving average (MA) model of order q is defined as:

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}, \tag{5.2}$$

where θ_j are coefficients, ϵ_{t-j} are lagged values of the error term ϵ_t and ϵ_t are independent and identically random variables distributed as $N(0, \sigma_u^2)$.

A general autoregressive moving average (ARMA) model of order (p, q) is

defined as:

$$y_t = c + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \cdots + \rho_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}, \quad (5.3)$$

where p and $q \in N$, y_t is the dependent variable, y_{t-i} are lagged values of the dependent variable y_t, ρ_i are their corresponding coefficients, ϵ_t is the error term (independent and identically random variables distributed as $N(0, \sigma_u^2)$), ϵ_{t-j} are lagged values of the error term, and θ_j are their coefficients. In other words, the dependent variable y_t is explained by a linear combination of its own lagged values and a combination of the current error term and its own lagged values.

ARMA model satisfies the stationarity condition. However, when the stationarity condition is not met, a generalisation of an ARMA model can be used in an Autoregressive Integrated Moving Average (ARIMA) model. ARIMA model differs from the ARMA model in an extended integrated part $I(d)$, where d is the order of integration. The integrated part allows the ARIMA model to deal with non-stationary series.

In this thesis, a simple AR(1) process for the baseline model is used. An AR(1) process is defined as:

$$y_t = c + \rho_1 y_{t-1} + \epsilon_t, \quad (5.4)$$

where ρ_1 is a coefficient, y_{t-1} is a lagged value of the dependent variable y_t and ϵ_t are independent and identically random variables distributed as $N(0, \sigma_\epsilon^2)$.

The choice of the process is based on the fact that many previous studies used the same model (e.g. Kholodilin *et al.* (2010), Pavlíček & Krištoufek (2015)), or its seasonal version (e.g. (Choi & Varian (2009b), Askitas & Zimmermann (2009)).

An Autoregressive (Integrated) Moving Average with additional explanatory variable(s) (ARMAX/ARIMAX) is formed when additional exogenous inputs are added to the ARMA/ARIMA model. We define ARAMAX (p, q) model as:

$$\begin{aligned} y_t = c + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \cdots + \\ + \rho_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}, \end{aligned} \quad (5.5)$$

where $p, q, r \in N, x_i$ are additional explanatory variables, β_i are their corresponding coefficients, y_t is the dependent variable, y_{t-i} are lagged values of the

dependent variable y_t, ρ_i are their corresponding coefficients, ϵ_t is the error term (independent and identically random variables distributed as $N(0, \sigma_u^2)$), ϵ_{t-j} are lagged values of the error term, and θ_j are their coefficients.

ARMAX model allows us to explain the prediction and the development of unemployment rate in the Czech Republic not just with its historical (lagged) values, but also with additional information, such as data about behaviour of job-seekers on the Internet.

5.2 Stationarity

A stochastic process $\{x_t : t = 1, 2, ...\}$ is stationary if for every collection of time indices $1 \leq t_1 \leq t_2 \leq ... \leq t_m$, the joint probability distribution of $(x_{t_1}, x_{t_2}, ..., x_{t_m})$ is the same as the joint probability of $(x_{t_1+h}, x_{t_2+h}, ..., x_{t_m+h})$ for all integers $h \geq 1$. A weak form of stationarity, covariance stationarity process, is defined as: A stochastics process $\{x_t : t = 1, 2, ...\}$ with finite second moment $[E({x_t}^2) < \infty]$ is covariance stationary if:

- $E(x_t)$ is constant,
- $Var(x_t)$ is constant,
- for any $t, h \geq 1, Cov(x_t, x_{t+h})$ depends only on h and not on t.[1]

In other words, a time series is weakly (covariance) stationary if the mean of such time series is constant, its variance is constant and finite, and the covariance between two periods depends only on the distance between these two periods. Covariance stationarity focuses only on the first two moments of a stochastics process and the covariance between x_t and x_{t+h}.

Stationarity of ARMA(p, q) process 5.1 depends its AR(p) part. MA(q) process is always stationary. AR(1) process is stationary if $|\rho| < 1$, an AR(p) process is stationary if the roots to z of:

$$1 - \rho_1 z - \rho_2 z^2 - ... - \rho_p z^p = 0, \tag{5.6}$$

are all in modulus larger than $1(|z| < 1)$. In other words, if the series contains a unit root, the series is non-stationary.

[1] Jeffrey M. Wooldridge (2016): *Introductory Econometrics. A Modern Approach.* Boston: Cengage Learning, sixth edition.

If time series are non-stationary, it can cause several problems in estimating models. To make processes stationary, series can be transformed by differencing. A process with a unit root is referred to as integrated of order one, $I(1)$, and the process become stationary after first differencing. A time series that is $I(1)$ is often said to be a difference-stationary process. A process with d unit roots is referred to as integrated of order $d, I(d)$, and the process can be made stationary by taking d differences.

Stationarity of macroeconomics series (e.g. unemployment rate) is hard to justify as the series are often subject to trends. However, Montgomery *et al.* (1998) and Koop & Potter (1999) pointed out that unemployment rate lying inside the unit circle should not show any presence of a unit root.

To test for stationarity, both formal and informal test exist. Informal tests examine correlation between y_t and y_{t-1}. In other words, they are based on testing deviation from the three conditions of covariance stationarity process (constant mean, constant and finite variance, covariance depending on a distance between two periods).

There are two basic formal statistical tests determining stationarity of time series: Augmented Dickey-Fuller (ADF) test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. The tests have the opposite null hypothesis and thus, they provide a complementary pair, which is ordinarily used for testing stationarity. Both tests are utilized in the paper *Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries* by Pavlíček & Krištoufek (2015). In our thesis, we also use both Augmented Dickey-Fuller Test and Kwiatkowski-Phillips-Schmidt-Shin test to determine whether stationarity of the series is presented as well.

5.2.1 Augmented Dickey-Fuller Test

As it is defined by Dickey & Fuller (1979), the Augmented Dickey-Fuller test is based on the Ordinary Least Squares (OLS) regression:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t, \quad (5.7)$$

where α is an intercept, βt is a time trend, and p represents the lag order. The ADF test analyses whether the time series contains a unit root. Its null hypothesis is non-stationarity, which differs from the KPSS's null hypothesis that is stationarity of the series.

The null hypothesis under which the series contains a unit root is found for:

$$H_0 : \ \gamma = 0,$$

against alternative hypothesis under which the series is stationary:

$$H_A : \ \gamma < 0,$$

The ADF test uses a modified t-distribution, and the ADF test statistics are computed as the usual t-statistics. The Augmented Dickey-Fuller statistic is equal to a negative number, which means that the lower the number appears, the stronger rejection of the H_0 hypothesis is.

5.2.2 Kwiatkowski-Phillips-Schmidt-Shin Test

As it is defined by Kwiatkowski *et al.* (1992), the Kwiatkowski-Phillips-Schmidt-Shin test's null hypothesis is opposite to that of the ADF test, i.e. KPSS test has the null hypothesis of stationarity. The test is based on the Ordinary Least Squares (OLS) regression of the series $\{y_t\}$:

$$y_t = \alpha + \beta t + k \sum_{i=0}^{t} \xi_i + \varepsilon_t, \tag{5.8}$$

where α is an intercept, βt is a time trend, and ξ_i represent independent and identically distributed random variables with zero mean and a unit variance. The KPSS test analyses whether the time series is stationary.

The null hypothesis under which the series is stationary is found for:

$$H_0 : \ k = 0,$$

against alternative hypothesis:

$$H_A : \ k \neq 0$$

.

The KPSS statistics is defined as:

$$KPSS = \frac{\sum_{t=1}^{n} S_t^2}{n^2 \widehat{\omega_T^2}}, \tag{5.9}$$

where $\widehat{\omega_T^2}$ is an estimator of the spectral density at frequency zero, and S_t is the partial sum of the residuals:

$$S_t = \sum_{i=1}^{t} \widehat{\varepsilon_i}. \tag{5.10}$$

5.3 Models Identification

In order to figure out potential problem with stationarity, all models used in this thesis work with the first difference of the series. We proceed with the first differences of the unemployment rate and the first logarithmic differences of the variables: total numbers of job vacancies published on the portal Jobs.cz in a given month and total numbers of reactions to the job postings published in a given month. We chose the logarithmic specification of the variables containing data from the portal Jobs.cz as the unemployment rates are produced in percent, and the combination of the percentage representation and the logarithmic transformation allows for a straightforward interpretation of a proportional relationship and each variable.

The basic relationship between the unemployment rate and data on number of job vacancies and responses to them from the Czech job search portal Jobs.cz, which we study in this thesis, is depicted in the equation:

$$\Delta UR_t = \alpha_0 + \alpha_1 \Delta \log\left(NumPos\right)_t + \alpha_2 \Delta \log\left(CanRes\right)_t + \varepsilon_t, \tag{5.11}$$

where ΔUR_t is the first difference of an unemployment rate at time t, $\Delta \log\left(NumPos\right)$ and $\Delta \log\left(CanRes\right)$ represent the first logarithmic difference of the number of the positions at Jobs.cz at time t and the first logarithmic difference of the candidates' responses to the posts at Jobs.cz at time t, respectively, and ε_t is an error term.

The general equation 5.11 is represented for the Czech Republic as a whole and for the 14 regions of the Czech Republic separately. The variable that differs across the 15 equations in total is the dependent variable ΔUR_t; the variables $\Delta \log\left(NumPos\right)$ and $\log\left(CanRes\right)$ remain the same.

In order to test heteroskedasticity of the linear regressions, we apply the Breusch-Pagan (BP) test. As it is defined by Breusch & Pagan (1979), the test has a form that assumes standard errors from the regressions are normally distributed. It examines whether the variance of the errors is dependent with explanatory variables, i.e. whether heteroskedasticity is present. In the

Breusch-Pagan test, we specify:

$$u^2 = \delta_0 + \delta_1 x_1 + ... + \delta_k x_k + error, \tag{5.12}$$

where u_t^2 is the error term of a linear regression, x_{tk} are explanatory variables, δ_k are their coefficients at time t. The null hypothesis of homoskedasticity is found for:

$$H_0 : \ \delta_1 = \delta_2 = ... \ = \delta_k = 0,$$

against alternative hypothesis that is defined as heteroskedasticity is present. The test is based on the F statistic or the LM statistic ($LM = n \cdot R^2_{\hat{u}^2}$) using $F_{k,\ n-k-1}$ distribution and χ^2_k distribution, respectively. If the p-value is below the chosen significance level, then the null hypothesis of homoskedasticity is rejected. One possibility how to correct heteroskedasticity is to use the heteroskedasticity and autocorrelation consistent (HAC) standard errors.

As we do not have more than 100 observations, when the sampling distribution tends to be normal, we need to test normality using the Shapiro-Wilk (SW) test. According to Shapiro & Wilk (1965), the normality test is sensitive to outliers and influence by sample size. The null hypothesis stands for the sample is normally distributed. The test statistic is defined as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_{(i)}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}, \tag{5.13}$$

where $y_{(i)}$ is the ith order statistic, $\bar{y}$ is the sample mean, a_i are coefficients, and n represents number of observations. If the p-value is less than the chosen significance level, then we reject the null hypothesis and have evidence that the tested sample is not normally distributed.

5.4 Nowcasting

Macroeconomic time series, such as unemployment rates, are commonly available with a substantial lag that hampers the effectiveness of real-time predictions. The lag occurs as most data sources and statistics used in economics are delayed due to data collection and processing, which both usually take approximately one month. Moreover, if the time series are available, some additional corrections to the reported values are presented. The past series, which is available immediately without any lag and that is strongly correlated with the

variable of interest, are very useful for predicting the present of the variable. This type of forecasting the present values is referred to as "nowcasting". In other words, if we want to nowcast the unemployment rate, we take the value from a month or more ago and estimate the current value without waiting. The base model is defined as:

$$\Delta UR_t = \alpha_0 + \sum_{i=1}^{L} \alpha_i \Delta UR_{t-i} + \vartheta_t, \tag{5.14}$$

where ΔUR_t is the first difference of an unemployment rate at time t, ΔUR_{t-i} represent the first difference of the lagged values of the unemployment rates, and ϑ_t is an error term.

As the variables containing data from the job search portal Jobs.cz might be potentially useful for nowcasting unemployment, we define the nowcasting model as:

$$\begin{aligned} \Delta UR_t = \beta_0 + \sum_{i=1}^{L} \beta_i \Delta\ UR_{t-i} + \sum_{j=0}^{L} \gamma_j \Delta\ \log\left(NumPos\right)_{t-j} \\ + \sum_{j=0}^{L} \delta_j \Delta \log\left(CanRes\right)_{t-j} + \varepsilon_t, \end{aligned} \tag{5.15}$$

where ΔUR_t is the first difference of an unemployment rate at time t, ΔUR_{t-i} represent the first difference of the lagged values of the unemployment rates, $\Delta\ \log\left(NumPos\right)_{t-j}$ and $\Delta \log\left(CanRes\right)_{t-j}$ represent the first logarithmic difference of the lagged values of the number of positions at Jobs.cz and the first logarithmic difference of the lagged values of the candidates' responses to the posts at Jobs.cz, respectively, and ε_t is an error term.

Both models contain a maximum lag L, where L is equal to 3, 6, 12. We selected these three maximum lags, because it allows evaluation of the quality of models considering the amount of information from the past. As the unemployment series are usually cyclical, we set the upper bound to 12 months.

Again, we observe equations for the Czech Republic and for the 14 regions of the Czech Republic separately. The variables that differs across all 15 equations are ΔUR_t and ΔUR_{t-i}; the variables $\Delta\ \log\left(NumPos\right)_t$ and $\Delta \log\left(CanRes\right)_t$.

5.5 Forecasting

The most studies from the Google Econometrics field (e.g. Krištoufek (2013), Preis *et al.* (2013)) aimed to nowcast economic indicators rather than to focus on quality statistics of models and significance of variables. Similarly, the main aim of this thesis is to provide easier and better prediction of movements in the unemployment rate in the context of the Czech economy, i.e. we compare forecasting ability of our models - the base model 5.14 and the nowcasting model 5.15.

As nowcasting aims for contemporary values, it is necessary to work only with a time series of forecast for the same horizon, i.e. with static one-step-ahead out-of-sample forecasts. The whole sample is divided into two subsamples. The first sample is used for estimation, the second one is used for forecast evaluation. The method of forecasting values that are already known is not real out-of-sample forecasting, since the full sample is actually known in advance. This method is called "pseudo out-of-sample". In this thesis, we select the data between 01/2010 and 08/2016 (80 observations) for estimating the models. For nowcasting performance, the series between 09/2016 and 04/2018 (20 observations) is chosen.

The first step in evaluation of the predictive ability of the models is to calculate forecast errors. These are defined as difference between the actual values of the dependent variable and its forecasted values:

$$e_{it} = y_t - \widehat{y_{it}}, \tag{5.16}$$

where $\widehat{y_{it}}$ is the forecasted value of the actual values of the dependent variable y_t, and e_{it} is the forecast error of model i in period t.

The most common measure of forecast accuracy is Mean Squared Error (MSE), which is defined as the average of squared forecast errors over the predicted period:

$$MSE_i = \frac{1}{T}\sum_{t=1}^{T} {e_{it}}^2, \tag{5.17}$$

where T is the length of out-of-sample forecast, and e_{it} is the forecast error of model i in period t. Squared root of MSE over the forecast period is called

Root Mean Squared Error (RMSE), and it is defined as:

$$RMSE_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T} {e_{it}}^2}. \tag{5.18}$$

Another measure of forecast accuracy is Mean Absolute Error (MAE), which is defined as the average of absolute values of the forecast errors over the predicted period:

$$MAE_i = \frac{1}{T}\sum_{t=1}^{T} |e_{it}| . \tag{5.19}$$

In comparison of squared and absolute errors, squared errors are usually applied to magnify higher errors. However, MAE/MSE statistics alone is not sufficient to assess whether forecasts of one model are significantly more precise or not.

When comparing accuracy of two competing forecasts, Diebold-Mariano (DM) test is conducted. The test is based on using aforementioned squared or absolute errors. Both errors (absolute and squared) will be utilized in our analysis. Diebold & Mariano (1995) proposed a test for comparing the predictive accuracy of two competing forecasts. If a set of forecasts is compared, one forecast is usually selected as a benchmark. The forecast errors from the two competing forecast measures are entered into a loss function $L(\cdot)$. It has different forms: $L(e_{it}) = {e_{it}}^2$ according to MSE, $L(e_{it}) = |e_{it}|$ according to MAE, etc.

The Diebold-Mariano test evaluates accuracy by creating a loss differential that is covariance stationary:

$$d_t = \ L(e_{1t}) - L(e_{2t}) , \tag{5.20}$$

where e_{1t} and e_{2t} are forecasts errors of models 1 and 2, respectively. In our case, the loss differential has a form based on Mean Absolute Error:

$$d_t = |e_{1t}| - |e_{2t}| , \tag{5.21}$$

and based on Mean Squared Error:

$$d_t = e_{1t}^2 - e_{2t}^2. \tag{5.22}$$

The null hypothesis under which the predictive accuracy of both models is

equal is found for:

$$H_0 : \; E\left(d_t\right) = 0 \;\Leftrightarrow\; E\left[L\left(e_{1t}\right)\right] = E\left[L\left(e_{2t}\right)\right] \tag{5.23}$$

against alternative hypothesis:

$$H_A : \; E\left(d_t\right) > 0 \;\Leftrightarrow\; E\left[L\left(e_{1t}\right)\right] > E\left[L\left(e_{2t}\right)\right]. \tag{5.24}$$

In other words, the loss differential is equal to zero under the null hypothesis, i.e. there are no quantitative differences between the forecasts of the models 1 and 2.

The Diebold-Mariano test statistic has a form:

$$DM \;=\; \frac{\bar{d}}{\sqrt{\frac{\widehat{LRV}_{\bar{d}}}{T}}}\,, \tag{5.25}$$

where $\bar{d}$ is the mean loss differential:

$$\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t, \tag{5.26}$$

and $\widehat{LRV}_{\bar{d}}$ is a consistent estimate of the asymptotic (long-run) variance of $\sqrt{T}\bar{d}$ defined as:

$$LRV_{\bar{d}} = \gamma_0 + 2\sum_{j=1}^{\infty}\gamma_j, \quad \gamma_j = Cov\left(d_t,\; d_{t-j}\right). \tag{5.27}$$

Under the null hypothesis, the DM statistic goes to a standard normal distribution, i.e.:

$$DM \;\rightarrow N\left(0,1\right).$$

Chapter 6

Empirical Results

In this chapter, the empirical results of our analysis of explanatory and forecasting power of job search portal data are described. We present outcomes of stationarity testing, which was done by 2 tests – Augmented Dickey-Fuller (ADF) test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, perform individual models' regression results, present outcomes from nowcasting models, and through the Diebold-Mariano test we describe forecasting results.

6.1 Stationarity

We use the ADF test and the KPSS test to examine stationarity of our data. We use both tests as they have opposite null hypothesis and thus, they provide a complementary pair. The results show us whether we should provide analysis of the original series or use some transformations of the series.

In Table 6.1, we present the results of the ADF and KPSS tests for the original variables of the unemployment time series of the Czech Republic and 14 regions separately as well as for their first differences. Since the unemployment series are already presented as percentage, we do not test their logarithmic transformation. The whole original series proves signs of non-stationarity, i.e. we do not reject unit roots of the original series. Vysočina Region from the original series fulfils ADF test, however, it does not meet conditions for the KPSS test. In order to make the series stationary, we take first differences of each series. For most of the cases, we reject unit roots of the series of the first differences. The only exception is the City of Prague, one of the 14 regions, where the tests are not completely complied.

In Table 6.2, we show the outcomes of testing stationarity for the original

variables of job search portal data, for the logarithmic transformations and their first differences. The original series are non-stationary as well as their logarithmic transformation. The variable of the candidates' responses to the posts at Jobs.cz fulfils KPSS test, nevertheless, it does not comply with requirements for ADF test. For the first differences of the series, the stationarity is not rejected. The results of testing stationarity of job search portal data are equal for the Czech Republic and each of the 14 regions.

In order to keep the models interpretable, we use the first differences of the unemployment rate and the first logarithmic differences of variables from job search portal Jobs.cz in our analysis.

Table 6.1: Stationarity testing (ADF test, KPSS test) for Czech Republic and 14 regions (Note: p-values are reported in the brackets)

Stationarity testing

	ADF test		**KPSS test**	
	Unempl.	**1^{st} difference**	**Unempl.**	**1^{st} difference**
Czech Republic	-1.916	-5.095	1.924	0.074
	[>0.1]	[<0.01]	[<0.01]	[>0.1]
Prague	0.760	-3.746	0.697	1.260
	[>0.1]	[0.024]	[0.014]	[<0.01]
Central Bohemian Region	-1.345	-4.985	1.828	0.178
	[>0.1]	[<0.01]	[<0.01]	[>0.1]
South Bohemian Region	-3.355	-4.948	1.851	0.029
	[0.066]	[<0.01]	[<0.01]	[>0.1]
Plzeň Region	-2.493	-4.727	1.993	0.033
	[>0.1]	[<0.01]	[<0.01]	[>0.1]
Karlovy Vary Region	-2.006	-5.340	1.966	0.085
	[>0.1]	[<0.01]	[<0.01]	[>0.1]
Ústí Region	-2.006	-5.182	1.963	0.113
	[>0.1]	[<0.01]	[<0.01]	[>0.1]
Liberec Region	-2.228	-5.024	2.021	0.051
	[>0.1]	[<0.01]	[<0.01]	[>0.1]
Hradec Králové Region	-2.072	-5.461	1.884	0.060
	[>0.1]	[<0.01]	[<0.01]	[>0.1]
Pardubice Region	-3.901	-6.317	2.007	0.021
	[0.017]	[<0.01]	[<0.01]	[>0.1]
Vysočina Region	-4.053	-5.138	1.979	0.023
	[<0.01]	[<0.01]	[<0.01]	[>0.1]
South Moravian Region	-2.862	-5.285	1.963	0.037
	[>0.1]	[<0.01]	[<0.01]	[>0.1]
Olomouc Region	-3.558	-5.327	1.993	0.025
	[0.040]	[<0.01]	[<0.01]	[>0.1]
Zlín Region	-3.203	-5.533	2.030	0.030
	[0.091]	[<0.01]	[<0.01]	[>0.1]
Moravian-Silesian Region	-2.135	-5.382	1.952	0.067
	[>0.1]	[<0.01]	[<0.01]	[>0.1]

Table 6.2: Stationarity testing (ADF test, KPSS test) for variables from the portal Jobs.cz (Note: p-values are reported in the brackets)

Stationarity testing for Czech Republic & each of the regions		
	ADF test	KPSS test
Jobs.cz - Number of Positions	-1.756	1.729
	[>0.1]	[<0.01]
- logarithm	-1.999	1.742
	[>0.1]	[<0.01]
- difference	-5.077	0.074
	[<0.01]	[>0.1]
- logarithmic difference	-5.356	0.054
	[<0.01]	[>0.1]
Jobs.cz - Candidates' Responses	-3.304	0.318
	[>0.1]	[>0.1]
- logarithm	-3.332	0.331
	[0.070]	[>0.1]
- difference	-7.409	0.032
	[<0.01]	[>0.1]
- logarithmic difference	-7.895	0.030
	[<0.01]	[>0.1]

6.2 Fundamental Models Performance

One of the aims of this thesis is to investigate the relationship between the Czech unemployment rate and job search on the Internet by users interested in changing/improving their current job positions or finding job vacancies in order to become employed. We examine the usability of the data from the Czech job search portal Jobs.cz in a set of models utilizing these data. The equations of basic relationship between the unemployment rate and data on a number of job vacancies and responses to them from the Czech job search portal Jobs.cz, which we study in this thesis, is depicted in the equation 5.11 for the Czech Republic as a whole as well as for the 14 regions separately. The general equation 5.11 differs across the 14 regions and the Czech Republic in the dependent variable (UR_t). Two explanatory variables (number of job postings published on the portal Jobs.cz in a given month, number of monthly reactions to the job postings) remain the same.

First of all, we test heteroskedasticity of the linear regressions and normality of the samples using the Breusch-Pagan (BP) test and the Shapiro-Wilk (SW) test, respectively. The results from both tests are shown in Table A.1

and Table A.2 in the Appendix section. The null hypothesis of homoskedasticity is rejected only for Ústí Region, Liberec Region and Moravian-Silesian Region. In other regions heteroskedasticity is not present.[1] The Shapiro-Wilk normality test shows that most samples are not normally distributed, except South Bohemian Region and Plzeň Region.[2]

Table 6.3 shows the results of the basic relationship for the Czech Republic. We observe that the explanatory variables of the data from the job search portal are quite different. The variable of number of job postings at Jobs.cz is equal to -1.808 with standard error 0.467, i.e. if we change dl_NumPos by 1 percent, we will get 1.808 percent less of $dURCR$. The $t-value$ is -3.869, which makes the variable statistically significant as it satisfies $|t - value| > 1.96$. The variable of candidates' responses to the job posting is equal to 0.088 with standard error 0.154 and $t - value$ 0.573. It means that if the variable dl_CanRes changes by 1 percent, the dependent variable $dURCR$ is expected to increase by 0.088 percent. This implies that only dl_NumPos is statistically significant for the Czech Republic model. Furthermore, the R-squared tells us the overall fit of the model. It gives values between 0 and 1. The closer it gets to 1, the more significant the model is. In our case, R-squared is equal to 0.162, which means that 16.2% in the variation in the unemployment rate is explained by this model.

The results of the basic relationship for 14 regions individually are presented in Table 6.4. The proportional relationship does not rapidly vary across the analysed 14 regions. We can notice that the variable dl_NumPos remains negative and statistically significant for all regions, and the variable dl_CanRes still has the positive sign and is statistically insignificant. The R-squared is ranged from 1.3% (Prague) to 21.6% (Vysočina Region).

Overall, the results of all equations of the basic relationship show that the changes in the unemployment rate are projected into the online job search-related terms, especially to the number of job vacancies at Jobs.cz.

[1] As most regressions have no evidence of heteroskedasticity and we want to unify the methodology, we use standard errors for all models.

[2] Even though the SW test rejects the null hypothesis of most sampling distribution, we estimate the models and comment the results as if the samples are normally distributed, i.e. we comment on the asymptotic properties of the estimates. Since the first 5 Gauss-Markov assumptions are satisfied and the number of observations is equal to 100, which we consider as a borderline for asymptotics.

Table 6.3: Fundamental model of the Czech Republic – Regression Results (Note: standard errors are reported in the brackets)

	Dependent variable:
	incdnt
dl_NumPos	-1.809***
	(0.465)
dl_CanRes	0.089
	(0.153)
Constant	-0.050*
	(0.028)
Observations	100
R^2	0.162
Adjusted R^2	0.145
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6.4: Fundamental models of 14 regions – Regression Results (Note: standard errors are reported in the brackets)

Fundamental models (Note: *p<0.1; **p<0.05; ***p<0.01)

Region	Dependent variable	dl_NumPos	dl_CanRes	Constant	Observations	R^2	Adjusted R^2
Prague	dURP	-0.125	0.066	-0.016	99	0.013	-0.007
		(0.176)	(0.058)	(0.011)			
Central Bohemian	dURCBR	-1.288***	0.131	-0.036*	99	0.134	0.116
		(0.347)	(0.115)	(0.021)			
South Bohemian	dURSBR	-2.411***	0.359	-0.039	99	0.122	0.104
		(0.662)	(0.219)	(0.040)			
Plzeň	dURPlR	-1.814***	0.223	-0.049*	99	0.165	0.148
		(0.424)	(0.140)	(0.026)			
Karlovy Vary	dURKVR	-1.667***	0.162	-0.073*	99	0.123	0.105
		(0.474)	(0.157)	(0.029)			
Ústí	dURUR	-1.307**	-0.078	-0.080**	99	0.077	0.058
		(0.611)	(0.202)	(0.037)			
Liberec	dURLR	-1.286***	-0.095	-0.071**	99	0.129	0.110
		(0.465)	(0.154)	(0.028)			
Hradec Králové	dURHKR	-1.854***	0.003	-0.047	99	0.163	0.146
		(0.510)	(0.168)	(0.031)			
Pardubice	dURPaR	-3.254***	0.459	-0.054	99	0.069	0.050
		(1.225)	(0.405)	(0.074)			
Vysočina	dURVR	-3.479***	0.149	-0.051	99	0.216	0.200
		(0.754)	(0.249)	(0.046)			
South Moravian	dURSMR	-2.150***	0.064	-0.055	99	0.164	0.147
		(0.563)	(0.186)	(0.034)			
Olomouc	dUROR	-2.711***	0.163	-0.069	99	0.139	0.121
		(0.752)	(0.248)	(0.045)			
Zlín	dURZR	-2.298***	0.051	-0.068*	99	0.181	0.165
		(0.574)	(0.189)	(0.035)			
Moravian-Silesian	dURMSR	-2.034***	-0.080	-0.059*	99	0.177	0.160
		(0.572)	(0.190)	(0.035)			

6.3 Nowcasting

Macroeconomic series are usually reported with a substantial lag, which hampers the effectiveness of real-time predictions. Such lag occurs due to data collection and processing that both usually take approximately one month. To solve this problem, we use the past series, which is available immediately without any lag and that is strongly correlated with the variable of interest. It helps us to predict the present of the variable. This method is referred to as "nowcasting".

The base model is depicted in the equation 5.14 and the nowcasting model, containing data from the job search portal Jobs.cz, is defined in the equation 5.15. Both equations are generalised for the Czech Republic and the 14 regions individually. In the models, we study joint significance of the variables and adjusted $R^2(\bar{R}^2)$ as a measure of the models' quality controlling for the number of explanatory variables. Both models contain a maximum lag L, where L is equal to 3, 6, 12.

Table 6.5 summarises the results of nowcasting for the Czech Republic. We observe that the inclusion of job search portal data substantially enhances the models. Looking at the adjusted R^2, we detect that added variables from Jobs.cz extended by lag order levels rapidly increase the $\bar{R}^2$. The adjusted R^2 tells us whether the added variables are significant or not. If there will be any insignificant variable, the $\bar{R}^2$ will be lower. Considering the number of lags, the results show that the greater the number of lags is, the more improved the models are. Joint significance is present not only for job-related data, but also for lagged unemployment rates. Furthermore, we can notice a seasonal pattern in the unemployment rate.

The results of nowcasting for 14 regions are divided into three tables according to the maximum lag. Table 6.6 presents results for the maximum lag equal to 3, Table 6.7 shows outcomes for $L = 6$, and Table 6.8 depicts results for $L = 12$. In tables, we observe that for all 14 regions, the extension of the series by job search portal data rapidly improves the models. One exception is Ústí region, where the $\bar{R}^2$ is negative for $L = 12$. Moreover, for Ústí Region and Moravian-Silesian Region, we detect that the base models with 3 and 6 lags are very weak, they even have negative values of $\bar{R}^2$. The base model with 12 lags of these regions has a positive value of $\bar{R}^2$, from which we can see that there is a seasonal (annual) pattern in the unemployment rates. In addition, it shows that every added lag enhances the models. This improvement we can

observe in the models of other regions. Moreover, we notice that lagged unemployment rates as well as job-related data are jointly statistically significant for all models.

To sum up all results of the nowcasting modelling of the series, the job search portal data are obviously important, as they enhance the models. Their usefulness is supported by a statistical significance of the variables job vacancies ($NumPos$) and reactions ($CanRes$) for the Czech Republic and all 14 regions.

Table 6.5: Nowcasting Summary for the Czech Republic (Note: p-values are reported in the brackets)

Nowcasting Summary (Czech Republic)		
Δu_{t-i} significance	L = 3	4.330 [<0.01]
	L = 6	3.085 [<0.01]
	L = 12	3.483 [<0.01]
$\Delta\log(NumPos)_{t-i} + \Delta\log(CanRes)_{t-i}$ significance	L = 3	7.424 [<0.01]
	L = 6	4.478 [<0.01]
	L = 12	1.560 [0.077]
$\bar{R}^2$ without variables from Job.cz	L = 3	0.094
	L = 6	0.118
	L = 12	0.255
$\bar{R}^2$ with variables from Jobs.cz	L = 3	0.416
	L = 6	0.434
	L = 12	0.383

Table 6.6: Nowcasting Summary for 14 regions - L = 3 (Note: p-values are reported in the brackets)

Nowcasting Summary (14 regions) L = 3

	Δu_{t-i} significance	$\Delta\log(NumPos)_{t-i}$ $+\Delta\log(CanRes)_{t-i}$ significance	$\bar{R}^2$ without variables Jobs.cz	$\bar{R}^2$ with variables Jobs.cz
Prague	2.208 [0.092]	2.983 [<0.01]	0.036	0.177
Central Bohemian Region	7.872 [<0.01]	6.254 [<0.01]	0.177	0.433
South Bohemian Region	16.082 [<0.01]	6.267 [<0.01]	0.320	0.531
Plzeň Region	16.060 [<0.01]	7.206 [<0.01]	0.320	0.557
Karlovy Vary Region	4.467 [<0.01]	4.557 [<0.01]	0.098	0.309
Ústí Region	0.474 [>0.1]	2.970 [<0.01]	-0.017	0.131
Liberec Region	1.210 [>0.1]	4.553 [<0.01]	0.007	0.239
Hradec Králové Region	2.976 [0.036]	8.507 [<0.01]	0.058	0.428
Pardubice Region	2.621 [0.055]	5.765 [<0.01]	0.048	0.325
Vysočina Region	7.021 [<0.01]	10.736 [<0.01]	0.158	0.542
South Moravian Region	5.050 [<0.01]	7.520 [<0.01]	0.112	0.431
Olomouc Region	5.383 [<0.01]	7.279 [<0.01]	0.121	0.429
Zlín Region	3.323 [0.023]	8.467 [<0.01]	0.068	0.432
Moravian-Silesian Region	0.453 [>0.1]	5.763 [<0.01]	-0.017	0.278

Table 6.7: Nowcasting Summary for 14 regions - L = 6 (Note: p-values are reported in the brackets)

Nowcasting Summary (14 regions) L = 6

	Δu_{t-i} significance	$\Delta\log(NumPos)_{t-i}$ $+\Delta\log(CanRes)_{t-i}$ significance	$\bar{R}^2$ without variables Jobs.cz	$\bar{R}^2$ with variables Jobs.cz
Prague	7.308 [<0.01]	3.719 [<0.01]	0.289	0.506
Central Bohemian Region	4.272 [<0.01]	3.998 [<0.01]	0.174	0.443
South Bohemian Region	9.462 [<0.01]	4.795 [<0.01]	0.353	0.598
Plzeň Region	8.994 [<0.01]	6.271 [<0.01]	0.340	0.643
Karlovy Vary Region	2.798 [0.015]	2.933 [<0.01]	0.104	0.317
Ústí Region	0.692 [>0.1]	1.745 [0.065]	-0.020	0.089
Liberec Region	1.170 [>0.1]	2.724 [<0.01]	0.011	0.226
Hradec Králové Region	2.611 [0.022]	5.357 [<0.01]	0.094	0.468
Pardubice Region	2.430 [0.032]	3.049 [<0.01]	0.084	0.312
Vysočina Region	4.999 [<0.01]	6.449 [<0.01]	0.205	0.577
South Moravian Region	3.012 [0.010]	4.847 [<0.01]	0.115	0.453
Olomouc Region	3.912 [<0.01]	4.461 [<0.01]	0.158	0.459
Zlín Region	2.276 [0.044]	4.846 [<0.01]	0.076	0.429
Moravian-Silesian Region	0.949 [>0.1]	3.043 [<0.01]	-0.003	0.245

Table 6.8: Nowcasting Summary for 14 regions - L = 12 (Note: p-values are reported in the brackets)

Nowcasting Summary (14 regions) L = 12

	Δu_{t-i} significance	$\Delta\log(NumPos)_{t-i}$ $+\Delta\log(CanRes)_{t-i}$ significance	$\bar{R}^2$ without variables Jobs.cz	$\bar{R}^2$ with variables Jobs.cz
Prague	7.308 [<0.01]	3.719 [<0.01]	0.289	0.506
Central Bohemian Region	4.272 [<0.01]	3.998 [<0.01]	0.174	0.443
South Bohemian Region	9.462 [<0.01]	4.795 [<0.01]	0.353	0.598
Plzeň Region	8.994 [<0.01]	6.271 [<0.01]	0.340	0.643
Karlovy Vary Region	2.798 [0.015]	2.933 [<0.01]	0.104	0.317
Ústí Region	0.692 [>0.1]	1.745 [0.065]	-0.020	0.089
Liberec Region	1.170 [>0.1]	2.724 [<0.01]	0.011	0.226
Hradec Králové Region	2.611 [0.022]	5.357 [<0.01]	0.094	0.468
Pardubice Region	2.430 [0.032]	3.049 [<0.01]	0.084	0.312
Vysočina Region	4.999 [<0.01]	6.449 [<0.01]	0.205	0.577
South Moravian Region	3.012 [0.010]	4.847 [<0.01]	0.115	0.453
Olomouc Region	3.912 [<0.01]	4.461 [<0.01]	0.158	0.459
Zlín Region	2.276 [0.044]	4.846 [<0.01]	0.076	0.429
Moravian-Silesian Region	0.949 [>0.1]	3.043 [<0.01]	-0.003	0.245

6.4 Forecasting

The main aim of the thesis is to provide easier and better prediction of movements in the unemployment rate in the context of the Czech economy. In the previous section, we present the results of the nowcasting modelling of the series. However, as our goal is to examine predictive power of job search data from the portal Jobs.cz, we are more interested in the out-of-sample performance. As we describe in the Methodology section, we divide the whole sample into two subsamples. The first sample, which consists of 80 observations (data between 01/2010 and 08/2016), is used for estimating models. The second sample, composed of 20 observations (data between 09/2016 and 04/2018), is used for forecast evaluation. For the comparison of forecasting ability of our models, we utilize the base model 5.14 and the nowcasting model 5.15. The models are then tested for significant differences in predictive accuracy using the Diebold-Mariano (DM) test. We determine that the DM test uses absolute errors in computing a loss function. The null hypothesis is that the forecast accuracy of both models is the same. For the alternative hypothesis we choose that the nowcasting model 5.14 with job search data is more accurate than the

base model 5.15. Ideally, we want to reject the null hypothesis and accept the alternative hypothesis.

In Table 6.9, the results of forecasting for the Czech Republic is summarised. We observe that the values are negative for all maximum lags. The null hypothesis of Diebold-Mariano test cannot be rejected even at 10 % level. Therefore, the improvement in forecast accuracy is statistically insignificant. Moreover, the outcomes do not change when we use absolute errors or squared errors.

The results of forecasting for 14 regions are again divided into three tables according to the maximum lag. Table 6.10 presents results for the maximum lag equal to 3, Table 6.11 shows outcomes for $L = 6$, and Table 6.12 depicts results for $L = 12$. In the tables, we observe that the outcomes do not differ across the regions and the maximum lags. Values of the DM test are negative for all regions, except Prague and South Bohemian Region for the maximum lag of 3 months. As in the Czech Republic, we utilize absolute and squared errors in the test. Nevertheless, the results are same for both cases. We find that we cannot reject the null hypothesis, thus the differences between the base and nowcasting models are statistically insignificant.

To sum up all results of the forecasting accuracy using the Diebold-Mariano test, they are the same for the Czech Republic and all 14 regions. None of the differences between the models is statistically significant and therefore, there is no improvement in forecast accuracy.

Table 6.9: Forecasting Summary for the Czech Republic (Note: p-values are reported in the brackets)

		Forecasting Summary (Czech Republic)	
Czech Republic		absolute errors	squared errors
Diebold-Mariano test	L = 3	-0.333	-1.000
		[>0.1]	[>0.1]
	L = 6	-1.347	-1.854
		[>0.1]	[>0.1]
	L = 12	-2.916	-2.950
		[>0.1]	[>0.1]

Table 6.10: Forecasting Summary for 14 regions - L = 3 (Note: p-values are reported in the brackets)

Forecasting Summary (14 regions) L = 3		
Diebold-Mariano test	absolute errors	squared errors
Prague	0.848	0.418
	[>0.1]	[>0.1]
Central Bohemian Region	-0.404	-0.642
	[>0.1]	[>0.1]
South Bohemian Region	0.310	-0.085
	[>0.1]	[>0.1]
Plzeň Region	-1.188	-1.025
	[>0.1]	[>0.1]
Karlovy Vary Region	-0.842	-1.335
	[>0.1]	[>0.1]
Ústí Region	-1.740	-2.374
	[>0.1]	[>0.1]
Liberec Region	-1.830	-2.451
	[>0.1]	[>0.1]
Hradec Králové Region	-1.723	-1.718
	[>0.1]	[>0.1]
Pardubice Region	-1.540	-2.092
	[>0.1]	[>0.1]
Vysočina Region	-1.992	-2.418
	[>0.1]	[>0.1]
South Moravian Region	-1.319	-1.786
	[>0.1]	[>0.1]
Olomouc Region	-2.363	-2.984
	[>0.1]	[>0.1]
Zlín Region	-1.241	-1.917
	[>0.1]	[>0.1]
Moravian-Silesian Region	-1.229	-2.385
	[>0.1]	[>0.1]

Table 6.11: Forecasting Summary for 14 regions - L = 6 (Note: p-values are reported in the brackets)

Forecasting Summary (14 regions) L = 6		
Diebold-Mariano test	absolute errors	squared errors
Prague	-2.112	-1.996
	[>0.1]	[>0.1]
Central Bohemian Region	-1.594	-1.740
	[>0.1]	[>0.1]
South Bohemian Region	-2.960	-2.392
	[>0.1]	[>0.1]
Plzeň Region	-2.607	-2.440
	[>0.1]	0.1]
Karlovy Vary Region	-1.091	-1.693
	[>0.1]	[>0.1]
Ústí Region	-2.530	-2.648
	[>0.1]	[>0.1]
Liberec Region	-2.806	-3.657
	[>0.1]	[>0.1]
Hradec Králové Region	-1.976	-1.839
	[>0.1]	[>0.1]
Pardubice Region	-0.859	-0.017
	[>0.1]	[>0.1]
Vysočina Region	-1.995	-2.590
	[>0.1]	[>0.1]
South Moravian Region	-1.785	-2.518
	[>0.1]	[>0.1]
Olomouc Region	-3.458	-3.080
	[>0.1]	[>0.1]
Zlín Region	-2.320	-2.678
	[>0.1]	[>0.1]
Moravian-Silesian Region	-1.691	-2.742
	[>0.1]	[>0.1]

Table 6.12: Forecasting Summary for 14 regions - L = 12 (Note: p-values are reported in the brackets)

Forecasting Summary (14 regions) L = 12		
Diebold-Mariano test	absolute errors	squared errors
Prague	-2.143	-2.123
	[>0.1]	[>0.1]
Central Bohemian Region	-2.102	-1.994
	[>0.1]	[>0.1]
South Bohemian Region	-2.493	-2.453
	[>0.1]	[>0.1]
Plzeň Region	-1.822	-1.816
	[>0.1]	[>0.1]
Karlovy Vary Region	-2.742	-2.744
	[>0.1]	[>0.1]
Ústí Region	-4.922	-3.871
	[>0.1]	[>0.1]
Liberec Region	-3.919	-3.344
	[>0.1]	[>0.1]
Hradec Králové Region	-3.062	-3.171
	[>0.1]	[>0.1]
Pardubice Region	-2.765	-2.312
	[>0.1]	[>0.1]
Vysočina Region	-2.205	-2.233
	[>0.1]	[>0.1]
South Moravian Region	-3.019	-2.710
	[>0.1]	[>0.1]
Olomouc Region	-3.921	-3.437
	[>0.1]	[>0.1]
Zlín Region	-3.622	-3.234
	[>0.1]	[>0.1]
Moravian-Silesian Region	-3.684	-3.137
	[>0.1]	[>0.1]

Chapter 7

Conclusion

In this thesis, we wanted to investigate the relationship between the Czech unemployment rate and job search on the Internet by users who are interested in changing/improving their current job positions or are unemployed and need to find some work. Thanks to the relationship, we could conclude whether online data improve unemployment prediction, which is needed to make effective government decisions. This thesis should have also provided easier and better prediction of movements in the unemployment rate, which is inaccurate as most data sources used in economics are commonly available only after a substantial lag.

Specifically, we worked with freely available data from the website of Integrated Portal of the Ministry of Labour and Social Affairs, where everyone can find detailed statistics of unemployment rates. Moreover, we focused on examination of the usability of data on numbers of job vacancies and responses to them gathered from the portal Jobs.cz, one of the most popular Czech job search portals. The thesis utilized the obtained data in a simple autoregressive model of the unemployment in the Czech Republic and extended it with extra variables containing data from the portal Jobs.cz. In addition to the augmented autoregressive model of the Czech Republic, we estimated the same models for 14 regions of the Czech Republic separately.

In the beginning, we defined models with a basic relationship between the unemployment rate and data on number of job vacancies and responses to them from the Czech job search portal Jobs.cz, where we studied the statistical significance of the variables numbers of job vacancies and responses to them. Then, we utilized nowcasting in order to solve a problem with the lag. We used two models – a base model with lagged unemployment rates and a

nowcasting model with current and lagged variables from Jobs.cz. Both models contain a maximum lag L, where L is equal to 3, 6, 12. We examined compound significance of variables and adjusted $R^2(\bar{R}^2)$ as a measure of the models' quality controlling for the number of explanatory variables. In the end, we divided the whole sample into two subsamples in order to use them for estimating models and for forecast evaluation. After that we compared forecasting ability of the models - the base model and the nowcasting model. The models were then tested for significant differences in predictive accuracy using the Diebold-Mariano (DM) test. In total, we estimated and compared 195 different models for the Czech Republic and the 14 regions individually.

We showed that data on job vacancies and reactions to them obtained from Jobs.cz improved the baseline models for the Czech Republic as well as for the 14 regions separately. However, this held only for the number of job vacancies as it is statistically significant in all the models. The responses to them were estimated to be statistically insignificant, when we added them to the base models. In the summary results of nowcasting, we observed that the inclusion of job search portal data substantially enhances the models. Looking at the adjusted R^2, we detected that added variables from Jobs.cz extended by lag order levels rapidly increased the $\bar{R}^2$ – for the Czech Republic, it reached up to 0.434, and for the 14 regions in total, it reached up to even 0.684. The usefulness of the job search portal data was supported by a statistical significance of the variables job vacancies and reactions to them. However, according to the Diebold-Mariano test, none of the differences between the base and nowcasting models were statistically significant. Therefore, we found no forecasting improvements of the unemployment rates in either the Czech Republic or in the regions.

To sum up, our results showed that the statistically significant job-related data enhanced the fundamental models as well as nowcasting models. It means that we were able to see their significance in the labour market and their usefulness in nowcasting. Nevertheless, we did not find any improvements in prediction of the Czech unemployment rate after adding the data from the portal Jobs.cz. It might be caused by the fact that the job search portal focuses only on a subset of the labour market, i.e. it requires job seekers to have tertiary education. Another issue may be that data from the portal do not cover only searches by unemployed people, but they take into account also people who intend to switch a job.

Our research could be improved in two potential ways. The first possibility

is to compare estimations of 14 regions separately with all regression equations using a procedure named Seemingly Unrelated Regression (SUR). It is a generalisation of the linear regression model that consists of several equations. Each equation has its own dependent variable and set of exogenous explanatory variables, which can be the same or different across the equations. Unlike estimating the set of equations separately, estimating all equations simultaneously with a generalised least squares (GLS) estimator leads to efficient estimates as it takes into account the correlation of the error terms. The second possibility is to re-estimate the models with renewed and enlarged data (e.g. using data from more widely oriented job search portal) and find out whether they confirm or reject their importance in prediction of unemployment rate.

Bibliography

Aastveit, K. A., K. R. Gerdrup, A. S. Jore, & L. A. Thorsrud (2011): "Nowcasting GDP in Real Time: A Density Combination Approach." *Journal of Business & Economic Statistics* **32**.

Aastveit, K. A. & T. Trovik (2012): "Nowcasting Norwegian GDP: The Role of Asset Prices in a Small Open Economy." *Empirical Economics* **42**: pp. 95–119.

Askitas, N. & K. F. Zimmermann (2009): "Google Econometrics and Unemployment Forecasting." *IZA Discussion Paper 4201, Institute for the Study of Labor (IZA)* .

Bai, J. & S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models." *Econometrica* **70**: pp. 191–221.

Banbura, M., D. Giannone, & L. Reichlin (2010): "Nowcasting." *Working paper series 1275, European Central Bank.* .

Beveridge, W. H. (1944): *Full Employment in a Free Society: a Report.* London: Allen & Unwin.

Breusch, T. S. & A. R. Pagan (1979): "A Simple Test for Heteroskedasticity and Random Coefficient Variation." *Econometrica* **47(5)**: p. 1287–1294.

Carrière-Swallow, Y. & F. Labbé (2013): "Nowcasting With Google Trends in an Emerging Market." *Journal of Forecasting* **32(7)**: pp. 289–298.

Choi, H. & H. Varian (2009a): "Predicting Initial Claims for Unemployment Benefits." *Google Inc* .

Choi, H. & H. Varian (2009b): "Predicting the Present with Google Trends." *Google Inc* .

Choi, H. & H. Varian (2012): "Predicting the Present with Google Trends." *The Economic Record* **8**: pp. 2 – 9.

Dickey, D. A. & W. A. Fuller (1979): "Distribution of the Estimators for Autoregressive Time Series With a Unit Root." *Journal of the American Statistical Association* **74(366)**: pp. 427–431.

Diebold, F. X. & R. S. Mariano (1995): "Comparing Predictive Accuracy." *Journal of Business and Economic Statistics* **20**: pp. 134–144.

D'Amuri, F. (2009): "Predicting Unemployment in Short Samples with Internet Job Search Query Data." *MPRA Paper* **(18403)**.

Ettredge, M., J. Gerdes, & G. Karuga (2005): "Using Web-based Search Data to Predict Macroeconomic Statistics." *Communications of the ACM* **48(11)**: pp. 87–92.

Giannone, D., L. Reichlin, & D. Small (2005): "Nowcasting: The Real-time Informational Content of Macroeconomic Data." *Journal of Monetary Economics* **55**: pp. 665–676.

Kholodilin, K. A., M. Podstawski, & B. Siliverstovs (2010): "Do Google Searches Help in Nowcasting Private Consumption?: A Real-Time Evidence for the US." *KOF Swiss Economic Institute Working Paper* **(256)**.

Koop, G. & S. M. Potter (1999): "Dynamic Asymmetries in U.S. Unemployment." *Journal of Business Economic Statistics* **17(3)**: pp. 298–312.

Krištoufek, L. (2013): "Can Google Trends Search Queries Contribute to Risk Diversification?" *Scientific Reports* **1310(1444)**.

Kwiatkowski, D., P. C.B.Phillips, P. Schmidt, & Y. Shin (1992): "Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root: How Sure Are We that Economic Time Series Have a Unit Root?" *Journal of Econometrics* **54**: pp. 159–178.

Modugno, M. (2011): "Nowcasting Inflation Using High Frequency Data." *International Journal of Forecasting* **29**: pp. 664–675.

Montgomery, A. L., V. Zarnowitz, R. S. Tsay, & G. C. Tiao (1998): "Forecasting the U.S. Unemployment Rate." *Journal of the American Statistical Association* **93(442)**.

Pavlíček, J. & L. Krištoufek (2015): "Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries." *FinMaP-Working Paper 34, Collaborative EU Project FinMaP – Financial Distortions and Macroeconomic Performance: Expectations, Constraints and Interaction of Agents.* .

Phillips, A. W. (1958): "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957." *Economica* **25**.

Platil, L. (2014): *Google Econometrics: An Application to the Czech Republic.* Master's thesis, Charles University, Prague.

Preis, T., H. S. Moat, & E. H. Stanley (2013): "Quantifying Trading Behavior in Financial Markets Using Google Trends." *Scientific Reports* **3**.

Scott, S. L. & H. R. Varian (2013): "Predicting the Present with Bayesian Structural Time Series." *International Journal of Mathematical Modelling and Numerical* **5(1)**: pp. 4–23.

Shapiro, S. S. & M. B. Wilk (1965): "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* **52(3-4)**: p. 591–611.

Simionescu, M. (2015): "The Improvement of Unemployment Rate Predictions Accuracy." *Prague Economic Papers* **2015(3)**: pp. 274 – 286.

Stevenson, B. (2008): "The Internet and Job Search." *NBER Working Paper* **(13886)**.

Suhoy, T. (2009): "Query Indices and a 2008 Downturn: Israeli Data." *Research Department, Bank of Israel* .

Tuhkuri, J. (2016): "Forecasting Unemployment with Google Searches." *ETLA Working Papers* **35**.

Wooldridge, J. M. (2016): *Introductory Econometrics. A Modern Approach.* Boston: Cengage Learning, 6th edition.

Wu, L. & E. Brynjolfsson (2009): "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities." *ICIS 2009 Proceedings* .

Yiu, M. S. & K. K. Chow (2011): "Nowcasting Chinese GDP: Information Content of Economic and Financial Data." *HKIMR Working paper* **4**.

Zacha, O. (2015): *Unemployment in the Czech Republic and Job Search on the Internet.* Bachelor's thesis, Charles University, Prague.

Čabla, A. & I. Malá (2017): "Modelling of Unemployment Duration in the Czech Republic." *Prague Economic Papers* **2017(4)**: pp. 438–449.

Appendix A

Appendix

Table A.1: Heteroskedasticity Testing (Breusch-Pagan test) (Note: p-values are reported in the brackets)

Heteroskedasticity Testing	
Breusch-Pagan test	Fundamental Models
Czech Republic	4.822
	[0.090]
Prague	1.564
	[>0.1]
Central Bohemian Region	3.435
	[>0.1]
South Bohemian Region	0.614
	[>0.1]
Plzeň Region	0.212
	[>0.1]
Karlovy Vary Region	8.918
	[0.012]
Ústí Region	10.550
	[<0.01]
Liberec Region	10.671
	[<0.01]
Hradec Králové Region	2.227
	[>0.1]
Pardubice Region	0.222
	[>0.1]
Vysočina Region	1.086
	[>0.1]
South Moravian Region	5.078
	[0.079]
Olomouc Region	3.163
	[>0.1]
Zlín Region	6.019
	[0.049]
Moravian-Silesian Region	9.469
	[<0.01]

Table A.2: Normality Testing (Shapiro-Wilk test) (Note: p-values are reported in the brackets)

Normality Testing	
Shapiro-Wilk test	Fundamental Models
Czech Republic	0.943
	[<0.01]
Prague	0.949
	[<0.01]
Central Bohemian Region	0.967
	[0.013]
South Bohemian Region	0.978
	[0.095]
Plzeň Region	0.978
	[0.096]
Karlovy Vary Region	0.908
	[<0.01]
Ústí Region	0.732
	[<0.01]
Liberec Region	0.821
	[<0.01]
Hradec Králové Region	0.949
	[<0.01]
Pardubice Region	0.643
	[<0.01]
Vysočina Region	0.969
	[0.018]
South Moravian Region	0.950
	[<0.01]
Olomouc Region	0.954
	[<0.01]
Zlín Region	0.937
	[<0.01]
Moravian-Silesian Region	0.790
	[<0.01]