MSc in Computer Science 2020-21

Project Dissertation

Project Dissertation title: Nowcasting unemployment with real time job vacancies data in the United States

Term and year of submission: Trinity Term 2021

Candidate Number: 1048799

Abstract

Machine learning has been used to analyse time series and macroeconomic factors for some time already. However, new, alternative data sources are still being looked for and new models developed to cater for them. Prediction of important indicators such as GDP is constantly improving by both public institutions as well as private sector. The knowledge of such can give advantage in policy making as well as on the financial markets. In this work we explore one of such indicators, that is the unemployment rate, and utilize a data set from the company Burning Glass Technologies which does online job offers scraping to improve the nowcasting of it. We also try to beat the results obtained from traditional time series analysis methods, namely Autoregressive models, by implementing the multivariate Bayesian Dynamic Linear Model with regressors obtained from the dataset. Finally in the answer we juxtapose the results, seeing the usefulness of the BDLM model and propose further improvements that can be done in the future.

Table of Contents

Chapter 1

Introduction

It is a dream for a politician or a hedge fund manager to just plug in available data to know immediately the details about the country's economic outlook. If we imagine this, it is a powerful tool for both policymaking, but also investment decisions. In this project we try to see if this scenario could be possible with the use of job vacancy data available online. This, scraped by Burning Glass Technologies, could be a powerful addition for nowcasting of unemployment. With this information, macroeconomic and investment –decisions could be better adjusted almost in real-time. The recent years with COVID-19 also prove how important it is to have reliable data at hand in making major decisions. Though some of such analysis and trials have been done, it is the first time that this data is used in such context in the US and with the use of Bayesian Dynamic Linear Models, which can make an important contribution to the field.

Nowcasting can be colloquially defined as the prediction of the very near past, present, and the very near future. In fact, the term itself originates from the weather forecast, as it referred to describing the weather conditions happening at the moment and in the immediate future. In fact, the very origins of it come

from the time when in 1860s Vice-Admiral Robert FitzRoy, famous for being the captain on ship cruising Charles Darwin, collected storm reports on the coast of England and issued the name forecast for the prediction on whether the bad weather was coming or not [1]. The term nowcasting itself was first coined in the 1980s by Professor Keith Browning to describe actions taken to predict short-range rainfall forecast [2]. Today we use it also in different branches such as economics or even healthcare. Therefore, for the purpose of this work, that is viewing nowcasting in the perspective of macroeconomic data, we will rather use one of the definitions found in Macmillan Dictionary which defines nowcasting as "the making of predictions about economic or political performance based on statistical analysis of currently available data" [3].

However, not only nowcasting has its origins in a relatively far past considering scientific achievements of today. One of the concepts that sparked the idea of weather forecasting is time series analysis. This concept of analysing data taken in equal time measures has its origins way earlier. In fact, it is hard to say exactly when, as people have been measuring change of things in time since the ancient times. However, we can definitely spot the beginnings of first serious mathematical models used for the analysis. This can be dated back to 1920s and 1930s and the workings of G. U Yule and J. Walker when a concept of moving average was introduced to remove periodic fluctuations in the time series, such as though related to seasonality of the year [4]. However, no main breakthrough happened until the ARMA model was described in 1951 in a Thesis by Peter Whittle, which was then highly popularized in 1970 book by George E. P. Box and

Gwilym Jenkins (so called Box-Jenkins's method) [5]. 20 years forward these topics became very popular in financial world of 1990s when more and more quantified hedge funds were found [6]. Suddenly the time series analysis models became a powerful tool for analysing stock market movements, which sparked again the debate of efficient and rational markets vs. predictive markets and beating the market. With time, we learned that people like Jim Simmons and Renaissance Technologies have proven that in fact, machine learning can allow us to beat the market and get extraordinary gains [7]. The same applied to macroeconomic indicators which served as recommendations for the next moves – they could be predicted. Soon the governments became interested in those by using various models on government held data [8].

Only recently, that is in 2010s, with another wave of interest in machine learning, did interest in predicting major economic variables became a truly growing topic. We see this in publications based on different government data, also in unemployment, such as work done by Federal Reserve Bank of Atlanta [9], the European Central Bank [10], the Reserve Bank of New Zealand in 2018 [11] or even recently in 2021 [12] which shows a growing trend in this field. However, there is definitely still plenty of room to explore. Even though unemployment is usually defined and calculated by counting people who are jobless, actively seeking work, and available to take a job, predicting this number can have various inputs influencing the model. Therefore, the quest for new regressors into the equation started within the idea of who has data, has power that comes with it. Recent examples show this trend by, for example, using the geo location data

coming from smartphones [13] or using more exotic and seemingly unrelated sources such as Google searches popularity [14]. All of them show some improvement in the calculations. That is why the innovation is this project is the usage of Burning Glass data set and job vacancy data to nowcast unemployment. For that, however, no ordinary method has been chosen. There is already literature using the AR model and the derivatives of it [15] therefore this project aims to use a less popular, Bayesian Dynamic Linear Model (BDLM). This relatively recent model, coming from 1980s and described thoroughly in a book by West and Harrison "*Bayesian Forecasting and Dynamic Models"* [16] allows for analysis, through the use of probabilities, even on the relatively small datasets. This is a case with the Burning Glass dataset as online job vacancy data was not significant before 2010 due to the level of development and popularization of the Internet. Hence, this project is a first known public attempt to use those methods on this data set for the US and prove or disprove the usability online job vacancy data, at least on this market. It is also within the new trend to use digital real-time data to predict macroeconomic factors. Therefore, the research question that will be answered is whether we can predict the unemployment rate based on online job vacancies data.

1.1. Motivation

The real-time prediction of macroeconomic data is a very important task in scientific literature [17]. It is because, similarly to the weather forecast, nowcasting is a very powerful tool that can influence the way lots of organizations,

countries, and individuals function. Especially with the unemployment rate it is an important problem to solve, as unemployment figures are one of the most important macroeconomic data used both by the private and the public sector. In finance it is sought for to see how the economy of a country is doing. In the public sector the government has to plan its fiscal, monetary and social policy, which is also influenced by the unemployment rate. Hence, unemployment rate plays a major role in those. A good example of that can be the Taylor rule which tells what the federal funds rate (interest rates) should be when inflation and employment level change. Therefore, again the unemployment rate it is an important factor in consideration of interest rates set by the Federal Reserve.

However, the data used by institutions often lags behind the reality due to the various factors and ways it is collected – it is often dependent on outside institutions which first need to conduct data gathering and calculations [18]. Therefore, the machine learning methods could allow for a more precise reflection of reality – the more real-time data it uses, the more accurate the model at the moment. This could be especially useful during structural breaks in the job market, such as seen during the COVID-19 pandemic. Then, in the beginning of 2020, almost any government-collected data proved futile, or really lagged in delivery due to the bureaucratic breakdown, whereas the online job vacancy data used in this project could be monitored almost in real-life time, seeing the dynamics of the job market and forecasting it based on those figures. If the project proves the usability of this data, this can be a great indicator for its further exploration and

fine-tuning of the results, as well as a catalyst for using real-time data as a powerful tool for macroeconomic factors determination.

1.2. Background, Aims and Goals

The overall goal of this project is to explore the potential use of online job vacancy data from Burning Glass Dataset to nowcast unemployment In the United States. The project will explore Autoregressive Model (AR) to first set a reference benchmark for nowcasting with the known methods and later to explore the possibilities of beating this benchmark with the use of this unique data and a Bayesian Dynamic Linear Model (BDLM). These techniques will be applied both on data from all fifty states and then on the general, federal level of the United States as a whole. In an event the above methods prove successful, they can be used to more accurately nowcast unemployment.

1.3 Project importance

Nowcasting, as it was mentioned before, is an important and growing field. Concentrating on macroeconomic data, we add value to the real life of thousands of people as proper analysis and ability to accurately predict its different quantitative aspects helps to dynamically adjust policies [19]. Unemployment in this case is especially something that policy makers want to track as it has influence on many important policy decisions. For example, it is now applied in

many Central Banks to adjust monetary policies in the economy [20]. However, it is also used to apply certain policies and justify actions of the government. The use of unemployment nowcasting can help predict trends in the GDP growth, which also helps predict inflation and business cycles. Arguably, the workings for this project could also help in the prediction of those. Furthermore, this work is based on the US data and, to the best of our knowledge, it is the first time in the academic literature that it is applied in this context to nowcast unemployment, where the benchmark is being challenged by the BDLM model.

1.4 Audience and potential beneficiaries

The target audience of the project are scholars interested in the topic of nowcasting and machine learning, as well as government bodies and think tanks which are interested in macroeconomic policy planning. In particular case of this project the potential beneficiary could be the Federal Reserve System (FED) which, as mentioned before, could use the outcome to predict the unemployment rate for next month and adjust its monetary policies to it. Another one is the policy makers, who can see the unemployment sooner and react with fiscal stimulus and other government measures. This was a case with COVID-19 pandemic where macroeconomic data was a key for proper policy making. Computer scientists can benefit from seeing the potential of using external data in linear predictions of real-life economic data and adjust it further for various regressors. Those working in hedge funds can see this useful to see the unemployment rate ahead of time and

predict its impact on the financial markets. Therefore, there can be plenty of beneficiaries coming from this project.

1.5 Scope of the project and outcomes

The project can be really complex and go beyond the intended activities, therefore it is important to set a clear scope and expectations for it. We define clear steps to follow through the process. Those include data research and background reading as the initial process to establish required context and state of the art in the literature. Then we analyse the scope of the methods used for the problem, as well as a justification for their use and performance. Next, we select, extract, and transform relevant records from the Burning Glass dataset. Based on our readings we design machine learning models for the problem and analyse the results they produce. Then we come up with a conclusion.

At the end of the project the outcome is to prove that the online job vacancy data indeed helps predict unemployment and that the selected model outperforms the AR benchmark model.

Chapter 2

Data

Currently, banking institutions often use very complex data with many indicators. An example is given in the Federal Bank of New York in the paper by Bok [21], which uses 36 indicators for its macroeconomic estimations of GDP. However, for the sake of project, we use limited inputs to only prove the usability of the given data, not try to compete with complexity of large systems possessed by banks – rather suggest implementation of new data into the existing systems. Therefore, for the project two data sets are used: private database of Burning Glass portal and publicly available Federal Reserve Economic Data (FRED) from the Federal Reserve Bank St. Louis.

2.1 Burning Glass Dataset

Burning Glass Dataset is a set provided by Burning Glass Technologies – a software company which scrapes for job postings online for job market analytics. It is probably the biggest go to place in this matter. Hence, the dataset on which the project was carried out constitutes of over 261 897 243 data entries in the form of job postings from the United States. Those job postings have been

scraped from all over the internet and form a tangible representation of online job offerings in the USA. Each job posting can have up to 57 columns of data. Not all of them are compulsory to fill out. They constitute of columns such as name of the position, salary, description, location, etc. The span of this data used for the assessment was from the earliest available date, that is from January 2010 to April 2021, which was the outside limitation.

The obvious limitation of this dataset is lack of access to offline job postings, such as adverts in the newspapers or on the front doors of local businesses. However, many of those find reflection online anyways and as the economy moves forward, most of its job postings are found online [22]. Because of the period analysed, that is from 2010, one can be assured of the increasing online presence of employers. Therefore, the dataset can be considered to form a representative form of job vacancies in the real economy.

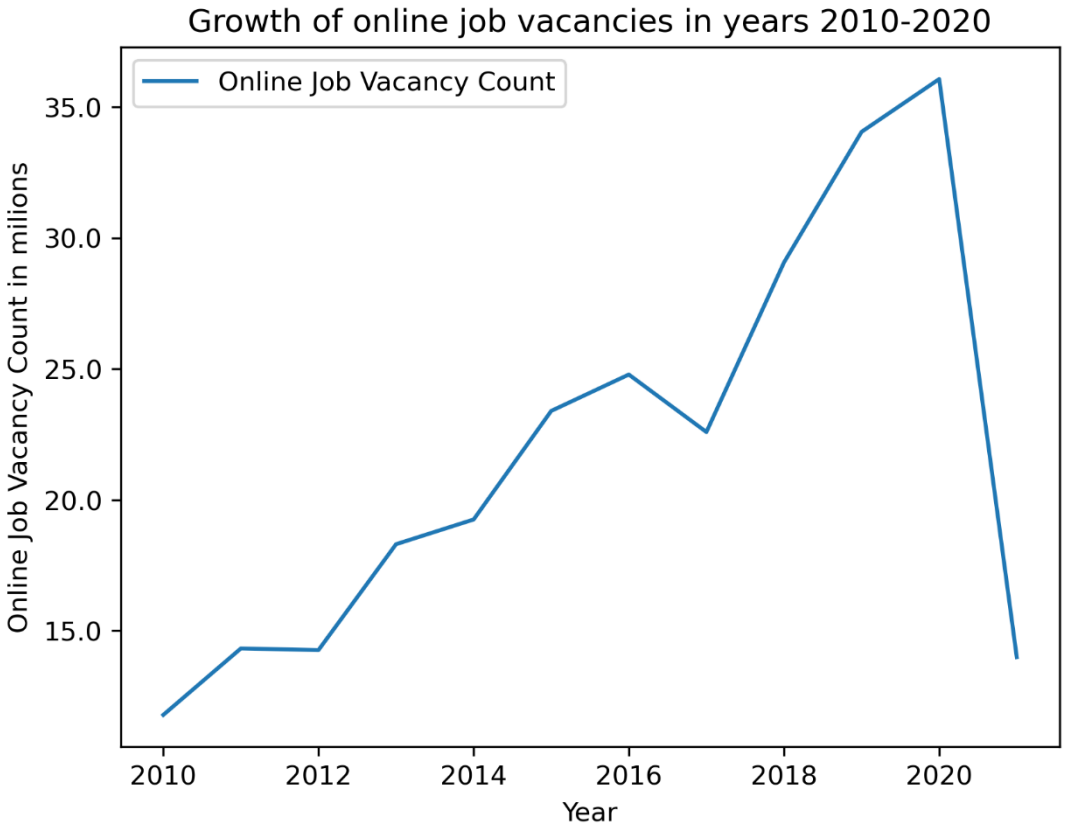

Figure 1. *Growth of online job vacancies in years 2010-2020 in millions*

As we see on Figure 1, the online job vacancies were increasing in years 2010-2020 at a high rate. It is worth noting that 2021 is not included as the year's full data is not available as of time's writing.

For the purpose of this project unique JobIDs of job offers were collected, grouped by state and by month. Hence the job counts for each state per month, as well as country as a whole was created. It will be referred to as jobs vacancy count.

Another operation performed on the data set was taking max and minimum annual salary from the available job postings (around 14% of the data has this

information) and taking the mean salary of it. As with the job IDs, this was done per state, per month, as well as for the country as a whole. This will be referred to as wages data.

2.2 “FRED” Dataset

Federal Reserve Economic Data of St. Louis is an online source with a multitude of macroeconomic data available, especially on the United States. It is a well-known resource used widely in the scientific world. From there data on unemployment rates for each of 50 states and for the country as a whole was downloaded. The span was set from January 2010 to April 2021, which was a limit set by the Burning Glass data dataset. In total it gives 136 months to analyse for all 50 states (total of 6800 observations). This data is accurate as it originates from a government source and is calculated upon the original sources of the data.

Chapter 3

Methods

For the analysis of data, I decided to use two linear machine learning models: Autoregressive Model (AR) and Bayesian Dynamic Linear Model (BDLM). Justification is that since it is a time-series analysis problem, the benchmark should be estimated by the most widely used, autoregressive model. As it is a univariate model, meaning it uses only one variable to estimate predictions, I decided to beat its performance with BDLM, which contrary to the AR method used in this project, works on probabilities and can be multivariate, meaning using multiple regressors for estimation, which will be required for using the job vacancy and wages data.

3.1 Autoregressive Model

Autoregressive model is a very popular technique for time-series analysis [23] and also the most naturally intuitive way of predicting the series. It gets the name from regressing on previous values from that same time series. It is univariate, meaning that the dependent variable is the effect of regressing its own lagged value by a

given time frame. Putting it simple, the dependent variable depends on the value of it in the past. It is expressed by an equation:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \in_t$$

Where y_t is the dependent variable, β_0 is a coefficient intercept, β_1 is a coefficient of the first lag, which is y_{t-1} and $\in_t$ is error.

The so-called order of an autoregression is the number of values used to predict the value of dependent variable in the past. So, for example, the previous model would be noted as AR(1). If we take into account two time units backwards, the equation would look like this:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \in_t$$

This would be called a second order autoregression, AR(2). These steps backwards, called lags, increase, we increase the order of auto regression. The general, AR(k) model formula for k-lags would be then:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_k y_{t-k} + \in_t$$

The time measurements could be days, months, years, etc. depending on the order used in the data. For example, in case of this project's unemployment rate, it is months. As everything is called on the value itself, this linear regression receives the prefix "auto".

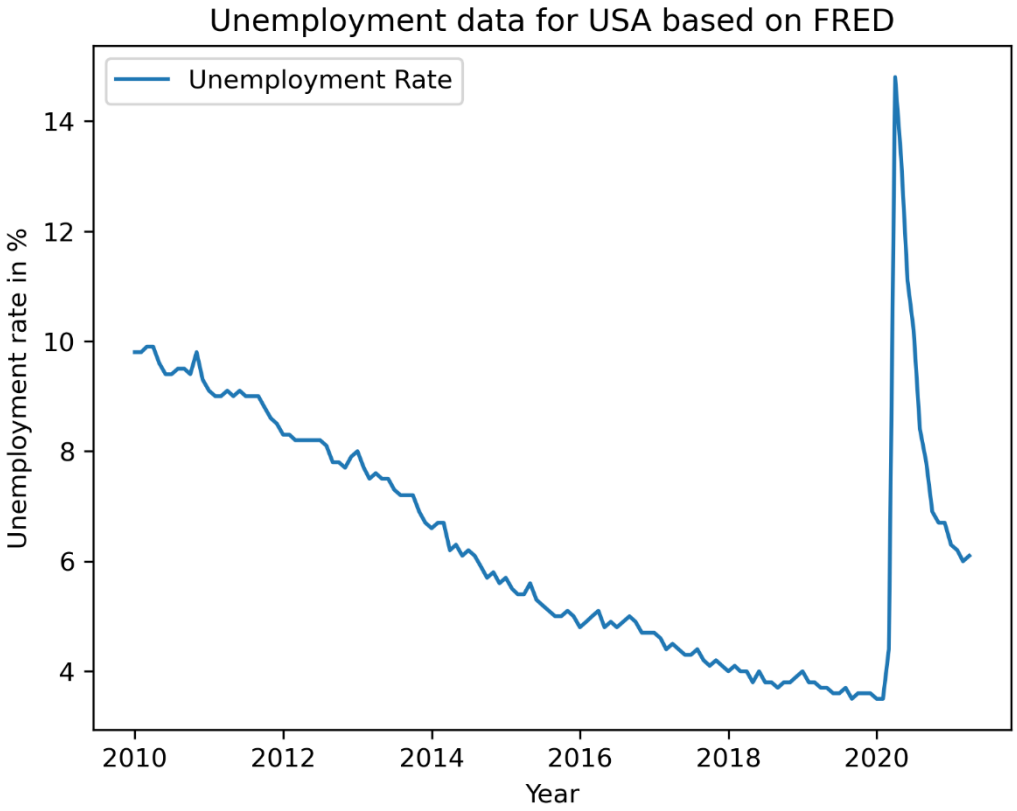

Figure 2. *Unemployment rate for the USA*

In case of this project's measurements, the dependent variable is the unemployment rate measured monthly. However, to make it useful for the model, the data has to be stationary, which means that its mean, variance and autocorrelation structure do not change over time. We can see from a graph of US unemployment rate (Figure 2) that it is definitely not a case.

However, we can make it stationary by differentiating the data, that is subtracting from the value its preceding value. This gives us instead of the list of unemployment rates, a list of changes in unemployment. We can see from the Figure 3 that now the data looks stationary and as such can be used on the AR model:

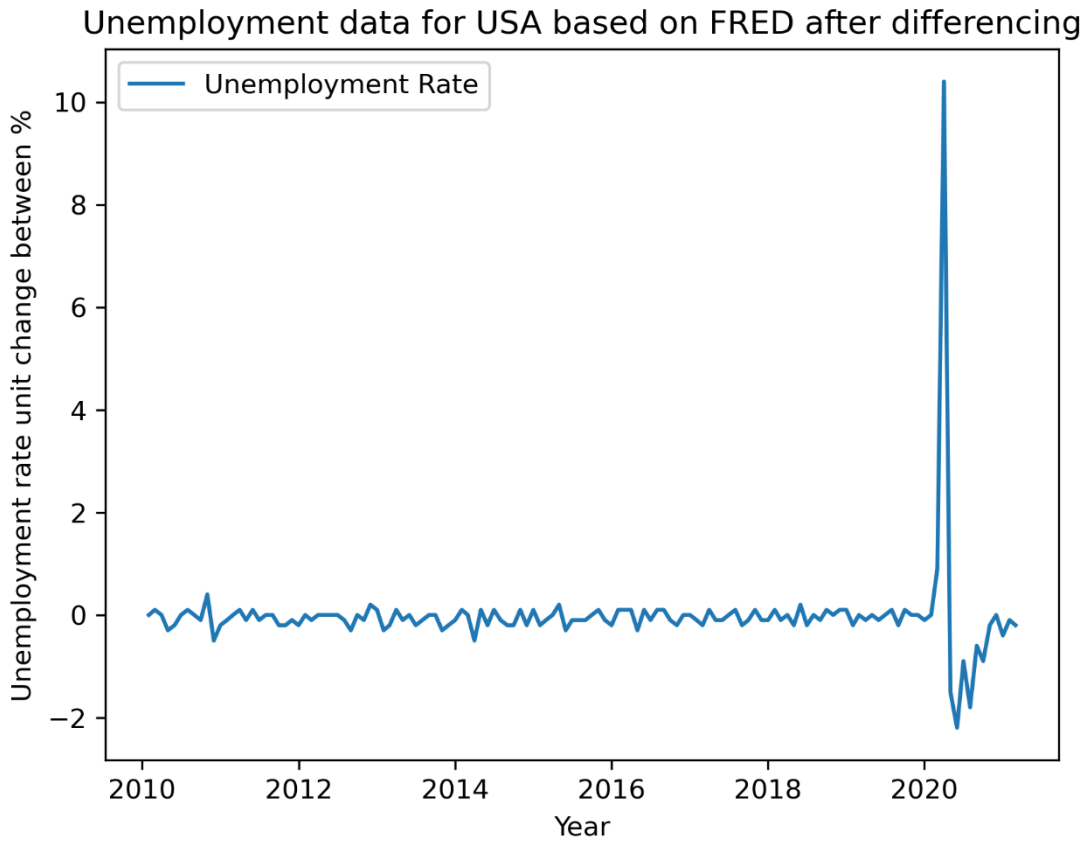

Figure 3. *Unemployment rate for USA based on FRED after differencing*

As we see, we still have “the Covid-19” bump, but it does not spoil the results on the data, which will be proven further.

3.2 Bayesian Dynamic Linear Model

The chosen method to beat the AR benchmark and prove the usefulness of the online job vacancy data is the Bayesian Dynamic Linear Model (BDLM). It bases on the Dynamic Linear Model but is ruled by Bayesian probabilities.

The Bayesian way offers a probabilistic approach to analyse time series and reduce uncertainty by incorporating past information. The BDLM is a special case of state space models, known also as Dynamic Linear Model [24], working by fitting the structural changes into the time series dynamically. This means that the parameters update as new information is added. Compared to the AR model, where all estimates are fixed, BDLM uses Maximum Likelihood estimation to make predictions. It also uses Monte Carlo Markov Chains to generate these estimates from distributions and classical recursive Kalman filter to estimate the model states.

Let us start with state space models first. Our goal is to build a probabilistic model to predict the dependent variable, given the history:

$$p(y_t|y_{t-1}, y_{t-2}, \dots)$$

where y_t is the dependent variable, and series $y_{t-1,}y_{t-2},\dots$ is the available history.

For that we build a state-space model and we need to provide some properties. At first, we use the Markov Property that future is independent from the past, given the present, which can be expressed as:

$$p(y_t|\mathcal{H})$$

where $\mathcal{H}$ is history.

Therefore, if we define $(Y_n) = (Y_0, Y_1, Y_2 \dots)$ to be a stochastic process in discrete time n=0,1,2… and discrete space S then (Y_n) has Markov property if for all times n and all states $y_0, y_1, y_2, \dots, y_n, y_{n+1} \in S$ we have:

$$P(Y_{n+1} = y_{n+1} | Y_n = y_n, Y_{n-1} = y_{n-1}, \dots, Y_0 = y_0)$$

$$= P(Y_{n+1} = y_{n+1} | Y_n = y_n$$

The second assumption that we have to make is that all the observations have an underlying dynamic, which we cannot see but we can show it mathematically as another hidden model, where x is a hidden state variable:

$$p(y_t | x_t)$$

Visually, we could see it as a solid line going beneath the prediction line as presented in Figure 4.

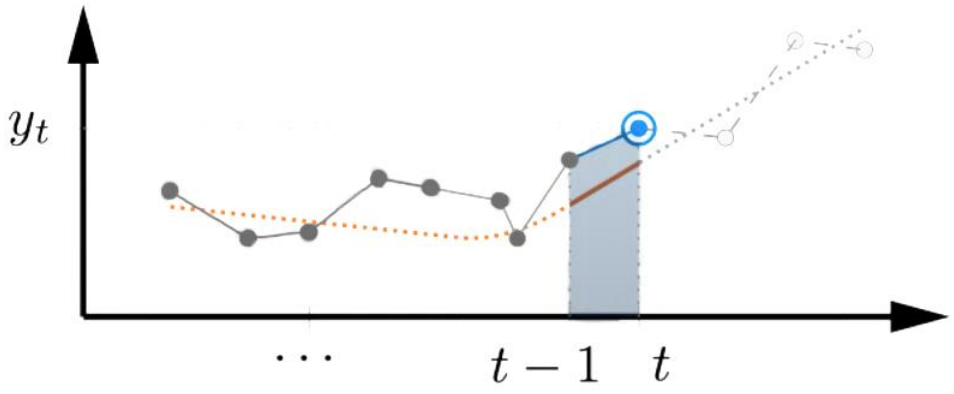

Figure 4. *BDLM hidden state variable graph*

With this theory we can define the goal of the SSM model, which is to predict y and x at time t, given the history $\mathcal{H} \equiv y_{t-1}$:

$$p(y_t, x_t | y_{t-1}, x_{t-1}) \rightarrow p(y_t, x_t)$$

Using the chain rule, we can decompose our model to two smaller models: the observation model and transition model:

$$p(y_t, x_t) = p(y_t|x_t) * p(x_t|x_{t-1})$$

where observation model shows how underlying dynamics can superpose to generate observations and transition model how are underlying dynamics evolving through time. This is of course repeated for all time steps.

For the model to be more readable, we represent the probabilities of observation and transition mode consecutively with functions:

$$y_t = h(x_t, v_t)$$

$$x_t = f(x_{t-1}, w_t)$$

For any probabilistic model we need to consider error terms, which are represented here by v_t and w_t .

Finally, we can present a BDLM model, which is a special case of the SSM model. In BDLM, all variables are Gaussian, meaning they are normally distributed, and all relationships are linear. We represent it by two linear equations based on our observation and transition model consequently:

$$y_t = Cx_t + v_t, \qquad v \sim \mathcal{N}(v; 0, R)$$

$$x_t = Ax_{t-1} + w_t, \qquad w \sim \mathcal{N}(w; 0, Q)$$

Then our model is defined by $M = \{C, A, R, Q\}$ with those variables serving as its definition. In BDLM, we define v as the gaussian distribution with mean 0 and

covariance matrix R and w as also the gaussian distribution with mean 0 but with covariance matrix Q.
C defines underlying dynamics (observations) that transform the model states. Vector x_t contains unobserved states of the system that evolve according to A. In case of time series, this will be data such as trend, seasonality, etc. Because it is a linear model, we assume C and A to be linear. Finally, R, Q represent error covariances.

These systems appeared under different names such as structural time series [25], state space approach [26] and finally DLM [27].

For this project, the predicted variable is the unemployment rate. It is tested on a univariate level (so with its previous values) as well as with regressors described in detail with each experiment in further sections.

3.3 Measurements of performance

For comparing the performance of the models, we will be using Mean Absolute Error (MAE), Mean Squared Error (MSE) and R^2 method. These are among the most common techniques for measuring accuracy of models for time series analysis [28].

3.3.1 MAE

Mean Absolute Error is the average of the absolute values of the deviation. It is represented by a formula:

$$MAE = \frac{\sum_{i=0}^{n} |y_i - \bar{y}_i|}{n}$$

It is perfect for telling how big the error in the actual forecast can be. Another advantage is that it is easy to interpret. It is also very good with data that is influenced by anomalies, such as the sudden increase of unemployment caused by COVID-19. If we do not predict such things to be the case in the future, MAE helps analyse the model better in this sense.

3.3.2 MSE

Mean Squared Error is the average of the square of the forecast error. It is a method commonly used for various assessments of precision and is represented by this equation:

$$MSE = \frac{\sum_{i=0}^{n} (y_i - \bar{y}_i)^2}{n}$$

In general, it is a good method when initially tuning the model, however, because we the square of the errors, some outliers might have huge impact on this value. Yet again, in case of COVID-19 anomaly on the job market MSE is affected strongly. Hence, even though useful in the process, MAE is better for measuring the actual performance of the models.

3.3.3 R^2

R^2 is another popular method of measuring error which tells us how much of the variance in the dependent variable can affect the variance in the independent variable. To calculate it for the project, the following equation has been used on the models:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where SS_{res} is the sum of squared residuals from the forecasted values and SS_{tot} is the sum of squared deviations of the dependent variable from the sample mean.

The results of R^2 that are high mean that variance in the model is similar to the real values, and if low, it means that there is not much correlation. It is worth noting that it does not represent the potential of accuracy of predictions in the future. In fact, it checks how well the model is fitted to the observed values. If the values are negative, that means that the chosen model fits worse than a horizontal line (does not follow the trend of the data). Also, the more features there are, the larger the R^2 value can be. Therefore, it might be not very representative for some calculations carried out later in this work.

Chapter 4

Calculations and Results

For evaluating the performance and accuracy of aforementioned methods, a set of tests was made. However, before that was carried out, first a check for correlation between the data and its general readiness was done.

4.1 Spearman's Correlation

Spearman's data correlation check was used to assess whether the experiment is worth pursuing. It is a non-parametric test that is used to measure the degree of association between two variables. It is calculated with the following formula:

$$p = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where p is the Spearman's rank correlation coefficient, $d_i$is the difference between the two ranks of each observation and n is the number of observations.

The data compared constituted of unemployment rate and cumulated job vacancy data. It was done for each state and the whole country.

State	Spearman's correlation
AK	-0.202
AL	-0.713
AR	-0.541
AZ	-0.598
CA	-0.635
CO	-0.596
CT	-0.414
DE	-0.416
FL	-0.697
GA	-0.742
HI	-0.159
IA	-0.671
ID	-0.484
IL	-0.570
IN	-0.672
KS	-0.605
KY	-0.743
LA	-0.317
MA	-0.543
MD	-0.594
ME	-0.415
MI	-0.755
MN	-0.601
MO	-0.650
MS	-0.618
MT	-0.572
NC	-0.676
ND	-0.309
NE	-0.497
NH	-0.569
NJ	-0.561
NM	-0.389
NV	-0.623
NY	-0.569
OH	-0.669
OK	-0.516
OR	-0.659
PA	-0.552
RI	-0.355
SC	-0.698
SD	-0.544
TN	-0.673
TX	-0.579
UT	-0.698
VA	-0.624
VT	-0.585
WA	-0.640
WI	-0.657
WV	-0.441
WY	-0.377
USA	-0.584

Figure 5. *Spearman's Correlation Results Table*

From the results we see that almost all of the states have a negative correlation, which can be considered strong (below -0.5). It is correct from what we see in the data as unemployment was declining over that period of time. This gives a justification for applying machine learning models with the analysed data.

4.2 Augmented Dickey-Fuller test

Before the AR model was finalized, the data, as mentioned in the methods description, had to be differentiated to make the unemployment rate a stationary variable. For this assessment first an Augmented Dickey-Fuller (ADF) test was performed. This is one of the most popular tests in this type of problem [29]. It is a unit root test meaning that it determines how strongly a time series is defined by a trend. The null hypothesis in this test is that the time series can be represented by a unit root, meaning that it is not stationary. The counter statement, needed to reject the null hypothesis is that the data is stationary. It is interpreted by the p-value from the test. As a threshold a p-value of 5% was used. Therefore, if the calculated p-value for data was above it, it failed to reject the null hypothesis and data is non-stationary. If the p-value was below 5% the null hypothesis was rejected, and data did not have a unit root and is stationary. Below are the results of the test before and after differencing.

State	**p-values**	**p after 1-differencing**	IL	0.047	0.000
			IN	0.073	0.000
AK	0.021	0.000	KS	0.063	0.000
AL	0.209	0.000	KY	0.241	0.000
AR	0.267	0.000	LA	0.084	0.000
AZ	0.242	0.000	MA	0.085	0.000
CA	0.206	0.000	MD	0.222	0.000
CO	0.420	0.000	ME	0.394	0.000
CT	0.455	0.001	MI	0.049	0.000
DE	0.213	0.000	MN	0.131	0.000
FL	0.227	0.009	MO	0.170	0.000
GA	0.337	0.000	MS	0.107	0.000
HI	0.150	0.000	MT	0.006	0.000
IA	0.005	0.000	NC	0.327	0.000
ID	0.229	0.000	ND	0.019	0.000

NE	0.110	0.000	SD	0.002	0.000
NH	0.002	0.000	TN	0.144	0.000
NJ	0.390	0.000	TX	0.090	0.000
NM	0.249	0.005	UT	0.062	0.000
NV	0.101	0.000	VA	0.096	0.000
NY	0.074	0.000	VT	0.005	0.000
OH	0.054	0.000	WA	0.016	0.000
OK	0.000	0.000	WI	0.047	0.000
OR	0.143	0.000	WV	0.001	0.000
PA	0.012	0.000	WY	0.079	0.000
RI	0.326	0.000			
SC	0.324	0.000	USA	0.066	0.000

Figure 6. *ADF Test Results*

As we see, after the differentiation the p-value was below 0.05 which allowed us to consider it as stationary and exclude the null hypothesis. The Augmented Dickey-Fuller test was passed for all states and the country.

To produce additional regressors, lags of vacancy data and unemployment rate were created from 1 to 12 lags each, giving a total of 24 additional regressors.

The lags have not been included on the wages data, as we will see shortly it did not provide accurate results, probably due to its sporadic appearance in the data set.

4.3 Results of models

For training of models, I used the span of first 100 months, then I used last 36 months for testing. As a measure of performance, I used Mean Absolute Error (MAE), Mean Squared Error (MSE) and R^2, as these are the most popular metrics in time series analysis and help compare the results not only within this project but also in other works. However, as the

measurement calculated for the improvement of performance (the last column) I decided to use only MAE as it is the most reflective error measurement for time series analysis [30].

For optimization, I used BIC which pointed out that AR with lag value of 5, out of 12 lags tested, had the best performance and will be used for comparison with BDLM model with the best performing regressors setup, which was simply using job vacancy data. The results are as follows:

State	BDLM MAE	BDLM MSE	BDLM R2	AR MAE	AR MSE	AR R^2	Improvement in %
AK	0.376	2.141	-4.467	0.379	1.715	-0.001	1%
AL	0.590	3.741	-0.013	0.598	4.156	-0.001	1%
AR	0.405	1.925	0.274	0.370	1.195	-0.004	-9%
AZ	0.492	3.418	-0.021	0.545	3.138	-0.004	10%
CA	0.606	4.844	0.391	0.628	4.080	-0.010	4%
CO	0.501	2.491	0.593	0.464	2.137	-0.012	-8%
CT	0.468	2.818	0.212	0.374	1.150	-0.027	-25%
DE	0.465	2.511	-0.095	0.562	2.856	-0.006	17%
FL	0.523	3.624	0.328	0.622	3.098	-0.004	16%
GA	0.515	3.555	0.213	0.512	2.634	-0.002	-1%
HI	0.438	1.506	0.482	1.032	13.430	-0.004	58%
IA	0.314	1.254	-0.148	0.483	2.260	-0.002	35%
ID	0.517	2.580	0.441	0.588	2.787	-0.001	12%
IL	0.602	3.936	0.042	0.666	5.109	-0.004	10%
IN	0.566	3.506	0.330	0.791	6.260	-0.001	28%
KS	0.370	1.576	-0.175	0.577	3.132	-0.001	36%
KY	0.575	3.605	0.124	0.750	6.423	-0.001	23%
LA	0.406	1.611	-1.301	0.442	2.171	-0.004	8%
MA	0.414	2.175	0.030	0.695	6.179	-0.003	40%
MD	0.366	1.977	-0.085	0.278	0.959	-0.012	-32%
ME	0.522	2.366	0.308	0.568	2.045	-0.005	8%
MI	0.733	5.313	0.205	1.139	13.776	-0.001	36%
MN	0.391	1.818	0.032	0.465	1.298	-0.005	16%

MO	0.486	2.864	0.234	0.535	2.615	-0.003	9%
MS	0.503	3.348	-0.091	0.609	3.453	-0.002	17%
MT	0.374	1.717	-0.409	0.483	2.235	-0.000	23%
NC	0.572	3.862	0.308	0.587	3.424	-0.003	3%
ND	0.201	0.469	-2.672	0.351	1.418	-0.003	43%
NE	0.252	0.737	-0.786	0.345	0.835	-0.000	27%
NH	0.340	1.339	0.132	0.775	5.799	-0.000	56%
NJ	0.462	3.005	0.287	0.813	5.772	-0.004	43%
NM	0.364	1.868	-4.025	0.400	1.137	-0.012	9%
NV	0.681	5.656	0.462	1.461	19.152	-0.002	53%
NY	0.421	2.453	0.143	0.627	4.760	-0.006	33%
OH	0.575	3.759	-0.039	0.713	4.661	-0.001	19%
OK	0.374	1.469	-1.232	0.564	3.226	-0.001	34%
OR	0.537	3.617	0.314	0.512	2.865	-0.006	-5%
PA	0.385	2.176	-0.427	0.644	4.245	-0.004	40%
RI	0.681	4.377	0.376	0.969	6.669	-0.003	30%
SC	0.595	4.292	0.324	0.494	2.640	-0.006	-20%
SD	0.249	0.729	-0.487	0.379	1.346	-0.000	34%
TN	0.535	3.302	0.239	0.702	5.267	-0.003	24%
TX	0.408	2.109	0.055	0.535	2.217	-0.007	24%
UT	0.439	2.024	0.318	0.443	1.935	-0.001	1%
VA	0.365	1.761	-0.122	0.478	2.467	-0.004	24%
VT	0.328	1.347	-0.070	0.717	5.307	-0.000	54%
WA	0.436	2.637	-0.060	0.689	4.162	-0.002	37%
WI	0.484	2.587	0.214	0.685	4.627	-0.001	29%
WV	0.491	2.472	-1.368	0.609	3.573	-0.000	19%
WY	0.445	1.788	-0.971	0.271	0.345	-0.014	-64%
USA	0.564	3.208	0.369	0.626	3.491	-0.004	10%

Figure 7. *Optimized Results of BDLM and AR models with online job vacancy data*

Only for 8 out of 50 states was AR performing better and it was by a tiny margin. Similarly for the country as a whole, BDLM overall performed better. The numbers were usually high

up by tens of percents. This is a proof that online job vacancy data can help predict unemployment. However, before this conclusion was made, further test had been carried out to see if other data or its combinations can improve this score.

In search for more regressors, the results above the wages data described before was added as a second regressor to the BDLM model. As a benchmark, the best performing AR model with lag of 5 was used. The results are as follows:

State	BDLM MAE	BDLM MSE	BDLM R2	AR MAE	AR MSE	AR R^2	Improvement in %
AK	0.428	2.823	-6.209	0.379	1.715	-0.001	-13%
AL	0.640	4.805	-0.301	0.598	4.156	-0.001	-7%
AR	0.462	2.520	0.050	0.370	1.195	-0.004	-25%
AZ	0.602	4.523	-0.351	0.545	3.138	-0.004	-10%
CA	0.736	6.411	0.193	0.628	4.080	-0.010	-17%
CO	0.790	7.549	-0.234	0.464	2.137	-0.012	-70%
CT	0.586	3.837	-0.073	0.374	1.150	-0.027	-57%
DE	0.566	3.307	-0.443	0.562	2.856	-0.006	-1%
FL	0.626	4.811	0.107	0.622	3.098	-0.004	-1%
GA	0.587	2.862	-0.188	0.512	2.634	-0.002	-15%
HI	0.485	1.958	0.327	1.032	13.430	-0.004	53%
IA	0.374	1.637	-0.499	0.483	2.260	-0.002	23%
ID	0.663	3.685	0.201	0.588	2.787	-0.001	-13%
IL	0.717	5.131	-0.248	0.666	5.109	-0.004	-8%
IN	0.691	4.619	0.117	0.791	6.260	-0.001	13%
KS	0.456	2.150	-0.603	0.577	3.132	-0.001	21%
KY	0.687	4.782	-0.162	0.750	6.423	-0.001	8%
LA	0.463	2.136	-2.051	0.442	2.171	-0.004	-5%
MA	0.496	2.864	-0.277	0.695	6.179	-0.003	29%
MD	0.446	2.602	-0.428	0.278	0.959	-0.012	-60%
ME	0.576	3.089	0.097	0.568	2.045	-0.005	-1%
MI	2.146	167.311	-24.034	1.139	13.776	-0.001	-88%
MN	0.489	2.420	-0.288	0.465	1.298	-0.005	-5%
MO	0.592	3.793	-0.015	0.535	2.615	-0.003	-11%

MS	0.612	4.433	-0.445	0.609	3.453	-0.002	0%
MT	0.494	2.440	-1.003	0.483	2.235	-0.000	-2%
NC	0.680	5.071	0.092	0.587	3.424	-0.003	-16%
ND	0.247	0.617	-3.837	0.351	1.418	-0.003	30%
NE	0.296	0.967	-1.343	0.345	0.835	-0.000	14%
NH	0.406	1.738	-0.127	0.775	5.799	-0.000	48%
NJ	0.554	3.963	0.059	0.813	5.772	-0.004	32%
NM	0.449	2.497	-5.715	0.400	1.137	-0.012	-12%
NV	0.786	7.544	0.282	1.461	19.152	-0.002	46%
NY	0.500	3.224	-0.127	0.627	4.760	-0.006	20%
OH	0.725	5.008	-0.384	0.713	4.661	-0.001	-2%
OK	0.458	1.942	-1.951	0.564	3.226	-0.001	19%
OR	0.640	4.805	0.089	0.512	2.865	-0.006	-25%
PA	0.482	2.925	-0.919	0.644	4.245	-0.004	25%
RI	0.985	10.267	-0.463	0.969	6.669	-0.003	-2%
SC	0.709	5.620	0.115	0.494	2.640	-0.006	-44%
SD	0.345	1.157	-1.360	0.379	1.346	-0.000	9%
TN	0.621	4.293	0.011	0.702	5.267	-0.003	12%
TX	0.486	2.778	-0.245	0.535	2.217	-0.007	9%
UT	0.503	2.639	0.111	0.443	1.935	-0.001	-14%
VA	0.427	2.319	-0.477	0.478	2.467	-0.004	11%
VT	0.404	1.792	-0.423	0.717	5.307	-0.000	44%
WA	0.531	3.486	-0.402	0.689	4.162	-0.002	23%
WI	0.576	3.371	-0.024	0.685	4.627	-0.001	16%
WV	0.577	3.270	-2.133	0.609	3.573	-0.000	5%
WY	0.569	2.490	-1.744	0.271	0.345	-0.014	-110%
USA	0.645	3.943	0.225	0.626	3.491	-0.004	-3%

Figure 8. *Results of BDLM and AR models with wages regressor*

In this configuration, only 22 states performed better with BDLM. Probably the fact that the wages data was available only for 14% of the job postings influences the accuracy of it,

which could still help if those numbers were appropriately distributed. However, it seems not to be the case with the online job postings.

In search for better performance than job vacancy, the wages regressor was rejected and replaced with 12 lags regressors of job vacancy data and 12 lags regressors of unemployment rate. This of course cause the dataset to diminish, which was imposed on the testing set, which now comprised of 24 instead of 36 months. The same, best performing AR model of 5 lags was selected.

State	BDLM MAE	BDLM MSE	BDLM R2	AR MAE	AR MSE	AR R^2	Improvement in %
AK	1.967	13.222	-10.072	0.379	1.715	-0.001	-419%
AL	2.990	36.887	-7.367	0.598	4.156	-0.001	-400%
AR	2.092	15.272	-4.449	0.370	1.195	-0.004	-465%
AZ	2.463	20.060	-5.308	0.545	3.138	-0.004	-352%
CA	3.057	32.629	-2.742	0.628	4.080	-0.010	-387%
CO	2.632	21.641	-2.655	0.464	2.137	-0.012	-467%
CT	2.218	17.075	-2.890	0.374	1.150	-0.027	-493%
DE	2.239	17.642	-4.286	0.562	2.856	-0.006	-298%
FL	3.336	46.660	-6.792	0.622	3.098	-0.004	-436%
GA	2.848	26.197	-4.370	0.512	2.634	-0.002	-456%
HI	3.624	106.428	-7.495	1.032	13.430	-0.004	-251%
IA	1.954	13.077	-7.129	0.483	2.260	-0.002	-305%
ID	2.297	16.469	-2.924	0.588	2.787	-0.001	-291%
IL	3.606	48.007	-6.648	0.666	5.109	-0.004	-441%
IN	3.403	50.930	-8.006	0.791	6.260	-0.001	-330%
KS	2.210	19.228	-9.200	0.577	3.132	-0.001	-283%
KY	2.795	28.266	-5.711	0.750	6.423	-0.001	-273%
LA	2.273	16.084	-7.432	0.442	2.171	-0.004	-414%
MA	2.683	31.377	-4.835	0.695	6.179	-0.003	-286%
MD	2.004	12.911	-4.723	0.278	0.959	-0.012	-621%
ME	2.157	14.832	-3.564	0.568	2.045	-0.005	-280%

MI	3.613	55.777	-5.176	1.139	13.776	-0.001	-217%
MN	1.760	9.814	-3.827	0.465	1.298	-0.005	-278%
MO	3.184	36.200	-9.010	0.535	2.615	-0.003	-495%
MS	2.803	25.198	-6.682	0.609	3.453	-0.002	-360%
MT	2.014	13.808	-8.134	0.483	2.235	-0.000	-317%
NC	2.968	28.930	-4.210	0.587	3.424	-0.003	-406%
ND	1.581	7.303	-5.877	0.351	1.418	-0.003	-350%
NE	1.456	6.829	-12.710	0.345	0.835	-0.000	-322%
NH	2.962	52.374	-13.103	0.775	5.799	-0.000	-282%
NJ	3.339	39.100	-4.220	0.813	5.772	-0.004	-311%
NM	2.049	12.912	-7.644	0.400	1.137	-0.012	-412%
NV	4.918	203.783	-11.186	1.461	19.152	-0.002	-237%
NY	2.572	20.992	-2.338	0.627	4.760	-0.006	-310%
OH	2.824	32.786	-7.271	0.713	4.661	-0.001	-296%
OK	2.049	15.311	-8.455	0.564	3.226	-0.001	-263%
OR	3.200	31.072	-5.261	0.512	2.865	-0.006	-525%
PA	2.511	20.594	-4.627	0.644	4.245	-0.004	-290%
RI	3.619	45.673	-4.363	0.969	6.669	-0.003	-273%
SC	2.772	25.679	-3.248	0.494	2.640	-0.006	-461%
SD	1.443	6.277	-7.209	0.379	1.346	-0.000	-281%
TN	3.054	37.711	-6.702	0.702	5.267	-0.003	-335%
TX	2.624	22.185	-6.215	0.535	2.217	-0.007	-390%
UT	1.843	10.538	-4.084	0.443	1.935	-0.001	-316%
VA	1.935	12.979	-4.507	0.478	2.467	-0.004	-305%
VT	2.795	45.411	-17.908	0.717	5.307	-0.000	-290%
WA	2.859	34.358	-8.654	0.689	4.162	-0.002	-315%
WI	2.699	27.568	-6.480	0.685	4.627	-0.001	-294%
WV	2.710	19.819	-8.659	0.609	3.573	-0.000	-345%
WY	1.487	7.694	-9.089	0.271	0.345	-0.014	-449%
USA	2.665	19.711	-3.409	0.626	3.491	-0.004	-326%

Figure 9. *Results of BDLM and AR models with 12 lags regressors*

Here, the results are worse for all analysed states and the US as a whole. They make the performance worse by hundreds of percent. Therefore, the lags for the BDLM do not necessarily improve the data. This is also connected with how the BDLM works – it is the aggregation of probabilities and past probabilities that not necessarily help the next variable predicted. There are, however, other ways we can use the data set to look for better outcomes.

We can think of seeing economy as connected vessels – all the states follow similar trends and patterns in terms of unemployment, mostly differing by scale and numbers. The country's reflects it. Therefore, we run another experiment in which there are 51 regressors of job vacancy for each state and the country as a whole. This gives us the most data-rich prediction, as it involves 6936 data entries, 136 per regressor. The test and train split are as usual in previous experiments. Then we juxtapose it with best performing AR model.

State	BDLM MAE	BDLM MSE	BDLM R2	AR MAE	AR MSE	AR R2	Improvement in %
AK	4.114	29.389	-74.040	0.432	2.000	-0.001	-852%
AL	5.162	41.954	-10.358	0.668	4.841	-0.002	-673%
AR	4.213	29.179	-10.006	0.406	1.388	-0.007	-938%
AZ	5.232	43.000	-11.840	0.617	3.657	-0.005	-748%
CA	6.295	63.424	-6.981	0.683	4.737	-0.012	-822%
CO	4.612	34.431	-4.630	0.523	2.490	-0.015	-782%
CT	4.934	39.518	-10.052	0.409	1.333	-0.030	-1106%
DE	4.367	29.865	-12.027	0.624	3.324	-0.004	-600%
FL	5.370	45.556	-7.452	0.702	3.610	-0.005	-665%
GA	5.357	47.125	-9.430	0.565	3.067	-0.005	-848%
HI	3.475	19.565	-5.723	1.145	15.621	-0.006	-203%
IA	3.082	15.290	-12.998	0.546	2.635	-0.002	-464%
ID	4.565	31.830	-5.899	0.656	3.242	-0.004	-596%
IL	5.527	49.037	-10.929	0.749	5.954	-0.006	-638%
IN	5.079	40.520	-6.745	0.893	7.296	-0.002	-469%

KS	3.444	19.330	-13.413	0.608	3.594	-0.002	-466%
KY	5.261	42.962	-9.435	0.851	7.488	-0.001	-518%
LA	3.766	25.266	-35.081	0.493	2.528	-0.006	-664%
MA	4.113	27.010	-11.043	0.752	7.177	-0.006	-447%
MD	3.994	26.459	-13.521	0.303	1.113	-0.017	-1218%
ME	4.297	29.519	-7.629	0.634	2.379	-0.005	-578%
MI	6.267	57.797	-7.648	1.227	15.871	-0.003	-411%
MN	3.670	20.714	-10.026	0.520	1.511	-0.008	-606%
MO	4.852	35.525	-8.507	0.611	3.049	-0.004	-694%
MS	5.222	45.085	-13.693	0.699	4.028	-0.003	-647%
MT	3.589	20.562	-15.880	0.543	2.604	-0.001	-561%
NC	5.396	47.519	-7.513	0.649	3.984	-0.005	-731%
ND	1.897	5.624	-43.083	0.374	1.643	-0.004	-407%
NE	2.381	9.163	-21.204	0.384	0.971	-0.002	-520%
NH	3.115	16.059	-9.414	0.863	6.754	-0.001	-261%
NJ	5.086	43.289	-9.275	0.853	6.533	-0.011	-496%
NM	3.931	27.066	-71.796	0.439	1.319	-0.017	-795%
NV	6.986	77.899	-6.416	1.671	22.332	-0.003	-318%
NY	4.655	36.068	-11.606	0.700	5.544	-0.008	-565%
OH	5.196	40.349	-10.153	0.804	5.432	-0.001	-546%
OK	3.229	16.650	-24.292	0.638	3.759	-0.002	-406%
OR	5.314	44.892	-7.509	0.585	3.340	-0.008	-808%
PA	4.214	30.247	-18.840	0.731	4.949	-0.003	-476%
RI	5.939	58.062	-7.271	1.073	7.752	-0.006	-453%
SC	5.567	49.147	-6.737	0.564	3.078	-0.008	-887%
SD	2.431	9.703	-18.796	0.419	1.567	-0.001	-480%
TN	4.993	39.457	-8.093	0.788	6.136	-0.004	-534%
TX	4.306	28.545	-11.797	0.602	2.582	-0.009	-615%
UT	4.265	25.672	-7.650	0.502	2.256	-0.002	-750%
VA	3.645	21.436	-12.652	0.529	2.871	-0.007	-589%
VT	3.147	15.169	-11.049	0.814	6.188	-0.001	-287%
WA	4.509	33.312	-12.395	0.769	4.845	-0.003	-486%
WI	4.327	29.313	-7.908	0.774	5.392	-0.002	-459%
WV	4.280	30.306	-28.040	0.677	4.160	-0.002	-532%

WY	3.324	17.913	-18.743	0.298	0.400	-0.013	-1015%
USA	3.700	28.287	-4.562	0.706	4.067	-0.006	-424%

Figure 10. *Results of BDLM and AR models with all states data as regressors*

As we see, the results are even worse than with lags. Probably the numbers, though similar in trend, do not resonate well with the model and do not count for its precision because of a high variety. Therefore, using different states unemployment data to predict only the state of interest has no justification.

However, one could argue that it is the specificity of the BDLM model that makes it more accurate than the AR model, not necessarily the extra regressor in form of job vacancy data. Therefore, in the next experiment we ran BDLM as univariate model, meaning basing the results only on the unemployment rate – the same data that AR model has. The results are juxtaposed below:

State	BDLM MAE	BDLM MSE	BDLM R2	AR MAE	AR MSE	AR R^2	% improvement
AK	0.602	2.007	-0.604	0.379	1.715	-0.001	-59%
AL	1.181	3.653	0.339	0.598	4.156	-0.001	-97%
AR	0.796	1.969	0.355	0.370	1.195	-0.004	-115%
AZ	0.975	3.317	0.165	0.545	3.138	-0.004	-79%
CA	1.525	6.614	0.336	0.628	4.080	-0.010	-143%
CO	1.196	3.877	0.421	0.464	2.137	-0.012	-158%
CT	1.137	3.477	0.285	0.374	1.150	-0.027	-204%
DE	0.970	3.450	0.091	0.562	2.856	-0.006	-73%
FL	1.315	4.867	0.325	0.622	3.098	-0.004	-111%
GA	1.143	3.465	0.411	0.512	2.634	-0.002	-123%
HI	1.312	9.181	0.218	1.032	13.430	-0.004	-27%
IA	0.655	1.647	0.092	0.483	2.260	-0.002	-36%
ID	1.028	2.929	0.425	0.588	2.787	-0.001	-75%
IL	1.292	5.727	0.149	0.666	5.109	-0.004	-94%

IN	1.238	4.723	0.301	0.791	6.260	-0.001	-57%
KS	0.726	1.982	0.086	0.577	3.132	-0.001	-26%
KY	1.023	3.878	0.256	0.750	6.423	-0.001	-36%
LA	0.715	2.149	-0.148	0.442	2.171	-0.004	-62%
MA	1.102	4.903	0.098	0.695	6.179	-0.003	-59%
MD	0.846	2.148	0.149	0.278	0.959	-0.012	-204%
ME	0.936	2.370	0.394	0.568	2.045	-0.005	-65%
MI	1.574	8.980	0.156	1.139	13.776	-0.001	-38%
MN	0.801	2.185	0.138	0.465	1.298	-0.005	-72%
MO	1.026	3.021	0.348	0.535	2.615	-0.003	-92%
MS	0.915	3.183	0.147	0.609	3.453	-0.002	-50%
MT	0.680	1.829	0.003	0.483	2.235	-0.000	-41%
NC	1.201	4.288	0.349	0.587	3.424	-0.003	-105%
ND	0.448	1.042	-0.041	0.351	1.418	-0.003	-28%
NE	0.391	0.700	-0.189	0.345	0.835	-0.000	-13%
NH	0.843	3.452	0.068	0.775	5.799	-0.000	-9%
NJ	1.266	6.017	0.204	0.813	5.772	-0.004	-56%
NM	0.669	1.984	-0.348	0.400	1.137	-0.012	-67%
NV	1.884	13.118	0.259	1.461	19.152	-0.002	-29%
NY	1.157	5.343	0.131	0.627	4.760	-0.006	-85%
OH	1.152	4.604	0.087	0.713	4.661	-0.001	-62%
OK	0.697	2.037	-0.140	0.564	3.226	-0.001	-24%
OR	1.198	4.045	0.340	0.512	2.865	-0.006	-134%
PA	0.847	3.654	-0.005	0.644	4.245	-0.004	-32%
RI	1.436	6.207	0.339	0.969	6.669	-0.003	-48%
SC	1.339	4.422	0.430	0.494	2.640	-0.006	-171%
SD	0.440	0.914	-0.072	0.379	1.346	-0.000	-16%
TN	1.166	4.071	0.283	0.702	5.267	-0.003	-66%
TX	0.915	2.874	0.149	0.535	2.217	-0.007	-71%
UT	0.859	2.219	0.299	0.443	1.935	-0.001	-94%
VA	0.869	2.332	0.132	0.478	2.467	-0.004	-82%
VT	0.770	2.461	0.080	0.717	5.307	-0.000	-7%
WA	0.977	3.594	0.101	0.689	4.162	-0.002	-42%
WI	1.002	3.311	0.239	0.685	4.627	-0.001	-46%

WV	0.770	2.757	-0.222	0.609	3.573	-0.000	-26%
WY	0.706	1.431	-0.340	0.271	0.345	-0.014	-161%
USA	1.090	3.934	0.226	0.626	3.491	-0.004	-74%

Figure 11. *Results of BDLM univariate model and AR model*

As we see, the results are worse than in case of BDLM with job vacancy regressor. In fact, they are much worse for every single state. In some cases, the performance drops over 200% which indicates that even though the model has more elaborate structure of work than autoregressive model, it does not mean it can have easily better performance. However, it also proves that the job vacancy data make the real difference in the calculations. Let us however discuss is in the bigger picture in the next chapter.

Chapter 5

Analysis and Discussion

5.1 Overall assessment

The thorough calculations conducted consisted of a fair number of models. The results measured in MAE, MSE and R^2 gave an outlook on models' performance. Interestingly, the graphs were of very similar tendencies, yet sometimes the results between states differed a lot. We could see those tendencies on graphs where all states would be displayed, however, this would make this project unreadable. Therefore, we could see an example of it on a smaller sample.

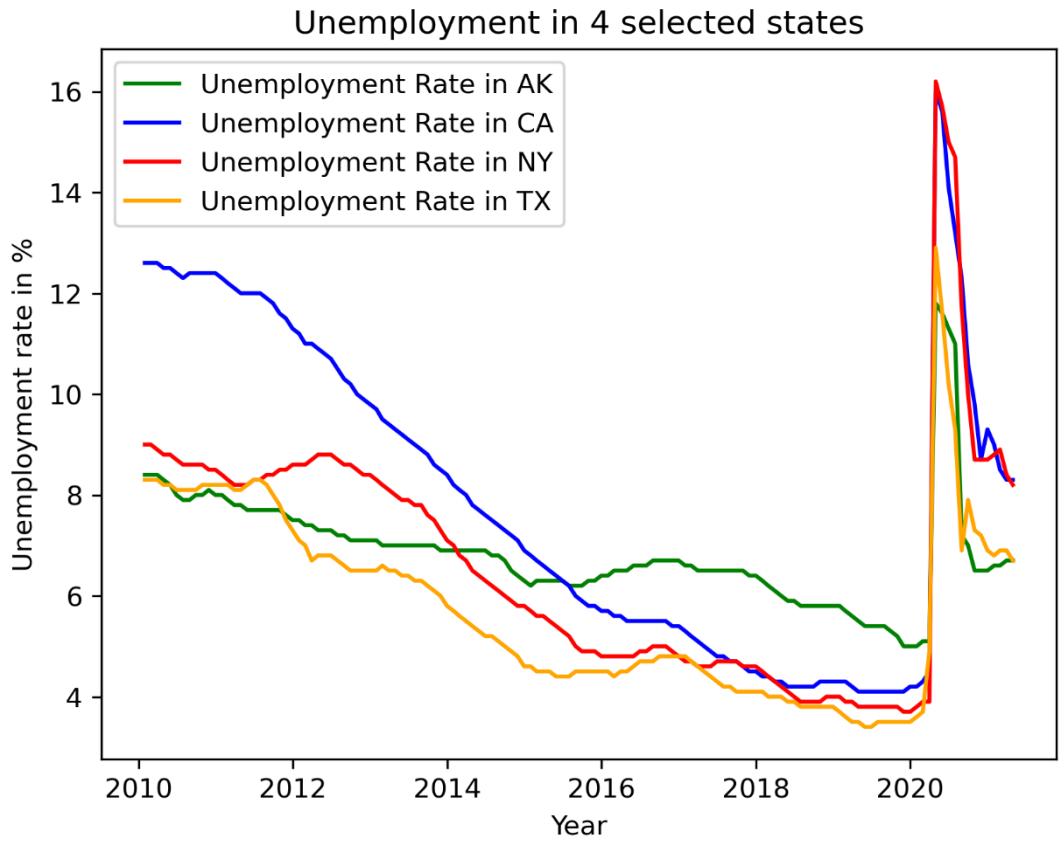

Figure 12. – *Unemployment rate in the state of AK, CA, TX, and NY.*

This is displayed in Figure 12 which presents unemployment rate from four states from different corners of the United States – Alaska, New York, Texas, and California. One can see a clear correlation between them, though economic activity of Alaska does not resemble that of for example New York. Yet, the results between those states sometimes differ by more than few percentage points. This can be a result of limited data probe and here probable additional information, such as inflation, GDP, etc. could help stabilize the score and make it more similar in terms of performance.

However, we can see on this example of those four states that the states follow in general the trends of the country, therefore we will focus mostly on the data for the USA. It will

help us generalize some concepts and also has the largest benefit for potential beneficiaries of this work.

It is worth noting, that the analysed period encompassed a very peculiar one in the unemployment rate history. Though it followed the usual cyclical trend of booming economy, that is as economy was booming the unemployment rate was falling down, according to the business cycle [25], it experienced a sudden and unexpected bump. This was the outbreak of COVID-19 when suddenly many businesses started firing employees. Though it is not worth getting into reasons of this behaviour, it is a statistically significant fact. We can see it reflect in the data where the MAE is relatively low for all the graphs but MSE is sometimes way higher as MSE represents the difference between the original and predicted values extracted by squared the average difference over the data set and MAE is the same difference but not squared but rather averaged. This helps to diminish the impact of "COVID bump" on the graph. However, it is worth noting that with this data available it is impossible for a linear model to predict this sudden increase in unemployment rate. Probably the input from other external sources, using NLP on the news or some outside hospitalization record could help predict it. However, with the data used for this project it should be treated as an anomaly which was amended to the best possible extent with fine tuning of the models.

The R^2 in the results helps us see whether the model fits correctly. The higher the value, the more accurate the model is (maximum is 1). We can see that the behaviour of BDLM and the use of probabilities helps make the model way closer to the real results than Autoregressive model, proving DLM's accuracy.

Surprisingly, more regressors was always the better choice. The wages data was rather not useful and decreased the accuracy of the results. However, the regressors of lagged

unemployment rate and job vacancy data had a positive impact on the model yet were not as accurate as simply using online vacancies.

Last but not least, the BDLM proven to work better when it was multivariate, compared to univariate. Meaning that the additional online job vacancy data was a pure reason for the precision of results.

5.2 Results Limitations

The presented results of course have their limitations which we have to be aware of in their analysis.

First it is the time span on which the data was analysed. There are accurate measurements of unemployment with modern techniques already in 1950s. However, it is also a necessary limitation as there is no meaningful number of online job vacancies outside of this time-zone.

Second, there are possibilities that online job vacancy data numbers were not complete. For sure they show a real trend of growth, however their number could differ in real world.

Third is connected to first. As data sample improves, one can show better results, otherwise the models risk overfitting. Though this was prevented with all possible measures when conducting the calculations, probably with time, as more data becomes available, even better results will be possible.

5.3 Possible future steps

There are some actions which could be done to further use the data set and provide clarifications.

First, would be to use the same methods with data split before Covid and during the pandemic. Though it limits the data set to two relatively small ones, it could be interesting to see the implications of the pandemic on the results as well as the adjustment of models in the dynamically changing months of COVID-19 outbreak and sudden changes in unemployment.

Another measure would be to try extracting more significant data from Burning Glass dataset. Though the main factors such as wages and vacancy count were already used in this project, one could try analysing the wordings of job offers and see if it provides any meaningful data for the analysis of unemployment.

Chapter 6

Conclusion

It can be clearly stated that the online job vacancy data is helpful in predicting the unemployment rate. In fact, it gets much better results than the results given by the univariate model of both AR and the BDLM itself. The Burning Glass data can be therefore a powerful resource of external data, outside of that monitored by the governments, to help predict the macroeconomic factor of unemployment rate. Moreover, it has proven useful in a challenging situation which is the COVID-19 crisis. The results of this work therefore can be seen as a good point for further exploration of this dataset. It is also yet another example of how data rich environments can enhance macroeconomic variables predictions and how valuable are the new sources of data to push the machine learning world forward.

Bibliography

[1] Moor, P. (2015). The birth of the weather forecast. BBC News. Available at: <https://www.bbc.com/news/magazine-32483678> Accessed 10th May 2021.

[2] Tabata, T. (2015). Re: What is the difference between 'forecasting' and 'nowcasting' which can be used for linear assets or asset health monitoring and control? Available at: <https://www.researchgate.net/post/What_is_the_difference_between_forecasting_and_nowcasting_which_can_be_used_for_linear_assets_or_asset_health_monitoring_and_control/5645c97e6225ffa5568b4567/citation/download>. Accessed 16th May 2021.

[3] Nowcasting (n.d.) In Merriam-Webster's collegiate dictionary. Available at: <https://www.macmillandictionary.com/dictionary/british/nowcasting>. Accessed 25th May 2021.

[4] Zoubier, L. (2017). A brief history of time series analysis. Stockholm University. Department of Statistics. Available at: <https://www.statistics.su.se/english/research/time-series-analysis/a-brief-history-of-time-series-analysis-1.259451>. Accessed 1st June 2021.

[5] Box, E., P., B., Jenkins, M., G. (1970) Time Series Analysis: Forecasting and Control. Holden-Day.

[6] Webb, A. (2021). The Rise of Quant Funds. Future Today Institute. Available at: <https://futuretodayinstitute.com/trend/the-rise-of-quant-funds/>. Accessed 10th June 2021.,

[7] Zuckerman, G. (2019). How Billionaire Jim Simmons Learned to beat the Market – And Bean Wall Street's Quant Revolution. Forbes. Available at: <https://www.forbes.com/sites/forbesdigitalcovers/2019/11/08/jim-simons-the-man-who-solved-the-market-gregory-zuckerman-book-excerpt/?sh=594742b913b6/>. Accessed 8th June 2021.

[8] Athey, S. (2018). The impact of machine learning on economics. The Economics of Artificial Intelligence, NBER volume, Forthcoming.

[9] Higgins, P. (2014). GDPNow: A model for GDP "Nowcasting". Working Paper 2014-7. Federal Reserve Bank of Atlanta.

[10] Cimadomo, J., Giannone, D., Lenza, M., Monti, F., Sokol, A. (2020). Nowcasting with Large Bayesian Vector Autoregressions. Working Paper Series 2453, European Central Bank.

[11] Richardson, A., Van Florenstein Mulder, T., Vehbi, T. (2018). Nowcasting New Zealand GDP using machine learning algorithms. Reserve Bank of New Zealand.

[12] Rea, D. and Malone, T. (2021). Nowcasting the current rate of unemployment using administrative data. Centre for Social Data Analytics Auckland University of Technology.

[13] Moriwaki D., (2020). Nowcasting Unemployment Rates with Smartphone GPS Data. Springer International Publishing. SN - 978-3-030-38081-6. ID - 10.1007/978-3-030-38081-6_3.

[14] Koop G., Onorante, L., (2013). Macroeconomic Nowcasting Using Google Probabilities. A chapter in Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part A, 2019, vol. 40A, pp 17-40 from Emerald Publishing Ltd.

[15] Claviera, O. (2019). Forecasting the unemployment rate using the degree of agreement in consumer unemployment expectations. J Labour Market Res (2019) 53:3 https://doi.org/10.1186/s12651-019-0253-4.

[16] West, M., Harrison J., (1989). Bayesian Forecasting and Dynamic Models. Springer-Verlag New York.

[17] Coulombely P. G., Leroux M., Stevanovicz D., Supernant S. (2019). How is Machine Learning Useful for Macroeconomic Forecasting? University of Pennsylvania & Université du Québec à Montréal.

[18] Federal Reserve Bank of New York (2021). Nowcasting Report. Available at: <https://www.newyorkfed.org/research/policy/nowcast.html.https://www.newyorkfed.org/research/policy/nowcast/overview.html/>. Accessed 15th June 2021.

[19] Bańbura, M., Giannone, D., Modugno, M., Reichlin, L. (2013). Now-casting and the real-time data flow. Working Paper Series 1564, European Central Bank.

[20] Doerr, S., Gambacorta L., Serena, M., J. (2021). Big data and machine learning in central banking. BIS Working Papers No 930. Monetary and Economic Department.

[21] Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. Annual Review of Economics Vol. 10:615-643.

[22] Carnevale A. P., Repnikov T., D. (2014). Understanding Online Job Ads Data. A Technical Report. Georgetwon University. Center on Education and the Workforce.

[23] Qiu, J., Jammalamadaka R., S., Nig, N. (2018). Multivariate Bayesian Structural Time Series Model, Journal of Machine Learning Research 19 (2018) 1-33.

[24] Harvey, A., C. (1991). Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, DOI 10.1017/CBO9781107049994.

[25] Durbin, T., Koopman, S. (2012). Time Series Analysis by State Space Methods, 2nd edition. Oxford Statistical Science Series, Oxford University Press, DOI 10.1093/ acprof:oso/9780199641178.001.0001.

[26] Petris, G., Petrone, S., Campagnoli, P. (2009). Dynamic Linear Models with R. Use R!, Springer.

[27] Bhattacharjee, J., (2019). Common metrics for Time Series. Medium. Available at: <https://joydeep31415.medium.com/common-metrics-for-time-series-analysis-f3ca4b29fe42/> Accessed 18th August 2021.

[28] Elliott, G., Rothenberg, T., Stock, J. (1996). Efficient Tests for an Autoregressive Unit Root. Econometrica, Vol. 64 No. 4, 813-836. doi:10.2307/2171846.

[29] Petzoldt, T., Van den Boogaart, K., Jachner, S. (2007). Statistical Methods for the Qualitative Assessment of Dynamic Models with Time Delay. Journal of Statistical Software. 22. 10.18637/jss.v022.i08.

[30] Christiana L., J., Eichenbaum M., S., Trabandt, M. (2016). Unemployment and Business Cycles. Econometrica, Vol. 84, No. 4 (2016), 1523–1569.