FROM CLICKS TO CAREERS: WHAT UKRAINE'S MOST POPULAR JOB SEARCH WEBSITE REVEALS ABOUT REGIONAL EMPLOYMENT PATTERNS

by

Petro Nochovkin

A thesis submitted in partial fulfilment
of the requirements for the degree of

MA in Economic Analysis

Kyiv School of Economics

2024

Thesis Supervisor: Professor Mihnea Constantinescu

Approved by: ____
Head of the KSE Defense Committee, Professor
____
____
____

Date ___________________________

Kyiv School of Economics

Abstract

FROM CLICKS TO CAREERS:
WHAT UKRAINE'S MOST POPULAR JOB SEARCH WEBSITE REVEALS ABOUT REGIONAL EMPLOYMENT PATTERNS

By Petro Nochovkin

Thesis Supervisor: Professor Mihnea Constantinescu

The ability to accurately and timely forecast labour market outcomes is essential for effective economic policymaking. However, the ongoing war in Ukraine has significantly limited the ability of statistical bodies to conduct traditional labour market surveys on the regional level. The polling conducted by the State Statistics Service as part of the labour market survey has been suspended for the unforeseen period. This has created a need for alternative methods of nowcasting labour market outcomes.

Web-scraping techniques offer a promising approach to nowcasting labour market outcomes in Ukraine. By studying data from employment websites, such as work.ua, it is possible to track changes in job postings and salaries. No research detailing their validity for the prediction of the changes in the numbers of employed or changes in average salaries for different regions of Ukraine has been published so far. If tested and proven efficient, the estimates can then be used to develop econometric models that can nowcast key labour market outcomes

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

DMFs. Dynamic factor models.

GDP. Gross Domestic Product.

CPI. Consumer price index.

CEA. Classification of Economic Activities.

ECB. European Central Bank.

ECM. Error Correction Model.

NBU. National Bank of Ukraine.

RSSSU. Regional State Statistics Services of Ukraine.

SSSU. State Statistics Service of Ukraine.

OLS. Ordinary least squares model.

TVREG. Time-variant regression models.

TVPLM. Time-variant panel regression model.

VC. Variant coefficients.

VAR. Vector autoregression.

Chapter 1

INTRODUCTION

The ability to accurately and timely forecast labour market outcomes is essential for effective economic policymaking. Yet, multiple barriers hinder the process of obtaining this kind of data: the traditional surveys take a lot of time to conduct and the necessary unbiased sampling requirements are often hard to comply with. The ongoing war in Ukraine has significantly limited the ability of statistical bodies to conduct traditional labour market surveys on the regional level. The application of the nowcasting methods might allow for obtaining higher frequency and more detailed data for forecasting, as compared to the official data, such as online polling being conducted by the State Statistics Service of Ukraine (SSSU, 2023). Alternative methods of forecasting labour market outcomes can suggest policy solutions now and today, which makes them more relevant and helpful.

Nowcasting techniques offer a promising approach to approximating changes in labour market outcomes in Ukraine. By obtaining web-scrapped data from employment websites, it is possible to track changes in job postings, job searches, and other labour market indicators. No research detailing their validity for the estimation of the unemployment rate or the changes in the numbers of employed population for various regions of Ukraine has been published so far. If tested and proven efficient, the estimates can then be supplemented with other nowcasting variables and further used to develop econometric models that can nowcast key labour market outcomes, such as the unemployment rate, for various regions of the country.

The research question tackled in this thesis considers the potential usefulness of using job search volumes, salary and vacancy indicators to nowcast official regional indicators. Since the existing research proved that such relationships were significant for official wage inflation prediction and changes in the unemployment rate, there are reasons to consider the possibility that job search websites might be useful for nowcasting other labour market variables as well.

The first hypothesis is that the number of resumes, vacancies and salaries from the job search website will be good predictors of the official labour market statistics. This is because previous research discovered only a slightly statistically significant relationship in the earlier data. However, earlier studies have been mostly attempting to predict the pre-COVID dynamics of change, while the popularity of the target job search website has not yet reached its full potential.

The second hypothesis is that the efficiency of the forecast will differ across different professions and regions of Ukraine - partially due to the differing popularity of online job search for different professions and regions.

The exact econometric technique related to the task of labour market nowcasting is confined to the application of time-variant fixed effect and ECM regression models: the primary reason for the use of this type of models is based on the estimates of their out-of-sample forecast performance. The exact version of the model best suited for nowcasting was determined from the comparison of statistical errors of the models' forecast. The final nowcast predicted statistically significant growth in salary for 9 out of 15 studied professions (based on the classification of economic activities). The entire manufacturing sector has been excluded from the forecast due to the lack of statistically significant relationship throughout the analysis of historical data. The nowcast for the employment was conducted based on the ECM-type model, which predicted changes in the regional labour market indicators in the regions of Ukraine with the highest accuracy.

C h a p t e r 2

LITERATURE REVIEW

The literature on economic nowcasting is quite extensive despite the relevant novelty of the research area. The most generalising study of the available nowcasting data sources and techniques can be confined to Buono et al. (2017), which provides a comprehensive overview of the most popular available options of nowcasting data for economic forecasting. However, since the nature of this paper is much more limited and confined to the labour market issues, there is a necessity to divide the further literature review into two separate sections: a section on the studies in nowcasting of the labour market and a section on the application of time-variant regressions for the purposes of nowcasting.

2.1. Empirical studies on the labour market nowcasting

With regards to the labour market, the earliest use of nowcasting appeared in the paper of Mclaren and Shanbhogue (2011) who used a volume of online searches to track labour and housing markets in the United Kingdom, as well as Askitas and Zimmermann (2009) who estimated the efficiency of internet search use for the estimation of the contraction in the German labour market during the 2008 Great Recession.

The most relevant study concerning labour market nowcasting is the recent ECB paper of the previously mentioned authors (Bańbura et al 2023). The paper describes the utilisation of bridge equations with economic sentiment indicators and unemployment rate data for different age groups. Furthermore, it provides a short review of the application of Dynamic Factor Models (DMFs) concerning quarterly unemployment forecasting based on monthly indicators. The bridge equations rely on several indicators, that are directly linked to the variable of interest. In the case of the DFMs, the variable of interest is driven by several common factors, which can summarise a large volume of data. Despite offering a nice outline with regards to the general structure of a nowcasting model, the application of its methodology to Ukrainian data is severely limited since the official employment information has not been collected since the start of the wartime period.

In order to look into the availability of possible solutions for the case of information deficit during the crisis period, another recent paper by Gomis et al. (2022) provides a good outlook of the ILO nowcasting model utilised for the estimation of unemployment during the COVID-19 pandemic, which is primarily based on the pooled survey data supplemented by other high-frequency variables, including Google mobility reports and job vacancies. In addition to offering some interesting insights concerning the use of time-variant regressions, the methodology can be helpful for the estimation of the hidden unemployment in Ukraine since it's been primarily focused on the estimation of changes in working hours.

With regards to the relevant literature and research in the same area done specifically for Ukraine, the existing studies for the war-time period mostly focus on nowcasting GDP as the economic variable with the best predictability with the use of nowcasting indicators Constantinescu et. al (2022). However, in terms of the most recent relevant research conducted for Ukraine in the context of the labour market, it is worth mentioning one promising paper by Faryna et al. (2021), which analyses official data on wage and unemployment from the State Statistics Service to find causal evidence between vacancy placement on the job search website and the official labour market statistics: the particular source of data utilised in this study was OLX.ua, one of the leading online advertisement platforms in Ukraine. The monthly data covered the period from 2016 to 2020, researchers attempted to analyse the inflationary impact of the job vacancy salaries on the official salary statistics. In order to reduce the likelihood of misreporting bias, the authors excluded the salary data below the 10th percentile (1,000 UAH) and trimmed the offered wages at the 0.5th and 99.5th percentiles by region and by category. Importantly, the paper concludes that despite the relatively short period of analysis online vacancy data can be used to approximate the official statistics on wage dynamics in Ukraine. When different types of heterogeneity are taken into account - in particular, sectoral or occupational heterogeneity - the accuracy of the forecast increases. Therefore, further research in this area, aimed at estimating the predictability of regional labour market indicators for different categories of economic activity can be hugely beneficial for the development of regional forecasting models as well as recovery and development plans. Thus, my paper will contribute to the body of literature and research in this particular area.

2.2. Time-varying regression models as a tool for nowcasting

Models with time-variant coefficients exhibit four beneficial factors, which make them particularly useful for nowcasting the labour market indicators. Firstly, they help to deal with the non-stationary behaviour of the long-term panel data observations for the target labour market indicators (mean and variance change over time). Secondly, they help to account for the technological and behavioural shifts in the labour market throughout the studied period, such as the one caused by COVID-19. Thirdly, it allows to factor in changing interactions between the publication of target online indicators (resumes/vacancies) and actual employment, which might change drastically during the periods of economic downturns and post-war recovery. Fourthly, models with time-variant coefficients adjust to changing dynamics, making the prediction intervals they produce more accurate. These intervals can provide more reliable estimates of uncertainty around forecasts, which is crucial for decision-making in economic planning and policy.

The flexibility of the time-variant models can be particularly useful in capturing sudden or gradual changes in how different predictors influence the response variable can be viewed as a potential solution for analysing the labour market in Ukraine throughout the calamity of the last 10 years - although not entirely new (first formalized by a random walk (Cooley and Prescott 1973; Schlicht 2020; Athans, 1974), the models have not been widely applied to the estimates of the labour markets. Most notable examples of their use belong to the areas of stock market performance analysis (Dangl et al. 2012), where the study evaluates predictive regressions that explicitly consider the time-variation of coefficients in a Bayesian framework, as well as the analysis of the forex market performance (Taveeapiradeecharoen and Aunsri 2020), where the use of large-dimensional vector autoregressive (VAR) models with time-variant coefficients is discussed.

The key assumption of a model with time-variant coefficients in all of these cases, however, is that the coefficients change only slowly over time (and therefore are highly auto-correlated).

The first detailed technical overview of the time-variant models used for estimations in the labour market is presented in Bańbura et al. (2013). Since this particular study dwells on the application of the bridge equations and MIDAS regressions for the utilisation of the high-frequency data for the explanation and forecasting of equally-spaced low-frequency data (monthly, quarterly or annual), it is not relevant to our case where the instruments used for applications are different.

The basic outline of the time-variant models is described in Schlicht (2020), which details the possible mathematical representation of time-variant coefficients in different versions of regression models. The paper demonstrates that the method of variant coefficients (VC) allows for various dynamics of time-varying coefficients, including random walks, smooth trends, and discrete jumps, as well as offers an alternative to penalized least squares estimation, potentially providing more accurate estimates for specific coefficient paths. This flexibility makes it suitable for modelling diverse real-world phenomena from areas of finance to climate change and labour markets, as mentioned earlier, with varying degrees of coefficient volatility.

Shamiri et al. (2021) introduced a monthly Nowcast of Employment by Region and Occupation (NERO) in Australia (combining the results from a series of linear regression models with the results of several ML regression, such as Random Forest (Breiman 2001), Gradient Boosting (Friedman 2001), and elastic net regression (Zou and Hastie, 2005). Having run the Random Forest, Gradient Boosting and Elastic Net models, initial estimates were combined or stacked based on their hyperparameters to produce a single, optimal set of nowcasted indicators. The performance of the model was evaluated using three measures – the MAPE, WAPE and RMSE. Upon the successful testing of the model, the results were followed by the publication of a methodology by Shamiri et al. (2022). The model has been adopted by the Australian Bureau of Statistics and has been used since to produce regular monthly reports on labour market indicators (JSA, 2023).

Another new and promising method of application for the time-varying coefficients with real panel data models has been described in a recent publication of The R Journal (Casas and Rubén 2022) The usage of time-variant regression models (TVREGs) will allow higher model interpretability by allowing the coefficients to vary smoothly or discretely over time. This class of models provides detailed insights into how relationships between variables evolve, identifies turning points and regime changes, and can be easily interpreted.

Both DFMs, MIDAS and time-variant regressions, as well as ML models, can be used to analyze dynamic relationships in time series data: all of them can be used to identify trends, regime changes, and causal relationships with different degrees of success for different labour market indicators. However, DFMs focus on identifying latent factors that drive the dynamics, while TVREGs focus on how individual relationships change over time. For this reason, the choice of time-variant regressions is preferred for the forecasting of labour market indicators, as in this case, their transformation more long-term dynamic factors (ageing society, quality of education), as well as short-term exogenous shocks (impact of war), come into play.

The literature reviewed underscores the advantages of using regression models with time-varying coefficients for forecasting macroeconomic variables in dynamic environments. To date, few studies have employed such models to analyse labour market indicators, specifically employment and wage data.

Notably, in the context of Ukraine and its regional dynamics, no existing labour market studies have utilised this methodological framework to examine the impacts of recent economic shocks. Consequently, testing and applying this type of model for economic nowcasting purposes not only holds great potential but also represents a significant contribution to global research in this field.

C h a p t e r 3

METHODOLOGY

3.1. Tagging economic activity and job searches

The first noteworthy problem with the data that needs to be addressed in the methodology section is the representability of work.ua data for the Ukrainian labour market. According to Factum Group, work.ua has been the most popular job search website in Ukraine for the last 9 years (Altunina 2015). The traffic data which can be accessed from Similarweb tracking platform supports this statement. Although this might not have been true for 2013, from which the initial vacancies data was collected, this should be more or less true since 2017, when the website has reached over 4 million weekly views: as the popularity of the job search websites grew over time, they became more and more representative of the general changes in the labour market. As can be seen from the official website host statistics (i.ua 2024), between 2017 and 2023 no significant change in the popularity of the website has been recorded. However, to correct this developing trend, a separate proxy variable capturing the trend of website growth based on the number of visits might be added to the estimated model.

As mentioned in the literature review, the nowcasting model based on online vacancies might exhibit heterogeneity caused by the pro-cyclical nature of the individual wages and composition bias, caused by their aggregation (Faryna et al. 2021). Although the more obvious solution to combat this problem would be to do data segmentation or provide the addition of more independent variables into the pooled model to increase the precision of the nowcasting. An alternative approach might be taken: the model in question might be estimated for different regions and different professions. The National Classification of Economic Activities (hereinafter referred to as CEA) used by the Statistics Services to report the official employment data would be an obvious choice. To compare it to the categories provided by the job search website work.ua, a tagging table has been constructed to match the categories from the CEA to work.ua categories.

Table 1: Tagging table for professional categories on work.ua and state CEA

CEA Code	**Vacancy category according to the State Classification of Economic Activities**	**work.ua categories (combined)**	**Type of tagging**	**Trustworthiness**
A	Agriculture, forestry and fisheries	Agriculture	Direct	Solid
X*	Total industry	Craft workers	Direct	Doubtful
B	Mining and quarrying	-	-	-
C	Processing industry	-	-	-
D	Supply of electricity, gas, steam and air conditioning	Service sector	Direct	Doubtful
E	Water supply; sewerage, waste management	Service sector	Direct	Doubtful
F	Construction	Construction, architecture	Direct	Solid
G	Wholesale and retail trade; repair of motor vehicles and motorcycles	Retail	Direct	Solid
H	Transport, warehousing, postal and courier activities	Transport, Logistics	Merged	Solid
I	Temporary accommodation and catering	Hotels, restaurants, tourism	Direct	Solid

CEA Code	Vacancy category according to the State Classification of Economic Activities	work.ua categories (combined)	Type of tagging	Trustworthiness
J	Information and telecommunications	IT, Telecommunications, Media, publishing	Merged	Solid
K	Financial and insurance activities	Finance, Accounting, Insurance	Merged	Solid
L	Real estate transactions	Real estate	Direct	Solid
M	Professional, scientific and technical activity	Design, creativity, Marketing, PR, Law	Merged	Solid
N	Activities in the field of administrative and auxiliary services	Middle management, HR, Sales, procurement, Secretariat	Merged	Solid
O	Public administration, defence; compulsory social insurance	Top management, Security	Merged	Doubtful
P	Education	Education, science	Direct	Solid
Q	Health care and provision of social assistance	Healthcare, pharma	Direct	Solid
R	Arts, sports, entertainment and recreation	Culture, Beauty, fitness, sports	Merged	Solid
S	Provision of other types of services	Other	Direct	Doubtful

* Letter used to denote aggregate column for the total industry workers, as reported by the Regional Statistical Services

Since in general, the categories provided by work.ua have been more detailed than the categories classified in the CEA, several of the job search categories were combined to comprise a unified CEA category. The "Trustworthiness" column provides the subjective estimate of the quality of such a merger: as can be seen from the table above, Sectors B 'Mining and quarrying' and C 'Processing industry' had to be excluded from our analysis entirely due to the lack of corresponding category in work.ua. Categories D 'Supply of electricity, gas, steam and air conditioning' and E 'Water supply; sewerage, waste management' suffered a similar fate due to the difficulties in finding evidence for tagging the services worker category to the jobs in the utilities sector. As for the remaining 16 CEA categories, 9 indicators were tagged to work.ua analogues directly and 7 were combined from several work.ua categories. out of the total of 16 categories, the subjective estimate of the quality of tagging was the following: 13 CEA codes were classified as solid, and 3 were classified as 'doubtful' due to the risk of missed vacancies (the corresponding CEA category is broader). The following research estimates will be primarily focused on the analysis of the 'solid' indicators to avoid potential discrepancies.

With regard to the regional selection, the choice was made in favour of the regions with the highest scores in the regional index of digital transformation, prepared by the Ministry of Digital Transformation in 2022. As can be seen from the chart below, 4 out of top 10 regions in the index are the oblast with the largest population. 2 of the leaders in population numbers did not make it to the top 10 of the regional sample, but were included in our analysis due to easy data availability.

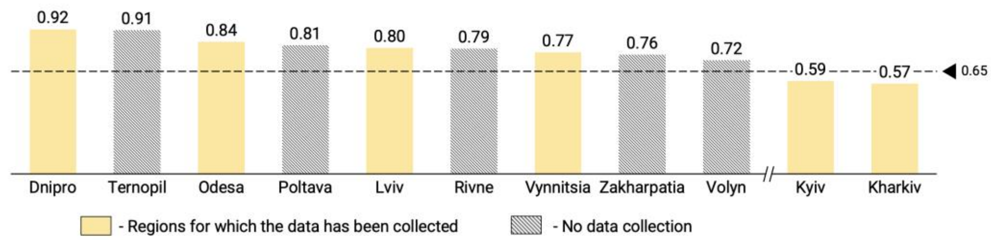

Figure 1. Top-10 regions with the highest access to the Internet (Ministry of Digital Transformation 2022)

Since the index consists not only of assessing the level of access to the Internet but also evaluates 5 other dimensions (Institutional capacity of digital transformation, development of centers for the provision of digital public services, accessibility of regional administrative sites, level of digital education, electronic document management, penetration of basic e-services, as well as industry digital transformation), it makes sense to use it only as an additional reference point.

Similarly to Faryna et. al (2021), we focus on forecasting the data for regions, which have the highest ratio of online vacancies to population (Kyiv (the capital and the region), Dnipropetrovsk region, Odesa region, Kharkiv region and Lviv region). However, in future, extensive research might be done to conduct estimates for other regions included in the top-10 of the highest ranking regions in the Index of Digital Transformation.

Another potential data problem that needs to be addressed in the methodology section of the paper is that for some regions the observational period is relatively short - as the indicators of the most popular Ukrainian job search websites demonstrate, work.ua has only become popular in 2017, as indicated by the number of monthly visits, which makes it logical not to use earlier data, although several Regional Statistics Services did provide their labour market data upon request starting as early as 2012. Most data prior to 2017 could not be accessed online, which limited the application of quarterly and annual data for forecasting purposes.

3.2. The setup of the TV- model

The basic concept of a time-varying regression model can be outlined as follows: Suppose dependent variable y is a linear function of some independent variables $x_1, x_2, \ldots, x_n$, and assuming that the number of observations is limited, with T denoting the time of the observations, the simple linear regression model can be written as:

$$y_t = a_1 x_{1,t} + a_2 x_{2,t} + \ldots + a_n x_{n,t} + u_t; \; t = 1,2,\ldots T. \qquad (1)$$

Where u_t denotes a standard error, and $a_1, a_2, \ldots, a_n$ denote the coefficients for different independent variables.

Suppose the coefficients $a_1, a_2, \ldots, a_n$ have the impact on the independent variable but assume that the coefficients change only slowly over time (as mentioned earlier, they are highly auto-correlated), so that:

$$a_{i,t+1} = a_{i,t} + v_{i,t} \qquad (2)$$

Where $v_{i,t}$ is a disturbance term with expectation 0 and variance ${\sigma^2}_i$. Note that the case of constant (time-invariant) coefficients is covered as well if ${\sigma^2}_{i,} = 0$ for some *i*. Then, the standard regression model can be rewritten as:

$$y_t = a_1 x_{1,t} + a_2 x_{2,t} + \ldots + a_n x_{n,t} + u_t;$$

$$E\{u_t\} = 0; \; E\{u^2{}_t\} = \sigma^2 \qquad (3)$$

$$a_{i,t+1} = a_{i,t} + v_{i,t}$$
$$E\{v_t\} = 0;\ E\{v^2{}_t\} = \sigma^2 \quad (4)$$

Since labour markets are inherently dynamic and subject to a wide range of temporal fluctuations, to accurately nowcast labour market indicators, it's essential to account for the changes in intertemporal relations between variables. For this reason, the usage of time-variant parameters has the potential to provide model flexibility and spatial variability needed to accurately nowcast labour market indicators. Time-variant parameters will allow the model to place more weight on recent data and adjust the importance of various features or indicators according to their relevance at any given time. The particular econometric model for nowcasting the annual labour market indicators for all of the Ukrainian regions on the basis of annually averaged indicators from the regional State Statistics Services (the period between 2013 and 2022) is based on a combination of TVPOLS regressions. To model the relationship between a high-frequency monthly variable of the job searched-based indicators and similarly high-frequency variables from the state statistics service for various regions (e.g., the unemployment rate, monthly change in the number of employed population), TVPLM regressions will be used as a type of time series regression to forecast future values of a variable based on its past values and the values of other variables.

3.3. Model selection

Multi-equation linear models with time-varying coefficients and time-varying coefficients panel data models will be used to estimate the relationship between the web-scraping-based indicators and the official labour market indicators for different regions. The preliminary assessment highlights the following three estimators as the most appropriate: 1) time-varying pooled ordinary least squares (TVPOLS), 2) time-varying random effects (TVRE), and 3) Time-varying fixed effects (TVFE).

The TVPOLS model is a straightforward extension of the classical Pooled Ordinary Least Squares with time-varying coefficients. In the TVPOLS setup, the linear regression model is:

$$Y_{it} = \beta_{0t} + \beta_{1t} X_{it} + u_{it} \tag{5}$$

Where Y_{it} is the dependent labour market variable for the region i at time t; X_{it} is the vector of independent job search variables, and β_{1t} is the vector of time-varying coefficients. In contrast, In TVRE, the model incorporates individual-specific random effects that vary over time:

$$Y_{it} = \alpha_{it} + \beta_{1t} X_{it} + u_{it} \tag{6}$$

Where α_{it} is the individual-specific random effect at time t. Going further, in TVFE, the model includes individual-specific fixed effects that vary over time:

$$Y_{it} = \gamma_{it} + \beta_{1t} X_{it} + u_{it} \tag{7}$$

Where γ_{it} is the individual-specific fixed effect at time t. These models are extensions of traditional regression models, with the main distinction consisting of the introduction of time-varying coefficients, random effects, or fixed effects, depending on the model type, which might be helpful for analysis of the generalised relationship between the job search website indicators and official labour market variables across all of the regions. The models will be estimated with the help of the tvReg package for R. Time-varying Generalized Least Squares (TVGLS) method is used to analyze and establish statistical relationships for individual regions. Here, 'generalized' refers to the model's capacity to handle different types of errors and data complexities, such as heteroskedasticity or autocorrelation, which can vary over time. By allowing parameters to change over time, TVGLS can account for the possibility that the way variables are related to one another might shift due to regional economic changes, professional differences, or other factors. On the other hand, time-varying Pooled Ordinary Least Squares (TVPOLS) is employed to examine the overall or average effect of variables across all regions and professions combined. While this approach also permits the relationship between variables to change over time, it assumes that the underlying dynamics are consistent across all regions—it pools the data together, ignoring region-specific effects or assuming they are constant. The aforementioned time-variant models are applied to various region samples to find interesting patterns in the relationships between the web-scrapped indicators and official statistics.

As mentioned earlier, to analyze within and between regional and professional variation, the application of the time-variant panel regression models (tvPLM) is needed, which extends traditional panel data models by allowing for time-varying coefficients. The basic mathematical setup for the tvPLM regression can be defined as follows:

$$Y_{it} = \Theta_{it} + \beta_{1t} X_{it} + u_{it} \tag{8}$$

Where θ_{it} is the individual-specific fixed effect (time-invariant), and β_{it} is the vector of time-varying coefficients for the i-th unit at time t. This model allows for the coefficients β_{it} to vary across both units and time, providing flexibility to capture time-varying effects at the individual level. The estimation of the tvPLM model involves considering how the coefficients change over time and across units. Common approaches include dynamic panel data methods, such as the Generalized Method of Moments (GMM) or Fixed Effects Instrumental Variable (FEIV) estimators. The choice of the specific estimation method may depend on the assumptions made about the presence of endogeneity, however, since this risk of endogeneity in the context of job search data is present - when predicting changes in employment based on the number of resumes on job search websites, heterogeneity refers to the differences in job-seeker characteristics (such as industry, experience level, and geographic location) and how these differences might influence employment trends differently during various economic conditions. For predictions about real salaries based on online vacancies, heterogeneity could arise from the variation in salary offerings across different sectors, regions, and job types. The pro-cyclical nature of wages means that the observed salary data may vary with economic cycles, potentially leading to composition bias if the mix of job postings shifts toward higher-paying or lower-paying industries during different economic phases. If not accounted for, it could lead to biases or inaccuracies in predictions, particularly if the model overgeneralizes from non-representative data.

One potential problem with using job search website data is that it may not be representative of the overall labour market. For example, more educated individuals or people with higher income may search more online leading so a biased composition of search. This could lead to an overestimation of the unemployment rate. As specified by Faryna et al. (2021), the important requirement for the data preparation for the analysis would be to ensure that the retained vacancies i) are full-time jobs, ii) provide average monthly salary indicators, and iii) provide wages listed in hryvnia (UAH). No data truncation has been done to avoid misreporting for some regions. In order to check the validity of the forecast, holdback data is taken for 2021 to make it the basis year for the comparison of the nowcasted data with the actual records. This allows us to employ the error metrics such as MAPE, WAPE and RMSE to estimate the validity of the forecast. The least absolute shrinkage and selection operator (LASSO) method as well as other elements of ML are used to select the most important category-level wage growth indices contributing to the official country-level index.

C h a p t e r 4

DATA OVERVIEW

4.1. Description of data sources, types and distributions

The primary data source for this thesis is the job search website of work.ua and the regional websites of the State Statistics Service and State Employment Service. The data from the employment websites has been scraped based on time snapshots available at the Web archive portal. The key four variables in this data set include information on the number of regional job postings, the number of resumes, and the average salary specified in the salaries and resumes across the regions.

A detailed description of the indicators can be found in the annexes: 9 low-frequency annual indicators (APPENDIX A) and 12 high-frequency labour market indicators (APPENDIX B) were tested for the nowcasting as dependent labour market variables. With regard to the independent variables, a quick overview of the web-scrapping availability of the work.ua portal demonstrated that 6 groups of variables are eligible for such choice (APPENDIX C). The summary statistics for the dataset can be seen in the table below.

Table 2. Descriptive statistics for the key dataset variables

vars	n	mean	sd	median	trimmed	mad	min	Max	kurtosis	SE
1	5490.0	3.5	1.7	3.5	3.5	2.2	1.0	6.0	-1.3	0.0
2	5490.0	8.0	4.3	8.0	8.0	5.9	1.0	15.0	-1.2	0.1
3	5490.0						Inf	-Inf		
4	5490.0	77304.1	190817.1	21842.5	33725.7	24366.5	607.0	1575551.0	33.4	2575.3
5	5490.0	141.6	14.3	144.0	143.2	10.4	13.0	178.0	10.2	0.2
6	366.0	200141.4	170481.1	142600.0	182128.5	158549.8	10764.5	797047.8	-0.2	8911.2
7	5490.0	10074.7	4334.6	9226.0	9555.6	3673.1	1430.0	41377.0	3.7	58.5
8	305.0	122.5	10.1	120.9	121.9	8.6	99.4	149.4	-0.1	0.6

9	305.0	101.6	7.3	101.4	101.6	4.7	81.1	123.2	1.1	0.4
10	305.0	112.0	5.9	110.5	111.3	3.8	97.4	134.0	2.1	0.3
11	305.0	100.9	7.2	100.8	100.9	4.4	80.4	122.2	1.2	0.4
12	5490.0	770.8	2556.6	178.0	288.1	200.8	0.8	32845.8	75.0	34.5
13	5490.0	22473.7	82146.0	5561.0	7790.6	7504.5	4.0	1109503.0	84.9	1108.7
14	5490.0	13400.3	3584.6	13257.3	13311.4	3680.6	0.0	30000.0	-0.2	48.4
15	5489.0	11607.0	3135.8	11447.3	11451.2	2927.3	4229.6	25625.8	1.4	42.3
16	5490.0	31.0	17.6	31.0	31.0	22.2	1.0	61.0	-1.2	0.2

The data for the number of vacancies obtained from the job search website of work.ua has been appended to the dataset with indicators from the official state statistics service data. The codes of professional classification of economic activities (CEA) and the data on regional location (oblast) are added as a factor variable. At the same time, with regard to the data of a similar economic nature (salaries), there is a sense to compare the distribution of two unofficial and one official data variable. The distribution of this data can be seen from the figure below.

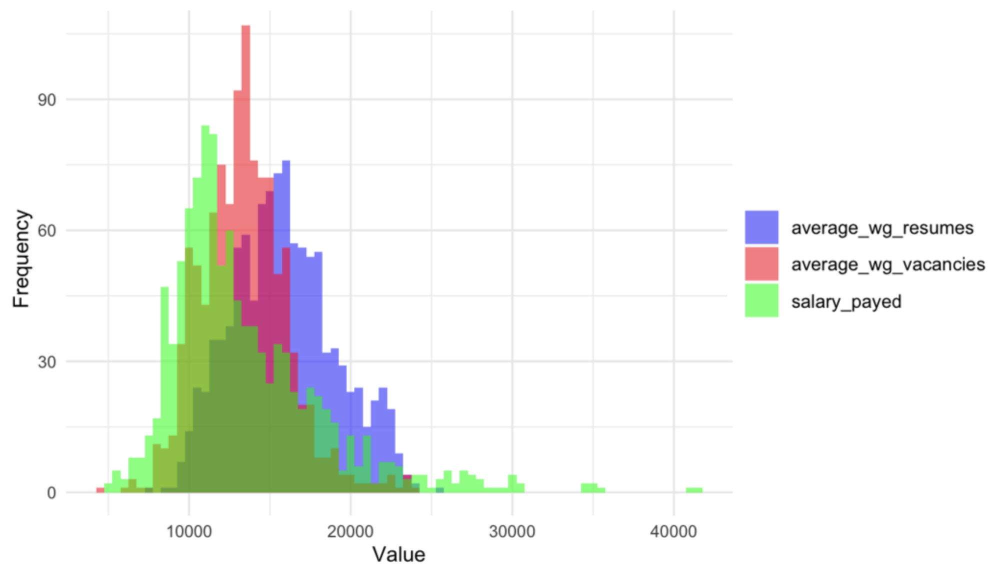

Figure 2. Distribution of the salary data from vacancies, resumes, and official statistics for 2021.

As can be seen from the plot, data for the key variables of interest – average wages reported in resumes and vacancies and official salary data provided by the

RSSS is mostly normally distributed. Although some outliers still exist the data from vacancies in work.ua is more normally distributed than the actual salary data due to the different data collection approaches (high-wage earners are likely to underreport some of their salaries in the official statistics, while job advertisements used to attract workers are likely to position higher salary offers and expectations). Nevertheless, the slight deviation in normality is not critical for our analysis since the majority of the following regression estimation assumptions still hold.

4.2. Visual review of key target variables

According to the basic correlation analysis, it becomes possible to discover that only a few variables have a high correlation with each other, as can be deduced from the correlation heatmap.

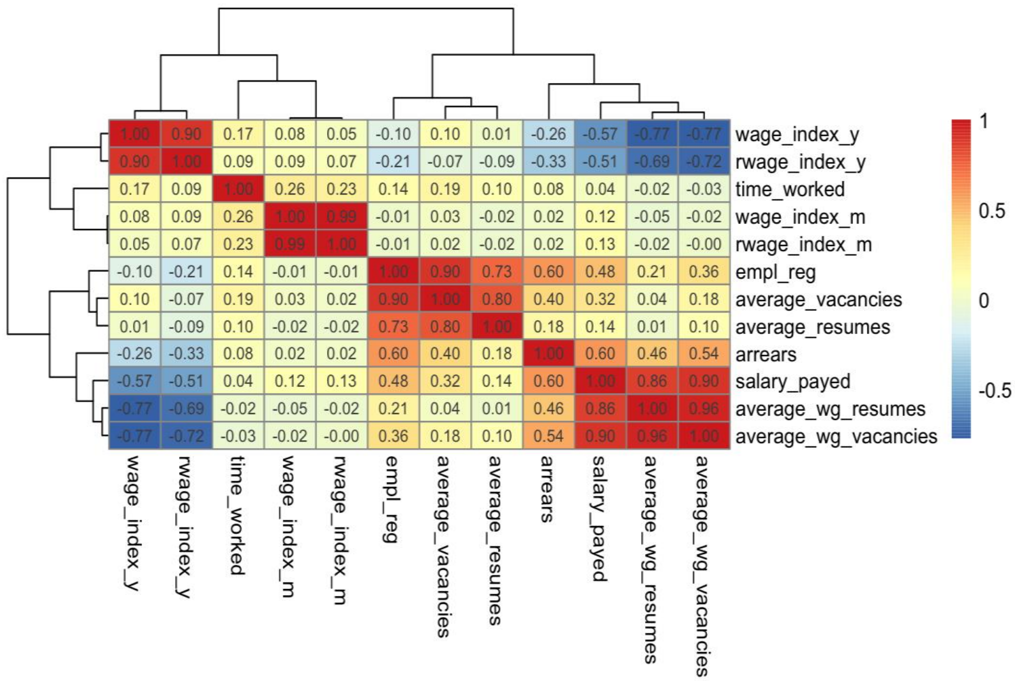

Figure 3. Heatmap for the main studied variables

As we can see from the plot, the correlation for the variables is indeed quite high: the most promising co-dependent variable pairs with positive correlation include officially registered employment (empl_reg) and number of resumes and vacancies from the job search website (average_vacancies and average_resumes); official salary indicators (salary_payed) and salaries from resumes and vacancies (average_wg_resumes, average_wg_vacancies). The clustering does not provide any notable insights here and can hence be discarded.

As for the interesting negative correlations, we also observe high correlation between the year-to-year wage indices and salaries from resumes and vacancies (average_wg_resumes, average_wg_vacancies). This is likely explained by the fact that annual wage indices essentially converted the nominal growth in the salary numbers into the real wage transformation, which has not been notably positive for the last couple of years in Ukraine.

In terms of statistical significance of the correlations, the Pearson correlation coefficient for the key variables has p-values lower than 0.05, which means that the correlation for these variables is statistically significant. The calculated Pearson coefficients are provided in the table below.

Table 3. Pearson correlation coefficients for the key variables.

Variable1	Variable2	Correlation	P_value
empl_reg	average_vacancies	0.900344	1.4E-133
empl_reg	average_resumes	0.734489	2.93E-63
salary_payed	average_wg_resumes	0.858496	1.3E-107
salary_payed	average_wg_vacancies	0.904692	6.2E-137
wage_index_y	average_wg_resumes	-0.77284	8.51E-62
wage_index_y	average_wg_vacancies	-0.77238	1.11E-61
rwage_index_y	average_wg_resumes	-0.68824	3.9E-44
rwage_index_y	average_wg_vacancies	-0.72231	1.87E-50

The further review of changes in time for the target variables demonstrated promising co-movement: as can be seen from the figure below, the change in official salaries is followed closely by the change in average salaries.

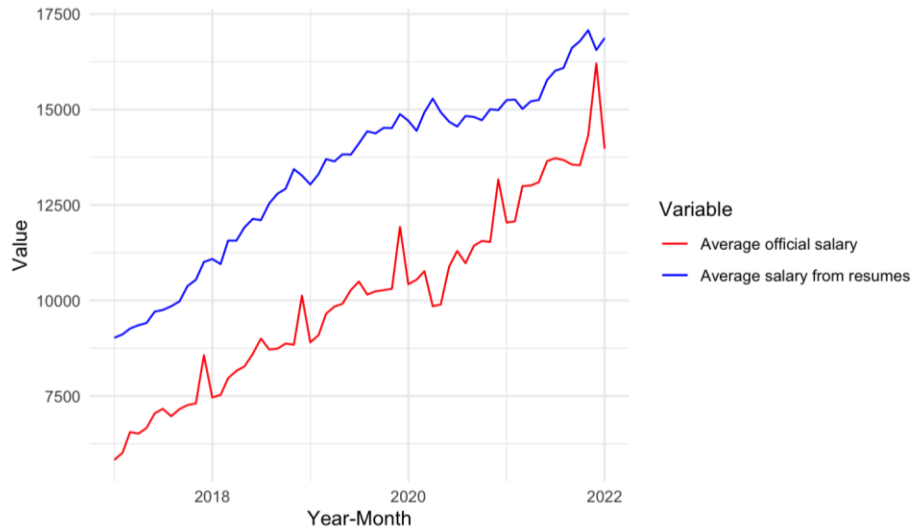

Figure 4. Growth of average salaries from resumes and official salaries reported by the SSSU[1]

From the visual analysis alone it is hard to discover whether there is a lead or lag available: for this purpose, more rigorous regression analysis will be done later to establish possible Granger causality. Since the variables are clearly cointegrated, we cannot simply use OLS - models, but will have to resort to models adapted for cointegration like ECM.

[1] (averaged across regions and professions).

C h a p t e r 5

ESTIMATION RESULTS

Simple OLS regression results for the key target variables, outlined during the stage of correlation analysis appeared to be promising: statistically significant relationships were identified for the officially registered number of employed (dependent variable) and number of resumes and vacancies (independent variables) published on the job search website. The reported R-squared (R^2) value was high in both cases: for employment, the coefficient of 0.825 suggests that approximately 82.5% of the variance in the dependent variable empl_reg (officially registered employment) is explained by the independent variables in the model (average_vacancies and average_resumes). Moreover, the coefficients for both average_vacancies and average_resumes are statistically significant at a 0.01 level, which implies a very low probability that these relationships are due to chance. The F Statistic is also significantly large and indicates that the overall model is statistically significant. For the salary data, however, the statistical significance of a simple OLS model was high only for the averaged totals, which indicated that we might need to test the causal relationship between the data contained in the average salaries from vacancies and average salaries from resumes (independent variables) and salary paid column (dependent variable) by controlling for different CEA codes (contained in the 'CEA code' column).

Furthermore, as noted earlier, the variables could be cointegrated: although in the case of non-stationary data properties like mean and variance change over time, individual time series are non-stationary, they might be cointegrated, meaning a linear combination of them is stationary. This indicates a long-term equilibrium relationship between the variables. The Engle-Granger two-step method is a commonly used approach to test for cointegration. Having conducted the Augmented Dickey-Fuller (ADF) test residuals for the linear models were proven to be cointegrated. Therefore, the application of simple OLS models was indeed not sensible.

Extending to an Error Correction Model (ECM)

Table 4. Regression results for the ECM model

	Dependent variable: diff_salary_payed
diff_average_wg_resumes	0.764***
	(0.055)
ecm_residuals	0.220***
	(0.042)
Constant	41.530
	(53.396)
Observations	304
R2	0.419
Adjusted R2	0.415
Residual Std. Error	927.146 (df = 301)
F Statistic	108.491*** (df = 2; 301)
Note:	*p<0.1; **p<0.05; ***p<0.01

Since cointegration between our salary variables was found, we can build an ECM to capture both short-term dynamics and long-term equilibrium. The application of the ECM model for the same aggregate salary variables demonstrated the following results: The t-value of 13.784) and the p-value (< 2e-16) indicate that this coefficient is highly statistically significant, meaning that changes in average wages from resumes have a significant impact on changes in official salary paid: for every one-unit increase in average wages from resumes, official salary paid increases by approximately 0.764 units. Significance of Predictors suggests that both the changes in average wages from resumes and ECM residuals are significant predictors of changes in official salary paid: The positive and significant coefficient for changes in average wages suggests that short-term changes in these wages collected from the job search website positively influence short-term changes in salary paid. At the same time, significant positive coefficient for model's residuals indicates that any disequilibrium in the long-term relationship between the predictors and predicted variables is partially corrected in the next period, implying a restoring force towards the long-term equilibrium. Overall, the model explains a moderate proportion of the variance in official salary paid, indicating that while it captures important relationships, there are other factors affecting its variations that are not included in the model.

Having created 12 different ECM-type models for the estimation of the different professions via created dummy variables, it was established that salaries from resumes work as better predictors compared to the salaries from vacancies. Models for professions of CEA code P (Education), R (Arts, sports, entertainment and recreation), and Q (Health care and provision of social assistance) demonstrated the worst performance, and models for CEA code K (Financial and insurance activities), I (Temporary accommodation and catering), and J (Information and telecommunications) performed best. The code output for the estimates of the models can be accessed in the APPENDIX E.

As can be seen from the figure below, there is a positive relationship between the variables for the professionals engaged in Financial and insurance activities: despite the fact that the application of the linear trend in this context might not be ideal, it still visualises the direct relationship between the independent and dependent variables of average salaries from resume and official salary from the RSSS respectively.

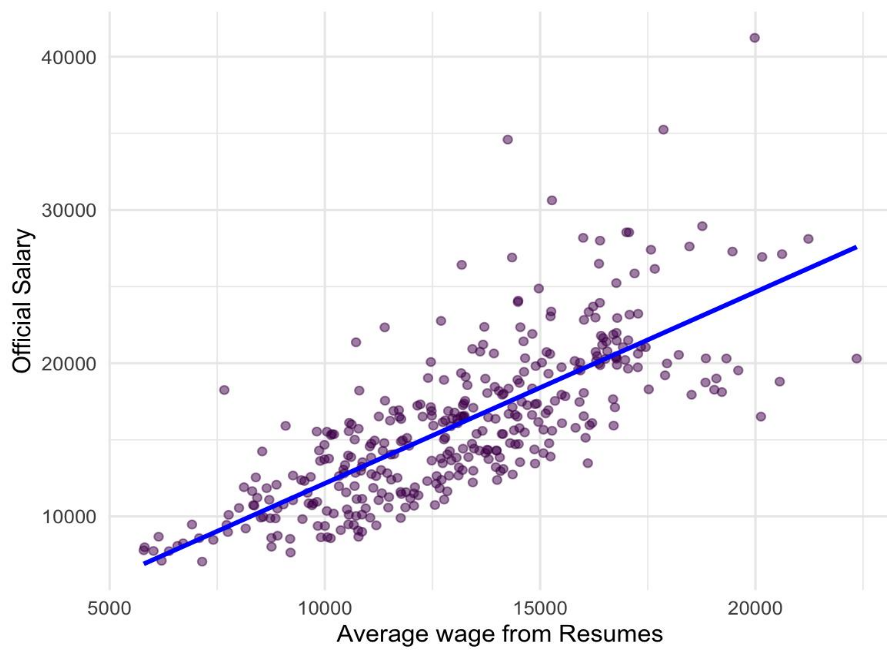

Figure 5. Fitted regression line for CEA Code K (Financial and insurance activities)

As the regression line does not fit the data ideally, the residual standard error for this estimate is 0.3081 on 5118 degrees of freedom, which suggest that the model is based on a large dataset and the typical prediction error is relatively small, indicating a reasonably good fit of the model to the data. Adjusted R-squared of the model is 0.4278, which means that despite the fact of the clear existence of breaks (several professions have been aggregated together) and the risk of overfitting due to using too many predictors (CEA codes) due to the relatively small number of observations, the variations in wages from job search indicators can explain ~42% of variation in the officially registered average salaries.

Further improvement of the regression model should be used to combat multicollinearity, which has so far been introduced by the inclusion of multiple CEA code dummy variables, especially if the data for different CEA codes is closely correlated. This can inflate the variance of the coefficient estimates and make them unstable.

Furthermore, if there are other important variables related to salary that are not included in the model, the estimates could be biased. To account for potential changes in the job search website popularity, the model should include the variable representing the website growth.

The second regression model for the same dataset focuses on the number of vacancies to predict changes in the registered number of employees. As expected, the regression model for the majority of codes demonstrated a positive relationship. The logic behind this observation can be attributed to the fact that the people who look for jobs are likely to find them in a short time. According to the poll conducted by the administration of the work.ua website itself, the average duration of job search in Ukraine for the majority of the vacancies on the website was less than one month (work.ua, 2023). The only professions which exceeded that time amount belonged to the categories O (Public administration and defence; compulsory social insurance), N (Activities in the field of administrative and auxiliary services), C (Processing industry), S (Provision of other types of services) from the classification of economic activities (CEA), which were not deemed significant of a forecast in the first place.

The significant effects across various professions have been identified for the professions of Agriculture (A), Construction (F), Wholesale and retail trade (G); Repair of motor vehicles and motorcycles, transport, warehousing, postal and courier activities (H), Temporary accommodation and catering (I), Information and telecommunications (J), Financial and insurance activities (K), Real estate transactions (L). Consequently, these professions have been used further to create a forecast for the changes in the labour market for 2021 as the last year with official data available. However, since R2 is often large in time series regressions, it is not an indicator of goodness of fit.

In order to proof the goodness of fit we check how the model performs in terms of forecasting on out of sample values. In the figure below one can see the out-of-sample forecast for the changes in the employment levels for 2021, created by the log version of the model.

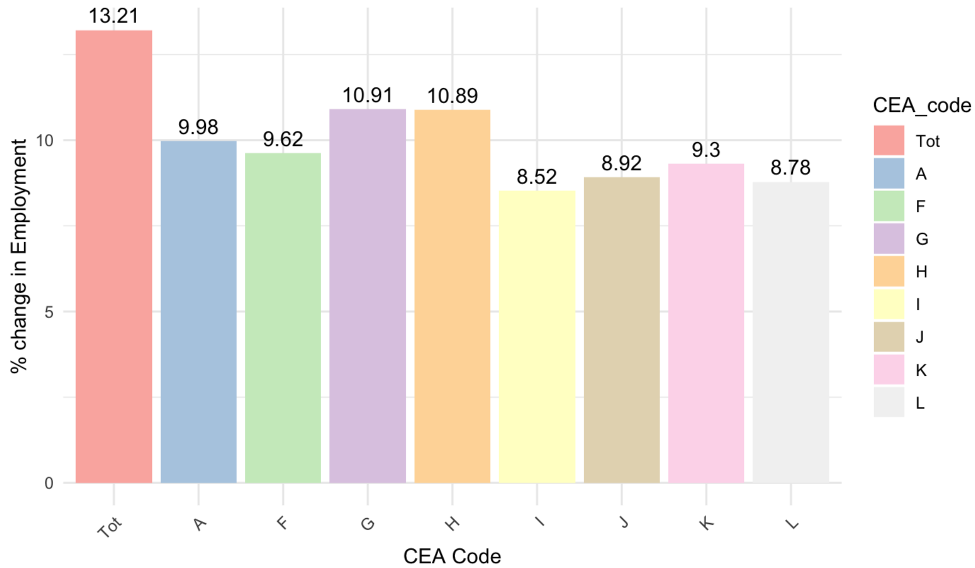

Figure 6. Change in forecasted employment by occupation (CEA codes)

As can be seen from the picture, the statistically significant forecast can only be provided for 8 out of 14 estimated professions. For them, the overall growth in official employment for 2021 across all the regions was estimated at the level of ~13.2%, which is significantly higher than the actual observations - just about 5%.

However, when looking at the individual observations, the results demonstrate that the actual change is very much close to the forecasted values: RMSE of the model is relatively low (0.412.

With regard to the forecasted changes in wages, by once again looking at the overall change for 2021, the general transformation looks similar to the forecasted change in employment: as can be seen from the figure below, the average increase in salaries (Tot) is expected to be equal to ~9%.

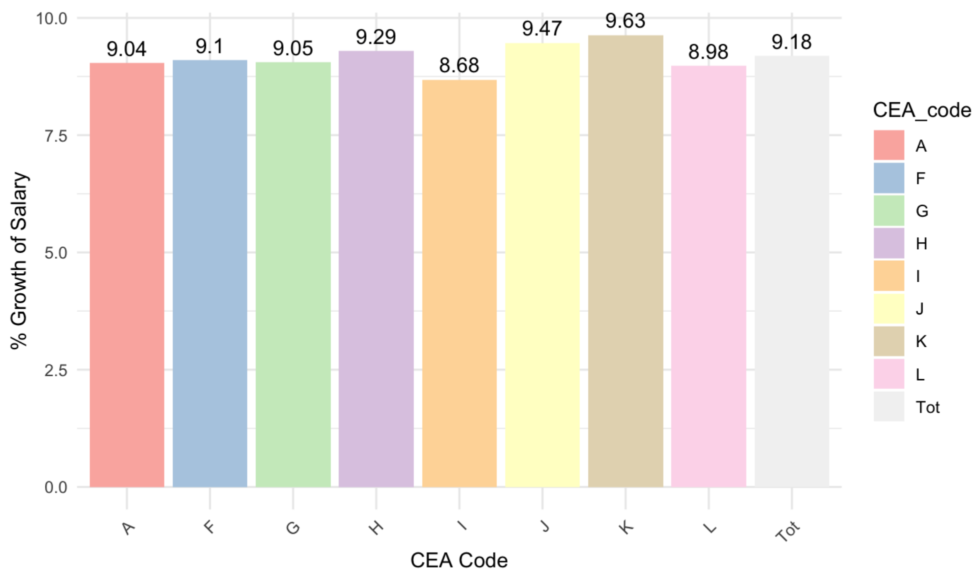

Figure 7. Change in forecasted salaries by occupation (CEA codes)

As can be seen from the figure above, the largest increase of ~9.6% is forecasted for the CEA code K "Financial and insurance activities", which demonstrated the best statistically significant results during the stage of separate model estimation. The model's RMSE is also surprisingly good - 0.215678, which makes it more trustworthy than the model for employment.

Comparison of the models' forecast significance demonstrates that the employment model yields higher RMSE than the Vacancies model: however, contrary to the expectations, the TVPLM model does not provide better predictions, which is likely explained by the forced averaging which cannot be avoided due to technical limitations of the corresponding function in R.

Table 5. Comparison of models' out of sample performance

Dependent variable	Model type	RMSE	MAPE	WAPE
Official Salaries	TVPLM	0.425	3.508	3.577
	OLS LOGs	0.216	1.806	1.805
Official Employment	TVPLM	2.267	19.433	20.015
	OLS LOGs	0.412	3.339	3.231

As can be seen from the table above, both for the Official Employment dependent variable and official salaries paid, the simple OLS model in logs has much lower values of RMSE, MAPE, and WAPE compared to the TVPLM. This suggests that the OLS provides a more accurate prediction of Official Employment than the TVPLM, as the errors between the predicted and actual values are on average smaller. In summary, for the data provided on Official Employment, the OLS model in logs outperforms the TVPLM across all three metrics.

C h a p t e r 6

CONCLUSIONS AND RECOMMENDATIONS

The statistical analysis of the labour market data from the job search website of work.ua and Regional State Statistics Service agencies demonstrated, that several variables from the official statistics can indeed be nowcasted using the proxies from the job search website. OLS with logs targeted variables of 8 professions demonstrated that the changes in the aggregate number of resumes from the job search websites act as a good proxy for the changes in the numbers of officially employed workers. The variables of average time worked, salary arrears and month-to-month salary indices appeared to be less efficient in the estimation than the actual statistical variables they were hypothetically built upon.

At the same time, the changes in the average salaries from wages and resumes on the work.ua website did not perform as good, potentially due to the effects of the shadow economy. In an attempt to combat this issue, testing different professions according to the CEA classification demonstrated that the highest statistical significance is demonstrated for the employees engaged in Financial and insurance activities.

With regard to best model selection, time-variant panel linear models, which utilised the VC method, did not perform as well as expected, likely due to the lack of training data and flaws of the official statistics.

In order to interpret and act upon providing recommendations about social policies and improve the prediction of short-term unemployment, inflation, and GDP changes, more research needs to be done. However, the model's utility for decision-making could be without this insight since at the moment the model is based on a specific dataset, and its findings may not generalize to other contexts or populations outside the study. In order to make its findings more trustworthy, basic testing should be done which would demonstrate the quality of the nowcast on the more fresh official statistical data collected since the start of the full-scale russian invasion of Ukraine in 2022, which has not been published yet. Additionally, starting in 2022, the quality of web-scraping data is expected to decrease (web archive does not provide snapshots regular enough to analyse the data with the sample frequency spacing).

However, the model's utility for decision-making could be without this insight since despite being based on a specific dataset, the model and its findings may generalize to other contexts or populations outside the study for the purposes of future research. Thus, the models can potentially be used by policy-makers to approximate changes in the labour market of Ukraine in the context of a deficit of official statistics.

WORKS CITED:

Altunina (2015), *Most Ukrainians regard Work.ua as the world's best job search site. Follow-up results, New Voice Business 2015*, https://biz.nv.ua/ukr/economics/bilshist-ukrajintsiv-vvazhajut-work-ua-krashchim-sajtom-poshuku-roboti-rezultati-doslidzhennja-65536.html

Askitas, Nikos and Zimmermann, Klaus, (2009), *Google Econometrics and Unemployment Forecasting, No 899, Discussion Papers of DIW Berlin*, DIW Berlin, German Institute for Economic Research, https://EconPapers.repec.org/RePEc:diw:diwwpp:dp899.

Bańbura, Marta, Irina Belousova, Katalin Bodnár, and Máté Barnabás Tóth. "*No time to slack: Nowcasting Employment in the Euro Area."* ECB Working Paper Series No. 2815, 2023. https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2815~1c1e2649d3.en.pdf

Breiman, L. Random Forests. *Machine Learning 45, 5–32* (2001). https://doi.org/10.1023/A:1010933404324

Casas, Isabel and Fernández-Casal, Rubén 2022. "*tvReg: Time-varying Coefficients in Multi-Equation Regression in R", The R Journal Vol. 14/1, March 2022 ISSN 2073-4859*. https://journal.r-project.org/articles/RJ-2022-002/

Constantinescu, Mihnea and Kappner, Kalle and Szumilo, Nikodem, *The Warcast Index: nowcasting economic activity without official data during the Ukraine war in 2022*. (May 27, 2022). Available at SSRN: https://ssrn.com/abstract=4147377 or http://dx.doi.org/10.2139/ssrn.4147377

Cooley, Thomas and Prescott, Edward, (1973), *An Adaptive Regression Model, International Economic Review, 14, issue 2, p. 364-71,* https://EconPapers.repec.org/RePEc:ier:iecrev:v:14:y:1973:i:2:p:364-71.

Dangl, T., and Halling, M. (2012). *Predictive regressions with time-varying coefficients. Journal of Financial Economics,* 106(1), 157–181. doi:10.1016/j.jfineco.2012.

Dario Buono, Gian Luigi Mazzi, George Kapetanios, Massimiliano Marcellino, And Fotis Papailias, 2017. *"Big data types for macroeconomic nowcasting", EURONA — Eurostat Review on National Accounts and Macroeconomic Indicators C82, E27, E37, E47:* https://cros-legacy.ec.europa.eu/system/files/euronaissue1-2017-art4.pdf

Faryna, Oleksandr, Pham, Tho, Talavera, Oleksandr et al., 2021: *Wage and unemployment: evidence from online job vacancy data. Journal of comparative economics. pp. 52-70.* ISSN 0147-5967, https://doi.org/10.1016/j.jce.2021.05.003

Gomis Roger, Paloma Carrillo, Steven Kapsos, Stefan Kühn and Avichal Mahajan. 2022. "*The ILO nowcasting model: Using high-frequency data to track the impact of the COVID-19 pandemic on labour markets" Statistical Journal of the IAOS 38 (2022) 815–830.* https://content.iospress.com/download/statistical-journal-of-the-iaos/sji220055?id=statistical-journal-of-the-iaos%2Fsji220055.

Labour Market Update - December 2023 - Jobs and Skills Australia: https://www.jobsandskills.gov.au/publications/labour-market-update-december-2023

Bańbura, Marta, Domenico Giannone, Michele Modugno and Lucrezia Reichlin. 2013. "*Now-casting and the real-time data flow" European Central Bank: Working Paper Series NO 1564:* https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1564.pdf

McLaren, Nick and Shanbhogue, Rachana, *Using Internet Search Data as Economic Indicators (June 13, 2011). Bank of England Quarterly Bulletin No. 2011 Q2, Available at SSRN:* https://ssrn.com/abstract=1865276 or http://dx.doi.org/10.2139/ssrn.1865276

Skakalska, Olena, Natalia Sizova, Oleksandra Bondar, 2023 - "*Number of days needed for recruiters to close vacancies in Ukraine*", work.ua website, URL: https://www.work.ua/articles/recruiting-system/3172/

Population survey from the State Statistics Service, 2023, Survey portal, State Statistics Service of Ukraine: https://survey.stat.gov.ua/

Schlicht, E. VC: a method for estimating time-varying coefficients in linear models. J. Korean Stat. Soc. 50, 1164–1196 (2021). https://doi.org/10.1007/s42952-021-00110-y

Shamiri, Samuel, Leanne Ngai, Peter Lake, Yin Shan, Amee McMillan, Therese Smith, 2021. "*Introducing NERO: A monthly Nowcast of Employment by Region and Occupation (NERO) in Australia". National Skills Commission, Australian Government,* July 2021:https://www.jobsandskills.gov.au/sites/default/files/2022-05/NSC%20-%20NERO%20-%20Methodology%20Paper.pdf

Shamiri, S., Ngai, L., Lake, P., Shan, Y., McMillan, A., Smith, T. and Sharma, K. (2022), *Nowcasting the Australian Labour Market at Disaggregated Levels. The Australian Economic Review*, 55: 389-404. https://doi.org/10.1111/1467-8462.12464

Taveeapiradeecharoen, P., & Aunsri, N. (2020). *A Large Dimensional VAR Model with Time-Varying Parameters for Daily Forex Forecasting. Wireless Personal Communications*. doi:10.1007/s11277-020-07531-8

APPENDIX A

Table 6. Statistical description for low-frequency independent variables

Variable name	Years	Provider	Frequency	Applicability
Labour force (15–70 y.o), thns ppl	2000 - 2019	Regional State Statistics Services	annual	Low (due to the small number of observations)
Employed (average), thns ppl	2000 - 2019	Regional State Statistics Services	annual	Low (due to the small number of observations)
Unemployed (ILO, 15-70) thns ppl	2000 - 2019	Regional State Statistics Services	annual	Low (due to the small number of observations)
Average duration of job search (#months)	2000 - 2019	Regional State Statistics Services	annual	Low (due to the small number of observations)
Average number of officially registered workers, thns ppl	2000 - 2019	Regional State Statistics Services	annual	Low (due to the small number of observations)
Labor force turnover ratio in	2000 - 2019	Regional State	annual	Low (due to the small number

Variable name	Years	Provider	Frequency	Applicability
% to the average registered number of full-time employees (by hiring)		Statistics Services		of observations)
Labor force turnover ratio in % to the average registered number of full-time employees (by dismissal)	2000 - 2019	Regional State Statistics Services	annual	Low (due to the small number of observations)
Average salary (nom UAH)	2000 - 2019	Regional State Statistics Services	annual	Low (due to the small number of observations)
Real salary (% previous year)	2000 - 2019	Regional State Statistics Services	annual	Low (due to the small number of observations)

APPENDIX B

Table 7. Statistical description for high-frequency independent variables

Variable name	Years[2]	Provider	Frequency	Applicability for nowcasting
Workforce	2019-2021	Regional State Statistics Services	quarterly	Medium (the small number of observations is still not big enough)
The average registered number of full-time employees by city and district	2019-2021	Regional State Statistics Services	quarterly	Medium (the small number of observations is still not big enough)
Time worked by full-time employees by cities of regional importance and districts	2019-2021	Regional State Statistics Services	quarterly	Medium (the small number of observations is still not big enough)

[2] time period varies for some regions due to the challenges with data availability on the websites of the State Statistics Service, which will require temporary exclusion of certain regions from the analysis until the full dataset is provided upon the request form

Variable name	Years[2]	Provider	Frequency	Applicability for nowcasting
Average monthly salary of full-time employees by district	2019-2021	Regional State Statistics Services	quarterly	Medium (the small number of observations is still not big enough)
Distribution of the number of full-time employees according to the amounts of wages charged to them and types of economic activity	2019-2021	Regional State Statistics Services	quarterly	Medium (the small number of observations is still not big enough)
Registered unemployment and number of vacancies	2019-2021	Regional State Statistics Services	monthly	High (the number of observations in sufficient to conclude a statistically significant causality)
The average registered number of full-time employees by type of economic activity	2019-2021	Regional State Statistics Services	monthly	High (the number of observations in sufficient to conclude a statistically significant causality)

Variable name	Years[2]	Provider	Frequency	Applicability for nowcasting
Worked time of full-time employees by type of economic activity	2019-2021	Regional State Statistics Services	monthly	High (the number of observations in sufficient to conclude a statistically significant causality)
Amount of salary arrears	2019-2021	Regional State Statistics Services	monthly	High (the number of observations in sufficient to conclude a statistically significant causality)
Growth/decrease rates of nominal and real wage index	2019-2021	Regional State Statistics Services	monthly	High (the number of observations in sufficient to conclude a statistically significant causality)
Average salary of full-time employees by type of economic activity	2019-2021	Regional State Statistics Services	monthly	High (the number of observations in sufficient to conclude a statistically significant causality)
Average monthly nominal salary by type of economic activity	2019-2021	Regional State Statistics Services	monthly	High (the number of observations in sufficient to conclude a statistically significant causality)

APPENDIX C

Table 8. Statistical description of the potential independent variables

Variable name	Years	Provider	Frequency	Applicability for nowcasting
Average expected salary, UAH	2013-2022	work.ua	monthly/quarterly/annual*	High (the number of observations in sufficient to conclude a statistically significant causality)
Average offered salary, UAH	2013-2022	work.ua	monthly/quarterly/annual*	High (the number of observations in sufficient to conclude a statistically significant causality)
Average number of resumes, units	2013-2022	work.ua	monthly/quarterly/annual*	High (the number of observations in sufficient to conclude a statistically significant causality)
Average number of vacancies, units	2013-2022	work.ua	monthly/quarterly/annual*	High (the number of observations in sufficient to conclude a statistically significant causality)

APPENDIX D

Table 9. Data to be used for sample regression analysis

Variable name	Transformation	Provider	Frequency
The number of registered full-time employees by type of economic activity (2018-2021)	levels/logs	Regional State Statistics Services	monthly
Worked time of full-time employees by type of economic activity (2018-2021)	levels/logs	Regional State Statistics Services	monthly
Average salary of full-time employees by type of economic activity (2018-2021)	levels/logs	Regional State Statistics Services	monthly
Growth/decrease rates of nominal and real wage index (2018-2021)	levels	Regional State Statistics Services	monthly

Variable name	Transformation	Provider	Frequency
Amount of salary arrears (2018-2021)	levels/logs	Regional State Statistics Services	monthly
Average expected salary, UAH (2018-2021)	levels/logs	work.ua	monthly/quarterly
Average offered salary, UAH (2018-2021)	levels/logs	work.ua	monthly/quarterly
Average number of resumes, units (2018-2021)	levels/1st difference	work.ua	monthly/quarterly
Average number of vacancies, units (2018-2021)	levels/1st difference	work.ua	monthly/quarterly/