BRILL

RESEARCH DATA JOURNAL FOR THE HUMANITIES AND SOCIAL SCIENCES 9 (2024) 1–14

BRILL OPEN

brill.com/rdj

Educational and Career Trajectories in Russia: Introducing a New Source and Datasets with a High Granularity

Danila Valko | ORCID: 0000-0002-8058-7539
Corresponding author
Institute of East European Studies, Freie Universität Berlin, Berlin, Germany
The Hannah Arendt Research Center, Berlin, Germany
d.v.valko@gmail.com

Mariia Vasilevskaia
The Hannah Arendt Research Center, Berlin, Germany
mariia.vasilevskaia@gmail.com

Maria Bunina
The Hannah Arendt Research Center, Berlin, Germany
Social Sciences Division, The University of Chicago, Illinois, USA
mwbunina@gmail.com

Mariia Kozlova
Department of Sociology, University of Victoria, Victoria, Canada
mariiakozlova@uvic.ca

Anna Maria Filippova
The Hannah Arendt Research Center, Berlin, Germany
Center For Independent Social Research, Yerevan, Armenia
annamariiafilippova@proton.me

Daria Rud
The Hannah Arendt Research Center, Berlin, Germany
daria.rud.moscow@gmail.com

Received 3 April 2024 | Revised 10 July 2024 | Accepted 26 July 2024 | Published online 4 October 2024

Published with license by Koninklijke Brill BV | DOI:10.1163/24523666-BJA10046
 | ISSN: 2452-3666 (online)

Abstract

Studying Russian society is challenging, especially during the period of the Russian military invasion. However, it takes on special significance during a period of economic and social transformation. Studying the career and educational trajectories of Russians in the context of East Studies offers a multifaceted perspective on the state of the job market and education sector and gives an understanding of the current situation of the country's economy and social structure. The lack of data with a high level of granularity is critical, especially for studying people with a focus on their career and educational trajectories. In this article, the authors respond to this request and present two datasets that can be useful for studying spatial and temporal patterns associated with people's life trajectories in the context of work and education. The authors utilised open data on CVs created or updated by employment portal users over the period 2015–2023 from the Federal Service for Labor and Employment (Rostrud) and prepared two cleaned datasets covering 83 regions of Russia. Dataset 1 is on the educational and career trajectories (N = 6,221,439) and Dataset 2 is on the activity of unemployed and job-seeking candidates (N = 7,662,089).

Keywords

educational and career trajectories – labour market – open data – Russia

– Related data sets "Educational and career trajectories in Russia" (Dataset 1) and "Activity of unemployed and job-seeking candidates in Russia" (Dataset 2) with URL www.zenodo.org/records/12727876 in repository "Zenodo"

1. Introduction

Studying Russian society can be challenging, mostly due to the lack of reliable data. Access to comprehensive and up-to-date data on Russian society, demography and labour market can be limited or restricted by the authoritarian regime, making it challenging for researchers (Cohen & Arieli, 2011). Existing data usually lack reliability or accuracy due to outdated collection methods, inconsistent reporting and methodology, or governmental control over information dissemination, leading to potential inaccuracies. Government control over data release and transparency can limit the availability of

detailed and transparent information regarding employment figures, wages, or workforce demographics. It is also known that official statistics from the Russian Federal State Statistics Service (Rosstat) produce inconsistencies in historical data because of changes in either definitions or methodologies used in data collection and reporting over time (Kapeliushnikov, 2023).

Apart from the labour market, there is a more broad and significant concept that covers and interconnects multidisciplinary research across individual life decision-making and labour policy development which is called an educational and career trajectory (Malik, 2019). It refers to the path an individual follows from their educational experiences through to their career choices and advancements over time (Hodkinson & Sparkes, 1997). It encompasses their educational pursuits, skill development, professional experiences, and the progression or changes in their career and life goals. Studying educational and career trajectories can provide valuable insights, but expanding research in this area requires high-quality, highly detailed data covering temporal and spatial aspects of the trajectories.

Studying educational and career trajectories is important for several reasons. Firstly, understanding how individuals move through educational systems and into careers helps policymakers design more effective education policies. It can highlight areas where improvements are needed, such as skill gaps, access to quality education, or alignment between education and job market needs. Secondly, trajectory studies shed light on disparities in educational and career opportunities, revealing barriers faced by certain demographic groups. This understanding is crucial in addressing inequalities and promoting diversity and inclusion in education and the workforce.

Thirdly, examining trajectories and candidates' activities allows for the evaluation of educational programs and institutions. It helps assess the effectiveness of different educational paths in preparing individuals for successful careers. It significantly impacts social mobility and individual well-being. Analysing these trajectories helps identify pathways for upward mobility and areas where interventions can improve outcomes for marginalised or disadvantaged groups. By studying past and current trajectories, researchers and policymakers can anticipate future challenges and opportunities in the job market, education sector, and overall economy, enabling proactive planning and interventions.

Existing studies show that in Russia career-educational choices are influenced by parental expectations, available opportunities, and the overall cultural and social context that shapes the system for distinguishing and

evaluating career-educational options (Bessudnov, 2016; Khavenson & Chirkina, 2019; Minina et al., 2020). While young people from lower classes experience a lack of opportunities to achieve their goals and are forced to lower their aspirations, their more privileged peers face class-based expectations that restrict their career and educational choice to one path only (Minina et al., 2020).

However, these studies do not appear to create a coherent picture of how particular class attributes interplay with cultural and political circumstances. There is a lack of studies that cover the effect of social policies in the labour or educational sectors, and the impact of foreign policies, for example, the beginning of large-scale invasion and mobilisation. For example, the case of the introduction of the Unified State Exam on the level of educational inequality and the quality of the labour force. Although there are some shreds of evidence that the Unified State Exam served as a democratising tool to decrease the barriers to higher education in Russia (Francesconi et al., 2019), educational performance proved to be highly dependent on the socioeconomic status of the family (Khavenson & Chirkina, 2019; Prakhov, 2016).

The educational attainment data now is limited, and it is especially rare to have educational trajectories data linked with career trajectories data. There is a unique longitudinal educational and career trajectories dataset that only captures the years 2012–2022 and the years after the war in Ukraine (TREC, 2022). Additionally, there is only a longitudinal monitoring survey covering the long period 1994–2022 (RLMS-HSE, 2023), which contains a basic questionnaire on a person's education and career and is used to study the dynamics of wage inequality and employment rates in Russia. However, these data do not contain precise indications of educational institutions, courses taken and company names. Thus, our data can supplement this dataset with highly detailed information. This will strengthen research on educational and career trajectories, as well as analysis of labour market activity in Russia over the past decades. These challenges related to data availability, reliability, and accessibility create obstacles for researchers, policymakers, and analysts seeking to understand and interpret the intricacies of Russian society comprehensively. Thus, in this article, we intend to present a new data source and two prepared datasets that partially cover these requirements, which we believe will be useful to the research community. First of all, it enhances the possibilities of researching social and political changes in Russia in the areas of education and work in conditions where the opportunities to collect data are limited. Second, it contributes to the fields of educational and career studies, especially those concerned with measuring the effects of state policies, crises, and wars.

2. Data and Methods

2.1. *Data Source*

The Ministry of Labor and Social Protection of the Russian Federation (Ministry of Labor of Russia) and The Federal Labor and Employment Service (Rostrud) introduced the Jobs in Russia (JR) portal[1] in 2009 as an official user register for unemployed citizens. This was the long-term result of President Medvedev's modernization program aimed at developing a digital society in Russia.

The project provides the opportunity to place and receive information about vacancies and applicants throughout the country free of charge. Information posting can be carried out by state authorities, employment services, personnel agencies, organisations, and citizens.[2]

In 2014, an update of the Federal Law "On Employment in the Russian Federation"[3] introduced a rule according to which employers must ensure the provision in the JR system of complete, reliable, and up-to-date information about the needs for employees and the conditions for their employment, as well as the availability of free jobs and vacant positions. According to the government decree, state structures and organisations were recommended to post information on their vacancies in the JR portal.

In 2020, citizens can register as unemployed and apply for benefits through the JR portal. This can also be done through public services and remotely, which has made this portal highly sought after by the public. For 2017–2020, information on more than 15 million personal files of citizens who applied for assistance in finding a suitable job was entered into the registers (Kopytok & Kuzmina, 2021).

To date, the JR portal provides information about vacancies in all regions of Russia, invitations from employers, CVs, and responses from job seekers, which is updated daily and can be freely downloaded in .xml, .json or .csv formats.[4] This is raw and semi-structured data that accumulates some updates and deletions from time to time.

2.2. *Data Processing*

We utilised archived raw CV updates from the JR portal as input data. The data was downloaded in.xml format for the last day of each month from January 2019 to December 2023.

1 The JR portal. www.trudvsem.ru (IP-locked to Russia).

2 Work of Russia (portal). TAdviser. www.tadviser.com/index.php/Product:Work_of_Russia_%28portal%29.

3 Law of the Russian Federation from 19.04.1991 № 1032-1 "On employment of the population in the Russian Federation".

4 Open Datasets. The JR portal. www.trudvsem.ru/opendata/datasets (IP-locked to Russia).

The processing ETL pipeline includes removing duplicated CVs and partially normalising the data using a hash function, and then uploading to a *Postgres* database (see code[5]). The main pipeline includes checking for errors in year of birth and other dates, as well as correcting numeric values. The normalization of textual information covers removing extra spaces, unnecessary prefixes, and non-alphabetic characters, reordering or removing excessive punctuation, and removing obvious duplicates. This pre-clean database was then split into two cleaned datasets:

– Dataset 1 on educational and career trajectories;
– Dataset 2 on the activity of unemployed and job-seeking candidates.

Dataset 1 is designed to study educational trajectories. It was obtained by aggregating all CVs for each candidate into a single record with a unique *id_candidate*, then normalising text information about the candidates' work experience and education, and finally excluding records that do not contain detailed information. Dataset 1 is also divided into three tables with general information, and detailed information on education and work experience. The tables can be joined by *id_candidate*.

Dataset 2 includes all records for the available period without detailed descriptions of working experience and education. It might be useful to study the general dynamics and candidates' activity on the JR portal and, therefore, on the labour market. The calculation of candidates' activity metrics is described in the following section.

3. Data Summary

3.1. *Dataset 1*

– **Educational and Career Trajectories in Russia deposited at Zenodo – DOI:**www.doi.org/10.5281/zenodo.10913325
  – Data files
  – Codebook
– **Temporal coverage:** 2015–2023
– **Geographical coverage:** 83 regions of Russia

Dataset 1 contains N = 6,221,439 personal records from candidate CVs for the 2019–2023 portal update period, so the earliest internal CV update is April 2015,

5 GitHub repository. www.github.com/tha-rc/rostrud_pipeline.

and the latest is October 2023. The dataset consists of only records containing work experience or education information and covers several variables: year of birth (median age is 38), gender (36.9% male), region of residence, maximum salary requested (median is 30,000 Rubles), positions requested, detailed work experience, and detailed education information.

3.2. *Dataset 2*

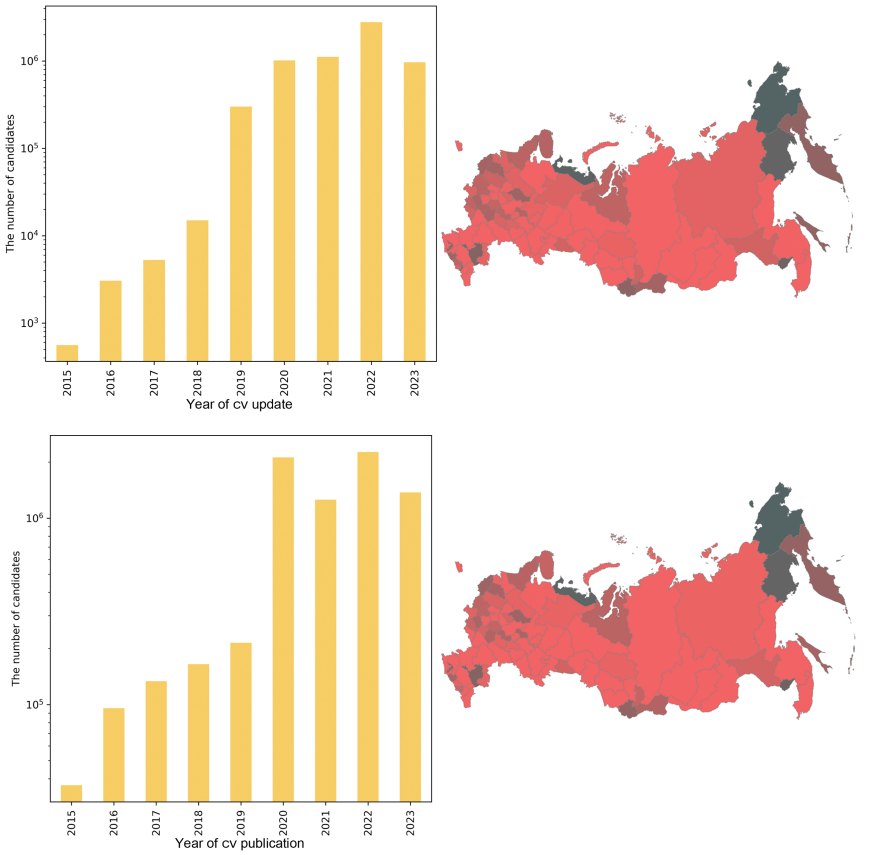
- **Activity of Unemployed and Job-seeking Candidates in Russia deposited at Zenodo – DOI:**www.doi.org/10.5281/zenodo.10913325
  - Data files
  - Codebook
- **Temporal coverage:** 2015–2023
- **Geographical coverage:** 83 regions of Russia

Dataset 2 contains N = 7,662,089 personal records from candidate CVs for the 2019–2023 portal update period, so the earliest CV was published in February 2015 and the latest in October 2023. The dataset covers several variables: year of birth (median age is 38), gender (38.7% male), level of education, region of residence, length of work experience, desired type of employment, maximum salary requested (median is 30,000 Rubles), dates related to resume creation, posting and updating, and related to candidate activity on the portal.

3.3. *Data Coverage*

The temporal and geographical coverage of the data is presented in Figure 1. The Figure illustrates the exceptional spatial data coverage. The data covers 83 regions of Russia (100% of the regions that were officially registered as 'Russian regions' as of 2008). Note that the regions represent the place of residence. JR does not provide information on the candidate's place of origin.

Our dataset comprehensively covers all regions of Russia and includes more than a million records per year since 2020. This extensive and detailed data is a significant advantage for researchers, providing a rich resource for in-depth analysis and robust conclusions. According to age, the data covers a representative sample of the Russian labour force (see Figure 2). Note that there was a pension reform in 2019, which led to the change in the retirement age: it is being gradually increased from 60 to 65 for men by 2023 and from 55 to 65 for women by 2026. The application for this change depends on the candidate's year of birth. In terms of gender, there is some bias in the data towards women – 60% female vs. 40% male, while according to aggregated official statistics, the average share of women in the labour force is about 49%.



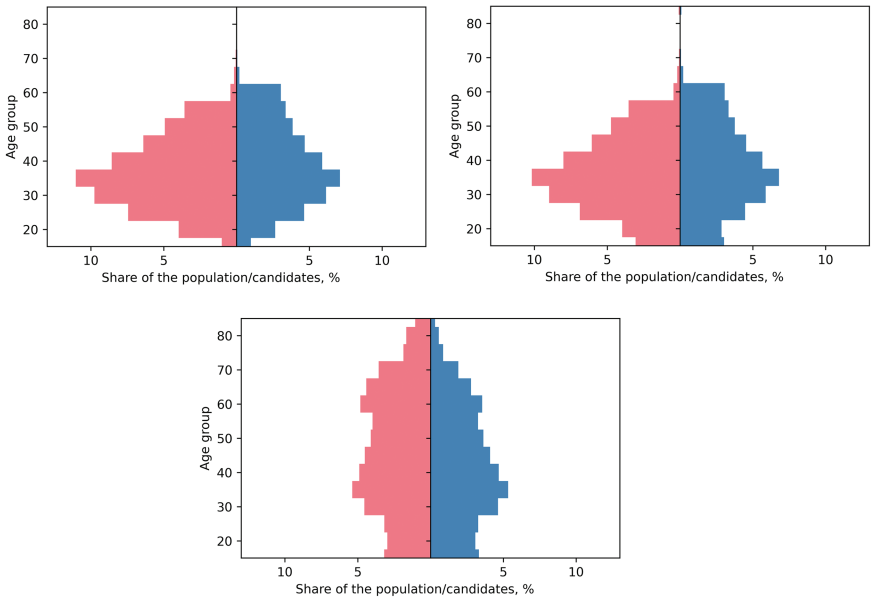
Note: Temporal coverage of data CV update/publication date: logarithmic scale. Geographical coverage: light colour – maximum 4.5%; dark colour – minimum 0.01% of the sample size.

FIGURE 1A-D Temporal and geographical coverage of data for CV updates and publication (left: dataset 1; right: dataset 2)

We assume that due to the somewhat patriarchal system of relations in Russia, women are more likely to work in less-qualified jobs, while men are more likely to obtain high-ranking positions (Klimova, 2012). These high-ranking positions are often achieved through promotion rather than through open platforms. This could explain why the percentage of male users on the platform is slightly lower.

3.4. *Data Limitations and Remarks*

As we mentioned earlier, the active use of the JR portal started in 2019–2020 due to changes in legislation, so it makes sense to use the activity statistics



Note: Top left: Dataset 1; top right: Dataset 2; bottom middle: demographic pyramid of the Russian population, estimate for 2023. Left axis: age group; lower axis: share of population/candidates (%); left side of the pyramid: female, right side: male.

FIGURE 2A-C Demographic pyramid of the data

from this period. However, educational and career trajectories can be explored for a longer period (1953–2028). Note that individuals across the datasets can be matched by *id_candidate*, which is unique.

Further, it must be noted that since using the portal is not obligatory, it does not represent the labour market of the whole country. The Russians who upload their CVs to the portal come from different regions; however, it does not mean that they depict the situation for the whole population of the country. Thus, the information contained in the datasets has some representativeness limitations.

Because the data records in CVs were manually filled out, descriptions of work experience and education contain some inaccuracies. In particular, in a small number of cases (<0.1%), the salary requested is specified in thousand Rubles instead of Rubles as in the general case. The names of companies and educational institutions may contain abbreviations, or they may be written in full. In some cases, it leads to the entries being duplicated, as it is difficult to exclude them from the records automatically.

Dates can also be presented in a fragmented way, as these fields in the CV form are not required. The year of birth for candidates with ages less than 14 or

greater than 85 was marked as not assigned. The year of education completion earlier than 1953, later than 2028, and earlier than the year of birth plus 10 was also marked as not assigned. In addition, for candidates aged over 65, there is ~0.02% of Dataset 1 records with education completion later than 2025. This is considered a rare case for Russia, but verifying whether this is an actual error is impossible, so we have kept such entries.

Finally, the text data presented in the datasets are in Russian and are written in the Cyrillic alphabet. Therefore, to use it for content analysis, some knowledge of Russian is required.

4. Research Potential

Studying workforce activity and career trajectories can provide valuable insights, but expanding research in this area requires high-quality, highly detailed data covering temporal and spatial aspects of the trajectories. However, this requirement seems to be very challenging, especially in the Russian context.

To date, we are aware of only two studies that have used data from the presented data source due to related technical difficulties. The first quantitative study was conducted by Vitovt Kopytok and Yulia Kuzmina from the Center for Advanced Governance. They published an analytical report on the impact of the coronavirus pandemic in 2020 on registered unemployment in Russia (Kopytok & Kuzmina, 2021). They showed that as a result of the introduction of the possibility to remotely apply for unemployment benefits and the increase in the amount of unemployment benefits, many more people have applied to employment services than in pre-pandemic periods. The structure of unemployment in Russia has changed due to the increase in the informal economy sector and long-term unemployed citizens. Additional payments to these categories helped to reduce the effects of the pandemic related to the fall in incomes of the population.

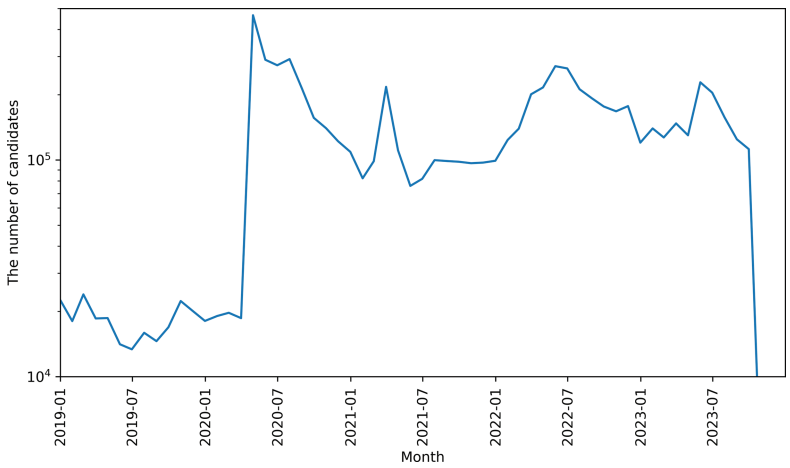
The second study by Suchkova et al. (2023) investigated the changes in the dynamics of applications for unemployment benefits in response to the abolition of regional restrictive measures during the first wave of COVID-19 spread in Russia. They showed that after the lifting of restrictive measures, the number of new applications for unemployment benefits did not decrease significantly. The result remained robust when an alternative measure was used. This research is interesting from the perspective of developing anti-crisis support measures for the population and employs the staggered difference-in-differences approach (Freedman et al., 2023).

Thus, Rostrud as the data source and our two prepared datasets can be utilised in a quantitative design and modern policy analysis (Angrist & Krueger, 1999), such as the difference-in-difference technique (Athey & Imbens, 2006; De Chaisemartin & D'Haultfoeuille, 2022), to examine several effects that follow peaks and dips in candidates' activity over the available period (see Figure 3). Some of the identified events that can be investigated are summarised in Table 1.

Another practical example is the development of new variables for analysing workforce activity. In order to help researchers study the activity of candidates on the JR portal and, consequently, on the labour market in Russia, we decided to identify the more active candidates. For this purpose, we applied clustering using the k-POD method (Chi et al., 2016), which is designed for the case when there are missing values in the data.

To perform clustering, we used only technical characteristics of CVs and candidate activity: the number of characters in the free-entry fields; the number of days between the CV creation date, and the date of its publication and update; the total number of CVs published; and the number of candidate responses to vacancies. So, we divided candidates into two clusters (see Figure 4): cluster 1: likely active candidates, and cluster 0: likely passive candidates.

Thus, taking into account political events, candidates' activity on the platform and their individual trajectories can provide a deep insight into social and political phenomena in Russian society.



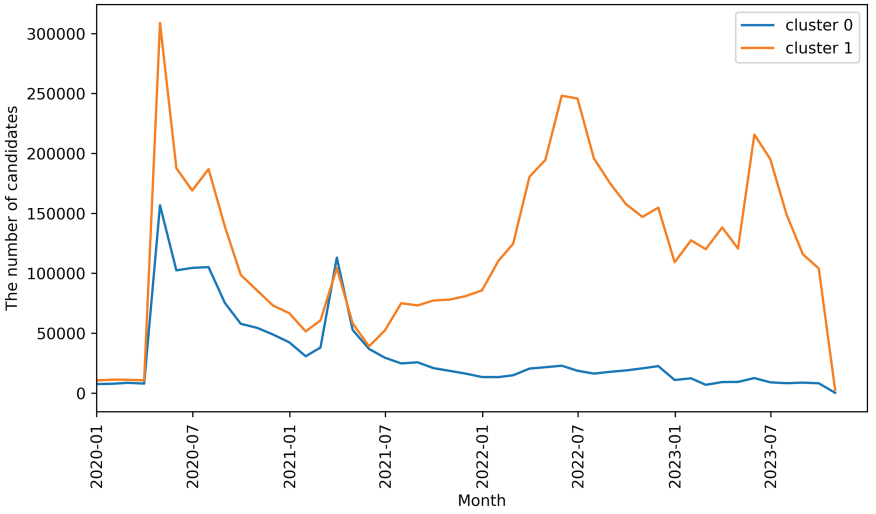
Note: Dataset 2: monthly aggregation, logarithmic scale.

FIGURE 3 The number of candidates that published their CVs by publication date

TABLE 1 Events and policies as an example

Date or period	Event or policy	Comment
May–June 2020	The opportunity to apply for unemployment benefits	From March 30 to May 11, 2020, Russia had a regime of non-working days with pay. This regime was introduced again at the federal level on May 4–7, 2021 and from October 30 to November 7, 2021, but then the regional authorities themselves could decide which industries to stop.[a]
February–March 2021	Introduction of anti-pandemic restrictions	
February 2022	Russian invasion of Ukraine	Russia's unemployment rate has fallen, leaving businesses struggling to find workers for the labour-intensive industries that dominate the country's economy.[b]

[a] *History of coronavirus restrictions in Russia* (2022, July 1). TASS. www.tass.ru/info/15101389
[b] Russia's war economy leaves businesses starved of labour. *Financial Times*. www.ft.com/content/dc76f0bb-cae2-4a3a-b704-903d2fc59a96



Note: Dataset 2: monthly aggregation; cluster 1: likely active candidates; cluster 0: likely passive candidates.
FIGURE 4 The number of candidates by activity clusters and CV publication date

5. Concluding Remarks

All raw data available in the source are distributed in the public domain and are completely free of charge based on the principles of use by the Russian Government.[6] The data that was obtained as a result of our scripts are distributed under the CC-BY-4.0 License. All code is provided under the MIT License and can also be used freely, see: www.github.com/tha-rc/rostrud_pipeline.

Acknowledgements

The authors are grateful for the support of Prof. Dr. Katharina Bluhm and the Institute for East European Studies, Freie Universität Berlin. We thank the reviewers for the careful and insightful review of the manuscript.

References

Angrist, J. D., Krueger, A. B. (1999). Empirical strategies in labor economics. In O. Ashenfelter, & D. Card (Eds.), *Handbook of Labor Economics* (vol. 3, Part A, Ch. 23, pp. 1277–1366). Elsevier. https://econpapers.repec.org/RePEc:eee:labchp:3-23.

Athey, S., & Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, *74*(2), 431–97.

Bessudnov, A. (2016). Sotsial'no-ekonomicheskoye i gendernoye neravenstvo pri vybore obrazovatel'noy trayektorii posle okonchaniya 9-go klassa sredney shkoly. [Socio-economic and gender inequality when choosing an educational trajectory after graduating from the 9th grade of secondary school]. *Educational Issues*, (1), 135–167.

Chi, J. T., Chi, E. C., & Baraniuk, R. G. (2016). k-POD: A method for k-means clustering of missing data. *The American Statistician*, *70*(1), 91–99.

Cohen, N., & Arieli, T. (2011). Field research in conflict environments: Methodological challenges and snowball sampling. *Journal of Peace Research*, *48*(4), 423–435.

De Chaisemartin, C., & D'Haultfoeuille, X. (2022). *Difference-in-differences estimators of intertemporal treatment effects*. NBER Working Paper No. 29873. www.nber.org/papers/w29873.

Francesconi, M., Slonimczyk, F., & Yurko, A. V. (2019). Democratising access to higher education in Russia: The consequences of the unified state exam reform. *European Economic Review*, *117*, 56–82.

6 www.github.com/tha-rc/rostrud_pipeline/blob/main/misc/protokol2016.pdf.

Freedman, S. M., Hollingsworth, A., Simon, K. I., Wing, C., & Yozwiak, M. (2023). *Designing difference in difference studies with staggered treatment adoption: Key concepts and practical guidelines*. NBER Working Paper No. 31842. www.nber.org/papers/w31842.

Hodkinson, P., & Sparkes, A. C. (1997). Careership: A sociological theory of career decision making. *British Journal of Sociology of Education*, *18*(1), 29–44.

Kapeliushnikov, R. I. (2023). The Russian labor market: Long-term trends and short-term fluctuations. *Russian Journal of Economics*, *9*(3), 245–270.

Khavenson, T., & Chirkina, T. (2019). Obrazovatelny vybor uchashihsya posle 9-go i 11-go klassov: Sravnenie pervichnyh i vtorichnyh effektov sotsialno-ekonomicheskogo polozhenia semji [Educational choice of Russian high school students in grades 9 and 11: A comparison of primary and secondary effects of the family's socioeconomic status]. *Journal for the Studies of Social Policies*, *17*(4), 1–20.

Klimova, A. (2012). Gender differences in determinants of occupational choice in Russia. *International Journal of Social Economics*, *39*(9), 648–670.

Kopytok, V., & Kuzmina, Yu. (2021). *COVID-19-era unemployment: What can administrative data tell us?* www.github.com/tha-rc/rostrud_pipeline/blob/main/misc/KopytokKuzmina2021.pdf.

Malik, V. (2019). The Russian panel study "Trajectories in education and careers". *Longitudinal and Life Course Studies*, *10*(1), 125–144.

Minina, E., Yanbarisova, D., & Pavlenko, E. (2020). Educational choice of Russian high school students in grade nine. *International Studies in Sociology of Education*, *29*(4), 326–343.

Prakhov, I. (2016). The barriers of access to selective universities in Russia. *Higher Education Quarterly*, *70*(2), 170–199.

RLMS-HSE (2023). *Russia longitudinal monitoring survey*. HSE. https://rlms-hse.cpc.unc.edu.

Suchkova, O. V., Stavniychuk, A. Y., Kalashnov, G. Y., & Osavolyuk, A. (2023). The effect of the removal of regional anti-COVID restrictive measures on the dynamics of applications for unemployment benefits in Russia. *Population and Economics*, *7*(2), 1–22.

TREC (2022). *Russian panel study "Trajectories in Education and Careers"*. HSE. http://TrEC.hse.ru.