Assessment of the possibility of using social network data in urban research

Anna Uskova[1*], *Julia* Salomatova[1], and *Nikita* Salomatov[1]

[1]Institute of Economics of the Ural Branch of the Russian Academy of Sciences, 29 Moskovskaya St., 620014 Ekaterinburg, Russia

Abstract. Widespread digitalization leads to the acceleration of socio-economic processes, which requires faster decision-making. Official statistics is the basis for managing the socio-economic development of a city or region. However, it has a number of limitations related to the timing of publication of information and details. At the same time, new data sources have been emerging in recent years. This study determines the possibility of using open data of the VKontakte social network in socio-economic research using the example of Russian million-plus cities. The main research methods were statistical, descriptive, and comparative analysis. The information base of the study was made up of depersonalized data of the VKontakte social network users, as well as statistical data. It is revealed that on the basis of depersonalized data from VKontakte, it is possible to characterize the gender and age composition of city residents, which is comparable with Rosstat data, to identify and evaluate labor migration. It is concluded that the data from the social network can be used to obtain analytical information in situations where there is no official statistics available. The data have high geographical detail and can act as an additional source of operational big data on the composition of the population and the quality of its life.

Key words: Big data; Social network "VKontakte"; Data from social networks; Pendulum migration; Socio-demographic portrait.

1 Introduction

The current turbulence of the world economy, instability, and social tension in society require a timely response to the ongoing processes, including from authorities at all levels. At the same time, official statistics, which is the basis for the analysis of the socio-economic situation and for the formation of economic and social policy, do not always meet such criteria as efficiency and sufficient detail. In this connection, the Russian Federal State Statistics Service (Rosstat) is constantly searching for new opportunities to use alternative sources of information, in particular, big data. Thus, according to the report on the results of Rosstat's activities for 2019 and tasks for 2020, pilot projects are being launched in certain statistical sectors (trade, tourism, consumer prices, demography) [https://clck.ru/35KeCA].

*Corresponding author: uskova.ay@uiec.ru

In particular, it is planned to work out the possibility of using big data generated by the Federal Tax Service in order to replace the forms of statistical observation, by the Pension Fund of the Russian Federation in order to generate statistics on wages and incomes of the population, by credit organizations in order to analyze the level of expenses and assess tourist flows, by cash registers in order to generate price statistics and retail trade, by mobile operators in order to estimate the population in the inter-census period, labor and pendulum migration.

Today, the existing digital technologies allow using the capabilities of social networks to obtain big data in many ways. Taking into account the coverage and activity of the population in social networks, these technologies are more representative and financially accessible than traditional opinion polls. In the article "Innovations in the field of data for development purposes" (https://www.un.org/ru/global-issues/big-data-for-sustainable-development/), published on the official website of the United Nations, it is noted that new elements obtained as a result of processing big data gained in real time can serve as an important addition to the data of official surveys and statistical information, thereby helping to analyze people's behavior and their experience. And their joint use allows for timely analysis of information at a qualitatively higher level. "Internet research is an interdisciplinary and multidisciplinary field of fundamental and applied research, combining various scientific disciplines" [1], such as sociology, political science, anthropology, economics, cultural studies, linguistics, law, journalism, etc. [2]. Social media data are a rich source of data because of their volume and diversity. According to the Global Digital 2023 report (https://www.web-canape.ru/business/statistika-interneta-i-socsetej-na-2023-god-cifry-i-trendy-v-mire-i-v-rossii/) as of January 30, 2023, there were 106.0 million users of social networks in the Russian Federation, which is 73.3% of the total population. At the beginning of 2023, 83.0% of all Internet users in the Russian Federation were present in at least one social network. 56% of the population of the Russian Federation are users of the VKontakte social network (hereinafter referred to as VKontakte), whose user traffic increased by more than a third in 2022.

The analysis of studies carried out on the basis of data from VKontakte allows us to identify the following research objects: posts published in communities (including comments to them, likes, reposts, tonality), subscribers' network connections, characteristics of participants in thematic communities, visual range. At the same time, the prevailing methods are the analysis of moods and content contained in the posts and comments of users of social networks. According to recent foreign studies, the wide range and depth of use of social network data cover many different subject areas [3]. Russian studies are devoted to such spheres of life as: education, ecology, tourism, healthy lifestyle, politics, charity, behavioral models of rural residents, quality of life. For example, Russian researchers determined the index of subjective (non-)well-being, based on the study of online activity of users of the most popular regional and urban communities in VKontakte [4]; the educational achievements of students were predicted, including on the basis of their digital footprint in VKontakte [5]. Gapich, Asatryan and Minkina consider the main conceptual approaches used in the study of the influence of visual content of social networks in the form of photographs, videos, graphics, drawings, demotivators, etc. on the processes of politicization and radicalization of youth [6]. Lakman and Timiryanova on the basis of VKontakte data revealed the spatial dependence of the age of the participants of the group "Abzakovo" on the remoteness of the city of residence from the ski resort [7].

At the same time, the data set contained not only in community posts, but also in the profile of VKontakte users is wide, thus, the potential of using VKontakte data for economic research has not yet been fully disclosed. This determines the relevance of this study, the purpose of which is to determine the possibility of using VKontakte data in

socio-economic research on the example of Russian million-plus cities. To achieve the purpose of the study, the following tasks were set:

- to determine the mechanism for obtaining depersonalized data from VKontakte user profiles;
- to assess the representativeness of the data obtained, including by comparing them with the official data of Rosstat;
- to test the use of VKontakte social network data in order to form and analyze the socio-demographic portrait of residents, identify attitudes to bad habits (smoking, alcohol), as well as identify and assess labor migration of the population.

2 Data and research methods

The analysis of social networks includes the development and evaluation of "[...] information tools and frameworks for collecting, monitoring, analyzing, summarizing and visualizing social network data, usually determined by the specific requirements of the target application" [8]. The fundamental difference between social media analysis and traditional business intelligence methods is that it uses real-time data, rather than exclusively structured and historical data, to gain insight into current issues, supporting effective decision-making [9]. Within the framework of the study, an approach was taken as the basis for the organization of the analysis process in social networks, which includes four separate stages: detection, collection, preparation and analysis of data [10].

1. The existing ways, methods and tools of data collection from VKontakte are analyzed: uploading via API, web scraping (parsing), emulation of the mobile application. The option of using the VKontakte API was chosen - an interface that allows you to get information from the database vk.com using HTTP requests to a special server. The syntax of queries and the type of data they return are strictly defined on the side of the service itself. VKontakte API methods are conditional commands that correspond to a particular database operation — getting information, writing or deleting. To get a list of users, the "users.get" method is used, in which you can specify parameters for getting people who fit certain criteria (gender, city of residence, age, etc.), and the fields that should be in the response (followers_count– number of subscribers, career, interests, etc.), containing information about the user. In addition to the above methods of data collection, access to data can be provided by the use of specialized platforms [11]. Some of them are commercial services with paid parsing services (TargetHunter, Segmento Target, Pepper.ninja, Cerebro, vk.barkov), or free with a limited amount of uploaded data - The social network data collection and processing platform SNLab, which was used during the collection of initial data for this study.

The empirical data for the study were data from the profiles of VKontakte users - residents of million-plus cities aged between 14 and 74 years, collected in March-April 2023. All data had been depersonalized in order to comply with the rules of personal data processing. Further, the data were aggregated according to the following characteristics of the population: gender (male, female), age, employment (studying at school, studying at university, working (city of work), attitude to bad habits – smoking and alcohol consumption. For the convenience of working with aggregated data, the data array was broken down by the following ages: 14-17 years, 18-44 years, 45-59 years, 60-74 years. The classification of the World Health Organization from 2016 was taken as a basis, in addition, the study introduced the category of children - 14-17 years old [12].

As part of the study, depersonalized data was collected from about 9.1 million VKontakte user profiles, which is on average 63% of the number of residents of the studied cities in a given age period - from 45% to 74%, depending on the city (Table 1). If we take the average values of the total number of residents of the city, then 51% of citizens have

profiles, which correlates with the data of the Global Digital 2023 report (56% of the population of the Russian Federation have a profile in VKontakte). Residents of Nizhny Novgorod have the most (59%), residents of Novosibirsk have the least (37%).

Table 1. Source data for cities.

City	Total residents aged 14-74, persons, as of 01.01.2023 (according to Rosstat)	Total users of the social network «VKontakte», persons	Ratio of users to the number of inhabitants (14-74 years old)	Ratio of users to the number of inhabitants (18-44 years old)
Volgograd	820 077	542 287	66%	102%
Voronezh	804 093	598 891	74%	100%
Ekaterinburg	1 205 547	618 718	51%	61%
Kazan	987 764	532 519	54%	60%
Krasnodar	817 553	531 602	65%	74%
Krasnoyarsk	865 444	586 510	68%	85%
Nizhny Novgorod	1 036 511	731 853	71%	97%
Novosibirsk	1 310 864	594 863	45%	53%
Omsk	882 931	600 286	68%	97%
Perm	834 389	535 164	64%	82%
Rostov-on-Don	940 282	636 332	68%	94%
Samara	889 270	578 395	65%	83%
Ufa	903 183	606 761	67%	75%
Chelyabinsk	936 353	599 226	64%	77%

Taking into account the size of the general population exceeding 13.2 million people, 99% probability that a random answer will fall into the confidence interval and 1% deviation of the average characteristics of the sample population from the average characteristics of the general population, the minimum required sample size is 16.6 thousand people. Thus, 8.3 million users who make up the sample within the framework of the study are sufficient to represent the characteristics of the general population with a given error and confirm the quantitative representativeness of the study data relative to Rosstat data.

Rosstat data on population size, gender and age composition of the population, federal statistical observations on socio-demographic problems were also used.

The main research methods of the collected data were statistical analysis of socio-demographic characteristics, comparative analysis, descriptive method and spatial statistics method, visualization of results.

3 Research results

3.1 Socio-demographic portrait of citizens

According to the study, the gender and age composition of VKontakte users living in million-plus cities differ from Rosstat data. First of all, main users of the social network are men aged 18-44. The maximum number of citizens covered by the social network is two age and gender groups: women aged 14-17 (98%) and men aged 18-44 (91%). Minimally - women and men aged 60-74 (26 and 27%, respectively).

Women make up an average of 55.1% of the city population, men – 44.9%, while among VKontakte users, women make up an average of 49.4%, and men – 50.6%.

This pattern may be associated with several reasons, including the use of Telegram, Odnoklassniki social networks by people aged 45-74 or not using them at all, the preference

of the female audience for social networks focusing on visual content. A larger percentage of men aged 18-44 may be associated with the pages of enterprises in the social network, when an enterprise, university department or other organization starts a page on behalf of the organization, but it does not lead it as a public page or group, but as a profile of a real person, and also with the fact that nonresident students of educational organizations can mark their place of residence in a social network in a million-plus city, while according to Rosstat they are registered at the place of permanent registration. It is also impossible to exclude the existence of fake accounts created by users (http://mediacentr.by/articles/30-skolko-vkontakte-fejkovykh-akkauntov.html).

3.2 Attitude to smoking and drinking alcoholic beverages

The study analyzed data from the "Life position" section of the profiles of VKontakte users who expressed their attitude to smoking (1,269 thousand) and alcohol consumption (1,271 thousand), which is 15.3% of the total number of users. Evaluation options involve choosing one of 5 degrees - from sharply negative to positive. For simplification, the following scheme was used: the relations "Neutral" and "Positive" were taken into account as "Positive", the rest - "Negative".

According to the study, 85.6% of VKontakte users who expressed their attitude to smoking are negative about it, and 14.4% are positive. The maximum loyalty to smoking among both sexes falls on the age category of residents aged 18-44 years, and the minimum – on 14-17 years. Among men, the maximum percentage of people loyal to smoking falls on the age category of residents 45-54 years old, and the minimum is 14-17 years old. Among women, the maximum percentage of loyal is in the age group of 18-44 years, and the minimum is in 14-17 years.

The study revealed cities with greater loyalty to smoking, for example, Novosibirsk - for both sexes, Volgograd and Omsk - for women. On the contrary, residents of Ekaterinburg and Ufa, regardless of gender, are minimally committed to smoking.

According to the study, 12.1% of VKontakte users who expressed their attitude to bad habits have a positive attitude to alcohol consumption, and 87.9% have a negative attitude. Residents of Novosibirsk of both sexes are as loyal to alcohol consumption as possible, while residents of Ekaterinburg, Nizhny Novgorod and Ufa, regardless of gender, are minimally committed to drinking. In Volgograd and Voronezh, women are as loyal to alcohol as possible, and in Kazan and Samara – men.

To assess the commitment of residents of the Russian Federation to a healthy lifestyle, Rosstat conducts a comprehensive survey of the living conditions of the population once every 2 years (about 60 thousand households and over 100 thousand citizens aged 15 years and more participate in the whole of the Russian Federation, urban and rural settlements with different populations, according to individual socio-demographic population groups).The prevalence of smoking and alcohol consumption was assessed as part of the observation.

According to the observation data in the whole of the Russian Federation, about 80% do not smoke or smoke occasionally, 20% smoke daily. The maximum percentage of smokers among both sexes is 35-44 years old (27.6%), and the minimum is 15-17 years old (1.1%). Among men, the maximum percentage of smokers falls on the age category of residents 45-54 years (46.4%), and the minimum – on 15-17 years (1.8%). Among women – from 0.3% of smokers in 15-17 years to 11% in 35-44 years, respectively. In the context of residents of cities with a population of 1 million or more persons aged 15 and over, the overall picture corresponds to the federal one, except that men smoke less than the national average, and women, on the contrary, smoke more (by 2 percentage points). According to the comprehensive monitoring data, in the Russian Federation at large, 14.3% consume strong

alcoholic beverages more than once a week. The maximum percentage of users among both genders falls on the age category of residents of 45-54 years (21%), and the minimum – 15-17 years (0.2%).

Thus, it can be concluded that the data on a healthy lifestyle from VKontakte is generally comparable to the data from Rosstat. At the same time, it should be taken into account that the data of Rosstat includes data from Moscow and St. Petersburg, whose population exceeds the number of cities under consideration. As for the quantitative representativeness of these data, they are 12 times higher than the sample of the integrated observation of Rosstat. At the same time, they provide a basis for conducting an analysis in the context of certain cities.

3.3 Labor migration within the region and between million-plus cities

The set of user data contained in VKontakte includes information such as the city of residence from the Contacts section and information about the user's career, containing the following fields: company name, country and city identifiers, city name, year of commencement and termination of work, position. Thus, the comparison of information from these fields makes it possible to identify and evaluate labor migration.

The Encyclopedia of Statistical Terms defines labor migration or migration of labor as the movement of the population across national borders, that is, external migration, or across administrative borders within the country, that is, internal, for the purpose of employment in the country (region) of entry (https://clck.ru/35KeNs). Currently, there are no official statistics regarding labor migration, including pendulum migration, questions about labor (pendulum) migration are received by statistical agencies only within the framework of population censuses, since 2002. However, the results of the population census give an idea only about labor pendulum migration in the whole country and by region (https://rosstat.gov.ru/vpn_popul). Rosstat also conducts selective monitoring of migrant labor every 5 years, the purpose of which is to compare the situation on the labor market of the population aged 15 years and older, depending on the migration status; the prevalence of attracting migrant workers (both foreign and Russian) by households and entrepreneurs; comparing the scale of attracting foreign workers to work in households migrants temporarily staying in Russia and Russian citizens (https://gks.ru/free_doc/new_site/imigr18/index.html). These data do not allow us to estimate the magnitude of labor migration in the context of municipalities and, moreover, settlements, and also do not give an idea of other types of pendulum migration. As part of this observation, Rosstat surveyed 130,000 households with a population aged 15 and over in all regions of the Russian Federation, which is only 0.24% of the total number of households.

In the absence of sufficient statistical data, pendulum migration is often investigated on the basis of survey/interviewing data of residents, passenger transportation data, Pension Fund data. In recent years, big data on population movement trajectories using geoinformation systems, localization of activity in mobile communication networks, and automobile traffic have also been actively used for this purpose [13, 14].

It is necessary to note the main problems associated with existing sources of information. For all their informativeness, sociological surveys/interviews are distinguished by their significant cost [13]. Big data generated by GPS trackers or mobile operators is more suitable for intraregional research, while such data does not give an idea of the socio-demographic characteristics of migrants. The limitation of the widespread use of mobile operators' data is also their high cost. A significant part of the above limitations can be overcome by the use of social media data, for example, VKontakte, which a) can be

obtained for free if appropriate competencies are available, b) allow one to form a socio-demographic portrait of people engaged in labor migration.

This study analyzes labor migration between million-plus cities, including Moscow and St. Petersburg, as well as intraregional labor migration between million-plus cities and their satellite cities based on data from 395.6 thousand users who indicated their place of work (4.3% of the total number of users), which is 18 times higher than the share of the Rosstat observation sample.

The coefficient of interregional attractiveness is calculated as the ratio of users-residents of million-plus cities who indicated a million-plus city as their place of work to the number of users-residents of city X who indicated one of the other million-plus cities as their place of work. As well as the coefficient of intraregional attractiveness - as the ratio of the number of users, residents of satellite cities who indicated the place of work of a given million-plus city to the number of users, residents of a given million-plus city who indicated the place of work of satellite cities. The results of the study showed that Volgograd is the least attractive among the million-plus cities for both intraregional and interregional labor migration, that is, more working citizens leave the city than enter (Table 2). Among the million-plus cities for interregional labor migration, Ekaterinburg, Kazan, Novosibirsk, Krasnoyarsk and Chelyabinsk are the most attractive. The data obtained, among other things, provide a logical explanation for Rosstat data on the dynamics of the permanent population in the studied cities.

Table 2. Information on the interregional and intraregional attractiveness of cities for migrant workers.

City X	Coefficient of interregional attractiveness, %	Coefficient of intraregional attractiveness, %	The ratio of the permanent population at the end of 2021 to 2010, %
Kazan	108,0	218,9	109,9
Volgograd	82,7	88,3	98,1
Voronezh	74,9	108,4	107,1
Ekaterinburg	137,8	188,3	110,0
Krasnodar	87,3	133,3	127,4
Krasnoyarsk	122,6	130,9	112,7
Nizhny Novgorod	80,8	159,3	99,3
Novosibirsk	126,4	204,4	110,7
Omsk	91,0	104,5	97,6
Perm	97,4	189,7	105,2
Rostov-on-Don	98,0	178,1	104,0
Samara	97,8	146,5	97,4
Ufa	80,3	159,1	106,7
Chelyabinsk	108,6	162,0	104,3

4 Conclusion

In this study, using the example of Russian cities with millions of people, it is shown that VKontakte data are quantitatively representative, have high geographical detail and breadth of coverage of the object of study.

The results of the analysis of users' attitudes to such harmful habits as smoking and alcohol consumption confirm that the data from the "Life position" section of the VKontakte social network user profile can be collected at any time and in any geographical reference, unlike statistical data published by Rosstat once every two years and do not give specifics in the context of subjects The Russian Federation and/or cities and localities.

Taking into account the availability of information about the year of birth in the user's profile, it is possible to conduct research in any age group breakdown.

As part of the analysis of users' employment, the possibility of studies of labor migration of the population, both interregional and intraregional, has been identified, the results of which, taking into account assumptions, can be used to assess the imbalance in the labor market of a particular city.

At the same time, it must be remembered that when working with social network data, there are a number of limitations and inaccuracies that the systems themselves impose. So, for example, a user may not update his profile, and he may be older (a person may not update his profession, place of work, even after changing them). Women are less likely to indicate their age on the Internet, which leads to their "rejuvenation". It turns out that with this method, Internet users are younger than in the census [15].

But even taking into account the existing limitations, VKontakte as a source of open data opens up new opportunities for analyzing the current socio-economic situation in cities.

References

1. Yu. Rykov, Russian Sociological Review **16,** 366 (2017) https://doi.org/10.17323/1728-192X-2017-3-366-394
2. V. Golbraikh, Monitoring Obshchestvennogo Mneniya: Ekonomicheskiei Sotsial'nye Peremeny **4,** 62 (2022) https://doi.org/10.14515/monitoring.2022.4.2140
3. C. Zachlod, O. Samuel, A. Ochsner, S. Werthmüller, Journal of Business Research **144**, 1064 (2022) https://doi.org/10.1016/j.jbusres.2022.02.016
4. E.V. Shchekotin, M.G.Myagkov, V.L. Goiko, V.V. Kashpur, G.Yu. Kovarzh, Monitoring of Public Opinion: Economic and Social Changes **1**, 78 (2020) https://doi.org/10.14515/monitoring.2020.1.05
5. V.V. Kashpur, E.Yu. Petrov, V.L. Goiko, A.V. Feshchenko, Tomsk State University Journal of Philosophy, Sociology and Political Science **64**, 140 (2021) https://doi.org/10.17223/1998863X/64/13
6. A.Eh. Gapich, S.S. Asatryan, O.V. Minkina, The Humanities and socio-economic sciences **12,** 38 (2021) https://doi.org/10.23672/k4225-9727-4779-i
7. I.A. Lakman, V.M. Timiryanova, Strategic Decisions and Risk Management, **2,** 170 (2021) https://doi.org/10.17747/2618-947X-2021-2-170-177
8. D. Zeng, H. Chen, R. Lusch, S.-H. Li, IEEE Intelligent Systems **25,** 13 (2010) https://doi.org/10.1109/MIS.2010.151
9. G.F. Khan: Seven Layers of Social Media Analytics: Mining Business Insights from Social Media; Text, Actions, Networks, Hyperlinks, Apps, Search Engine, and Location Data. CreateSpace Independent Publishing Platform (2015)
10. S. Stieglitz, M. Mirbabaie, B. Ross, Ch. Neuberger, International Journal of Information Management **39,** 156 (2018) https://doi.org/10.1016/j.ijinfomgt.2017.12.002
11. A.L. Blaginin, Eh.R. Saifulin, A.Yu. Sarkisova, From the experience of organizing automated data collection at Tomsk University. In Big data and society's problems: collection of articles (2022), pp. 34-46
12. Computer program «Obtaining data on labor migration based on the profiles of users of the social network «VKontakte» (certificate number: 2023619437)

13. A.A. Sokolova, Economy of the North-West: Problems and Prospects **1,** 52 (2022) https://doi.org/10.52897/2411-4588-2022-1-52-66
14. P.A. Dyachkova, N.L. Mosienko, World of Economics and Management **21,** 205 (2021) https://doi.org/10.25205/2542-0429-2021-21-4-205-228
15. I.B. Orlova, E.V. Fomin, National security / nota bene **3,** 48 (2020) https://doi.org/10.7256/2454-0668.2020.3.33274