***An approach to identifying social protest potential based on a city-to-city similarity metric***

***Introduction.*** Nowadays, the evaluation of city-to-city similarity is a relevant approach that can be applied to various tasks, including infrastructural, architectural, and other areas [1, 2, 3, 4]. Interest in this research direction is mainly driven by the possibility of using similarity assessment to address urban planning problems [4]. At the same time, social conflicts that arise during the implementation of construction projects are of great importance in the field of urban planning [5, 6]. Therefore, forecasting social risks (such as protests and conflicts) is a relevant task for urban planning. Undoubtedly, this problem is even more relevant for socio-economic policymaking [7–11].

Motives and factors of modern social protests are under careful investigation due to their significant impact on the socio-political situation, which, in turn, is closely connected with socio-economic processes [7–11]. The relevance of this research area is beyond doubt, since certain mechanisms of protests remain debatable or even unknown [7, 8]. The growing number of recent studies on this topic also serves as evidence of strong interest.

The main goal of this research is to develop a method for assessing the potential for social conflict in a given municipality (city, town, or village). The core of the approach is estimating the similarity not only of the socio-economic environment, but also the demographic structure (human capital) to characteristics that could show the predisposition to certain types of social conflict. In other words, similarity between two municipalities, where one of which has experienced a particular social conflict, can serve as an indicator of a likely comparable social response to the same type of stimulus.

***Related works.*** There are various approaches to investigating social conflicts. So far, the most common approach has been to examine the problem within the context of sociological and psychological research [7, 8]. However, it should be noted that forecasts based only on sociological data do not demonstrate high accuracy [7].

Also one of the common approaches is based on mathematical modeling of collective behavior [8]. It mainly involves the development of various agent-based systems that allow the analysis of protest dynamics through imitation processes. For example, an epidemiological model is often used as a representation of protest movements [8].

In recent years, machine learning approaches have also gained popularity in this field of research [9, 10, 11]. These studies are generally based on analyzing public sentiment through the monitoring of social networks and mass media [9, 10, 11].

In study [9], protests are investigated through the analysis of the stylistic features of mass media articles covering labor protests that took place in Romania (2018–2025). Within the context of the present research, particular interest lies in the evidence supporting the spatio-temporal hypothesis of protest diffusion, meaning that a protest in one location increases the occurrences of protests in neighboring areas [9]. This conclusion resonates with the effectiveness of the epidemiological approach to analyzing protest dynamics. However, the present research proposes a method that considers at least the “spatial” component in terms of the socio-economic and demographic space, which is often influenced by geographical closeness but is not determined by location alone.

The study [10] used social network data to forecast civil unrest events. A modified neural network was proposed, trained on Twitter data (now X.com) using the example of the 2019 protests in Hong Kong. The effectiveness of this approach was demonstrated by comparing it with various other machine learning algorithms [10].

In [11], protests in Latin America (2012–2014) were examined using a logistic regression model that was trained on data not only from the social network Twitter (now X.com), but also from other sources, including news feeds, political event databases, the TOR network, and economic indicators. It is important to note that only currency exchange rates were used as the economic indicator [11]. While this is a valuable factor, it is not sufficient to capture the full socio-economic condition of a territory within a country.

Despite the evolution of protests in terms of motives (shifting away from purely materialistic reasons), the significance of socio-economic factors remains under debate - “misery matters” [12, 13]. Study [12] examines the correlation between GDP and protest intensity, while [13] investigates how both general and individual socio-economic factors influence the likelihood of respondents to participate in protests. Moreover, the findings of [13] provide evidence for the hypothesis that shared-grievances are more important than individual-grievances. This evidence supports the idea of the present research, because the city-to-city (or any other municipal entity) similarity assessment is more about “environment matters” rather than “misery matters”. Therefore, this research proposes an approach detached from subjective data, assessing similarity not only through socio-economic indicators but also by demographic structure (human capital).

***Data.*** The analyzed dataset includes socio-economic and demographic data for over 1800 municipal entities in Russia (districts, cities, and villages) for the period 2014 to 2022, collected from the Federal State Statistics Service (ROSSTAT). Most records (around 80%) correspond to municipal districts, as separate data for small cities and villages are generally unavailable in open access. In total, the dataset includes more than 9000 examples, since each municipality–year combination is treated as an independent observation.

Table 1 presents the selected socio-economic indicators. These factors were chosen based on their demonstrated effectiveness in migration forecasting study []. Therefore, they provide a reliable basis for assessing the socio-economic conditions of a municipality.

1. Average number of employers in organizations - avgemployers (ppl.)

2. Average salary - avgsalary (rub.)

3. Shopping area - shoparea (sq.m.)

4. Number of seats in café, bars and restaurants - foodseats (num.)

5. Retail turnover - retailturnover (thnd. rub.)

6. Live area per capita - livarea (sq.m.)

7. Number of sporting venues - sportsvenue (num.)

8. Number of services (barbershops, repairs) - servicesnum (num.)

9. Length of roads - roadslen (km.)

10. Number of livestock - livestock (num.)

11. Productivity of land (vegetables) - harvest (centners)

12. Agricultural production - agrprod (thnd. rub.)

13. Number of healthcare organizations - hospitals (num.)

14. Number of places in preschool organizations - preschool (num.)

15. Volume of self-produced goods - factoriescap (thnd. rub.)

All socio-economic indicators (except average salary) were normalized on a per capita basis to enable fairer similarity evaluation across municipalities and to provide a better assessment in terms of population well-being (see Table 1).

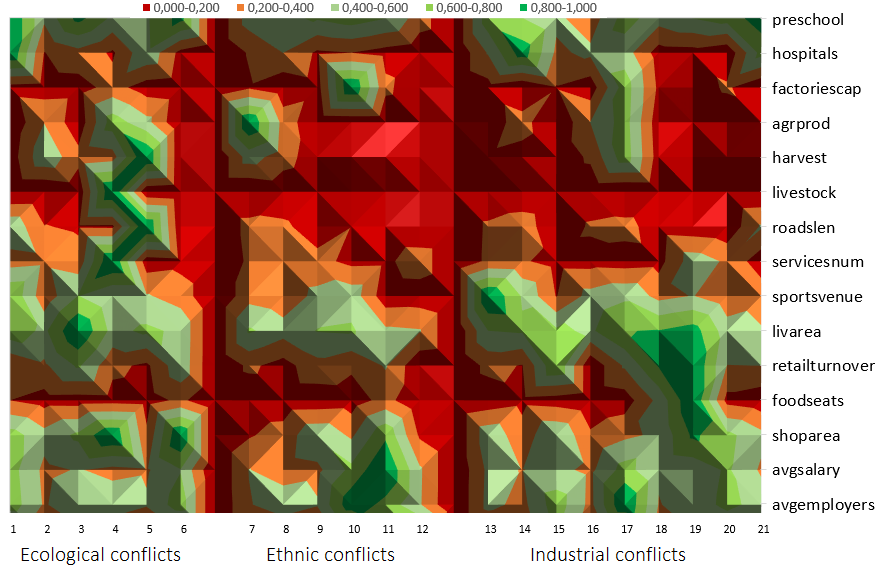
To represent the population structure of the municipalities, demographic data were collected for 14 cohorts (from ‘0–4’ to ‘65–69’), disaggregated by gender for each year. Unfortunately, data for cohorts older than ‘65–69’ were missing for many municipalities. It is important to note that the demographic data were also normalized as the proportion of each cohort (age group) in the total population of a given municipality, which makes it possible to evaluate human capital in terms of structure rather than quantity.

***Set of Social Conflicts.*** The socio-economic and demographic characteristics of a municipality at the time it experiences a protest could be used to identify a predisposition to certain social conflicts or general social tension. Thus, the set of municipalities (or conflict cases) against which similarity is assessed forms one of the crucial components of the proposed approach.

A set of the most notable social conflicts in Russia from 2013 to 2023 has been collected. It includes 21 cases of social protests in different municipalities, each involving mass gatherings of people (over 100 participants) and receiving wide media coverage. These cases can be grouped into three categories: six ecological conflicts (Shies 2018, Poltavskaya village 2022, etc.); six ethnic conflicts (Chemodanovka 2019, Makhachkala 2023, etc.); and nine industrial conflicts (Kushtau 2020, Anapa 2021, etc.).

The focus on high-profile conflicts is motivated by their potential to highlight the socio-economic and demographic indicators that predispose municipalities to social tension. In other words, these cases can be viewed as “local extrema” of an abstract function representing social conflict.

During the comparison of the socio-economic conditions of these local extrema (conflict cases), certain trends can be observed. Figure 1 shows that in the entertainment (“servicesnum”, “sportsvenue”, “retailturnover”, “foodseats”) and in the production sectors (“factoriescap”, “agrprod”, “harvest”, “livestock”) municipalities that experienced ethnic conflicts exhibit the lowest values (many red and few green zones). This indicates that municipalities with ethnic conflicts were in a more depressed socio-economic environment than those experiencing industrial or ecological conflicts. In particular, municipalities with ecological conflicts appear to be the most balanced (see Fig. 1).



Some trends and dependencies were also identified in the analysis of the demographic structure of these municipalities. Figure 2 illustrates certain differences between categories of conflicts. For example, municipalities that experienced ethnic conflicts have the highest proportion of young people in the population (cohorts “20–24,” “25–29,” and “30–34”; see Fig. 2). In the “25–29” cohort, the proportion of both males and females is nearly double that in municipalities with ecological conflicts.

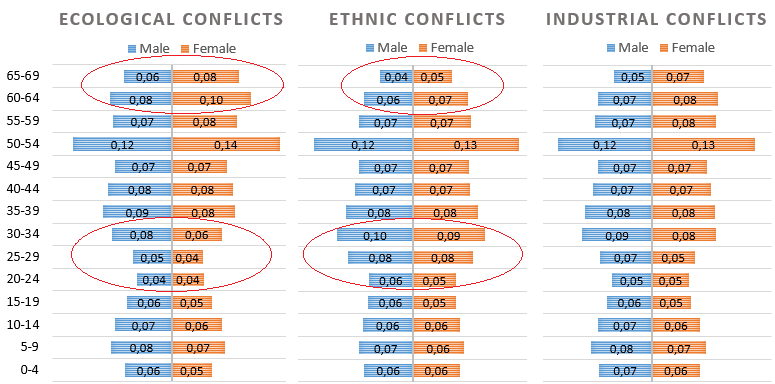


Рис. 2. Усредненные значения демографических структур МО по каждому типу конфликта

Municipalities with ecological conflicts show the lowest proportion of young people, but at the same time have the highest proportion of older population (cohorts “60–64” and “65–69”; see Fig. 2). In contrast, the demographic structure of municipalities with ethnic conflicts demonstrates the lowest proportion of people in these cohorts. For example, compared to ecological conflicts, in the “60–64” cohort municipalities with ethnic conflicts have about 25% fewer men and 30% fewer women, while in the “65–69” cohort the difference is even greater – about 33% fewer men and nearly 38% fewer women. The population structure of municipalities with industrial conflicts falls somewhere in between.

Thus, notable differences between municipalities with certain types of conflicts can be observed both in socio-economic conditions and in demographic structure. This provides strong evidence for the potential use of these similarities as a measure of the risk of social conflict.

***Similarity metric.*** Each municipal entity (municipal district, city, village etc.) in the data defined by the socio-economic and demographic components. The socio-economy condition could be represented as a tuple:

where is the value of the specific socio-economic indicator (see Section 3, Table 1)

The demographic structure can be represented by two separate tuples, indicating the proportions of females and males in the population:

here is the number of females in a given cohort, is the total number of females in the population and *k* is the number of cohorts. The division enables calculation of the proportion of each cohort within the population, allowing assessment of the demographic structure. The same approach is used to represent the male population structure:

where is the number of male in the cohort, is the total number of male in the population.

For evaluating the similarity between two municipal entities x and y, it is proposed to use the combined sum of the mean squared deviations of socio-economic and demographic indicators:

where – is the value of socio-economic feature of municipality x*, n* is the number of socio-economic features, and is the proportion of female/male in the certain cohort of municipality x*, k* isthe number of cohorts and is the weighting coefficient that could balance the socio-economic and demographic component.

First, it is necessary to evaluate the applicability of the similarity metric (4). This involves estimating the similarity of a given municipality to all other municipal entities in the dataset. Table illustrates the case of Michurinsk City in 2022. As expected, in the majority of cases, the proposed metric (4) identifies the same municipality observed in different time periods as the most similar to a given municipality. This confirms the reliability of the similarity metric. It should be mentioned that in the case of Michurinsk City the set of top similar municipalities includes not only its geographical neighbor Morshansk City (both located in Tambov Oblast), but also the geographically distant Vyshnevolotsky Dist. (Tver Oblast), demonstrating similarity despite the lack of spatial closeness.

The next step is to examine the potential of using the similarity metric in the context of the main idea, that a municipality similar to another, which has experienced a specific type of “social turbulence”, may demonstrate a comparable social response to a more or less similar stimulus.

As an example, the similarity metric (4) is applied to identify the most similar municipal entities to Krasnoarmeysky District, where Poltavskaya Village experienced large-scale protests against a landfill in 2022 [link to media sources???]. The results are partly presented in Table~\ref{sim-2}, where all municipal entities are sorted by their similarity to Krasnoarmeysky District. The metric values are normalized to the range from 0 to 1.

The Figure~\ref{sim-3} presents a graph that illustrates all values of metric (4) from the assessment of Krasnoarmeysky District. The graph demonstrates that a similarity value from 0 to 0.1 is reached by approximately the 300th example, while the range from 0.1 to 0.2 is only crossed around the 5245th example (Troitsky District, see Table~\ref{sim-2} and Figure~\ref{sim-3}). This result shows a certain degree of homogeneity across Russia, at least in terms of per capita characteristics. In other words, a large number of municipal entities do not differ substantially from one another, which is generally expected within a single country. At the same time, the shape of the curve can serve as an indicator of uniqueness: the sharper its initial rise, the fewer similar municipalities exist.

The results of socio-economic and demographic similarity to Krasnoarmeysky Dist. can be further considered in the context of identifying potential social conflicts. In the case of Kurganinsky Dist. (first place in similarity, see Table~\ref{sim-2}), media analysis revealed no evidence of significant conflicts. It should be noted that, within the framework of the proposed approach, not only similarity but also the presence of a comparable triggering factor is essential.

In the analysis of Tikhoretsky Dist. (2022), third place in similarity (see ), a fairly large-scale social protest was identified (see Fig. 4, left). In 2022, the protest occurred in Alexeyevskaya Village in response to a water supply outage and involved over 200 participants. A considerable number both in the context of Russian protests and relative to the village’s population of just over 3,000 residents.



Рис. 4. Митинги в Тихорецком муниципальном районе в 2022 году (слева) и 2024 (справа)

However, in this case, one can only speak of a social tension in the context of similar socio-economic and demographic conditions, while the primary interest lies in the possibility of a comparable reaction to the same trigger. Such a case occurred in Kamenny Village (population ~1,000, see Fig. , right), where local residents recorded video demanding to deal with the landfill. Nevertheless, this event took place in Tikhoretsky Dist. in 2024, which is not included in the dataset.

When analysing Belorechensky Dist. for 2019 (26th place, see Table~\ref{sim-2}), a social protest related to a landfill was identified, although it actually occurred in 2021 (see Figure~\ref{belo}). The protest involved approximately 500 participants. Belorechensky Dist. of 2021 is also presented in the dataset and is in the 50th place in similarity (see Table~\ref{sim-2}). This example demonstrates that municipalities with similar socio-economic and demographic conditions can indeed exhibit comparable reactions to specific social triggers.

******

Рис. 5. Митинг в Белореченске за закрытие мусорного полигона в 2021 году

It is evident that the similarity assessment shows potential to be useful in the analysis and evaluation of the predisposition of certain municipal entities to specific social conflict. However, this approach is unable to analyze municipalities based on the similarity (4) to indicators that are typical for the majority of social conflicts (see Table), which can be interpreted as an estimation of the overall social tension.

In other words, using the similarity approach as the basis, it is possible to develop a method, which allows assessing the existence of municipal entities that can be close to several municipalities that experienced social conflicts of both the same and different types.

***Method for assessing the relative risk of social conflict.*** A ranking system is proposed to evaluate the overall potential for social conflict. This system will assign each municipal entity in the dataset a ranking score based on their individual similarity to specific municipality that serves as example of social conflict (21 case, see Table). The sum of all these scores will considered as a measure of social risk (protest, conflict).

In general, the proposed method can be described as follows:

- Preparation of the necessary datasets (socio-economic and demographic data for municipal entities) and selection of conflict cases;

- Assessment of the similarity (4) of all municipal entities to a specific conflict-affected entity;

- Ranking of all municipal entities according to their level of similarity;

- Repeating steps 2 and 3 for each conflict case;

- Calculation of the sum of ranking scores as an overall measure of social conflict risk.

It is clear that in addition to the accuracy of the similarity metric, the applied set of socio-economic and demographic indicators, and the quality of the data, the ranking system will also have a crucial impact on the overall effectiveness. There are numerous ways to perform the ranking according to the level of similarity.

For example, the simplest approach is to apply a progressive ranking of similarity, ranging from 1 (most similar, first place) to 0 (least similar, last place). However, due to a certain degree of homogeneity observed in the country (see Fig.~\ref{sim-3}), such a scoring system will heavily favor municipal entities that are more or less similar to a large number of others.

In the logic of the hypothesis of this study, the most similar municipal entities may share the same degree of predisposition. Therefore, a top-ranking system should be considered in this case. An initial analysis of the graphs suggests that it is reasonable to assume the first 300 most similar municipal entities as comparable (see Fig.~\ref{sim-3}). Consequently, the proposed ranking system assigns scores as follows: 1 point for municipal entities ranked from 1 to 100 by the similarity metric (4); 0.5 points for places from 101 to 200; and 0.2 points for places from 201 to 300. However, it should be noted that both the threshold (top 300) and the corresponding score values remain subject to further optimization.

***Result analysis.*** The complete distribution of the ranking scores is presented in the diagram (see Fig.~\ref{scores}). Almost half of the municipal entities in the dataset (4035 out of more than 9000) obtained a nonzero score, indicating some degree of closeness (4) to at least one conflict-affected entity. However, the majority have scores of 1 or less (3049 cases, see Fig.~\ref{scores}), while only 986 entities reached the relatively higher cumulative score. According to the ranking system, the maximum possible total score is 21 (1 point for each conflict). However, in practice this is highly unlikely, since it would mean that a single municipal entity is simultaneously very similar to all 21 others, which is not possible given the way conflict-affected entities were selected (see Section 3.2).

Table 5 shows part of the output of the developed method, including the first five municipal entities with the highest total ranking scores. The highest total score of 4.5 was obtained by Tuapsinsky District in 2022 and 2021. It is not surprising that Tuapsinsky District received 1 point (by entering the top 100 in similarity) with conflict-affected municipal entities №18 (Anapa City) and №19 (Gelendzhik City), since all of them are located in Krasnodar Krai and are geographically close (see Table~\ref{rank-tab}). However, Table~\ref{rank-tab} also shows that Tuapsinsky District received 0 points with Krasnoarmeysky District (case №2), meaning it did not even enter the top 300 in similarity, despite the fact that Krasnoarmeysky District is also in Krasnodar Krai.

From this point of view, Novoaltaysk City (third place, see Table 5) becomes even more interesting, as it received 0.5 points with №18 (Anapa City) and 0.2 points with №19 (Gelendzhik City), despite being more than 4,000 kilometers away. These results once again confirm one of the theses of this study (see Section 2), namely that the socio-economic and demographic space is not necessarily determined by geographical closeness.

While the assessment of similarity to a single municipal entity (see Section 4.1) reflects the predisposition to a specific type of social conflict (a comparable response to a similar trigger), the overall ranking assessment, as noted earlier, can be considered as an evaluation of the general potential for social conflict. Accordingly, the analysis considers conflicts of any kind, if they involve mass gatherings of people (over 100 participants).

During the analysis of the Tuapsinsky Dist. (1–2 places, 4.5 points, see Table~\ref{rank-tab}), a political protest was identified in Tuapse City (population ~60,000) in 2021, organized in support of Alexey Navalny (see Figure~\ref{Tuapse}). In 2022, no evidence of social conflict was recorded in the media. It is important to note that the score shows potential, but the occurrence of a social conflict still requires a trigger. In 2024, the Tuapsinsky Dist. experienced a significant social conflict in Dzhubga (population ~7,000), where several hundred people participated in a protest against the new urban development plan (the so-called "Genplan", Figure~\ref{Tuapse}). The particularly interesting point here is that the cause of the protests in Anapa (case №18) and Gelendzhik (case №19) in 2021 (see Table~\ref{conf-tab}) was also the "Genplan", with which the Tuapsinsky Dist. demonstrates a high similarity (4).

In the city of Novoaltaysk (population ~70,000, third place, see Table), two significant social conflicts were identified, which occurred in 2018 and 2019. While data for 2019 are not available in the dataset, 2018 data are present, and the overall social risk score for that year is also high (2.9 points, 75th place). The 2018 protest was related to the new retirement reforms, whereas in 2019 residents protested the closure of the maternity hospital (see Figure).

In the Kandalakshsky District, a significant social conflict was recorded in 2017 (4th place, 4.2 points, see Table~\ref{conf-tab}), where several hundred people protested in Kandalaksha City against the unfair tax system (see Figure~\ref{...}). The city of Apatity was also analyzed, and a conflict was identified in 2017, where over 100 researchers protested, demanding increased funding for science (see Figure~\ref{...}). The overall social risk score for Apatity City in 2017 is indeed high – 3.9 points (13th place).

The analysis illustrates that the proposed method could indeed be useful, at least as an additional tool for the evaluation of social risk or tension. The identified protests occurred in relatively small settlements by Russian standards (population less than 100,000) and received limited media coverage, so the detection of these protests in such areas can be considered a promising sign of the effectiveness of the method.

Feature importance and forecasting the dynamics. The developed method provides a deterministic evaluation of the potential for social conflict based on socio-economic and demographic indicators for a specific year. Consequently, it is not possible to calculate the evaluation for a future year (or for any period with missing data). Forecasting models could be applied to address this limitation.

There are various fields of machine learning applications and forecasting of social conflicts is no exception []. Therefore, it is reasonable to investigate of possibility to use this approach in the context of the proposed method for assessment the potential of social conflict. On the one hand, implementation of machine learning will allow evaluating the importance of socio-economic and demographic factors, and from the other hand, it could predict the dynamics of changes in the potential.

Since the dataset provides socio-economic and demographic data of municipal entities across specific years, a straightforward application of machine learning would be to forecast the overall social risk score (see Table~\ref{rank-tab}, column ‘Total’) for the subsequent year based on the socio-economic and demographic condition of the previous year. It is necessary to note that while the proposed ranking system is simple and easy to interpret from a human perspective, it limits the effectiveness of model training. For this reason, the top 300 should be represented as a continuous scale from 1 to 0, where the 1st place corresponds to 1 and the 301st place to 0. This adjustment would not cause any substantial changes in the overall social risk assessment, but the smoother scoring would improve the predictive performance of the model.

It is proposed to develop two independent models that will forecast the potential of social conflict: one based only on socio-economic indicators and the other based only on demographic indicators. This approach allows a clearer focus on the importance of features, since using both sets of factors in a single model could overlap their individual significance.

The final dataset for model training includes approximately 7000 examples. Records from 2022 were excluded, as the model requires an assessment of the potential of social conflicts for the following year. Cases with missing data for the subsequent year within 2014–2022 were also removed.

The Random Forest method was chosen for the initial experiments. It is a simple and widely used approach that also provides the estimation of feature importance based on the Gini impurity. To enhance the robustness of both the feature importance estimates and the accuracy of the model, the training process was repeated 50 times, and the average values were calculated.

***Reference***

[1] A Similarity Approach to Cities and Features. 2022. URL: <https://www.researchgate.net/publication/358646626_A_Similarity_Approach_to_Cities_and_Features>

[2] Analysis of the Uniqueness and Similarity of City Landscapes Based on Deep Style Learning. 2021. URL: <https://doi.org/10.3390/ijgi10110734>

[3] Exploring venue-based city-to-city similarity measures. 2013. URL: <https://www.researchgate.net/publication/257212810_Exploring_venue-based_city-to-city_similarity_measures>

[4] The scenario method in urban planning. 2022. URL: <https://www.researchgate.net/publication/276321581_The_scenario_method_in_urban_planning>

[5] Are land use conflicts a barrier to sustainable city development? Evidence from Chattogram District of Bangladesh. 2025. URL: <https://www.researchgate.net/publication/388926768_Are_land_use_conflicts_a_barrier_to_sustainable_city_development_Evidence_from_Chattogram_District_of_Bangladesh>

[6] Conflicts in urban peripheries in Europe. 2023. URL: <https://www.researchgate.net/publication/373236871_Conflicts_in_urban_peripheries_in_Europe>

[7] Davydov D. A. (2022) Dynamics of Mass Protest Actions in Modern Russia: An Event Study. Monitoring of Public Opinion: Economic and Social Changes. No. 5. P. 72–93. https:// doi.org/10.14515/monitoring.2022.5.2199 (In Russ.)

[8] Petrovskii S, Shishlenin M, GlukhovA (2025) Understanding street protests: from amathematical model to protest management. PLoS ONE 20(4): e0319837. <https://doi.org/10.1371/journal.pone.0319837>

[9] Nicula, Alexandru-Sabin and Cretan, Remus and Simionescu, Mihaela and Oancea, Bogdan and Dragan, Alexandru, The City as Stage: Labour Protest, Sentiment, and Machine Learning. DOI: <http://dx.doi.org/10.2139/ssrn.5357390>

[10] Iyda, J.J., Geetha, P. An improved deep belief neural network based civil unrest event forecasting in twitter. Appl Intell 53, 5714–5731 (2023). https://doi.org/10.1007/s10489-022-03746-3

[11] Korkmaz, G., Cadena, J., Kuhlman, C.J. et al. Multi-source models for civil unrest forecasting. Soc. Netw. Anal. Min. 6, 50 (2016). <https://doi.org/10.1007/s13278-016-0355-8>

[12] Korotayev, A. V., Sawyer, P. S., & Romanov, D. M. (2021). Socio-Economic Development and Protests: A Quantitative Reanalysis. Comparative Sociology, 20(2), 195 222. <https://doi.org/10.1163/15691330-bja10030>

[13] Azedi, A. (2024). From shared grievances to collective action: A multilevel study of economic adversity and protest. *International Sociology*, *40*(1), 3-32. <https://doi.org/10.1177/02685809241292076> (Original work published 2025)