Supplementary materials for "Quantifying world geography as seen through the lens of Soviet propaganda"

M.V. Tamm[1,#], M. Oiva[2], K.D. Mukhina[1], M. Mets[2], M. Schich[3]
[1] *School of Digital Technologies, Tallinn University, Tallinn, Estonia;*
[2] *School of Humanities, Tallinn University, Tallinn, Estonia;*
[3] *Baltic Film and Media School, Tallinn University, Tallinn, Estonia;*
[#] *This is the corresponding author*

(Дата: 18 октября 2024 г.)

I. DATA PREPARATION

A. Dataset characterization

The corpus of Soviet Newsreel "News of the Day" (Новости дня / Хроника наших дней] was downloaded from Russian footage archive Net-Film[1] with permission of the owners, it was previously introduced and discussed in [2]. The Daily News journal was the main newsreel journal produced by the Central Film Studios of Documentary Film in Moscow. The corpus includes almost all issues of this newsreel from 1954 to January 1992 (except for the year 1965), as well as a few surviving issues from 1944 to 1953. In Fig. 1a the number of issues per year is presented. Starting from 1954 the newsreels have been saved systematically, and the newsreel production have peaked with 72 reels in 1954 and 65 in 1955. For thirty years, in 1956-1986 the usual annual number of newsreels was stable at 48-52 issues, meaning approximately one issue per week. Starting from 1987 the annual number of newsreels dropped to 26 issues (1 issue in 2 weeks). Overall, the corpus includes more than 1700 short films of usually 9-10 minutes length.

The films are complemented with metadata, including the information on the production date, the crew, and the short outlines. The newsreels and metadata are in Russian; three members of the research team (MT, MO and KM) are fluent in Russian and thus were able to perform data cleaning, preparation and preliminary analysis.

Typically, each newsreel is split into several (usually 5-10) short news stories. These stories are typically well separated (e.g., by a black screen between them) and are topically unrelated. There is a small fraction (around 3%) of single-topic issues (year-end, celebration-related, dedicated to party congresses, etc.) which either consist of a single story or a sequence of very short stories (up to 15 in 10 minutes) filmed in different places but united by a single topic (e.g., "working women in the USSR"). Finally, the dataset includes 30 double issues, i.e., two consequential issues united into a single film on a single topic. These are dedicated mostly to big political events, 18 out of 30 double issues are in years 1990-91.

B. Stories and outlines

We use an outline of a story as an elementary unit of analysis. We mostly use the outlines available in the complementary metadata. We made an extensive random check and found that the outlines are of satisfactory quality, with a very small number of mistakes: the fraction of outlines with typos in place names was significantly below 5%, and we only once (out of several hundreds checked) been able to find a film outline in which one of the stories was missing. In the vast majority of cases the format of outlines allowed automatic splitting into stories. Exceptions where (i) around 1% of newsreels where there were typos in the numbering of stories within a newsreel which we corrected manually, (ii) around 3% of of newsreels (most of them from years 1989-92) which had a different format of outlines: instead of a contents summary it included description of camera movement, wide shots vs close-ups, etc; for these roughly 50 newsreels we have rewritten the outlines to match the format of the rest.

Overall the dataset consists of 12 707 story outlines (on average 7.5 per newsreel), in Fig. 1b their distribution by year is presented, the full list of the outlines is provided in [10]. It is seen that the huge majority of the dataset (97.5%) corresponds to 1954-1986, i.e., the period between the death of Stalin and the early years of perestroika. Interestingly,

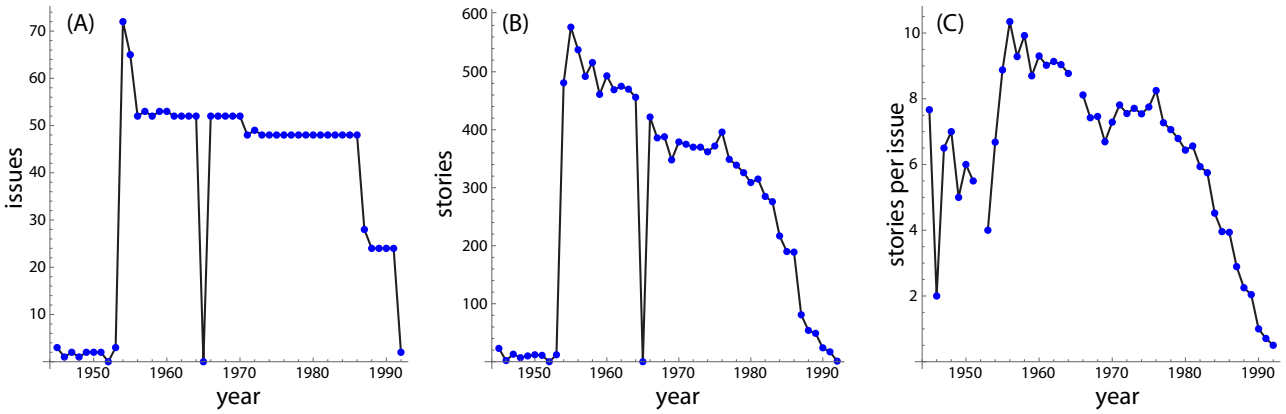

Рис. 1: Temporal structure of the newsreel corpus used (A) number of issues in the dataset per year, (B) number of stories per year, (C) mean number of stories per issue.

the number of stories per newsreel issue trends down with time, especially after 1975. Median date of a story is 1968 and 50% of stories belong to the period 1960-76 with 25% dated before and 25% after this period.

C. City population

For estimates of the city population we use the USSR censuses for Soviet cities and UN and (if needed) national data for foreign cities.

In the case of USSR population is an extremely important variable (see Fig. of the main text), and USSR census is a relatively consistent and reliable dataset. As a proxy of the population we use an average fraction of population of the USSR living in a given city averaged over three censuses of 1959, 1970 and 1979 [**?**]. The list of cities of interest include all 309 cities with population more than 0.03% of the population of the USSR, except Moscow. For the purposes of models that include additional variables apart from the population one, we further enrich the list to make sure that 5 largest cities of each union-level republic is included. This increases the size of the dataset to 328 cities. It is done to avoid too small grouping of cities and contrast capitals of Union-level republics with non-capital cities of the same republics. We use population of cities "including other urban dwellings answering to the city council" since it correlates with the number of mentions slightly better than the population of city proper. Note, however, that large discrepancy between population of city proper, and population including other urban dwellings is especially common for coal-mining towns. As a result, their observed underrepresentation (see Table II of the main text and auxiliary tables) might be partly due to this decision.

Unfortunately, due to varying standards of the national statistical bodies, there is no equivalent universal dataset for population of the cities worldwide (note, however, that huge discrepancies between population of metropolitan areas and cities proper is less common in 1950s-70s than in the modern period). In the absence of such a dataset we use, wherever possible, the 1970 estimate from the 2018 World Urbanization Prospects Report of the UN Population Division[4]. For the cities, for which such estimate is not available, we use data from national statistical bodies. In case there is no data for 1970, we approximate population linearly between two closest censuses before and after 1970.

These complications do make the population figures for foreign cities somewhat ambiguous. However, we found that for foreign cities population plays much smaller role in determining city mentions than in the case of Soviet cities. Indeed, if for Soviet cities, according to the geography model, a city from the most popular region (North-East) is mentioned similarly as a 4.8 times larger city in the least popular region (West Urals), for foreign cities a city in the most popular region (Austria and Finland) is equivalent to a city 30 times larger from the least popular one (third world). We therefore expect that minor ambiguities in the population variable for cities from different countries are not particularly relevant.

The list of cities of interest initially consists of 135 cities with population above 1 mln in 1970 and is further enriched to allow for the fact that capitals, cities in Europe and in socialist countries are mentioned more frequently. To do that, we include all capitals and all cities in Europe and in China with population above 0.5 mln, European capitals and cities in non-European socialist countries with population above 0.25 mln, and all cities in European socialist

countries, Austria and Finland with population above 0.1 mln. The resulting enriched dataset includes 310 cities. Of these only 113 are mentioned at least once, but recall that our approach allows to extract information from cities with zero mentions.

In order to roughly estimate the mentions of cities outside the aforementioned close lists we use slovnet[5], a Python library dedicated to analysing Russian language, to extract named entities from the story outlines. By analysing the output of slovnet we found that there are some places outside the cities of interest lists, which are mentioned extensively, including, for example, Tynda (the end point of Baikal-Amur railroad, an important construction project of 1970-80s), Mikhailovskoye (birthplace of Alexander Pushkin) and Zvezdny Gorodok (a place where Soviet astronauts were trained) inside the USSR, and Geneva (location of many important international negotiations) outside it.

Creating a full clean list of places mentioned in the dataset implies very significant manual work and is not needed from the point of view of the methodology presented here. The task is especially daunting in the case of places inside the USSR, in part because they are mentioned more, in part because of the large number of places with coinciding names, places named after prominent communist politicians, which are easy to confuse with mentions of those politicians themselves and other entities (streets, plants, collective farms) named after them, etc. That is why we only produced this analysis for foreign cities. The results are summarized in the table I. Thus, the cities of interest constitute more than 60% of all foreign places mentioned in the dataset, and contribute more than 90% to all mentions of foreign places.

Dataset	No cities	Cities with non-zero mentions	No mentions
Full	...	180	879
Above 1M	135	62	598
All cities of interest	310	113	792

Таблица I: Mentions of cities of interest as compared to mentions of all cities outside the USSR.

D. City mentions

For each city in the list we obtained and cleaned the corresponding list of mentions. In order not to miss any relevant mentions, for each city the story outlines were searched for matching substring(s) covering all possible Russian word forms derivative from the city name (these substrings where selected from Wictionary [**?**] and pymorphy library[7] and supplemented by the authors' knowledge of Russian grammar). For cities whose names names has changed during the Soviet period (Мариуполь/Жданов, Волгоград/Сталинград, Ленинград/Петроград, etc) all forms of the name were checked.

The resulting lists of matches where classified manually into relevant and irrelevant mentions. This stage is reasonably fast for the dataset of this size (roughly 2 weeks of work) but is not scalable for larger datasets like, e.g. full corpora of TV news or newspapers for a period of similar length. However, (i) this work must be way easier for analytical languages like English, Chinese or French, (ii) there is a strong evidence (see, e.g. [8]) that such tasks can now be automated with reasonably high precision using large language models.

One particular complication typical for the Soviet period is that in many cases multiple entities are named after the same prominent person, so that additional research is needed to disentangle them. One illustrative example is the difference between Gorky train line ("Горьковская железная дорога") and Gorky metro line ("Горьковская линия метро") in Moscow: both are ultimately named after the writer Maxim Gorky; however, the former is named after the city of Gorky (now Nizhny Novgorod) which is in turn named after the writer, while the latter is called after Gorky street in Moscow (which is named after the writer) and is unrelated to the city of Gorky.

We used the following classification of city mentions:

Type 1 - direct mention of the city as a location of filming or of city-dwellers;

Type 2 - mentions of entities (plants, universities, football teams, etc) located in the city and having city name or city-derivative adjective in their name (Moscow State University - Московский государственный университет, Dynamo Kyiv - Киевское Динамо, ...);

Type 3 - mentions of the area surrounding the city, which can take the form of mention of the city name with specification "рядом с"(near), "неподалеку от"(not far from), etc., administrative divisions (oblasts, etc) named after

their center city, as well as informal geolocation names like "Подмосковье"(Moscow region), "Рижское взморье"(Riga seacost), etc.;

Type 4 - mention of the objects and entities named after the city but not located in or near it, like Warsaw pact or Paris commune shoe factoryö

Type 5 - irrelevant: there is an automatic matching but it is a coincidence, due to random homonymy or similar origin of the name like in the Gorky example above.

Occasionally, an outline of a story mentions a single city multiple times. Such a multi-mention is counted as a single mention, and is assigned the type with the smallest number. For example, a phrase "В Варшавском аэропорту прошла торжественная встреча делегаций, прибывших в Варшаву на саммит стран Варшавского договора" (A ceremonial reception for the delegations arriving in Warsaw for the Warsaw Pact countries' summit took place at Warsaw Airport), which includes Type 1 mention (прибывших в Варшаву) includes type 1 mention (Warsaw per se), type 2 mention (Warsaw airport) and type 4 mention (Warsaw pact), and is counted as a single type 1 mention.

All mentions of the cities in the cities of interest list are manually classified into these 5 types, the resulting tables are publicly available at [9].

II. DETAILED RESULTS OF THE MODELS

Together with this supplementary text we provide two supplementary tables in the .xslx format, containing the detailed information on the run of all studied models for the Soviet cities [11] and for the foreign cities[12]. Below we give the detailed outline of the structure of these files and the information contained in them.

A. Soviet cities models

I. *Raw data on mentions and population.* Master table contains full information on the contemporary Cyrillic name(s) of the cities in the cities of interest list, their population at each of the three censuses, and the number of mentions of each city in the dataset.

II. *Results for the population-only model.* Pop_only_pval table contains the results of the population-only model, including comparison of actual mentions of each city with corresponding predicted mentions, and individual p-value of each city. Thus, 24 cities are over-mentioned with $p < 0.001$ and 6 cities are similarly undermentioned. Tallinn, Bratsk, Riga, Sevastopol, Yalta, Rustavi, Vilnius, Cherepovets, Minsk and Volzhsky form the top 10 of most significantly overmnentioned cities. Conversely, Ufa, Perm, Donetsk, Dnipro, Horlivka, Kemerovo, Kazan, Novokuznetsk, Barnaul and Baku are the top 10 most significantly undermentioned ones.

Notably, both in terms of population and in terms of mentions St. Petersburg is a significant outlier: it is 2.2 times larger than second largest city in the dataset (Kyiv), and accumulates 3.6 times more mentions than the second most mentioned one (also Kyiv). It is known that such outliers can significantly influence the results of the fitting. We found that, indeed, there are some minor but notable changes in the results of the model optimization over the whole dataset and over the same dataset but without St. Petersburg (see the three last columns of the Pop_only_pval table). First, the optimal value of the scaling exponent is slightly smaller $a = 1.24 \pm 0.05$ instead of $a = 1.33 \pm 0.04$ for the full dataset, which is borderline significant (note, however, that a is statistically significantly larger than 1 in both cases). Second, the ordering of the most over- and under-mentioned cities slightly changes. In particular, St. Petersburg and Volgograd replace Cherepovets and Volzhsky in the list of most overmentioned cities, with St. Petersburg becoming the most significantly overmentioned one. In turn, Dzerzhinsk and Chita replace Kazan and Baku in the list of the most undermentioned ones. This changes are, however, relatively minor (except when discussing St. Petersburg itself). Therefore, we decided to keep the whole sample. Note nevertheless that results for St. Petersburg should be interpreted with a certain caution. Moreover, we have checked that if omission of any other city from the dataset does not change the results in statistically significant way.

III. *Results of the geography, specialization and full models.* For each of these three models we provide four tables, specifying

(i) the list of variables used, including population, flags designating that a city belongs to a certain geographic group, and flags designating specializations present in the cities;

(ii) log of the optimization process: which merges of geographical regions (omissions of the specialization variables) where attempted in which particular order, what were the results of loss function optimization, and whether attempts where accepted or not;

(iii) table of the resulting values of parameters and their confidence interval in the final version of the model;

(iv) values of actual and predicted numbers of mentions for each city, and corresponding p-values.

Apart from that, we provide two summary tables, specifying

(i) the list of seed geographical regions, their definitions, and which macro-regions they are allocated to by the optimized geography model and by the optimized full model;

(ii) the list of specializations studied, and whether they are statistically significant.

IV. *Model comparison.* Finally, we provide a table with comparative summary of the models, which includes information on the number of outliers with p-values below 0.0001, 0.001, 0.01 and 0.05, as well as R^2 and normalized deviation $\langle\sigma\rangle$ defined as

$$\langle\sigma\rangle = \left(\frac{1}{K}\sum_{i=1}^{K}\frac{(N_i - M_i)^2}{M_i}\right)^{1/2} \tag{1}$$

where N_i is the number of mentions of i-th city, M_i is the corresponding expected number, and K is the total number of cities in the dataset. Note that for a set of Poisson random variables with expected values $\{M_i\}$ $\langle\sigma\rangle$ is expected to converge to 1. Thus, $\langle\sigma\rangle$ has the meaning of "how large are the observed deviations from expectations as compared to the situation when such deviations are due purely to random noise".

It can be see that on all metrics both geography and specialization models are a significant improvement on the population-only model, while full model is a significant improvement on them both. On balance, it can be argued that geography model explains the data slightly better than specialization one, however note that geography model has 16 relevant parameters (scaling exponent and expression levels in 15 regions), while specialization model has only 9 (scaling exponent, residual expression level, and boost factors for 7 relevant specializations). Meanwhile, it is striking that full model has a significantly larger explanatory power than the geography one despite having just 15 relevant parameters.

In terms of particular metrics, note that switch from population-only to full model allows to eliminate large outliers almost completely (from 19 to 3 cities with $p < 0.0001$) and to reduce the number of moderate outliers from 69 cities with $p < 0.05$ for the population-only model to 41 for the full model (note that in the dataset of $K = 328$ cities one expects roughly 16 such outliers for purely random reasons, so the number of excess outliers is reduced by a factor of 2). Other natural metrics, such as $1 - R^2$ and $< \sigma > -1$ tell the same story: the full model allows to explain 50%-60% of variation unexplained by the population-only model.

B. Foreign cities model

The table with the results of the foreign cities model has a similar structure. It contains

(i) the master list of the cities of interest with their population, and associated variables (flag indicating the city is a capital, population of the country, geographical location), all populations used are as of 1970, with UN Population Division 2018 World Urbanization Report being the main source of data, and national census authority data used in the cases a city is absent from it;

(ii) the list of seed geographical areas used, and their assumed proximity (i.e., for which areas merger was assumed possible); note that (i) contrary to the Soviet cities model proximity here is understood politically rather than geographically, i.e., socialist countries form a complete graph in terms of proximity, Australia and Canada are connected, etc.;

(iii) model optimization log (i.e., sequence of simplifications attempted and whether they where accepted or not);

(iv) model optimization result, with values of all parameters and corresponding confidence intervals;

(v) model expectation for individual cities vs actual numbers of mentions, and corresponding p-values.

Notably, the way the formula

$$\log m_{F,i} = \log c + a \left[\log P_i + \frac{\mathbb{I}_{i,cap}}{2} \log \frac{P_{c,i}}{P_i} \right] + \sum_\alpha \log k_\alpha \mathbb{I}_{i,\alpha} \tag{2}$$

for the expected number of mentions allows for a capital status of a city is itself a result of optimization. We start in face with a more general assumption

$$\log m_{F,i} = \log c + a \log P_i + s\mathbb{I}_{i,cap} + b\mathbb{I}_{i,cap} \log \frac{P_{c,i}}{P_i} + \sum_\alpha \log k_\alpha \mathbb{I}_{i,\alpha} \tag{3}$$

implying that there capital status of a city might give either a constant (via parameter s) or population-dependent (via parameter b) boost to representation. It turned out that the second mechanism is enough to describe the observed data, i.e., assumption $s \neq 0$ does not pass the significance test. Furthermore, it turns out that $b \approx a/2$ and the assumption $b \neq a/2$ does not pass the significance test either.

Partly due to the sparseness of the dataset, there is not a single city with $p < 0.0001$. There are 6 cities with $p < 0.001$, 5 of them are overmentioned, 1 is undermentioned, with clear individual reasons in all cases. The overmentioned cities are Accra (capital of the first decolonized African country), Hiroshima (nuked in 1945), Santiago (attention related to the pro-Socialist activities of the Allende government and the subsequent anti-Allende coup), New York (location of the UN) and Stockholm (Sweden's traditional neutrality, as opposed to the USSR-guaranteed post-WWII neutrality of Finland and Austria, puts it into intermediate place between those two and the rest of Western Europe). Conversely, Madrid – the capital of a heavily anti-communist Franco regime – is strongly undermentioned.

There are additionally two minor comments related to individual places.

Berlin. It is almost impossible to disentangle mentions of East and West Berlin. Indeed, (i) many mentions of Berlin in the dataset refer to the pre-World War II history, (ii) in many cases both sides of the divide are mentioned in a single story. For definiteness, we decided to use the population figure corresponding to the entirety of Berlin, and to treat it as capital of East Germany. We accept that this choice is imperfect but no better options seems available. However, readers should be aware that different choices will result in slight differences in the fitting results for East Germany.

Albania. Similarly, classification of Albania should be treated with caution: there is a single Albanian city (Tirana) in the dataset, and all its mentions happen before 1957, i.e., before Albania-Soviet split;

---

[1] https://www.net-film.ru/en/

[2] M. Oiva, K. Mukhina, V. Zemaityte, A. Karjus, M. Tamm et al., A framework for the analysis of historical newsreels. *Hum. Soc. Sci. Comm.*, **11**, 1-15 (2024).

[3] https://www.demoscope.ru/weekly/ssp/ussr59_reg2.php, https://www.demoscope.ru/weekly/ssp/rus59_reg2.php, https://www.demoscope.ru/weekly/ssp/ussr70_reg2.php, https://www.demoscope.ru/weekly/ssp/rus70_reg2.php, https://www.demoscope.ru/weekly/ssp/ussr79_reg2.php, https://www.demoscope.ru/weekly/ssp/rus79_reg2.php

[4] 2018 Revision of World Urbanization Prospects, UN Population Division, https://population.un.org/wup/

[5] https://github.com/natasha/slovnet#ner-1

[6] Русский викисловарь (Russian wiktionary), https://ru.wiktionary.org/wiki/

[7] https://pypi.org/project/pymorphy2/

[8] A. Karjus, Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence, arXiv:2309.14379 (2023).

[9] Here will be the link to a repository with the classification of city mentions.

[10] Link to the file Daily.news.outlines.by.story.csv to be added here.

[11] Link to the file USSR.citilist.4publication.xls to be added here.

[12] Link to the file Foreign.citilist.clean.xls to be added here.