

Supplementary Materials for

City representation in the Soviet propaganda: quantifying biases of the Soviet worldview

Mikhail V. Tamm, Mila Oiva, Ksenia D. Mukhina, Mark Mets, Maximilian Schich

9 Data preparation

Dataset characterization

12 The corpus of Soviet Newsreel “News of the Day” (Новости дня / Хроника наших
13 дней] was downloaded from Russian footage archive Net-Film[1] with permission of the
14 owners, it was previously introduced and discussed in Ref [2]. The “News of the Day”
15 journal was the main newsreel journal produced by the Central Studios of Documentary Film
16 in Moscow. The corpus includes almost all issues of this newsreel from 1954 to January 1992
17 (except for the year 1965), as well as a few surviving issues from 1944 to 1953. Figure S1
18 illustrates the contents of two exemplary newsreels.

In Figure S2a 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

23 from 1987 the annual number of newsreels dropped to 26 issues (1 issue in 2 weeks). Overall,
24 the corpus includes more than 1700 short films of usually 9-10 minutes length.

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

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

37

38 Stories and outlines

39 We use an outline of a story as an elementary unit of analysis. We mostly use the
40 outlines available in the complementary metadata. We made an extensive random check and
41 found that the outlines are of satisfactory quality, with a very small number of mistakes: the
42 fraction of outlines with typos in place names was significantly below 5%, and we only once
43 (out of several hundreds checked) been able to find a film outline in which one of the stories
44 was missing. In the vast majority of cases the format of outlines allowed automatic splitting
45 into stories. Exceptions where (i) around 1% of newsreels where there were typos in the
46 numbering of stories within a newsreel which we corrected manually, (ii) around 3% of

47 newsreels (most of them from years 1989-92) which had a different format of outlines:
48 instead of a contents summary it included description of camera movement, wide shots vs
49 close-ups, etc.; for these roughly 50 newsreels we have rewritten the outlines to match the
50 format of the rest.

51 Overall the dataset consists of 12 707 story outlines (on average 7.5 per newsreel), in
52 \fig{newsreel}b their distribution by year is presented, the full list of the outlines is provided
53 in [3]. It is seen that the huge majority of the dataset (97.5%) corresponds to 1954-1986, i.e.,
54 the period between the death of Stalin and the early years of perestroika. Interestingly, the
55 number of stories per newsreel issue trends down with time, especially after 1975. Median
56 date of a story is 1968 and 50% of stories belong to the period 1960-76 with 25% dated
57 before and 25% after this period.

58 The choice of outlines as a data source as opposed to using, the automatic transcripts
59 of the narrator's voice is due to their much higher quality: at the time when data preparation
60 was done the automatic transcript software for Russian language produced large number of
61 mistakes, especially in the names of persons and geographic locations, which is essential for
62 this work. This approach clearly has its limitations. For example, it excludes cases where the
63 place of filming is not explicitly mentioned, and it takes no account for the screen time
64 dedicated to a geographic location or to the related aspects of visual aesthetics[10]. Without
65 doubt, the progress in AI technologies will soon make it possible to go beyond these
66 limitations.

67

68 City population

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

71 In the case of USSR population is an extremely important variable (see Fig. 3A of the
72 main text), and USSR census is a relatively consistent and reliable dataset. As a proxy of the
73 population we use an average fraction of population of the USSR living in a given city
74 averaged over three censuses of 1959, 1970 and 1979 [4]. The list of cities of interest include
75 all 309 cities with population more than 0.03% of the population of the USSR, except
76 Moscow. For the purposes of models that include additional variables apart from the
77 population one, we further enrich the list to make sure that 5 largest cities of each union-level
78 republic is included. This increases the size of the dataset to 328 cities. It is done to avoid too
79 small grouping of cities and contrast capitals of Union-level republics with non-capital cities
80 of the same republics. We use population of cities ``including other urban dwellings
81 answering to the city council'' since it correlates with the number of mentions slightly better
82 than the population of city proper. Note, however, that large discrepancy between population
83 of city proper, and population including other urban dwellings is especially common for coal-
84 mining towns. As a result, their observed underrepresentation (see Table II of the main text
85 and auxiliary tables) might be partly due to this decision.

86 Unfortunately, due to varying standards of the national statistical bodies, there is no
87 equivalent universal dataset for population of the cities worldwide (note, however, that huge
88 discrepancies between population of metropolitan areas and cities proper is less common in
89 1950s-70s than in the modern period). In the absence of such a dataset we use, wherever
90 possible, the 1970 estimate from the 2018 World Urbanization Prospects Report of the UN
91 Population Division[5]. For the cities, for which such estimate is not available, we use data
92 from national statistical bodies. In case there is no data for 1970, we approximate population
93 linearly between two closest censuses before and after 1970.

94 These complications do make the population figures for foreign cities somewhat
95 ambiguous. However, we found that for foreign cities population plays much smaller role in

96 determining city mentions than in the case of Soviet cities. Indeed, if for Soviet cities,
97 according to the geography model, a city from the most popular region (North-East) is
98 mentioned similarly as a 4.8 times larger city in the least popular region (West Urals), for
99 foreign cities a city in the most popular region (Austria and Finland) is equivalent to a city 30
100 times larger from the least popular one (third world). We therefore expect that minor
101 ambiguities in the population variable for cities from different countries are not particularly
102 relevant.

103 The list of cities of interest initially consists of 135 cities with population above 1 mln
104 in 1970 and is further enriched to allow for the fact that capitals, cities in Europe and in
105 socialist countries are mentioned more frequently. To do that, we include all capitals and all
106 cities in Europe and in China with population above 0.5 mln, European capitals and cities in
107 non-European socialist countries with population above 0.25 mln, and all cities in European
108 socialist countries, Austria and Finland with population above 0.1 mln. The resulting
109 enriched dataset includes 310 cities. Of these only 113 are mentioned at least once, but recall
110 that our approach allows to extract information from cities with zero mentions.

111 In order to roughly estimate the mentions of cities outside the aforementioned close
112 lists we use slovnet[6], a Python library dedicated to analyzing Russian language, to extract
113 named entities from the story outlines. By analyzing the output of slovnet we found that there
114 are some places outside the cities of interest lists, which are mentioned extensively,
115 including, for example, Tynda (the end point of Baikal-Amur railroad, an important
116 construction project of 1970-80s), Mikhailovskoye (birthplace of Alexander Pushkin) and
117 Zvezdny Gorodok (a place where Soviet astronauts were trained) inside the USSR, and
118 Geneva (location of many important international negotiations) outside it.

119 Creating a full clean list of places mentioned in the dataset implies very significant
120 manual work and is not needed from the point of view of the methodology presented here.

121 The task is especially daunting in the case of places inside the USSR, in part because they are
122 mentioned more, in part because of the large number of places with coinciding names, places
123 named after prominent communist politicians, which are easy to confuse with mentions of
124 those politicians themselves and other entities (streets, plants, collective farms) named after
125 them, etc. That is why we only produced this analysis for foreign cities. The results are
126 summarized in table S1. Thus, the cities of interest constitute more than 60% of all foreign
127 places mentioned in the dataset and contribute more than 90% to all mentions of foreign
128 places.

129

130 City mentions

131 For each city in the list, we obtained and cleaned the corresponding list of mentions.
132 In order not to miss any relevant mentions, for each city the story outlines were searched for
133 matching substring(s) covering all possible Russian word forms derivative from the city name
134 (these substrings were selected from Wictionary [7] and pymorphy library[8] and
135 supplemented by the authors' knowledge of Russian grammar). For cities whose names
136 names has changed during the Soviet period (Mariupol/Zhdanov, Volgograd/Stalingrad,
137 Leningrad/Petrograd, etc) all forms of the name were checked.

138 The resulting lists of matches were classified manually into relevant and irrelevant
139 mentions. This stage is reasonably fast for the dataset of this size (roughly 2 weeks of work)
140 but is not scalable for larger datasets like, e.g. full corpora of TV news or newspapers for a
141 period of similar length. However, (i) this work must be way easier for analytical languages
142 like English, Chinese or French, (ii) there is strong evidence (see, e.g. [9]) that such tasks can
143 now be automated with reasonably high precision using large language models.

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

152 We used the following classification of city mentions:

153 Type 1 - direct mention of the city as a location of filming or of city-dwellers;
154 Type 2 - mentions of entities (plants, universities, football teams, etc) located in the city and
155 having city name or city-derivative adjective in their name (Moscow State University -
156 Московский государственный университет, Dynamo Kyiv - Киевское Динамо, ...);
157 Type 3 - mentions of the area surrounding the city, which can take the form of mention of the
158 city name with specification "рядом с" (near), "неподалеку от" (not far from), etc.,
159 administrative divisions (oblasts, etc) named after their center city, as well as informal
160 geolocation names like "Подмосковье" (Moscow region), "Рижское взморье" (Riga
161 seacost), etc.;
162 Type 4 - mention of the objects and entities named after the city but not located in or near it,
163 like Warsaw pact or Paris commune shoe factory;
164 Type 5 - irrelevant: there is an automatic match but it is a coincidence, due to random
165 homonymy or similar origin of the name like in the Gorky example above.

166 Occasionally, an outline of a story mentions a single city multiple times. Such a multi-
167 mention is counted as a single mention and is assigned the type with the smallest number. For

168 example, a phrase “В Варшавском аэропорту прошла торжественная встреча делегаций,
169 прибывших в Варшаву на саммит стран Варшавского договора” (A ceremonial reception
170 for the delegations arriving in Warsaw for the Warsaw Pact countries' summit took place at
171 Warsaw Airport), which includes Type 1 mention (прибывших в Варшаву) includes type 1
172 mention (Warsaw per se), type 2 mention (Warsaw airport) and type 4 mention (Warsaw
173 pact), and is counted as a single type 1 mention.

174 All mentions of the cities in the cities of interest list are manually classified into these 5
175 types. For consistency, all annotations used in the further analysis, are done by MT. To check
176 the reliability of human annotation two other Russian-speaking members of the team (MO
177 and KM) made test annotation of 407 story outlines related to 12 selected cities (Baku,
178 Izhevsk, Helsinki, Kaunas, Kursk, Lviv, Novgorod, Paris, Ryazan, Sofia, Tomsk, Tula)
179 according to the following instruction:

180 ***

181 Annotation instruction

182 For each story in the list separately

183 i) Find all mentions of the city and city-named entities in the text of the outline.

184 ii) if the city or city dwellers are mentioned directly, classify as 1 ["Москвичи

185 вышли на парад", "Новосибирск. Ловля лосося", "на шоссе Киев-Краснодар...", "матч
186 Динамо (Тбилиси)"]

187 iii) if not already classified, but there is an entity mentioned which is named after the
188 city and located in it, classify as 2 ["Горьковский автозавод", "Московский
189 кинофестиваль", "Бакинский ансамбль народных танцев"]

190 iv) if not already classified but there is a mention of the region centered in the city, or
191 the vicinity of the city, or of the entity named after the region, classify as 3 ["уборка свеклы

192 в колхозах Винницкой области", "соревнования под Красноярском", "Калининская
193 атомная электростанция в Удомле"]

194 v) if not already classified but there is a mention of the entity named after the city but
195 located elsewhere/nowhere, classify as 4 ["Казанский вокзал в Москве", "Фабрика имени
196 Парижской коммуны", "страны Варшавского договора"].

197 vi) else, if mention is simply homonymy or mistake, classify as 5.

198 If possible, try to figure out where the mentioned entities were located. If in doubt or
199 borderline classify explicitly agricultural entities ("Кишиневский экспериментальный
200 совхоз") as located in the vicinity of the city (i.e., classify as 3), and all the other (industrial,
201 cultural, etc) ones ("Сталинградская ГЭС") as located within a city (i.e., classify as 2).

202 When classifying, take into account, not only where the event is taking place but also
203 where the mentioned entity is located: "матч Динамо (Киев) в Тбилиси", "выступление
204 шахтера шахты X (Ленинск-Кузнецкий) на всесоюзной партийной конференции в
205 Москве", "На Ленинградский завод моторов закончено производство 218й турбины для
206 Красноярской ГЭС" are counted as mentions of Kyiv, Leninsk-Kuznetsky and Krasnoyarsk,
207 respectively (they are also counted, of course, as mentions of Tbilisi, Moscow and St
208 Petersburg).

209 ***

210 The full results of this annotation are available at [3]. Table S1 summarizes the most
211 important results, showing that both precision and recall of the annotation used (if the result
212 of the alternative annotators is considered a ground truth) is around 95%. A more detailed
213 analysis of discrepancies shows that they are mostly due to human error (more or less equally
214 distributed between annotators) and partly to different treatment of borderline cases. The

215 tables of mentions for these representative cities, marked-up by two annotators
216 independently, are available in the supplementary Annotation.Comparison.zip Archive

217

218

219 **Detailed results of the models**

220

221 Together with this supplementary text we provide two supplementary tables in the .xlsx
222 format, containing the detailed information on the run of all studied models for the Soviet and
223 foreign cities [3]. Below we give the detailed outline of the structure of these files and the
224 information contained in them. We also provide multiple comments on various aspects of the
225 results.

226

227 Soviet cities models

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

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

239 *The role of censoring.* We checked how different choices in the level of censoring the
240 cities by population influence the results of the model. The corresponding results are
241 provided in Table S3. Clearly, although including more cities reduces the confidence
242 intervals for the parameters, the confidence intervals strongly overlap for censoring at 0.03%,
243 0.05% and 0.1% of the population of the USSR.

244 *The influence of Moscow.* As mentioned in the main text, two properties of Moscow –
245 being the capital of the USSR and being the host city of the “Novosti Dnya” newsreel
246 production – make it incomparable to other cities of the USSR. As a result, Moscow is
247 mentioned roughly 5 times more than expected from population only model. Therefore, it is
248 not surprising that its inclusion shifts the scaling data dramatically (see Table S3): the loss
249 function puts a lot of weight on fitting this one big outlier to the detriment of the fitting the
250 rest of the data. On the other hand, the only imperfect comparison available is to the capitals
251 of the foreign cities, where we found that capital effects can be estimated by replacing the
252 city population by the geometric mean of the populations of the city and the corresponding
253 country. If this renormalized population is used for Moscow, it turns out that it is in fact
254 undermentioned by a factor of roughly 2 as compared to the prediction of the population-only
255 model, which might indicate that capital effect work differently here and/or that significant
256 fraction of stories are located in Moscow by default without explicit mention in the outlines.
257 In any case, Moscow is a completely unique case and we exclude it from further
258 consideration.

259 *The influence of St. Petersburg on the fit.* After Moscow is excluded, St. Petersburg is
260 the second significant outlier both in terms of population and in terms of mentions: it is 2.2
261 times larger than second largest city in the dataset (Kyiv). Since it does not have the unique
262 properties of Moscow, we keep it in the dataset, but check how much this single point
263 influences the results of the fitting. We found that, indeed, there are some minor but notable

264 changes in the results of the model optimization over the whole dataset and over the same
265 dataset but without St. Petersburg (see the three last columns of the Pop_only_pval table).
266 First, the optimal value of the scaling exponent is slightly smaller $\alpha = 1.24 \pm 0.05$ instead
267 of $\alpha = 1.33 \pm 0.04$ for the full dataset, which is borderline significant (see Table S3).
268 Second, the ordering of the most over- and under-mentioned cities slightly changes. In
269 particular, St. Petersburg and Volgograd replace Cherepovets and Volzhsky in the list of most
270 overmentioned cities, with St. Petersburg becoming the most significantly overmentioned
271 one. In turn, Dzerzhinsk and Chita replace Kazan and Baku in the list of the most
272 undermentioned ones. These changes are, however, relatively minor (except when discussing
273 St. Petersburg itself). Therefore, we decided to keep the whole sample. Note nevertheless that
274 results for St. Petersburg should be interpreted with a certain caution. Moreover, we have
275 checked that if omission of any other city from the dataset does not change the results in a
276 statistically significant way.

277 *Time evolution of the population-only model.* The data we study spans several decades of
278 Soviet history. It is natural to ask how much the observed patterns of mentioning cities
279 change throughout this period. Our ability to study this is somewhat limited due to the
280 sparseness of the data. However, we provide here the results of the population-only model
281 run on the data from three eleven-year periods: 1954-64, 1966-76 and 1977-87 (recall that
282 1965 is missing from the dataset, and more than 97% correspond to the 1954-86 interval). We
283 use the population data from the 1959, 1970 and 1979 censuses, respectively, as a measure of
284 city population, and use 0.05% population cut-off for the first two periods and 0.06% for the
285 third, so that there are no cities above the cut-off which are not included in our 328-city
286 dataset. Note that using the whole dataset without cut-off would have been methodologically
287 wrong. For example, cities, which are small in the earliest period but subsequently become
288 large enough to be included in the dataset do not form a representative sample of small cities.

289 The scatters plots of mentions versus population for each period are presented in Figure S3
290 and the parameters of corresponding models are summarized in Table S3. There are several
291 important observations to be made. First, the overall number of mentions systematically
292 decreases with time in agreement with the decreasing number of stories per year (compare
293 Figure S2B). Second, for each period separately the number of mentions does scale with
294 population size as predicted by the population model. In all cases the scaling exponents are
295 above one with high confidence, indicating the presence of agglomeration effects. However,
296 the scaling exponent trends down with time, i.e. in later period the distribution of mentions
297 becomes less skewed towards larger cities. Third, on a single-city level there exist multiple
298 different scenarios. Some of the most “popular” cities, e.g., Sevastopol and Tallinn, are
299 overmentioned throughout each period separately. Mentions of some others, e.g.,
300 Krasnoyarsk, Qaragandy, Vladimir, are more localized in time (in case of Krasnoyarsk this is
301 clearly connected to the construction of Krasnoyarsk hydroelectric dam). Fourth, the most
302 dramatic change between the first period and the later two is related to the status of Kyiv.
303 Indeed, in 1954-64 Kyiv is clearly the third most important city in the USSR hierarchy: it is
304 mentioned significantly more than population-based expectation and has almost double the
305 number of mentions of the fourth-most-mentioned city (which, interestingly, is Odesa, i.e.,
306 another Ukrainian city). Conversely, both in the 1966-76 and in the 1977-87 periods Kyiv is
307 mentioned less than expected based on its population, and, despite remaining the third largest
308 city in the USSR, is mentioned less than some smaller cities. The mentions of Odesa drop
309 even more dramatically. One possible explanation for this change might be related to
310 importance of Ukraine and Kyiv. Interestingly, this change coincides to a well-known shift
311 from promotion of Ukraine as second-most-important republic of the USSR during N.
312 Khruschev era to comparative neglect and insidious Russification in the later period [11,12].

313 III. *Results of geography, specialization and full models.* For each of these three models
314 we provide four tables, specifying
315 (i) the list of variables used, including population, flags designating that a city belongs to a
316 certain geographic group, and flags designating specializations present in the cities factors
317 (sheets Geo_variables, Spec_variables and Full_variables).
318 (ii) log of the optimization process: which merges of geographical regions (omissions of the
319 specialization variables) where attempted in which particular order, what were the results of
320 loss function optimization, and whether attempts where accepted or not (sheets
321 Geo_clustering_log, Spec_clustering_log and Full_clustering_log);
322 (iii) table of the resulting values of parameters and their confidence interval in the final
323 version of the model, including the lists of optimized geographical regions, their composition,
324 and corresponding boost factors (sheets Geo_confidence, Spec_confidence and
325 Full_confidence);
326 (iv) values of actual and predicted numbers of mentions for each city, and corresponding p-
327 values (sheets Geo_expectations, Spec_expectations and Full_expectations).

328 Apart from that, we provide two summary tables, specifying
329 (i) the list of seed geographical regions, their definitions, and which macro-regions they are
330 allocated to by the optimized geography model and by the optimized full model (sheet
331 Seed_regions);
332 (ii) the list of specializations studied, and whether they are statistically significant (sheet
333 Specializations).

334 *Seed specializations and choice between them.* The initial list of specializations is provided in
335 table S5 for a quick reference. Generally, we start with feeding into the model a wide set of
336 variables, compatible with several alternative hypotheses, and then let the optimization

337 evolve and choose one option out of many. Below we discuss three particular instances of
338 this approach.

339 *Sub-republican autonomies and administrative units.* We start with distinguishing 4 classes
340 of cities: capitals of autonomous republics inside and outside Russia proper, capitals of non-
341 national sub-republican units (oblasts and krai's) and cities located inside autonomous
342 republics but having no capital status. The model optimization process algorithm attempts to
343 both (i) merge these classes of cities in different combination and (ii) discard them (i.e.,
344 essentially merge the classes with a “dummy class” of cities with no administrative function
345 and located outside autonomies). In this case the optimization resulted in discarding all
346 classes except for capitals of autonomous republics inside Russia proper, which turned out to
347 statistically significantly reduce the representation. Note, however, that the size of the
348 “capitals of autonomous republics outside Russia proper” class is very small (just 3 cities:
349 Batumi, Sukhumi, and Nukus), making the inference in this case somewhat less reliable.

350 *Ports and recreation cities.* We start with 6 classes for port cities located ashore of various
351 masses of water (Arctic and Pacific oceans, Azov, Baltic, Black and Caspian seas). We also
352 introduce a class of cities specializing in recreation in order to check the hypothesis that
353 predominantly recreational cities (e.g., Sochi, Yalta, Jurmala) are represented differently than
354 predominantly military or trade ports (e.g., Sevastopol, Novorossiysk, Kaliningrad). Once
355 again, in the model optimization state the city classes corresponding to the shores of different
356 seas might be either merged or discarded, and the “recreation” class can be either preserved
357 (meaning that there is statistically significant difference between recreational and non-
358 recreational cities) or discarded. The result of optimization in this case is a bit unexpected: it
359 turns out that there are two significantly different classes of seas: overrepresented Black,
360 Baltic and Pacific on one side, and not overrepresented Arctic, Azov and Caspian on the
361 other. This difference might possibly be rationalized by noting that the first set of seas is

362 relatively more “outward looking” (that is, related to international transportation,
363 international relations and corresponding history) than the second. Moreover, there is no
364 significant difference between recreational seaside cities and military/trade ports.

365 *Hydroelectricity*. We started with two classes of cities with hydroelectric power dams, one
366 corresponding to huge dams with power above 2 GW and medium-sized dams of 0.5-2 GW.
367 It turned out that only huge dams lead to a statistically significant increase in representation.
368 Notably, all 4 dams in question were built during the studied period, and it is mostly the
369 construction stage that is being covered in the newsreels. However, the same is true for most
370 of the medium-sized dams and leads, in their case, to no observed representation effect.

371 *Validation of specializations*. The industrial specializations, which we found relevant,
372 particularly hydroelectricity and metallurgy, is well-known and reflected in Soviet culture on
373 multiple levels, from heroic Komsomol songs to E. Evtushenko’s flagship epic poem
374 “Bratskaya ges” (“The Bratsk Station”) to sarcastic mentions in the openly anti-Soviet
375 sources, like in this famous song by Y. Aleshkovsky:

376 И пусть в тайге придётся сдохнуть мне,
377 Я верю: будет чугуна и стали
378 На душу населения вполне.

379 (“And even if I have to kick off in taiga, I believe: there will be enough cast iron and steel per
380 capita”).

381 However, it is interesting to check that this is not a coincidence and the specializations which
382 the model finds to be relevant are indeed represented in the data. We made a direct check of
383 the topics of stories mentioning 7 representative cities (Odesa, Krasnoyarsk, Tbilisi,
384 Cherepovets, Kazan, Oskemen and Donetsk) and counted the stories directly related to the
385 relevant city specializations, the results are provided in Table S6. In most cases (e.g.,
386 mentions of Cherepovets steelworks) the counting is very straightforward, except for the

387 number of “capital status” related stories are an estimate from below, as we counted only the
388 most unquestionable ones (stories directly mentioning Georgia in case of Tbilisi and
389 Tatarstan instead of Kazan). The table shows that indeed the specialization-related stories
390 contribute quite significantly to mentions of corresponding cities. Moreover, the fraction of
391 such stories is seemingly higher for the representation-boosting specializations.

392 *Representation of regions in the full model.* We find (see Figure 4D and Table 2 of the main
393 text) six contingent regions, in which representation significantly differs from that in the rest
394 of the country. Importantly, there are more deviations down than up from this default (“rest
395 of the country”) level, i.e. this level itself is slightly (about 20%) elevated above the average
396 over the whole of USSR. We relate overexpression of Moscow region to its geographical
397 accessibility and Northern Kazakhstan to its importance in the virginlands reclaiming
398 narrative. The slightly elevated “rest of the country“ region can be split into several groups of
399 locations with different rationale for importance. We relate interest in Eastern Siberia and Far
400 East with the exoticity of those places and narrative of expansion into wild lands („stroyki
401 kommunizma“), in the South (Northern Caucasus, Lower Volga, Georgia and Blacksoil
402 region) with its better climate and recreational attractiveness. Western part of the USSR
403 (Baltic coast, Belarus, Moldova and Western Ukraine) seems to be of importance because of
404 general Eurocentric bias of the Soviet worldview. This bias simultaneously explains the
405 systematic neglect of the South-Eastern republics of the USSR (Central Asia, Armenia,
406 Azerbaijan and Kazakhstan, except for its Russian-speaking North). Central parts of Russia
407 proper (West Urals, and, to a lesser extent, Middle Volga, East Urals and West Siberia) seem
408 to suffer from what we call “neglect of intermediate situations”: these parts are quite far away
409 from Moscow, but still not virgin and exotic enough to warrant additional interest.

410 *Ukraine.* The most puzzling and interesting phenomenon is a very significant
411 underrepresentation of the region covering most of Central and Eastern Ukraine, as well as

412 Rostov region of Russia. Without doubt, the study of the role of this region in the Soviet
413 worldview and its development in time (note the drastic fall in mentions of Kyiv and Odesa
414 with time, see Fig. S3) is of extreme interest and importance, especially in the view of recent
415 Russian aggression against Ukraine. Here we formulate a hypothesis about possible
416 explanation. We conjecture that underrepresentation of Eastern Ukraine and Russia-Ukraine
417 borderlands might be another manifestation of the “neglect of intermediate situations”
418 pattern. The population of these regions was mixed, and identities of its residents formed a
419 continuum spectrum from purely Russian to purely Ukrainian, including people speaking
420 Russian but self-identifying as Ukrainian, bilinguals, speakers of Russian-Ukrainian pidgin
421 language (“Surzhyk”), etc. This complexity resulted in Eastern Ukraine (except, maybe,
422 purely Ukrainian-speaking Western part) to fall in-between of the standard Soviet
423 nomenclature of nationalities. In turn, Central and Southern Ukraine (i.e., most notably Kyiv
424 and Odesa) should have been seen as more properly-Ukrainian in the 1950s and 1960s, where
425 it was [11,12] ideologically fashionable to celebrate Ukraine-ness as something distinct
426 (although inseparably united with Russia), and as more in-between (i.e., similar to Eastern
427 Ukraine) in 1970s and 1980s, the age of tacit Russification of Ukraine.
428 That is to say, we suggest that for Soviet ideologues might have felt that Eastern (and later
429 also Central and Southern) Ukraine are perplexingly “neither fully Eastern European nor fully
430 Russian” and, as such, better left without discussion.

431 IV. *Model comparison*. Finally, we provide a table with comparative summary of the
432 models, which includes information on the number of outliers with p -values below 0.0001,
433 0.001, 0.01 and 0.05, as well as R^2 and normalized deviation $\langle\sigma\rangle$ defined as

$$434 \quad \langle\sigma\rangle = \left(\frac{1}{K} \sum_i \frac{(n_i - m_i)^2}{m_i} \right)^{1/2}$$

435 (S1)

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

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

450 In terms of particular metrics, note that switch from population-only to full model allows
451 to eliminate large outliers almost completely (from 19 to 3 cities with $p < 0.0001$) and to
452 reduce the number of moderate outliers from 69 cities with $p < 0.05$ for the population-only
453 model to 41 for the full model (note that in the dataset of $K=328$ cities one expects roughly
454 16 such outliers for purely random reasons, so the number of excess outliers is reduced by a
455 factor of 2). Other natural metrics, such as $(1 - R^2)$ and $(\langle \sigma \rangle - 1)$ tell the same story: the
456 full model allows to explain 50%-60% of variation unexplained by the population-only
457 model.

458

459 Foreign cities model

460 The table with the results of the foreign cities model has a similar structure. It contains
461 (i) the master list of the cities of interest with their population, and associated variables (flag
462 indicating the city is a capital, population of the country, geographical location), all
463 populations used are as of 1970, with UN Population Division 2018 World Urbanization
464 Report being the main source of data, and national census authority data used in the cases a
465 city is absent from it;
466 (ii) the list of seed geographical areas used, and their assumed proximity (i.e., for which areas
467 merger was assumed possible); note that (i) contrary to the Soviet cities model proximity here
468 is understood politically rather than geographically, i.e., socialist countries form a complete
469 graph in terms of proximity, Australia and Canada are connected, etc.;
470 (iii) model optimization log (i.e., sequence of simplifications attempted and whether they
471 were accepted or not);
472 (iv) model optimization result, with values of all parameters and corresponding confidence
473 intervals;
474 (v) model expectation for individual cities vs actual numbers of mentions, and corresponding
475 p-values.

476 *Seed geographical areas.* The choice of initial geographical areas, as well as area-
477 dependent censoring of city population is data-driven. The seed areas include, separately, all
478 13 countries recognized as “socialist” in contemporary Soviet discourse (both Comcon and
479 non-Comcon); Finland and Austria, whose high representation has been observed in the data;
480 and USA, Canada, Japan and Australia, for which we hypothesized that their representation
481 might be different from neighbouring countries. The rest of the world was split on continental
482 level into Africa, Asia, Europe and Latin America.

483 *Capital status.* The way the formula

$$484 \quad \log m_{F,i} = \log c + a \left(\log P_i + \frac{1}{2} I_{i,cap} \log \frac{P_{i,c}}{P_i} \right) + \sum_{\alpha} I_{i,\alpha} \log k_{\alpha} \quad (S2)$$

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

487 $\log m_{F,i} = \log c + a \log P_i \log P_i + b I_{i,cap} \log \frac{P_{i,c}}{P_i} + s I_{i,cap} + \sum_{\alpha} I_{i,\alpha} \log k_{\alpha}$ (S3)

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

493 *Outliers.* Partly due to the sparseness of the dataset, there is not a single city with $p <$
494 0.0001. There are 6 cities with $p < 0.001$, 5 of them are overmentioned, 1 is
495 undermentioned, with clear individual reasons in all cases. The overmentioned cities are
496 Accra (capital of the first decolonized Sub-Saharan African country and, as such, the focal
497 point of the anticolonial movement in the late 1950s – early 1960s), Hiroshima (nuked in
498 1945), Santiago (attention related to the pro-Socialist activities of the Allende government
499 and the subsequent anti-Allende coup), New York (location of the UN) and Stockholm
500 (Sweden's traditional neutrality, as opposed to the USSR-guaranteed post-WWII neutrality of
501 Finland and Austria, puts it into intermediate place between those two and the rest of Western
502 Europe). Conversely, Madrid – the capital of a heavily anti-communist Franco regime – is
503 strongly undermentioned.

504 *In-country city hierarchy.* In most cases the dataset is too sparse to probe the
505 representation of the city hierarchy inside countries, except for the most over-represented
506 ones. We summarize the data for non-capital cities of the 4 countries in the Capitalist I and
507 Socialist I groups (Mongolia and Albania had no non-capital cities above 0.1 mln in 1970) in
508 Table S8. It is notable that the model prediction of how mentions are split between the capital
509 and other big cities seems to be consistently good.

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

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

520 *Mongolia*. Similarly to Albania, there is a single Mongolian city (Ulaanbataar) in the
521 dataset, unlike Tirana, the mentions of Ulaanbataar are evenly distributed through the dataset.
522 Mongolia is notable as the only non-European country which is mentioned on par with the
523 most mentioned European ones. It might be explained by a combination of the ideological
524 conformity of the Mongolian regime, its close proximity to the Soviet Union and competition
525 with China for the influence in Mongolia. However, given the sparseness of the dataset this is
526 a relatively low-confidence result which needs further confirmation.

527

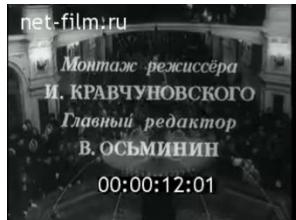
528 **References**

- 529 1. <https://www.net-film.ru/en/>
- 530 2. M. Oiva, K. Mukhina, V. Zemaityte, A. Karjus, M. Tamm, T. Ohm, M. Mets, D.
531 Chavez Heras, M. Canet Sola, H.H. Juht, M. Schich, A framework for the analysis of
532 historical newsreels. *Hum. Soc. Sci. Comm.* **11**, 1-15 (2024).
- 533 3. Supplementary tables for this paper are available at
https://github.com/thummm/soviet_newsreels/
- 534 4. City population according to the 1959 USSR census available at
https://www.demoscope.ru/weekly/ssp/ussr59_reg2.php,
https://www.demoscope.ru/weekly/ssp/rus59_reg2.php; according to the 1970
535 USSR census available at https://www.demoscope.ru/weekly/ssp/ussr70_reg2.php,
536 https://www.demoscope.ru/weekly/ssp/rus70_reg2.php; according to the 1979 USSR
537 census available at https://www.demoscope.ru/weekly/ssp/ussr79_reg2.php,
538 https://www.demoscope.ru/weekly/ssp/rus79_reg2.php (in Russian).
- 539 5. 2018 Revision of World Urbanization Prospects, UN Population Division,
<https://population.un.org/wup/>
- 540 6. <https://github.com/natasha/slovnet#ner-1>
- 541 7. Русский викисловарь (Russian wiktionary), <https://ru.wiktionary.org/wiki/>
- 542 8. <https://pypi.org/project/pymorphy2/>
- 543 9. A. Karjus, Machine-assisted mixed methods: augmenting humanities and social
544 sciences with artificial intelligence, *Humanities and Social Sciences
545 Communications*, 12, 277 (2025).
- 546 10. M. Oiva, T. Ohm, K. Mukhina, M. Canet Sola, M. Schich, Soviet View of the
547 World. Exploring Long-Term Visual Patterns in “Novosti dnia” Newsreel Journal
548 (1945-1992), *Journal of Cultural Analytics*, 9, 4 (2024).
- 549
- 550
- 551
- 552

559

560 Figures

561



Novosti dnya 24/1954

1. The first session of the Supreme Soviet of the USSR of the fourth convocation in the Kremlin.

2. Tractor columns are going along the winter roads of Kazakhstan.

3. Boiler and fan plant in Tula. The inventor, turner Chekalin, works with a new cutter.



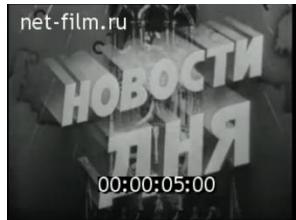
4. The librarian of a mountain village in the Sayan Mountains delivers books to tafalar reindeer breeders.

5. Jewelry trade in the village of Krasnoe, Kostroma region, jewelers at work.

6. New kindergarten in Tbilisi.

7. Construction of a funicular in the city of Chongqing in Southwest China, the townspeople travel in the funicular.

562

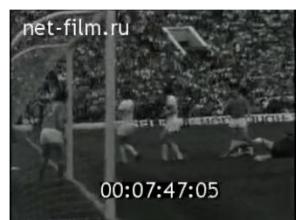
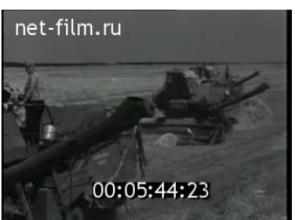


Novosti dnya 30 / 1970

1. A view of oil rigs on Lake Samotlor in the Tyumen region. Oil workers are working at a well.

2. The electricians from Volgoelectrostroy in Gorky are raising the power line support.

3. Production processes at the Vladimir Ilyich Moscow Electromechanical Plant.



4. Moscow city. Speech by Deputy Chairman of the Council of Ministers of the USSR Z. N. Nuriev at the X International Congress of Soil Scientists at the Rossiya Hotel.

5. Agricultural processes in the collective farm "Lenin's Way" in the Peschanokopsky district of the Rostov region.

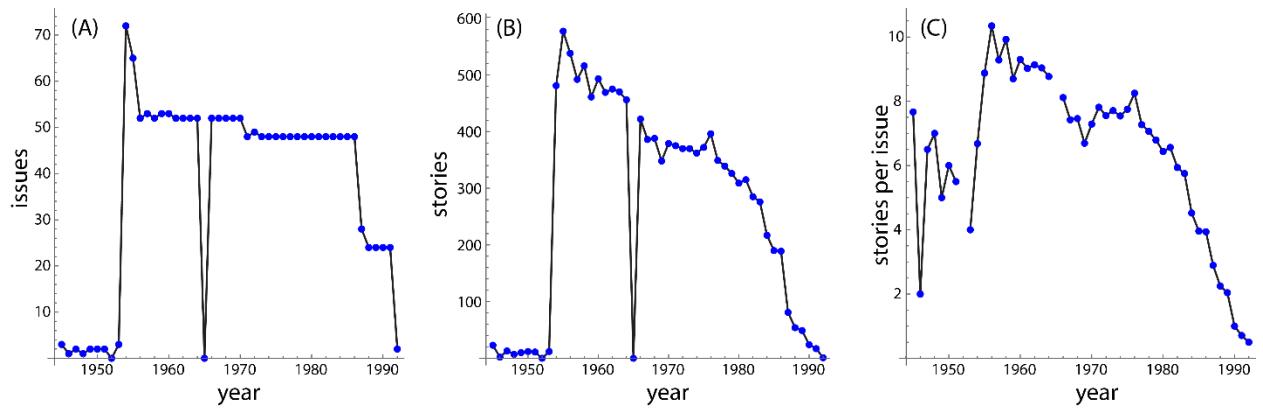
6. Builders work in a subway mine. A meeting of builders at the opening of the Belyaevsky station of the Moscow Metro.

7. Moments of the final game of the USSR Cup on football between the teams "Dynamo" Kiev and "Zarya" Voroshilovgrad.

563

564 **Fig. S1.**

565 Snapshots from two exemplary newsreels, issue 24 of 1954 (top two rows) and issue 30 of
566 1970 (bottom two rows), with one snapshot per story. Snapshots are accompanied with English
567 translations of the corresponding outlines, mentions of the cities are given in bold.



568

569

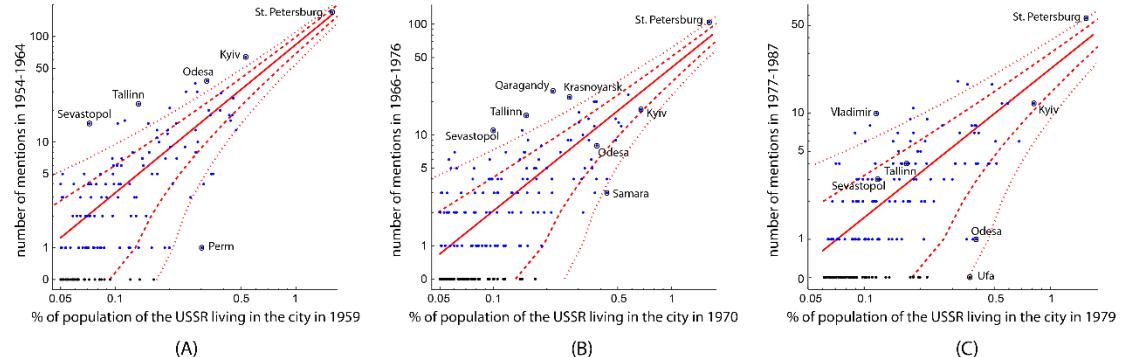
570 **Fig. S2.**

571 Temporal structure of the newsreel corpus used (A) number of issues in the dataset per year,

572 (B) number of stories per year, (C) mean number of stories per issue.

573

574



575

576

577 **Fig. S3.**

578 Scatter plots of the number of mentions vs population for the cities in the USSR for 3 periods
 579 of equal length: (A) mentions in 1954-64 vs population as of 1959 census; (B) mentions in
 580 1966-76 vs population as of 1970 census; (C) mentions in 1977-87 vs population as of 1979
 581 census. Red lines are best power-law fits with characteristics summarized in Table S3, dashed
 582 and dotted lines correspond to confidence intervals with $p = 0.05$ and $p = 0.001$,
 583 respectively. Cities with zero mentions (black dots) are shown out of scale. Selected cities are
 584 outlined, see discussion in the text.

585 *(Note for the referees: we provide this picture in higher resolution in a separate
 586 supplementary Figure.Time.pdf file).*

587

588

589 **Tables**

590

| Main classification, done by MT, used for further analysis | Test alternative classification, done by KM and MO | |
|--|--|------------------------|
| | Relevant (types 1, 2) | Irrelevant (types 3-5) |
| Relevant (types 1, 2) | 227 | 9 |
| Irrelevant (types 3-5) | 12 | 159 |

591

592 **Table S1.**

593 Results of the classification consistency test.

594

595

| Dataset | Number of cities | Cities with non-zero mentions | Number of mentions |
|------------------------|------------------|-------------------------------|--------------------|
| Full | ... | 180 | 879 |
| Above 1 mln | 135 | 62 | 598 |
| All cities of interest | 310 | 113 | 792 |

596

597 **Table S2.**

598 Mentions of cities of interest as compared to mentions of all cities outside the USSR.

599

600

| Cut-off | Number of cities | a | c |
|---|------------------|-------------------|-----------------|
| > 0.03% | 308 | 1.33 ± 0.04 | 1.34 ± 0.13 |
| > 0.03% + additionally at least 5 cities per republic | 328 | 1.33 ± 0.04 | 1.35 ± 0.13 |
| > 0.05% | 188 | 1.32 ± 0.05 | 1.38 ± 0.15 |
| > 0.1% | 81 | 1.37 ± 0.07 | 1.19 ± 0.22 |
| > 0.03% + Moscow | 309 | 1.82 ± 0.03 | 0.56 ± 0.06 |
| > 0.03% + Moscow with renormalized population | 309 | 1.175 ± 0.015 | 1.75 ± 0.10 |
| > 0.03% - St. Petersburg | 307 | 1.24 ± 0.05 | 1.53 ± 0.13 |

601

602 **Table S3.**

603 Parameters of the population-only model as function of the level of censoring cut-off.

604 Influence of Moscow and St. Petersburg is also shown. a is the scaling exponent, c is the

605 expected number of mentions for a city with 0.03% of population of the USSR

606

607

| Period | Cut-off | Number of cities | a | c |
|---------|---------|------------------|-----------------|-----------------|
| 1954-64 | 0.05% | 151 | 1.41 ± 0.07 | 0.61 ± 0.10 |
| 1966-76 | 0.05% | 194 | 1.29 ± 0.08 | 0.44 ± 0.08 |
| 1977-87 | 0.06% | 176 | 1.18 ± 0.11 | 0.36 ± 0.08 |

608

609 **Table S4.**

610 Parameters of the population-only model fitted separately for three periods of equal length:

611 1954-64, 1966-76 and 1977-87.

612

613

| Specialization | No of cities | Comments | Outcome |
|---|--------------|--|---|
| Capitals of Union level-republics | 14 | | Relevant, increases representation |
| Capitals of national autonomous republics inside Russia | 17 | | Relevant, decreases representation |
| Capitals of national autonomies outside Russia | 3 | | Irrelevant |
| Other cities located within national autonomies | 10 | | Irrelevant |
| Capitals of non-national regions (oblast or krai) | 112 | As of 1970 | Irrelevant |
| Port on the Black Sea | 12 | Location near the sea regardless of specialization (military, commerce, recreational, etc) | Relevant, increases representation, joined with Baltic and Pacific |
| Port on the Baltic Sea | 8 | See above | Relevant, increases representation, joined with Black Sea and Pacific |
| Port on the Pacific Ocean | 4 | See above | Relevant, increases representation, joined with Black Sea and Baltic |
| Port on the Azov Sea | 5 | See above | Irrelevant |
| Port on the Caspian Sea | 5 | See above | Irrelevant |
| Port on the Arctic coast | 4 | See above | Irrelevant |
| Resort city | 11 | Specialization in recreation regardless of seaside or inland location | Irrelevant |
| Hydroelectricity, big | 5 | Plants of >2 GW | Relevant, increases representation |
| Hydroelectricity, medium | 10 | Plants of 0.5-2 GW | Irrelevant |
| Steelworks | 18 | Full cycle only | Relevant, increases representation |
| Non-ferrous metallurgy | 18 | | Relevant, increases representation |
| Automobile plant | 14 | | Irrelevant |
| Coal mining | 38 | | Relevant, decreases representation |
| Newsreel-producing film studio | 26 | | Irrelevant |

614 **Table S5.**

615 Initial set of specializations used in the specialization model, with optimization outcomes for
616 each of them. In order to avoid overfitting, only those 7 specializations, which are found to be
617 relevant in the specialization model, are used in the full (specialization plus geolocation)
618 model.

619

| City name | Contemporary city name, Cirilllic | Specialization | Mentions | Mentions related to specialization | Fraction related to specialization |
|-------------|-----------------------------------|-----------------------------|----------|------------------------------------|------------------------------------|
| Odesa | Одесса | Seaside | 48 | 35 | 73% |
| Krasnoyarsk | Красноярск | Hydroelectricity | 47 | 25 | 53% |
| Tbilisi | Тбилиси | Republic capital | 38 | 10 | 26% |
| Cherepovets | Череповец | Steelwork | 18 | 18 | 100% |
| Kazan | Казань | Autonomous republic capital | 17 | 4 | 24% |
| Oskemen | Усть-Каменогорск | Non-ferrous metallurgy | 11 | 8 | 73% |
| Donetsk | Донецк/Сталино | Coal | 12 | 4 | 33% |

620

621

622 **Table S6.**

623 Fractions of stories related to relevant city specialization for several selected cities.

624

625

| Model | Population | Geolocation | Specialization | Full |
|------------------------------|------------|-------------|----------------|-------|
| No of cities | 308 | 328 | 328 | 328 |
| No of relevant parameters | 2 | 16 | 9 | 15 |
| R^2 | 0.905 | 0.952 | 0.945 | 0.964 |
| $\langle \sigma \rangle - 1$ | 3.08 | 1.89 | 2.12 | 1.53 |
| Cities with $p < 0.0001$ | 19 | 5 | 8 | 3 |
| Cities with $p < 0.0001$ | 30 | 14 | 13 | 7 |
| Cities with $p < 0.05$ | 69 | 58 | 52 | 41 |

626

627 **Table S7.**

628 Goodness of fit characteristics of the models for the cities in the USSR: coefficient of
 629 determination R^2 , excess variation as compared to the Poisson distribution $\langle \sigma \rangle - 1$ (formula
 630 S1) and number of cities with significant deviations from the prediction of the model.

631

632

633

| Country | Cities | Mentions | Expected mentions |
|----------------|--|----------|-------------------|
| Austria | Vienna | 43 | 48.1 |
| | Graz, Linz, Salzburg, Innsbruck | 4 | 4.7 |
| Bulgaria | Sofia | 29 | 30.4 |
| | Plovdiv, Varna, Burgas, Ruse, Stara Zagora | 2 | 4.8 |
| Czechoslovakia | Prague | 51 | 47.2 |
| | Brno, Ostrava, Bratislava, Plzen, Kosice | 10 | 7.8 |
| Finland | Helsinki | 26 | 16.7 |
| | Turku, Tampere | 1 | 2.0 |

634

635 **Table S8.**

636 Mentions of non-capital cities in the four most over-represented countries compared to the
637 model prediction and to the expectation for similar-sized cities in the Capitalist II (“over
638 Europe”) country group.

639