Online Appendix

Every table and graph in this paper, the R code that generated them, and the Latex source code of this manuscript are available in the GitHub repository of this project at `https://github.com/martin-kosiik/Geopolitics-of-Repressions-updated`.

A Historical Background

A.1 German–Soviet Relations

The relations between Weimar Germany and Soviet Union can be characterized as neutral or even cooperative. Both countries were somewhat isolated in the international system dominated by the Western powers (Great Britain, France, USA) and sought to find allies. The good relations were first established by the Treaty of Rappalo in 1922 in which both countries renounced the territorial and financial claims against each other and agreed to secret military cooperation (Gatzke, 1958) and then reaffirmed by the Treaty of Berlin in 1926. Furthermore, a trade treaty was signed between the two countries in 1925 (Morgan, 1963).

Hitler was named chancellor on 30 January 1933 and effectively become a dictator on 24 March 1933 by the passing of the Enabling Act which gave him the power to enact laws without approval of the parliament. The relations with Soviet Union quickly turned hostile for several reasons. First, Hitler called in *Main Kampf* for Germany to obtain *Lebensraum* (living space) in the east, presumably at the expense of the Soviet Union and he often spoke of Judeo-Bolsheviks (Haslam, 1984, p. 6). Moreover, Hitler's anti-communist was one of factors contributing to his political success as he presented himself as the only leader strong enough to prevent a Communist revolution in Germany. This was not only empty rhetoric as Hitler soon after his rise to power banned the German Communist Party and started to persecute its members (Evans, 2004, chapter 5). Hostility also manifested itself in the German-Soviet relations as the military cooperation between the two countries was canceled in August 1939 and trade treaties were not extended.

The opposition to fascism led to change in policy of the Communist International (Comintern) with appointment Georgi Dimitrov as its general secretary in 1934. The Communist parties in democratic countries were now encouraged to form coalitions (Popular Fronts) with social democratic parties to prevent rise of fascism, in contrast to the previous aggressive and uncompromising approach. This policy was affirmed by the Seventh World Congress of the Comintern in 1935 (Haslam, 1979).

The newly formed Popular Front coalitions won elections and entered government in some European countries including France and Spain. In Spain however, the coup of nationalists against the new government in 1936 sparked a civil war. The Soviet Union heavily supported the republican government, while Germany supplied the nationalists which further increased the tensions between the two countries. As a response, Japan and Germany signed the Anti-Comintern Pact in 1936 in which they committed to co-operate for defense against communistic disintegration. Meanwhile in the Soviet Union, many

people were persecuted for alleged cooperation with Germany including leading general Mikhail Tukhachevsky.

The orientation of German foreign policy began to shift in spring of 1939. Until that point, Hitler hoped that he could ally with Poland in a war against the Soviet Union or that Poland would at least allow the passing of German troops (Weinberg, 2010, chapter 26). But Poland repeatedly refused the German offers for closer relations such as to join the Anti-Comintern Pact and thus Hitler changed the strategy and in April 1939 ordered the German army to began planing for the invasion of Poland (Kotkin, 2017, p. 621). However, France and Great Britain granted security guarantees to Poland in March 1939 to deter German aggression. Hitler thus tried to negotiate neutrality of the Soviet Union to avoid simultaneously facing Western powers, Poland and the Soviet Union in war. Soviet neutrality was potentially beneficial for Stalin too. A long and costly war would weaken both the capitalist and the fascist enemies of the Soviet Union. Moreover, Stalin believed that conditions of war could bring about socialist revolutions in those countries just as in Russia in 1917. After brief negotiations, on 23 August 1939 the Molotov-Ribbentrop pact was signed between Germany and the USSR which guaranteed non-belligerence between the two countries. In addition, a secret protocol of the treaty marked the German and Soviet spheres of influence in Eastern Europe.

The pact of the two former ideological enemies caused great shock and astonishment both among Party officials and ordinary people. Victor Kravchenko (1947, p. 332), a Soviet official who later defected to the US, described in his memoir the disbelief upon hearing about the pact

> There must be some mistake, I thought, and everyone around me seemed equally incredulous. After all, hatred of Nazism had been drummed into our minds year after year. The big treason trials [...] have rested on assumption that Nazi Germany and its Axis friends [...] were preparing to attack us.

Another party official later recalled that "it left us all stunned, bewildered, and groggy with disbelief" (Robinson and Slevin, 1988, p. 137).

Nazi Germany attacked Poland on 1 September 1939 from the west and shortly after that, on 17 September, the Red Army invaded the eastern part of the country. As was agreed in the pact, Poland was partitioned between Germany and the Soviet Union. However, the mistrust between the two countries was still present as evidenced by a violent clash of German and Soviet troops near Lwów on 20 September (Kotkin, 2017, p. 685)

Hitler enjoyed major success in the first years of the war. By summer 1940, German forces defeated French army and annexed Denmark and Norway. But German industry

was severely lacking raw materials needed in war effort against Britain and the US which, according to some historians, motivated Hitler to invade the resource-rich USSR (Tooze, 2008). The German attack on the Soviet Union on 22 June 1941 ended 2 years of fragile cooperation. Although Stalin received numerous warnings by his intelligence about the impeding German attack, he was generally dismissive of them as British efforts to embroil him in war with Germany (Kotkin, 2017, chapter 14).

The Eastern Front became the bloodiest theater of World War II with more than 10 million soldiers killed in combat and another 3.3 million of Soviet prisoners of war starved to death by Germans (Snyder, 2011, p. 155). Moreover, the Eastern Front was site of the worst atrocities committed on the civilian population, most notably the Holocaust.

After the surrender of Germany in May 1945 its territory was partitioned into 4 occupation zones (American, British, French, and Soviet). Various industrial disarmament programs were put in place in all occupation zones to limit and control the German military capacity. Thus, in the post-war period militarily weak Germany no longer presented a geopolitical threat as it did before. Instead, the rivalry of the Soviet Union and the United States became the new main source of tensions in the international relations.

To summarize, there were several events in the period from 1921 to 1960 that fundamentally altered the Soviet-German relations. First, Hitler's rise to power, which was definitely consolidated by the passing of The Enabling Act on 23 March 1933, brought in heightened hostilities and tensions into the Soviet-German relations. Another turning point was the Molotov–Ribbentrop Pact signed 23 August 1939 which started a brief period of limited cooperation between the two countries. On 22 June 1941, the German invasion of the Soviet Union officially terminated the pact marking the beginning of one of the most bloody conflicts of World War II. The war finally ended on 8 May 1945 with unconditional surrender of Germany.

A.2 Soviet Political Repressions

The Soviet Union had large and powerful coercive apparatus. The Soviet secret police (which was throughout the years named the Cheka, OGPU, NKVD, MVD and the KGB)[10] employed at its height (1937–1938) 270,730 persons (Gregory, 2009, p. 2). The political repressions were usually carried under Article 58 of the Criminal Code. The Article 58 punished counter-revolutionary activities which included treason, espionage, counterrevolutionary propaganda, agitation and failure to report any of these crimes. In practice, this broad definition meant that anyone regarded as politically inconvenient could be arrested and prosecuted.

[10]We will refer to the Soviet secret police as the NKVD in this text since this was the name of the agency for the largest part of the period of our interest

During the mass operations, the central office of the NKVD would typically set quotas for the number of arrests which the regional branches were supposed to reach and exceed (Gregory, 2009, chapter 6). The local NKVD officer had to decide themselves who to target to meet the quotas.

The sentences were in most cases issued extrajudicially by so-called "troikas", three-person committees composed of a regional NKVD chief, a regional party leader, and a regional prosecutor. The NKVD chief usually dominated the process as party leaders sometimes feared that they themselves would be targeted (Snyder, 2011, p. 82). Only rarely was a person acquitted from his charge. The most common sentences for political crimes in the Stalinist period were execution and prison term in a labor camp (Gulag) (Gregory, 2009, p. 21). A term in the Gulag of less then 5 years was considered lighter sentence in these cases.

With the rise in repressions in the 1930s, the Gulag system significantly expanded. At its height, it consisted of at least 476 distinct camp complexes each containing hundreds of prisoners. The Gulag system offered the Soviet state cheap source of labor that produced substantial amount the country's coal, timber, and gold supply. The mortality of prisoners was high due to heavy work, malnutrition, and cold climate (Applebaum, 2003).

The death of Stalin in 1953 marked a start of decline in political repressions in the USSR. The new Soviet leader, Nikita Khrushchev, denounced Stalin and the mass repressions of his period in his speech *On the Cult of Personality and Its Consequences* in 1956. The suppression of dissent continued in the Khrushchev and Brezhnev era but in much milder form. Khrushchev gradually dismantled the Gulag system, granted amnesty to many political prisoners and started the process of rehabilitation of victims of the Stalinist period although they were limited to only some categories of victims and offences (Applebaum, 2003; Dobson, 2009).

A.3 Ethnic Minorities in the USSR

The Soviet Union was from its inception a multi-ethnic state. According to the 1926 Census, the Russians made up only half of the total population.[11] Among other large ethnic group were Ukrainians, Belorussians and Kazakhs. A significant fraction of citizens of the USSR belonged to ethnic groups with their own independent states including Polish, German, Estonian, Latvian, Lithuanian, Finish, and Greek minorities. The Bolshevik elites were aware of the multi-ethnic nature of their newly formed state and wanted to avoid a perception of the Soviet Union as a project of Russian imperialism. Furthermore,

[11] Full data on population of the USSR by ethnicity from the 1926 Census is available at `http://www.demoscope.ru/weekly/ssp/ussr_nac_26.php`. Population numbers for only the 38 ethnic groups featured in our dataset is provided in table 10 in the appendix.

the Bolsheviks hoped that they could exert political influence in countries with cross-border ethnic ties to Soviet diaspora nationalities by promoting the interests of minorities in the USSR. [12]

As a consequence, the Soviet policy towards its ethnic minorities in the 1920s was largely accommodating (Martin, 2001). The languages and culture of minorities were promoted and minorities were encouraged to enter local governments and party structures (so-called *korenizatsiya* policy). Some minority groups were well represented even in the NKVD (Gregory, 2009, p. 25). In some cases Autonomous Soviet Socialist Republics (ASSR) were established (including Volga German ASSR) which had given the regional minorities certain degree of independence.

This attitude changed drastically in the 1930s. First, the *korenizatsiya* policy started to be reversed in the 1932. From 1934, the NKVD started to deport ethnic minorities from the state frontier zone in Eastern Europe. This involved forced resettlement of 30 000 of Ingermanland Finns and tens of thousands of Poles and Germans to Kazakhstan and West Siberia (Polian, 2003, p. 95). In 1937 and 1938, the NKVD conducted mass operations specifically targeted at minorities with cross-border ethnic ties. Poles, Latvians, Germans, Estonians, Finns, Greeks, Chinese, and Romanians were arrested in large numbers as supposed spies and saboteurs of foreign governments. More than 320 000 people were arrested in the national operation out of which about 250 000 were executed (Martin, 1998, p. 855).

The persecutions further escalated with the World War II. Following the German invasion into the Soviet Union in 1941, Stalin ordered deportation of about 430 000 Soviet Germans (most of them living in Volga German ASSR) into Kazakhstan and Siberia (Polian, 2003, p. 134). Similar "preventive" deportation followed for Finns and Greeks as well. Between 1943-1944, forced resettlement of another six ethnic groups (Karachais, Kalmyks, Chechens, Ingushetians, Balkars, and Crimean Tatars) were carried out for alleged or actual cooperation of some of these minorities with the German troops (even if many more served in the Red Army).

[12] Martin (2001) refers to this argument as the *Piedmont Principle*.

B Imputation of Missing Data

B.1 Inferring Ethnicity from Names

In this section, we explain our method for predicting ethnicity of an individual from his or her names. Using names for imputing ethnicity has several advantages. First, full name is available for every individuals in the dataset. Second, names have been shown to be highly predictive of ethnicity in a variety of applications (Mateos, 2007; Hofstra et al., 2017; Hofstra and Schipper, 2018).

Given the high number of predictors, we need a model that is not computationally demanding but at the same time achieves reasonable level of prediction accuracy. Naive Bayes classifier meets these criteria and has been for this reason used in wide range of applications including text classification (Gentzkow et al., 2019).

B.1.1 Naive Bayes Classifier

Let $\boldsymbol{x} = (x_1, x_2, x_3)$ be features used for predicting ethnicity, that is person's first, last, and patronymic (given after father's first name) names. Using Bayes theorem, we can express the probability that particular observation belongs to ethnic group E_k given its features as

$$p(E_k \mid \mathbf{x}) = \frac{p(E_k)\ p(\mathbf{x} \mid E_k)}{p(\mathbf{x})}, \tag{B.1}$$

in other words, the posterior probability is proportional to the product of prior probability and likelihood. Assuming conditional independence of features allows us to substitute $p(\mathbf{x} \mid E_k)$ such that we get

$$p(E_k \mid \mathbf{x}) = \frac{p(E_k)\ \prod_{i=1}^{3} p(x_i \mid E_k)}{p(\mathbf{x})}. \tag{B.2}$$

All terms in this equation now can be estimated from the data: the prior probability $p(E_k)$ as a proportion of E_k in the data, $p(x_i \mid E_k)$ as a proportion of people with name x_i in the ethnic group E_k and $p(\mathbf{x})$ simply calculated such that the sum of $p(E_k \mid \mathbf{x})$ for all k is one. The Naive Bayes classifier then chooses the ethnicity with the highest posterior probability as its prediction, that is

$$\hat{y} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}}\ p(E_k) \prod_{i=1}^{3} p(x_i \mid E_k). \tag{B.3}$$

One potential issue is that whenever a likelihood of a certain feature is estimated to be zero then the posterior probability is always zero regardless of the prior or the likelihoods of other features. For example, suppose that a person has a typical German

first name but a rare surname which does not appear in the training set at all. Then the useful information contained in the first name will be completely ignored since the zero likelihood of the surname will override any other value and we will end up with the posterior probability of zero for all ethnic groups.

To address this problem, we apply Laplace smoothing. For every ethnicity, let c_j be number of people with a name j and N be total number of member of that ethnic group in the data. Without applying any smoothing, we would estimate the likelihood $p(x_i \mid E_k)$ simply as a relative frequency, i.e. $\hat{\theta}_j = \frac{c_j}{N}$. With Laplace smoothing, we estimate the likelihood $\hat{\theta}_j$ as

$$\hat{\theta}_j = \frac{c_j + \alpha}{N + \alpha d} \qquad j = 1, \ldots, d \tag{B.4}$$

where parameter $\alpha > 0$ is a smoothing parameter. This ensures that for any finite value of N, $\hat{\theta}_j$ will never be exactly zero. In our model, relatively small value of $\alpha = 0.005$ turned out to be sufficient and was chosen.

It is important to note that the conditional independence assumption often does not hold in the data and the estimated posterior probabilities therefore have to be taken with a grain of salt. However, our main goal is the best out-of-sample accuracy of the model's predictions. In this respect, Naive Bayes classifier have been shown to perform well in many applications, despite its often violated assumptions (Domingos and Pazzani, 1997).

B.1.2 Adjusting for Unbalanced Prediction Accuracy

To reliably asses the out-of-sample performance of our model, we used 10-fold cross-validation on the data with non-missing ethnicity. That is, the data is first randomly split into 10 groups. A model is fitted to 9 group and the remaining group is used to test the model's performance. This process is then repeated 9 times until every group has been tested. Using this method, the resulting overall accuracy of our model is 79.3%. However, we are also interested in how this varies by ethnicity. For this reason we calculate sensitivity and specificity for each ethnic group.[13] The results, provided in table 3 in the appendix, show that the sensitivity differs significantly by ethnicity. Some ethnic groups with distinctive names such as Chinese or Japanese are classified with accuracy higher than 90% while for other ethnicities such as Chuvash or Udmurt it is about 10%. This severe imbalance in sensitivity and specificity across ethnic groups could potentially cause bias in the imputations.

We develop adjustments that try to correct for these biases in the model's predictions. Let P_{it} be the number of people with predicted ethnicity i arrested at time t, R_{it} be the

[13]Sensitivity measures the proportion of observations in the class (in our case ethnicity) that are correctly identified by the model as such (i.e. number of true positives divided by all positives). Specificity measures the proportion of observations *not* in the class that are correctly identified as such (i.e. number of true negatives divided by all negatives).

actual number of people with ethnicity i arrested at time t, α_i and β_i be sensitivity and specificity of our classifier for ethnic group i and N_t be the total number of arrests at time t. Then the predicted arrests of a given ethnicity are sum of true positives and false positives, that is

$$P_{it} = \alpha_i R_{it} + (N_t - R_{it}) \cdot (1 - \beta_i)\,. \tag{B.5}$$

We are interested in R_{it} but we only directly observe P_{it} and N_t. However using simple algebra, R_{it} can be expressed as

$$R_{it} = \frac{P_{it} - N_t\,(1 - \beta_i)}{\alpha_i + \beta_i - 1}. \tag{B.6}$$

We will refer to this method of correcting predictions as parsimonious adjustment. The parameters α_i and β_i are not known to us but we can use their estimates from the cross-validation on the training data. This assumes that the these parameters do not differ significantly for the training and test data. But this might not be the case. Suppose, for example, that Armenians are often misclassified as Chechens and that the number of Armenians in the data with missing ethnicity is disproportionately higher than in the data with information on ethnicity. Then the cross-validated specificity for Chechens in the training set will underestimate the specificity in the test set because it does not take into account the higher proportion of Armenians.

Fortunately, we can address this potential bias with a more complex model. First for all ethnic groups i and j, we define the misclassification rate b_{ij} as share of people with ethnicity j that are classified as i. Notice that for $i = j$, the misclassification rate is simply prediction accuracy for ethnicity i. It follows from the definition of the terms that predicted number of arrests for ethnic group i at time t, P_{it}, is equal to

$$P_{it} = \sum_{j=1}^{K} b_{ij} R_{jt} \qquad i = 1, \ldots, K. \tag{B.7}$$

This equation can be expressed in matrix form as

$$\mathbf{P}_t = \mathbf{B} \cdot \mathbf{R}_t, \tag{B.8}$$

where $\mathbf{P}_t = (P_{1t}, \cdots, P_{Kt})$, $\mathbf{R}_t = (R_{1t}, \cdots, R_{Kt})$, and $\mathbf{B} = (b_{ij})_{i=1,\ldots,K,\, j=1,\ldots,K}$. To express $\mathbf{R}_t$, we just apply basic linear algebra

$$\mathbf{R}_t = \mathbf{B}^{-1} \cdot \mathbf{P}_t. \tag{B.9}$$

We will call this method the full (confusion) matrix adjustment. Compared to the parsi-

monious adjustment (in equation B.6), this correction no longer assumes that the test set sensitivity and specificity be accurately estimated from the training set. The full matrix adjustment makes only somewhat weaker assumption that the train and test set misclassification rates are not substantially different. On the other hand, the estimates of misclassification rate will likely be noisier (have higher variance) compared to the estimates of specificity and sensitivity because they are based on fewer observations.

One final issue that we encountered when applying these adjustments to the actual data was that some predicted values of $\mathbf{R}_t$ were negative. We decided to replace all negative values with zero in order to preserve this basic feature of the data. Finally, we scaled all values such that the total number of arrests would stay unchanged after the adjustment and rounded it to the nearest integer. The comparison of total number of arrests for each adjustment and ethnic group is provided in table 2. A graph showing the change in arrests in time by ethnicity and imputation adjustment is provided in figure 2.

B.2 Imputing Missing Date of Arrest

Our strategy for imputing the missing arrest dates is to predict it from the date of trial. For this reason, we model the number of days between arrest and trial and fit it to a subset of the data for which both dates are known. It is reasonable to expect that the average number of day from arrest to trial could vary considerably throughout the years. Hence we use the year of trial as a predictor for our model.

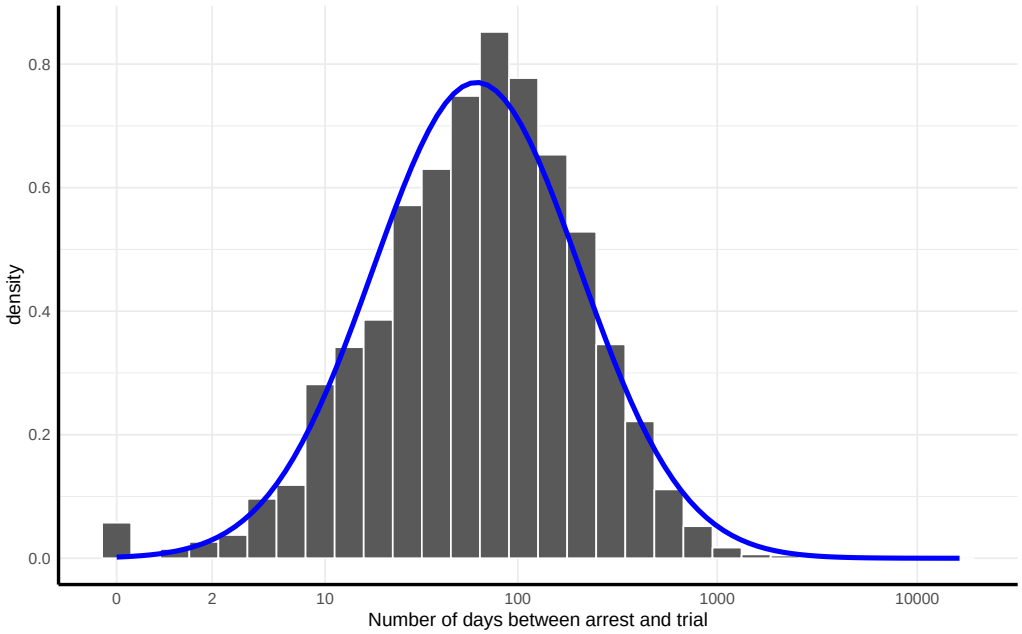
We begin by examining the data with both dates available. The histogram for number of days between arrest and trial (on scale of $\log_{10}(1+x)$) is shown in figure A1. First, we can see that there is fairly large variance in the variable with number of days ranging from 0 to more than 1000.[14] Second, the transformed data seems to be following the normal distribution except for the density at 0 which is much higher than the normal model would predict. Moreover, the zero values are making the estimated mean of the normal distribution lower than would be appropriate for the positive values resulting in poor fit as can be seen in figure A1.

To address this problem, we model the zero and positive values separately in a two-stage process using a method described in Gelman and Hill (2006, p. 537-538). Let y be the number of days between arrest of an individual and his or her trial and X be a set of dummy variables indicating the year of trial. We also define I^y as an indicator variable that equals 1 if $y > 0$ and 0 otherwise and y^{pos} to be only positive values of y (i.e. $y^{\text{pos}} = y$ if $y > 0$). In the first stage, we predict I^y using logistic regression

$$\Pr\left(I_i^y = 1\right) = \text{logit}^{-1}\left(X_i\alpha\right). \tag{B.10}$$

[14] Zero, of course, corresponds to both arrest and trial being in the same day.

Figure A1: Histogram for Number of Days between Arrest and Trial



Notes: The x-axis is shown on $\log_{10}(1+x)$ scale.

In the second stage, simple log-linear regression is applied to predict only the positive values y^{pos}

$$\log\left(y_i^{\text{pos}}\right) \sim \mathrm{N}\left(X_i\beta, \sigma\right). \tag{B.11}$$

We then fit the first model to the data where the exact dates of both arrest and trial are available and the second model to the subset of the same data for which $y > 0$. The results of both of these models are provided in table 6 in the appendix. The years of trial appear to be important predictors both in the first stage and even more in the second stage. However, the unexplained variance is still high making up about 77% of the total variability in the dependent variable in the second model.

We proceed to apply the fitted models to the missing data to get the predicted probability of y being positive and the mean value of y if it is positive. For each observation with missing date of arrest X_i, we then randomly draw from the Bernoulli distribution with $\text{logit}^{-1}\left(X_i\hat{\alpha}\right)$ as its parameter to obtain $\hat{I}_i^y$. We also draw from the normal distribution with mean $X_i\hat{\beta}$ and exponentiate the result to get $\hat{y}_i^{\text{pos}}$. Finally, the predicted number of days is calculated simply as $\hat{y}_i = \hat{I}_i^y \cdot \hat{y}_i^{\text{pos}}$.

The histogram of the imputed values is provided in figure 4 in the appendix. The resulting distribution highly resembles the distribution in figure A1 including the fraction of zero values indicating our model captures the actual data fairly well.

Nevertheless, another complicating factor is that for significant number of observations we do not have the exact date of trial but only year. In particular, while the year of trial

is recorded for all 903 455 observations where the date of arrest is to be imputed, the month of trial is missing for 369 393 of them and the day for 390 174. To fill in the missing month, we take a random sample from all months with probability equal to the relative frequency of the months of trial in the non-missing data for the years from 1921 to 1960. Even simpler method is used to impute the missing days where we just randomly choose a day within given month with uniform probability.[15] The imputed months and days of the trials are therefore only weakly informed guesses, nevertheless they enable us to carry on with the analysis.

The final step is to calculate the imputed date of arrest by subtracting the predicted number of days from the date of process (i.e. we go back in time by given number of days). Since we conduct the analysis with annual observations, we ignore predicted month and day of arrests keep only information on year. The number of arrests for each ethnicity by year (including the imputed years) is then counted for the period from 1921 to 1960 which forms our final dataset.

The resulting time series of all arrests with imputed years is plotted in figure 5 in the appendix. Arrests with imputed dates seem to follow similar trends with the labeled data although there is slight divergence at the beginning and end of the series and in the 1930s.

International Relations Controls

The summary of major changes in international relations of the Soviet Union with states that have significant minorities in the USSR that we use as control variable in our default difference-in-differences specification (equation 4.1) are provided in table 1. In particular, we created a separate dummy variable for every combination of state and phase of geopolitical relations. The blank cells indicate that no special dummy variable covers that years (i.e. there was no significant change in relations with the given country in that year). For example in case of Hungarian ethnicity, we created one dummy variable for the period of war with the USSR (from 1941 to 1944) and another one covering the whole post-1945 period. We briefly describe the relevant history below to explain why we choose such classification. For more detailed information, consult Weinberg (2005) or other general overview of World War II.

In case of Japan, the Soviet Union and Japan engaged in minor border clashes near Mongolia from 1935. However, these skirmishes escalated into large scale conflict in 1938 with Battle of Lake Khasan. The war ended in 1939 with a decisive Soviet victory at Battles of Khalkhin Gol. This defeat deterred Japan from further conflict with the Soviet

[15] Every date, however, has to be consistent with the calendar. This means that for January we take a sample of numbers from 1 to 31, for February from 1 to 28 and so on.

Union (Haslam, 1992). The two countries remained in peace until August 1945 when the USSR invaded Manchuria.

The Soviet Union invaded Finland in November 1939. This conflict ended (which became know as the Winter War) in March 1940. This peace did not last long since Finland joined the German invasion into the USSR in June 1941. Hungary was another country that allied with Germany in war against the Soviet Union.

Poland was attacked by both Germany and Soviet Union in September 1939. By the end of the month the Polish army was defeated and the Polish territory was partition between Germany and Soviet Union along the line that was agreed in the Molotov-Ribbentrop pact. This changed with German invasion in 1941 when the German army gained the control of the whole Poland. Poland stayed under German occupation for 4 years and most of its territory was liberated by the Red Army by January 1945.

Some cases are difficult to classify. For instance, China was in embroiled in a civil war from 1927 to 1949. Although the Soviet Union sometimes supported certain Chinese warlords, it is hard to identify some major changes in the relations with the Soviet Union and hence China does not appear on this list. Greece was occupied by Italy and Germany from 1941 to 1944 but the Soviet Union was not directly involved in Greece and thus Greece also do not feature on the list.

Nonetheless, this table provides only very coarse classification of changes in geopolitical relations and that is why we perform additional checks such as excluding all ethnic groups with independent states from analysis.

References

Applebaum, Anne (2003), *Gulag: A History*, New York: Doubleday.

Dobson, Miriam (2009), *Khrushchev's Cold Summer: Gulag Returnees, Crime, and the Fate of Reform after Stalin*, Ithaca, New York: Cornell University Press.

Domingos, Pedro and Pazzani, Michael (1997), "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss", *Machine Learning* 29 (2), pp. 103–130, DOI: 10.1023/A:1007413511361.

Evans, Richard J. (2004), *The Coming of the Third Reich*, New York: Penguin Press.

Gatzke, Hans W. (1958), "Russo-German Military Collaboration During the Weimar Republic", *The American Historical Review* 63 (3), pp. 565–597, DOI: 10.2307/1848881.

Gelman, Andrew and Hill, Jennifer (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge, UK: Cambridge University Press.

Gentzkow, Matthew, Kelly, Bryan T. and Taddy, Matt (2019), "Text as Data", *Journal of Economic Literature* (forthcoming), DOI: 10.1257/jel.20181020.

Gregory, Paul R. (2009), *Terror by Quota: State Security from Lenin to Stalin*, New Haven: Yale University Press.

Haslam, Jonathan (1979), "The Comintern and the Origins of the Popular Front 1934-1935", *The Historical Journal* 22 (3), pp. 673–691.

Haslam, Jonathan (1984), *Soviet Union and the Struggle for Collective Security in Europe 1933-39*, London: Palgrave.

Haslam, Jonathan (1992), *The Soviet Union and the Threat from the East, 1933-41*, Basingstoke: Palgrave Macmillan.

Hofstra, Bas, Corten, Rense, Tubergen, Frank van and Ellison, Nicole B. (2017), "Sources of Segregation in Social Networks: A Novel Approach Using Facebook", *American Sociological Review* 82 (3), pp. 625–656, DOI: 10.1177/0003122417705656.

Hofstra, Bas and Schipper, Niek C de (2018), "Predicting ethnicity with first names in online social media networks", *Big Data & Society* 5 (1), DOI: 10.1177/2053951718761141.

Kotkin, Stephen (2017), *Stalin: Waiting for Hitler, 1929-1941*, London: Penguin Press.

Kravchenko, Victor (1947), *I Chose Freedom: The Personal and Political Life of a Soviet Official*, London: Robert Hale Limited.

Martin, Terry (1998), "The Origins of Soviet Ethnic Cleansing", *The Journal of Modern History* 70 (4), pp. 813–861, DOI: 10.1086/235168.

Martin, Terry (2001), *The Affirmative Action Empire: Nations and Nationalism in the Soviet Union, 1923-1939*, Ithaca, New York: Cornell University Press.

Mateos, Pablo (2007), "A review of name-based ethnicity classification methods and their potential in population studies", *Population, Space and Place* 13 (4), pp. 243–263.

Morgan, R. P. (1963), "The Political Significance of German-Soviet Trade Negotiations, 1922-5", *The Historical Journal* 6 (2), pp. 253–271.

Polian, Pavel (2003), *Against Their Will: The History and Geography of Forced Migrations in the USSR*, Budapest: Central European University Press.

Robinson, Robert and Slevin, Jonathan (1988), *Black on Red: My 44 Years Inside the Soviet Union*, Washington, D.C: Acropolis Books.

Snyder, Timothy (2011), *Bloodlands: Europe between Hitler and Stalin*, London: Vintage.

Tooze, Adam (2008), *The Wages of Destruction: The Making and Breaking of the Nazi Economy*, New York: Penguin Books.

Weinberg, Gerhard L. (2005), *A World at Arms: A Global History of World War II*, 2 edition, Cambridge, UK: Cambridge University Press.

Weinberg, Gerhard L. (2010), *Hitler's Foreign Policy 1933-1939: The Road to World War II*, New York: Enigma Books.

Table 1: Major Changes in Relations with the USSR

	State				
Year	Baltic states	Finland	Japan	Hungary	Poland
1921					
1922					
1923					
1924					
1925					
1926					
1927					
1928					
1929					
1930					
1931					
1932					
1933					
1934					
1935					
1936					
1937			War		
1938			War		
1939		War	Neutrality		War
1940	Annexation	War	Neutrality		Soviet occupation
1941	Nazi occupation	War	Neutrality	War	Nazi occupation
1942	Nazi occupation	War	Neutrality	War	Nazi occupation
1943	Nazi occupation	War	Neutrality	War	Nazi occupation
1944	Post-war	War	Neutrality	War	Nazi occupation
1945	Post-war	Post-war	War	Post-war	Post-war
1946	Post-war	Post-war	Post-war	Post-war	Post-war
1947	Post-war	Post-war	Post-war	Post-war	Post-war
1948	Post-war	Post-war	Post-war	Post-war	Post-war
1949	Post-war	Post-war	Post-war	Post-war	Post-war
1950	Post-war	Post-war	Post-war	Post-war	Post-war
1951	Post-war	Post-war	Post-war	Post-war	Post-war
1952	Post-war	Post-war	Post-war	Post-war	Post-war
1953	Post-war	Post-war	Post-war	Post-war	Post-war
1954	Post-war	Post-war	Post-war	Post-war	Post-war
1955	Post-war	Post-war	Post-war	Post-war	Post-war
1956	Post-war	Post-war	Post-war	Post-war	Post-war
1957	Post-war	Post-war	Post-war	Post-war	Post-war
1958	Post-war	Post-war	Post-war	Post-war	Post-war
1959	Post-war	Post-war	Post-war	Post-war	Post-war
1960	Post-war	Post-war	Post-war	Post-war	Post-war

Additional Tables

Table 2: Total arrest by ethnicity and imputation adjustment, 1921-1960

	Arrests			
Ethnicity	Only Labeled	Labeled + Unadj. Imput.	Labeled + Parsimon. Adj.	Labeled + Full-matrix Adj.
Russian	550 349	1 064 741	1 041 329	1 041 325
Belorussian	67 615	85 525	70 394	72 390
Polish	61 221	85 257	73 056	79 973
German	60 798	168 422	164 827	171 712
Ukrainian	54 398	91 812	94 256	96 626
Kazakh	37 125	46 541	46 340	42 775
Tatar	32 098	72 422	72 933	70 825
Jewish	31 047	43 704	42 313	42 788
Latvian	15 442	21 626	19 398	18 624
Chinese	9 693	11 507	10 642	10 490
Estonian	9 402	15 562	13 865	13 676
Chuvash	8 910	14 894	29 582	23 177
Bashkir	8 428	17 879	21 780	19 366
Finnish	8 347	14 609	13 151	13 393
Mordvin	6 011	12 646	29 026	27 908
Buryat	5 679	6 735	6 651	6 711
Mari	5 385	7 485	12 989	12 666
Lithuanian	4 651	5 474	5 259	5 626
Karelian	4 174	9 900	13 078	11 872
Korean	4 060	8 821	11 477	11 841
Komi	3 616	5 832	6 704	5 825
Ossetian	3 236	3 722	3 446	3 445
Udmurt	3 090	4 469	11 694	10 178
Armenian	2 937	4 851	4 732	4 732
Kabardian	2 733	4 437	3 833	3 946
Greek	2 246	24 504	26 316	26 914
Khakas	2 221	8 136	7 094	7 139
Altai	1 894	2 475	2 588	2 632
Yakut	1 706	3 323	3 486	3 359
Georgian	1 621	3 048	2 651	2 451
Moldovan	1 391	2 780	4 162	4 077
Kalmyk	1 294	2 169	2 072	2 015
Japanese	1 231	14 574	10 927	10 922
Uzbek	1 061	4 044	8 301	8 798
Hungarian	1 018	1 611	1 645	1 556
Bulgarian	1 015	2 477	3 721	3 408
Balkar	861	4 741	3 837	3 270
Chechen	696	8 511	11 716	12 839

Table 3: Naive Bayes Performance Measures by Ethnicity

Ethnicity	Sensitivity	Specificity
Altai	0.475	1.000
Armenian	0.799	1.000
Balkar	0.972	0.999
Bashkir	0.476	0.997
Belorussian	0.503	0.975
Bulgarian	0.365	1.000
Buryat	0.772	1.000
Estonian	0.695	0.996
Finnish	0.789	0.998
Georgian	0.560	0.999
German	0.878	0.988
Greek	0.695	0.995
Hungarian	0.316	0.999
Chechen	0.554	0.999
Chinese	0.922	0.997
Chuvash	0.102	0.995
Japanese	0.967	0.996
Jewish	0.867	0.997
Kabardian	0.881	0.999
Kalmyk	0.846	1.000
Karelian	0.155	0.995
Kazakh	0.833	0.999
Khakas	0.827	0.998
Komi	0.233	0.998
Korean	0.491	0.999
Latvian	0.673	0.995
Lithuanian	0.560	0.999
Mari	0.194	0.999
Moldovan	0.271	0.999
Mordvin	0.162	0.997
Ossetian	0.835	1.000
Polish	0.790	0.980
Russian	0.886	0.869
Tatar	0.817	0.995
Udmurt	0.075	0.999
Ukrainian	0.427	0.976
Uzbek	0.310	0.999
Yakut	0.184	0.998

Table 4: Descriptive Statistics of Arrests from 1921 to 1960 by Ethnicity, Part 1

	Only labeled data					Labels + Ethnicity imputations (no adj.)				
Ethnicity	Mean	St.dev.	Min	Max	Total	Mean	St.dev.	Min	Max	Total
Altai	42	144	0	901	1 663	44	147	0	924	1 742
Armenian	55	112	0	524	2 210	63	127	0	614	2 516
Balkar	21	63	0	370	841	24	68	0	401	970
Bashkir	199	480	0	2 071	7 964	215	513	0	2 282	8 585
Belorussian	1 558	3 291	4	18 768	62 316	1 690	3 577	4	20 458	67 584
Bulgarian	17	47	0	224	680	20	53	0	245	793
Buryat	141	428	0	2 192	5 629	145	435	0	2 217	5 792
Estonian	200	675	1	3 435	7 998	247	798	1	4 066	9 874
Finnish	162	654	0	3 237	6 493	183	699	0	3 415	7 328
Georgian	30	69	0	320	1 220	38	81	0	369	1 513
German	693	1 662	0	8 658	27 713	872	2 048	1	10 227	34 878
Greek	36	131	0	612	1 453	71	187	0	957	2 844
Hungarian	24	93	0	562	956	29	103	0	618	1 149
Chechen	16	29	0	110	624	33	53	0	249	1 303
Chinese	229	1 085	0	6 882	9 179	250	1 185	0	7 518	9 990
Chuvash	209	430	0	2 455	8 364	242	500	0	2 877	9 666
Japanese	30	95	0	547	1 216	91	183	0	891	3 654
Jewish	526	1 299	1	7 267	21 043	603	1 448	2	8 199	24 119
Kabardian	66	186	0	1 061	2 630	68	189	0	1 083	2 707
Kalmyk	6	13	0	58	245	8	14	0	58	300
Karelian	98	411	0	2 352	3 938	147	513	0	2 963	5 865
Kazakh	885	1 953	0	9 740	35 401	988	2 164	0	10 742	39 534
Khakas	32	98	0	487	1 264	48	131	0	662	1 920
Komi	85	189	0	1 137	3 395	101	226	0	1 358	4 050
Korean	93	362	0	2 203	3 712	100	379	0	2 300	4 001
Latvian	353	1 273	0	6 753	14 126	406	1 424	0	7 557	16 237
Lithuanian	101	255	0	1 365	4 028	105	263	0	1 392	4 211
Mari	60	120	0	549	2 391	63	126	0	586	2 521
Moldovan	29	49	0	211	1 162	33	56	0	259	1 328
Mordvin	130	258	0	1 377	5 197	156	313	0	1 711	6 248
Ossetian	21	34	0	158	830	23	36	0	161	907
Polish	1 077	2 722	0	14 023	43 088	1 190	2 983	0	15 460	47 598
Russian	11 786	27 149	46	157 725	471 450	14 807	33 665	54	196 301	592 263
Tatar	688	1 406	0	6 275	27 539	764	1 562	0	7 098	30 560
Udmurt	72	135	0	781	2 864	77	146	0	849	3 071
Ukrainian	1 160	2 668	10	14 694	46 384	1 329	3 025	12	16 819	53 175
Uzbek	26	58	0	268	1 059	69	157	0	746	2 752
Yakut	39	68	0	348	1 571	52	86	0	426	2 084

Table 5: Descriptive Statistics of Arrests from 1921 to 1960 by Ethnicity, Part 2

	Labels + Arrest date imputations					Labels + Arrest date + Ethnicity imput. (no adj.)				
Ethnicity	Mean	St.dev.	Min	Max	Total	Mean	St.dev.	Min	Max	Total
Altai	47	146	0	903	1 894	62	161	0	955	2 475
Armenian	73	140	0	665	2 937	121	184	0	863	4 851
Balkar	22	64	0	375	861	119	226	0	1 058	4 741
Bashkir	211	508	0	2 100	8 428	447	1 133	0	5 548	17 879
Belorussian	1 690	3 459	5	19 637	67 615	2 138	4 076	8	22 668	85 525
Bulgarian	25	54	0	245	1 015	62	96	0	347	2 477
Buryat	142	431	0	2 201	5 679	168	451	0	2 244	6 735
Estonian	235	756	1	3 872	9 402	389	949	1	4 832	15 562
Finnish	209	718	0	3 534	8 347	365	852	0	3 991	14 609
Georgian	41	86	0	383	1 621	76	128	0	577	3 048
German	1 520	3 568	2	20 096	60 798	4 211	10 367	2	63 686	168 422
Greek	56	148	0	687	2 246	613	1 134	0	4 727	24 504
Hungarian	25	97	0	584	1 018	40	115	0	670	1 611
Chechen	17	32	0	143	696	213	434	0	2 225	8 511
Chinese	242	1 121	0	7 111	9 693	288	1 247	0	7 879	11 507
Chuvash	223	449	0	2 500	8 910	372	736	0	3 328	14 894
Japanese	31	95	0	550	1 231	364	773	0	4 199	14 574
Jewish	776	1 761	1	8 475	31 047	1 093	2 333	3	10 318	43 704
Kabardian	68	194	0	1 113	2 733	111	232	0	1 197	4 437
Kalmyk	32	105	0	620	1 294	54	147	0	837	2 169
Karelian	104	433	0	2 471	4 174	248	636	0	3 579	9 900
Kazakh	928	2 035	0	10 065	37 125	1 164	2 370	0	11 537	46 541
Khakas	56	138	0	552	2 221	203	494	0	2 296	8 136
Komi	90	195	0	1 169	3 616	146	283	0	1 534	5 832
Korean	102	384	0	2 270	4 060	221	496	0	2 406	8 821
Latvian	386	1 358	0	7 181	15 442	541	1 594	5	8 463	21 626
Lithuanian	116	279	0	1 518	4 651	137	303	1	1 655	5 474
Mari	135	280	0	1 451	5 385	187	370	0	1 535	7 485
Moldovan	35	55	0	225	1 391	70	102	0	409	2 780
Mordvin	150	291	0	1 550	6 011	316	635	0	2 510	12 646
Ossetian	81	192	0	1 160	3 236	93	210	0	1 245	3 722
Polish	1 531	3 315	0	15 503	61 221	2 131	4 178	0	18 510	85 257
Russian	13 759	30 374	53	173 860	550 349	26 619	52 769	63	237 714	1 064 741
Tatar	802	1 631	0	6 741	32 098	1 811	4 264	1	20 929	72 422
Udmurt	77	142	0	814	3 090	112	193	0	948	4 469
Ukrainian	1 360	2 997	17	16 484	54 398	2 295	4 324	24	21 486	91 812
Uzbek	27	58	0	268	1 061	101	196	0	912	4 044
Yakut	43	71	0	351	1 706	83	127	0	479	3 323

Table 6: Arrest Date Imputation - Model Results

	Dependent variable:	
	I^y	$log(y^{\text{pos}})$
	logistic	*OLS*
	(1)	(2)
(Intercept)	-0.771^{***} (0.161)	0.888^{***} (0.025)
Year of Trial - 1922	1.630^{***} (0.586)	1.038^{***} (0.033)
Year of Trial - 1923	0.955^{*} (0.510)	0.976^{***} (0.039)
Year of Trial - 1924	2.128^{**} (1.004)	1.122^{***} (0.043)
Year of Trial - 1925	1.019^{*} (0.586)	1.112^{***} (0.043)
Year of Trial - 1926	0.508^{*} (0.284)	0.972^{***} (0.027)
Year of Trial - 1927	0.346 (0.234)	0.904^{***} (0.024)
Year of Trial - 1928	-0.260^{**} (0.123)	0.422^{***} (0.016)
Year of Trial - 1929	-0.209^{**} (0.101)	0.307^{***} (0.012)
Year of Trial - 1930	0.012 (0.103)	0.754^{***} (0.012)
Year of Trial - 1931	0.166 (0.111)	0.695^{***} (0.013)
Year of Trial - 1932	0.299^{***} (0.106)	0.463^{***} (0.012)
Year of Trial - 1933	0.193 (0.139)	0.457^{***} (0.016)
Year of Trial - 1934	-0.198^{*} (0.116)	0.602^{***} (0.014)
Year of Trial - 1935	-0.206^{*} (0.112)	0.874^{***} (0.014)
Year of Trial - 1936	0.858^{***} (0.099)	-0.298^{***} (0.011)
Year of Trial - 1937	1.116^{***} (0.101)	0.486^{***} (0.012)
Year of Trial - 1938	1.845^{***} (0.151)	2.010^{***} (0.013)
Year of Trial - 1939	1.560^{***} (0.162)	1.579^{***} (0.013)
Year of Trial - 1940	0.463^{***} (0.113)	0.705^{***} (0.013)
Year of Trial - 1941	0.282^{***} (0.109)	0.641^{***} (0.013)
Year of Trial - 1942	0.234^{**} (0.114)	0.833^{***} (0.013)
Year of Trial - 1943	-0.422^{***} (0.118)	0.804^{***} (0.015)
Year of Trial - 1944	0.175 (0.127)	0.936^{***} (0.015)
Year of Trial - 1945	0.264^{*} (0.142)	1.182^{***} (0.016)
Year of Trial - 1946	0.164 (0.161)	0.987^{***} (0.018)
Year of Trial - 1947	0.231 (0.179)	0.860^{***} (0.020)
Year of Trial - 1948	0.810^{***} (0.192)	0.735^{***} (0.017)
Year of Trial - 1949	0.512^{***} (0.186)	0.953^{***} (0.019)
Year of Trial - 1950	0.532^{***} (0.188)	0.908^{***} (0.019)
Year of Trial - 1951	-0.080 (0.197)	0.844^{***} (0.024)
Year of Trial - 1952	0.077 (0.269)	0.619^{***} (0.031)
Year of Trial - 1953	0.003 (0.589)	1.680^{***} (0.070)
Year of Trial - 1954	-0.526 (0.462)	2.253^{***} (0.071)
Year of Trial - 1955	-0.713^{*} (0.425)	1.324^{***} (0.071)
Year of Trial - 1956	0.950^{***} (0.367)	0.683^{***} (0.029)
Year of Trial - 1957	0.595 (0.367)	0.813^{***} (0.034)
Year of Trial - 1958	-0.232 (0.333)	1.036^{***} (0.044)
Year of Trial - 1959	-0.716 (0.516)	1.042^{***} (0.087)
Year of Trial - 1960	4.292^{***} (0.094)	3.697^{***} (0.011)
Observations	812,592	805,800
R^2		0.235
Adjusted R^2		0.235
Log Likelihood	$-38{,}300.970$	
Akaike Inf. Crit.	76,681.930	
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 7: Difference-in-differences results

	Model		
	(1)	(2)	(3)
β_{1922}	0.257 (0.158)	0.273 (0.184)	−0.111 (0.177)
β_{1923}	0.409** (0.188)	0.560*** (0.216)	0.284 (0.178)
β_{1924}	−0.116 (0.213)	0.046 (0.268)	−0.862*** (0.191)
β_{1925}	−0.302 (0.224)	−0.230 (0.286)	−1.624*** (0.221)
β_{1926}	0.453* (0.235)	0.523* (0.316)	−1.589*** (0.201)
β_{1927}	0.272 (0.185)	0.321 (0.239)	−0.459** (0.199)
β_{1928}	0.202 (0.260)	0.152 (0.338)	−0.809*** (0.260)
β_{1929}	0.780*** (0.270)	0.675** (0.321)	0.481* (0.282)
β_{1930}	0.850*** (0.305)	0.650* (0.358)	0.571* (0.328)
β_{1931}	0.936*** (0.316)	0.695* (0.362)	0.706** (0.338)
β_{1932}	−0.184 (0.224)	−0.329 (0.279)	−0.339 (0.222)
β_{1933}	0.434* (0.241)	0.292 (0.291)	−0.114 (0.227)
β_{1934}	0.919*** (0.274)	0.871** (0.346)	1.272*** (0.304)
β_{1935}	0.916*** (0.231)	0.959*** (0.294)	0.961*** (0.239)
β_{1936}	0.265 (0.207)	0.274 (0.265)	0.526** (0.232)
β_{1937}	0.165 (0.167)	0.202 (0.212)	0.393* (0.204)
β_{1938}	0.799*** (0.210)	1.130*** (0.232)	1.162*** (0.229)
β_{1939}	1.778*** (0.238)	1.985*** (0.276)	2.022*** (0.218)
β_{1940}	3.540*** (0.231)	3.571*** (0.249)	3.620*** (0.210)
β_{1941}	4.007*** (0.211)	4.006*** (0.201)	4.066*** (0.191)
β_{1942}	2.041*** (0.245)	2.129*** (0.253)	2.085*** (0.238)
β_{1943}	1.380*** (0.337)	1.235*** (0.378)	0.893** (0.359)
β_{1944}	1.346*** (0.361)	1.239*** (0.403)	1.418*** (0.341)
β_{1945}	2.912*** (0.250)	2.940*** (0.275)	2.432*** (0.265)
β_{1946}	2.582*** (0.252)	2.556*** (0.280)	2.445*** (0.288)
β_{1947}	2.419*** (0.256)	2.382*** (0.278)	2.556*** (0.242)
β_{1948}	2.127*** (0.264)	2.251*** (0.278)	2.247*** (0.237)
β_{1949}	1.681*** (0.290)	1.803*** (0.298)	2.025*** (0.260)
β_{1950}	1.973*** (0.262)	2.049*** (0.289)	2.065*** (0.270)
β_{1951}	2.426*** (0.260)	2.447*** (0.306)	2.642*** (0.243)
β_{1952}	2.806*** (0.259)	2.774*** (0.311)	2.706*** (0.248)
β_{1953}	2.699*** (0.238)	2.596*** (0.262)	2.765*** (0.248)
β_{1954}	2.991*** (0.285)	2.940*** (0.342)	2.941*** (0.263)
β_{1955}	1.540*** (0.278)	1.595*** (0.331)	1.496*** (0.279)
β_{1956}	0.116 (0.259)	0.129 (0.321)	0.294 (0.254)
β_{1957}	0.106 (0.257)	0.136 (0.309)	0.142 (0.228)
β_{1958}	0.071 (0.228)	0.059 (0.273)	0.138 (0.215)
β_{1959}	−0.199 (0.246)	−0.177 (0.287)	−0.550** (0.249)
β_{1960}	−0.948*** (0.250)	−1.053*** (0.291)	−1.665*** (0.252)
Eth. with ind. state excluded	No	Yes	No
Only rehabilitated ind.	No	No	Yes
Geopol. relations controls	Yes	No	Yes
Ethnicity-spec. time trends	No	No	No
Observations	1,520	1,120	1,520
Adjusted R^2	0.850	0.845	0.836

Notes: Cluster-robust standard errors are in the parentheses. The coefficients from model (1) are plotted in the figure 1, from model (2) in the figure 7, and from model (3) in the figure 8. For additional information, refer to the notes of the respective figures. *p<0.1; **p<0.05; ***p<0.01

Table 8: Difference-in-differences results - Ethnicity-specific time trends

	Model		
	(1)	(2)	(3)
β_{1922}	0.257 (0.158)	0.287* (0.158)	0.287* (0.158)
β_{1923}	0.409** (0.188)	0.470** (0.185)	0.470** (0.185)
β_{1924}	−0.116 (0.213)	−0.025 (0.214)	−0.025 (0.214)
β_{1925}	−0.302 (0.224)	−0.180 (0.228)	−0.180 (0.228)
β_{1926}	0.453* (0.235)	0.606** (0.239)	0.606** (0.239)
β_{1927}	0.272 (0.185)	0.455** (0.191)	0.455** (0.191)
β_{1928}	0.202 (0.260)	0.415 (0.271)	0.415 (0.271)
β_{1929}	0.780*** (0.270)	1.024*** (0.294)	1.024*** (0.294)
β_{1930}	0.850*** (0.305)	1.124*** (0.320)	1.124*** (0.320)
β_{1931}	0.936*** (0.316)	1.240*** (0.334)	1.240*** (0.334)
β_{1932}	−0.184 (0.224)	0.151 (0.254)	0.151 (0.254)
β_{1933}	0.434* (0.241)	0.799*** (0.257)	0.799*** (0.257)
β_{1934}	0.919*** (0.274)	1.315*** (0.289)	1.315*** (0.289)
β_{1935}	0.916*** (0.231)	1.342*** (0.246)	1.342*** (0.246)
β_{1936}	0.265 (0.207)	0.722*** (0.220)	0.722*** (0.220)
β_{1937}	0.165 (0.167)	0.652*** (0.196)	0.652*** (0.196)
β_{1938}	0.799*** (0.210)	1.315*** (0.225)	1.315*** (0.225)
β_{1939}	1.778*** (0.238)	2.343*** (0.249)	2.343*** (0.249)
β_{1940}	3.540*** (0.231)	4.102*** (0.227)	4.102*** (0.227)
β_{1941}	4.007*** (0.211)	4.591*** (0.185)	4.591*** (0.185)
β_{1942}	2.041*** (0.245)	2.654*** (0.221)	2.654*** (0.221)
β_{1943}	1.380*** (0.337)	2.024*** (0.302)	2.024*** (0.302)
β_{1944}	1.346*** (0.361)	1.986*** (0.317)	1.986*** (0.317)
β_{1945}	2.912*** (0.250)	3.596*** (0.190)	3.596*** (0.190)
β_{1946}	2.582*** (0.252)	3.295*** (0.210)	3.295*** (0.210)
β_{1947}	2.419*** (0.256)	3.162*** (0.204)	3.162*** (0.204)
β_{1948}	2.127*** (0.264)	2.901*** (0.209)	2.901*** (0.209)
β_{1949}	1.681*** (0.290)	2.486*** (0.237)	2.486*** (0.237)
β_{1950}	1.973*** (0.262)	2.808*** (0.214)	2.808*** (0.214)
β_{1951}	2.426*** (0.260)	3.291*** (0.184)	3.291*** (0.184)
β_{1952}	2.806*** (0.259)	3.702*** (0.198)	3.702*** (0.198)
β_{1953}	2.699*** (0.238)	3.625*** (0.175)	3.625*** (0.175)
β_{1954}	2.991*** (0.285)	3.948*** (0.219)	3.948*** (0.219)
β_{1955}	1.540*** (0.278)	2.527*** (0.224)	2.527*** (0.224)
β_{1956}	0.116 (0.259)	1.134*** (0.212)	1.134*** (0.212)
β_{1957}	0.106 (0.257)	1.154*** (0.191)	1.154*** (0.191)
β_{1958}	0.071 (0.228)	1.149*** (0.165)	1.149*** (0.165)
β_{1959}	−0.199 (0.246)	0.910*** (0.186)	0.910*** (0.186)
β_{1960}	−0.948*** (0.250)	0.191 (0.214)	0.191 (0.214)
Ethnicity-spec. time trends	None	Linear	Quadratic
Eth. with ind. state excluded	No	Yes	No
Geopol. relations controls	Yes	Yes	Yes
Observations	1,520	1,520	1,520
Adjusted R^2	0.850	0.871	0.871

Notes: Cluster-robust standard errors are in the parentheses. The coefficients from these models are plotted in the figure 11. For additional information, refer to the notes of the figures 11. *p<0.1; **p<0.05; ***p<0.01

Table 9: Difference-in-differences results - Ethnicity Imputation Adjustments

	Model		
	(1)	(2)	(3)
β_{1922}	0.257 (0.158)	0.263* (0.151)	0.227** (0.110)
β_{1923}	0.409** (0.188)	0.477** (0.207)	0.338** (0.141)
β_{1924}	−0.116 (0.213)	−0.035 (0.212)	−0.038 (0.135)
β_{1925}	−0.302 (0.224)	−0.293 (0.222)	−0.140 (0.178)
β_{1926}	0.453* (0.235)	0.486** (0.231)	0.211 (0.150)
β_{1927}	0.272 (0.185)	0.278 (0.182)	0.143 (0.135)
β_{1928}	0.202 (0.260)	0.107 (0.252)	0.023 (0.163)
β_{1929}	0.780*** (0.270)	0.644*** (0.249)	0.410** (0.170)
β_{1930}	0.850*** (0.305)	0.699** (0.278)	0.476** (0.200)
β_{1931}	0.936*** (0.316)	0.850*** (0.313)	0.557*** (0.210)
β_{1932}	−0.184 (0.224)	−0.376* (0.222)	−0.317* (0.169)
β_{1933}	0.434* (0.241)	0.281 (0.206)	0.186 (0.165)
β_{1934}	0.919*** (0.274)	0.871*** (0.291)	0.627*** (0.203)
β_{1935}	0.916*** (0.231)	0.820*** (0.201)	0.602*** (0.147)
β_{1936}	0.265 (0.207)	0.173 (0.234)	0.050 (0.183)
β_{1937}	0.165 (0.167)	0.096 (0.174)	0.011 (0.148)
β_{1938}	0.799*** (0.210)	0.757*** (0.227)	0.626*** (0.178)
β_{1939}	1.778*** (0.238)	1.944*** (0.245)	1.385*** (0.168)
β_{1940}	3.540*** (0.231)	3.774*** (0.255)	3.059*** (0.175)
β_{1941}	4.007*** (0.211)	4.194*** (0.215)	3.463*** (0.153)
β_{1942}	2.041*** (0.245)	1.986*** (0.243)	1.702*** (0.190)
β_{1943}	1.380*** (0.337)	1.457*** (0.401)	1.031*** (0.296)
β_{1944}	1.346*** (0.361)	1.260*** (0.387)	1.002*** (0.298)
β_{1945}	2.912*** (0.250)	2.987*** (0.259)	2.544*** (0.204)
β_{1946}	2.582*** (0.252)	2.575*** (0.253)	2.263*** (0.214)
β_{1947}	2.419*** (0.256)	2.389*** (0.252)	2.066*** (0.215)
β_{1948}	2.127*** (0.264)	2.039*** (0.265)	1.749*** (0.211)
β_{1949}	1.681*** (0.290)	1.523*** (0.281)	1.311*** (0.227)
β_{1950}	1.973*** (0.262)	1.944*** (0.249)	1.695*** (0.199)
β_{1951}	2.426*** (0.260)	2.345*** (0.251)	2.020*** (0.220)
β_{1952}	2.806*** (0.259)	2.831*** (0.267)	2.365*** (0.210)
β_{1953}	2.699*** (0.238)	2.697*** (0.235)	2.296*** (0.201)
β_{1954}	2.991*** (0.285)	3.042*** (0.297)	2.657*** (0.250)
β_{1955}	1.540*** (0.278)	1.551*** (0.283)	1.233*** (0.236)
β_{1956}	0.116 (0.259)	0.106 (0.263)	−0.180 (0.220)
β_{1957}	0.106 (0.257)	0.120 (0.256)	−0.168 (0.224)
β_{1958}	0.071 (0.228)	0.092 (0.231)	−0.191 (0.219)
β_{1959}	−0.199 (0.246)	−0.172 (0.251)	−0.452** (0.220)
β_{1960}	−0.948*** (0.250)	−0.918*** (0.259)	−1.246*** (0.230)
Ethnicity imputation adjust.	Full-matrix	Parsimonoius	None
Ethnicity-spec. time trends	None	None	None
Eth. with ind. state excluded	No	No	No
Geopol. relations controls	Yes	Yes	Yes
Observations	1,520	1,520	1,520
Adjusted R^2	0.850	0.845	0.900

Notes: Cluster-robust standard errors are in the parentheses. The coefficients from these models are plotted in the figure 12. For additional information, refer to the notes of the figures 12. *p<0.1; **p<0.05; ***p<0.01

Table 10: Pre-treatment characteristics of ethnic groups in the USSR

Ethnic group	Total population	Ling. similarity to Russian	Urbanization rate	Ind. state
Altai	39 062	0	0.30	0
Armenian	1 567 568	1	35.45	0
Balkar	33 307	0	1.23	0
Bashkir	713 693	0	2.12	0
Belorussian	4 738 923	4	10.32	0
Bulgarian	111 296	3	6.26	0
Buryat	237 501	0	1.05	0
Estonian	154 666	0	23.00	1
Finnish	134 701	0	10.55	1
Georgian	1 821 184	0	16.93	0
German	1 238 549	1	14.92	1
Greek	213 765	1	21.21	1
Hungarian	5 476	0	63.33	1
Chechen	318 522	0	0.98	0
Chinese	10 247	0	64.87	1
Chuvash	1 117 419	0	1.60	0
Japanese	93	0	76.34	1
Jewish	2 599 973	1	82.43	0
Kabardian	139 925	0	1.27	0
Kalmyk	129 321	0	1.29	0
Karelian	248 120	0	2.91	0
Kazakh	3 968 289	0	2.18	0
Khakas	45 608	0	1.08	0
Komi	375 871	0	2.56	0
Korean	86 999	0	10.52	0
Latvian	141 703	2	42.31	1
Lithuanian	41 463	2	63.16	1
Mari	428 192	0	0.84	0
Moldovan	278 905	1	4.86	0
Mordvin	1 340 415	0	2.19	0
Ossetian	272 272	1	7.86	0
Polish	782 334	3	32.75	1
Russian	77 791 124	5	21.32	1
Tatar	2 916 536	0	15.48	0
Udmurt	504 187	0	1.21	0
Ukrainian	31 194 976	4	10.54	0
Uzbek	3 904 622	0	18.66	0
Yakut	240 709	0	2.20	0

Note:
Total population and urbanization rate of the ethnic group in the USSR is taken from 1926 census. The linguistic similarity to Russian is measured by the number of common nodes in the language tree (cladistic similarity). Independent state equals one if the ethnic group was a core group in an independent country that existed in the interwar period.

Table 11: Synthetic German minority weights, Only ethnicities without ind. state

Ethnic group	W-Weight
Tatar	0.53
Jewish	0.19
Korean	0.13
Ukrainian	0.11
Khakas	0.03
Chuvash	0.01

Table 12: Synthetic German minority weights, Only rehabilitated individuals

Ethnic group	W-Weight
Polish	0.32
Tatar	0.32
Korean	0.27
Mari	0.05
Greek	0.03

Additional Figures

Figure 2: Number of Predicted Arrests by Ethnicity, Year, and Prediction Adjustment

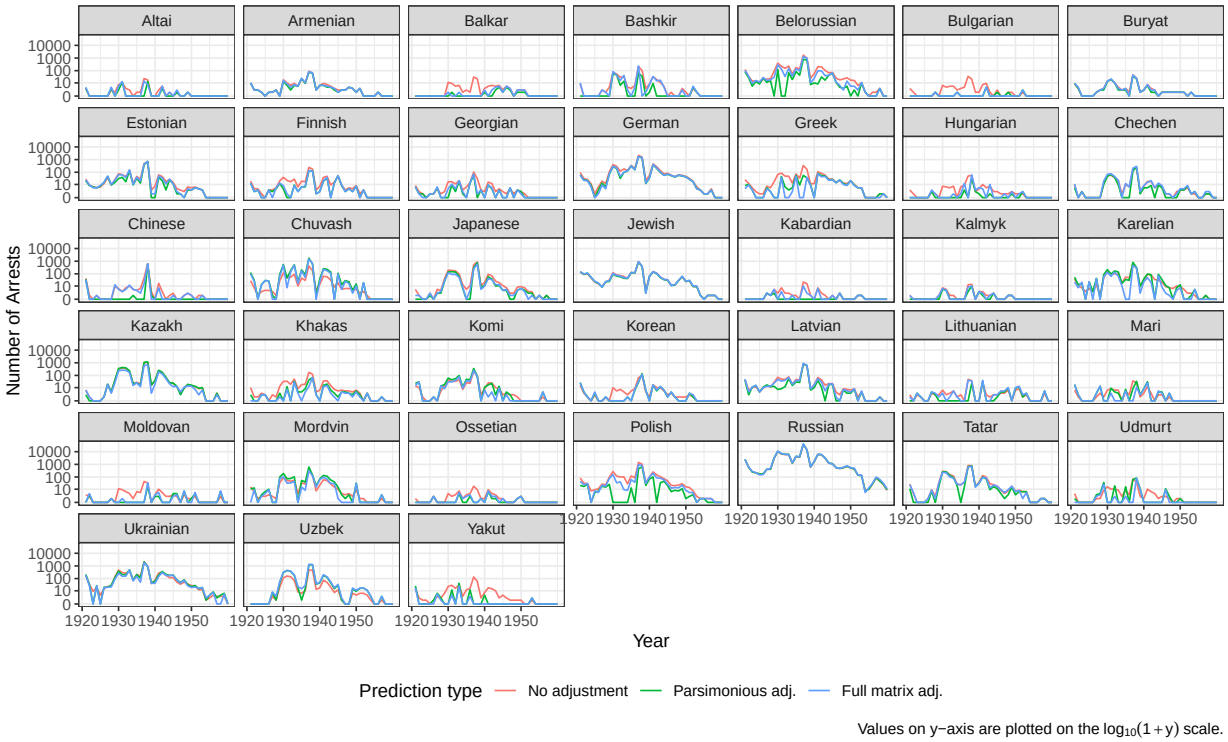


Figure 3: Histograms of Arrests by Ethnicity and Year

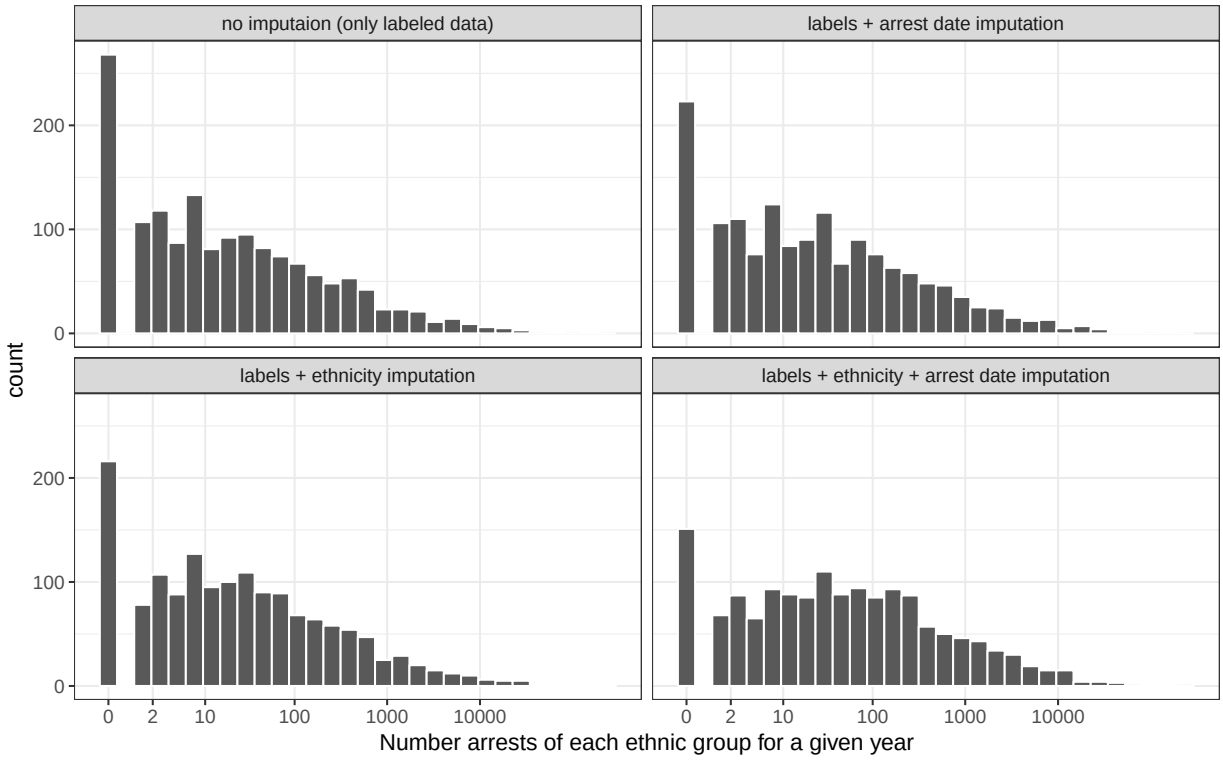


Figure 4: Histogram of Imputed Number of Days between Arrest and Process

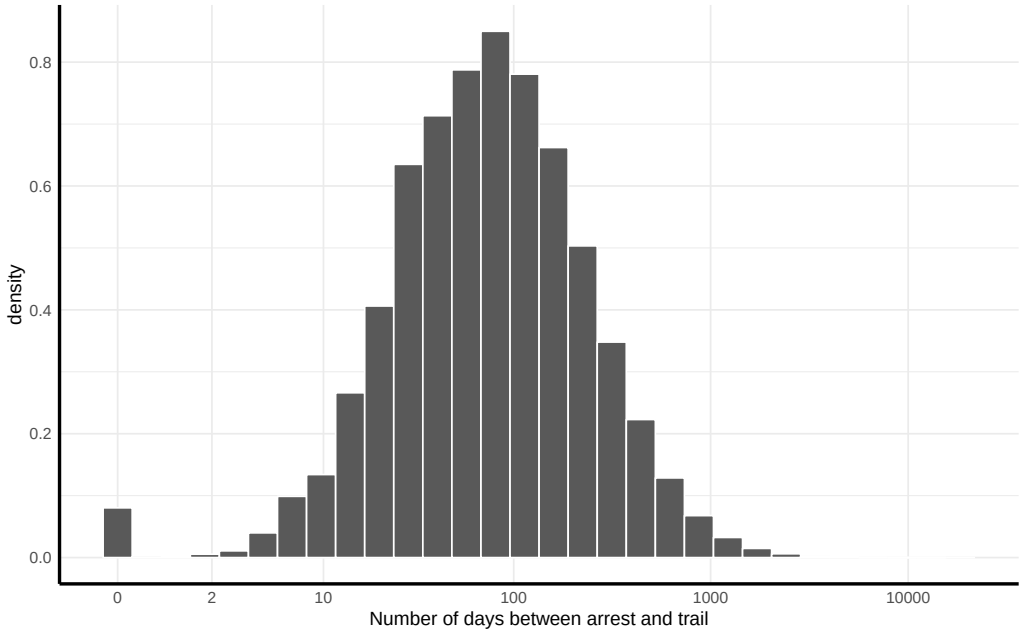


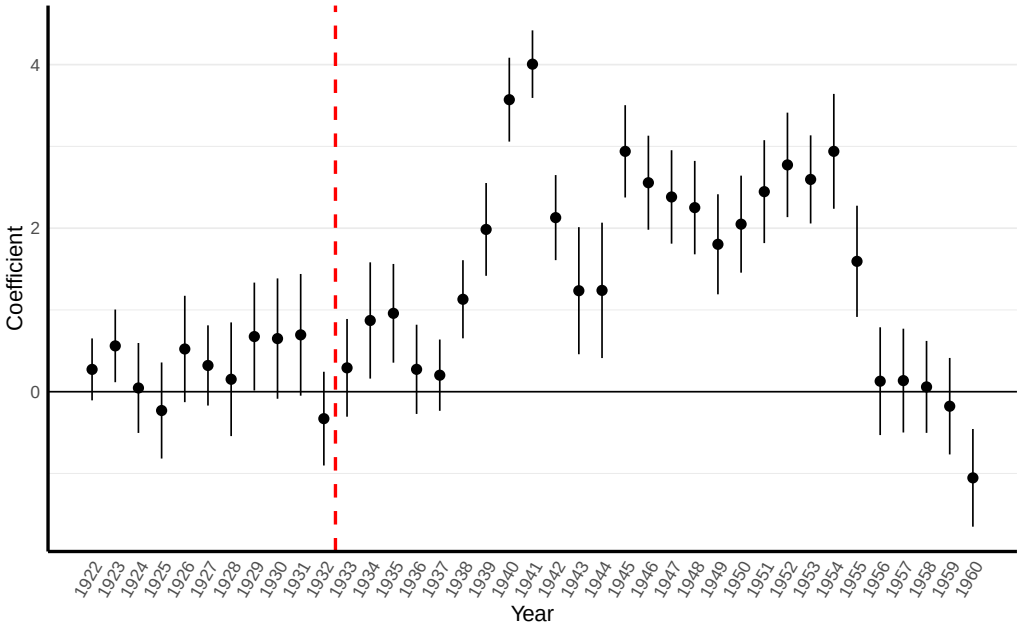
Figure 5: Time Series of Arrests

Note: Values on the y-axis are plotted on the $\log_{10}(1+y)$ scale.

Figure 6: Map of the Soviet Union in 1926

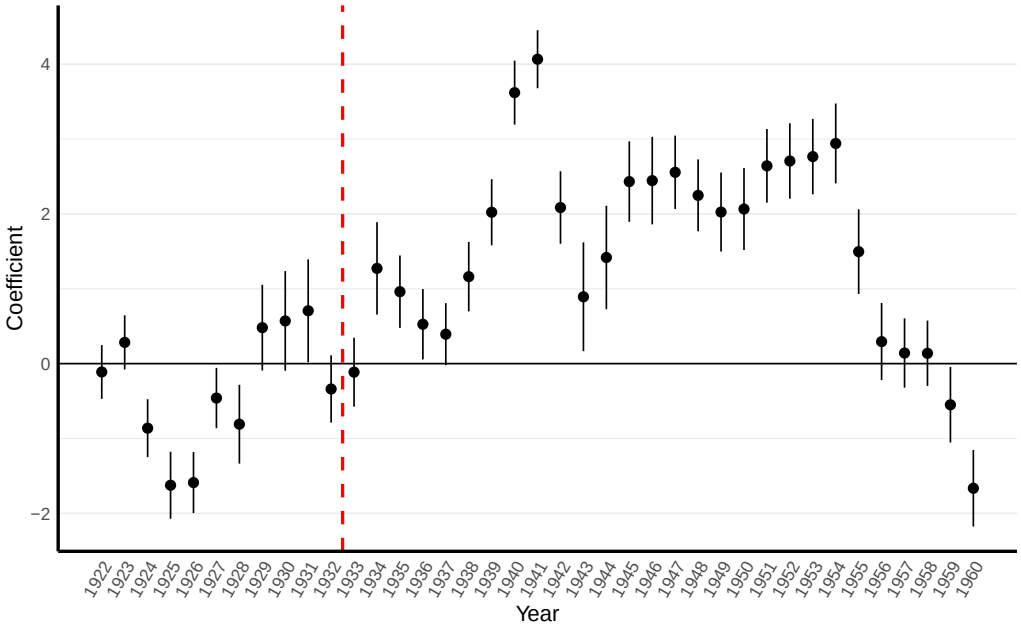
Notes: The shaded area shows the 250 km border buffer. The regions within are governorates (*guberniye*) and Autonomous Soviet Socialist Republics. The source of the map is Sablin et al. (2018).

Figure 7: Dynamic DiD, Only Ethnicities without Independent State



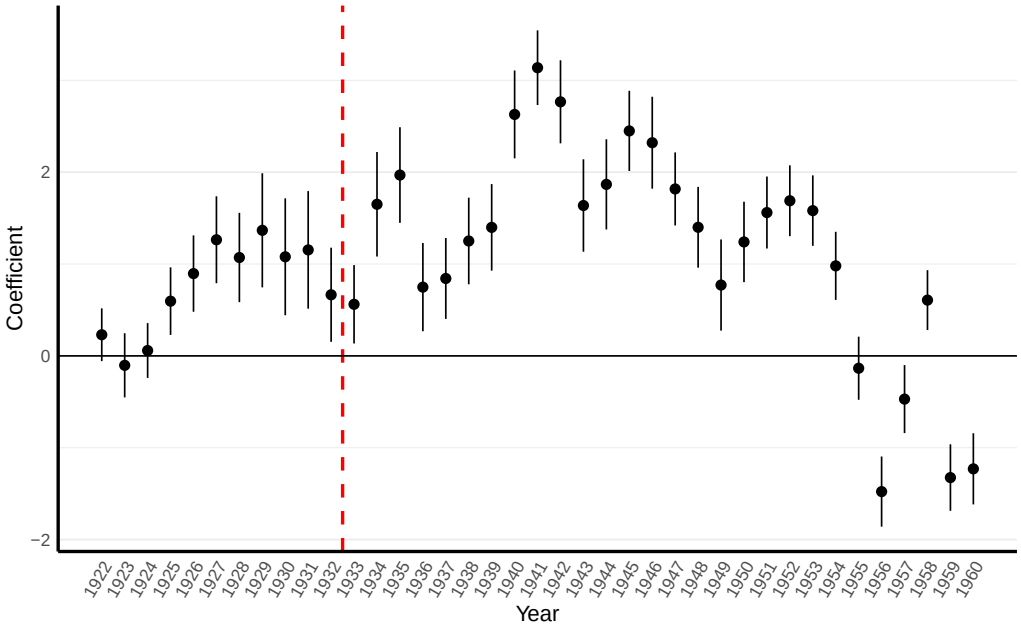
Notes: Only ethnic groups without independent state are included in the control group. Ethnicity and date of arrest were imputed. Full matrix adjustment was applied on ethnic group imputations. The quadratic ethnicity-specific time trends are included. Standard errors are clustered on the level of ethnicity and are based on cluster robust estimator by Pustejovsky and Tipton (2018). Error bars show 95% confidence intervals.

Figure 8: Dynamic DiD, Only Rehabilitated Individuals



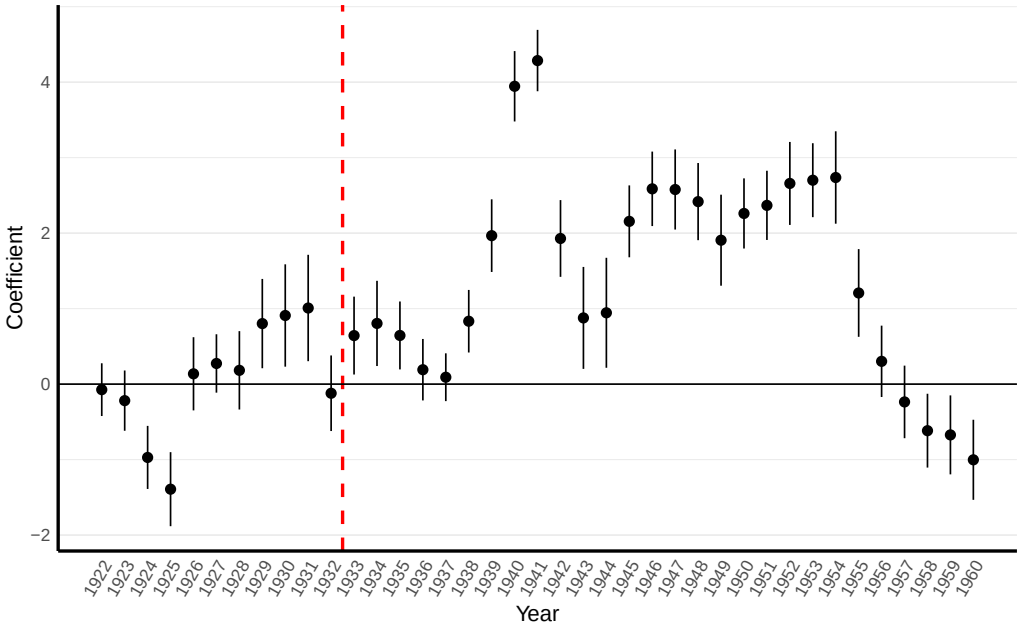
Notes: All 38 ethnic groups are included. Ethnicity and date of arrest were imputed. Full matrix adjustment was applied on ethnic group imputations. Controls for major changes in relations with the USSR are included. Standard errors are clustered on the level of ethnicity and are based on cluster robust estimator by Pustejovsky and Tipton (2018). Error bars show 95% confidence intervals.

Figure 9: Dynamic DiD, Only Border Regions



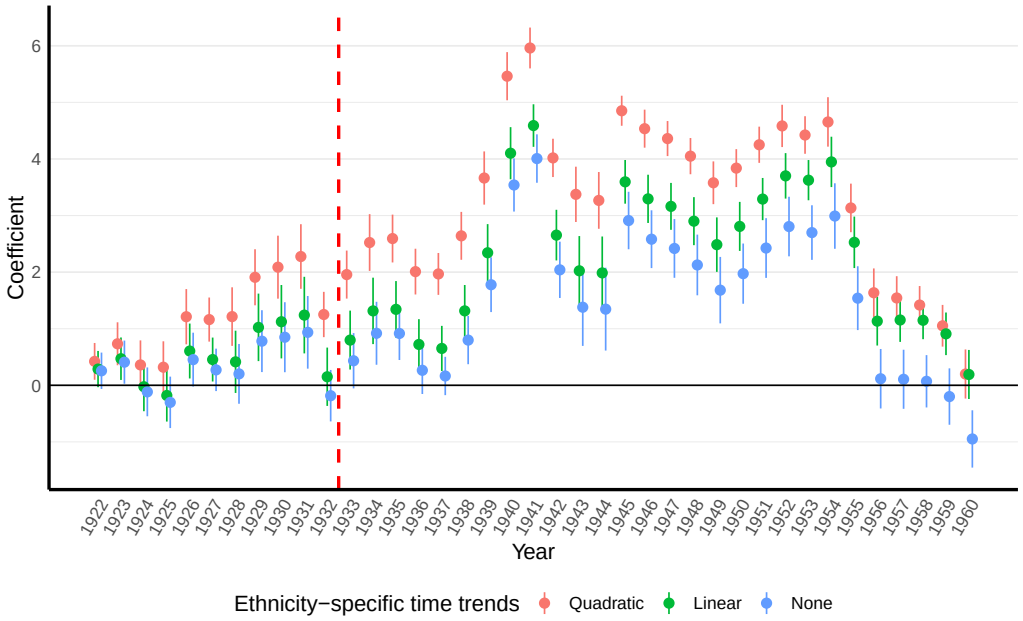
Notes: All 38 ethnic groups are included. Ethnicity and date of arrest were imputed. Full matrix adjustment was applied on ethnic group imputations. Controls for major changes in relations with the USSR are included. Standard errors are clustered on the level of ethnicity and are based on cluster robust estimator by Pustejovsky and Tipton (2018). Error bars show 95% confidence intervals.

Figure 10: Dynamic DiD, Border Regions Excluded



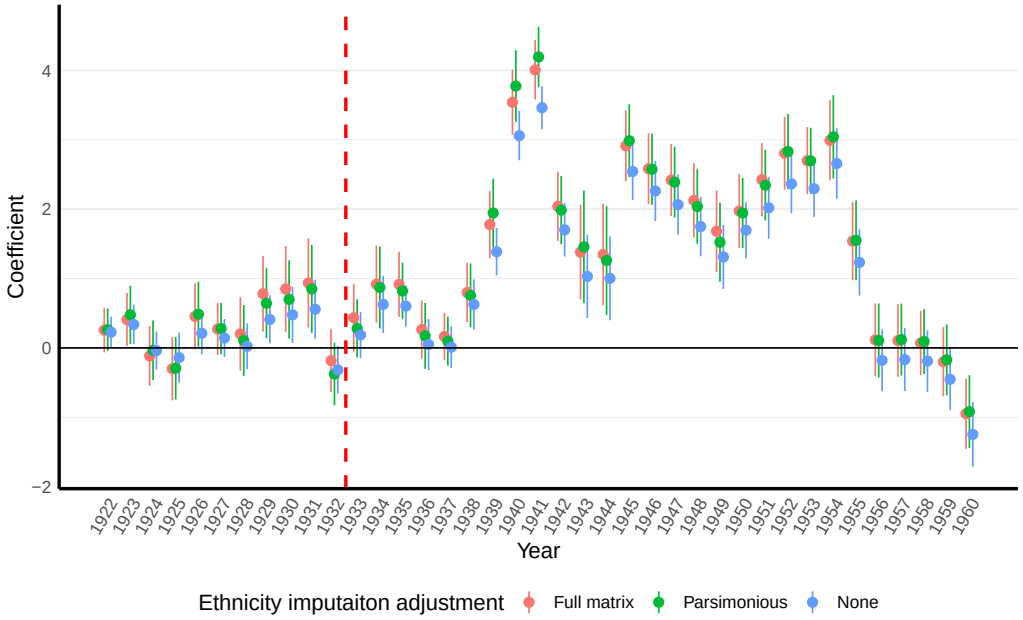
Notes: All 38 ethnic groups are included. Ethnicity and date of arrest were imputed. Full matrix adjustment was applied on ethnic group imputations. Controls for major changes in relations with the USSR are included. Standard errors are clustered on the level of ethnicity and are based on cluster robust estimator by Pustejovsky and Tipton (2018). Error bars show 95% confidence intervals.

Figure 11: Comparison of Ethnicity-specific Time Trends for DiD



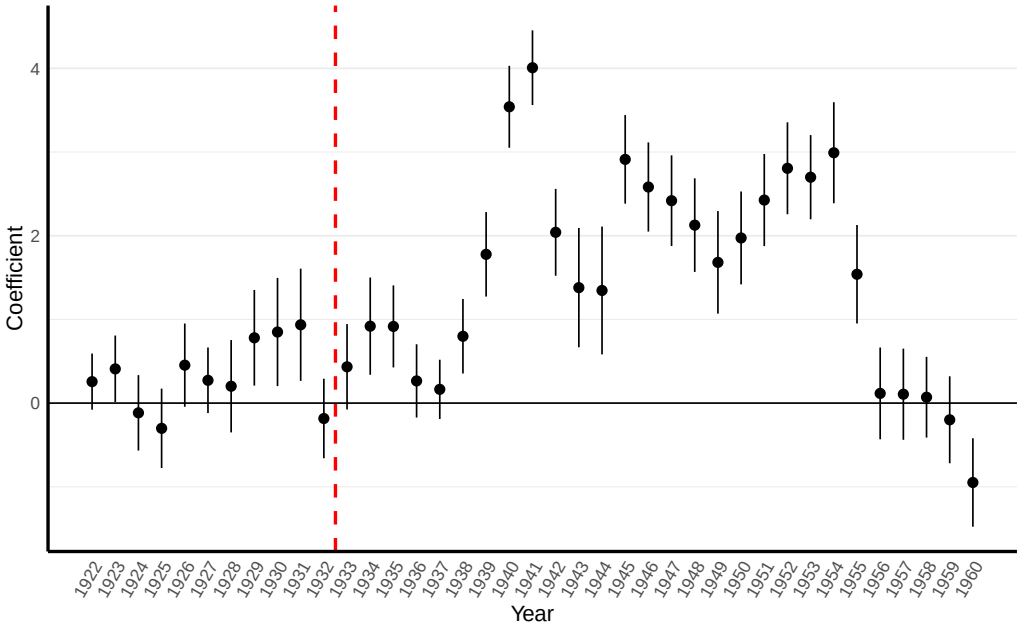
Notes: All 38 ethnic groups are included. Ethnicity and date of arrest were imputed. Full matrix adjustment was applied on ethnic group imputations. Standard errors are clustered on the level of ethnicity and are based on cluster robust estimator by Pustejovsky and Tipton (2018). Error bars show 95% confidence intervals.

Figure 12: Comparison of Ethnicity Imputation Adjustments for DiD



Notes: All 38 ethnic groups are included. Ethnicity and date of arrest were imputed. Standard errors are clustered on the level of ethnicity and are based on cluster robust estimator by Pustejovsky and Tipton (2018). Error bars show 95% confidence intervals.

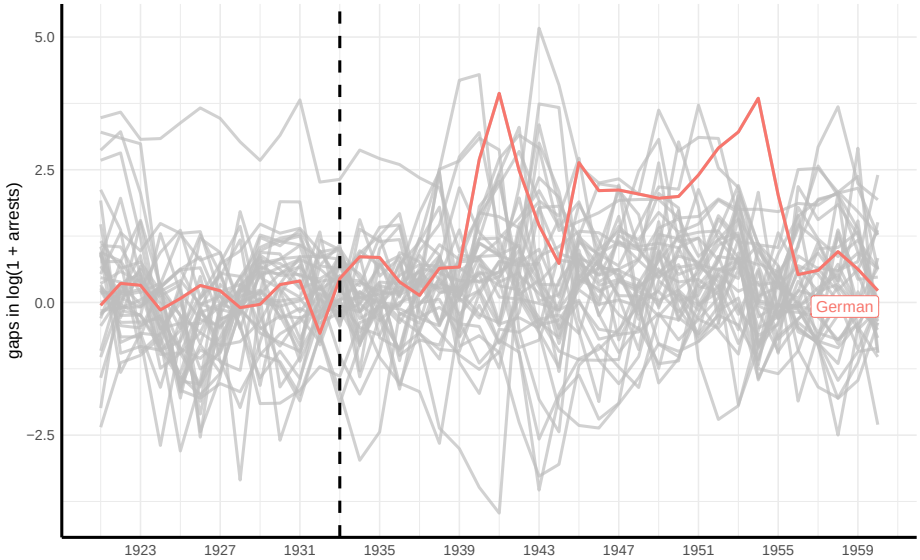
Figure 13: Dynamic DiD, Stata Standard Errors



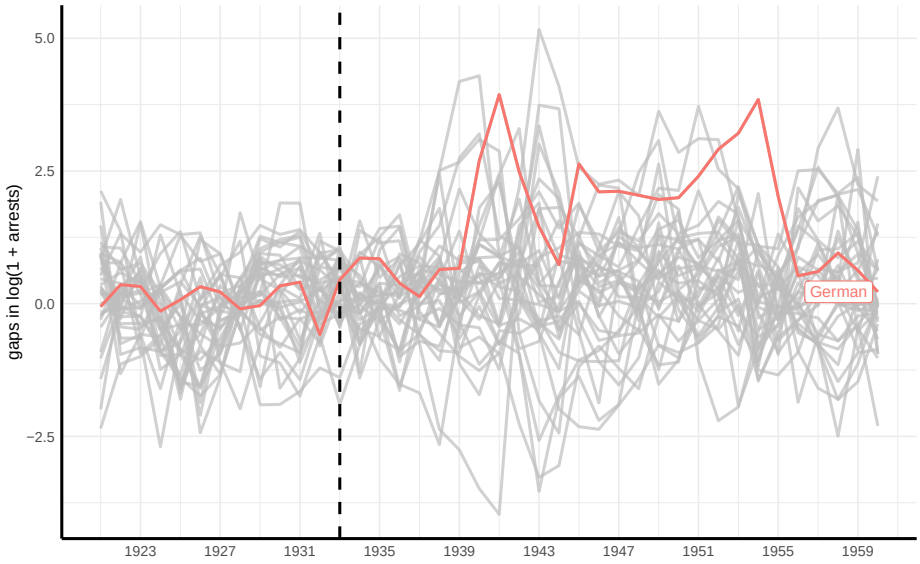
Notes: All 38 ethnic groups are included. Ethnicity and date of arrest were imputed. Full matrix adjustment was applied on ethnic group imputations. We used Stata standard errors clustered on ethnicity. Error bars show 95% confidence intervals.

Figure 14: Gaps between synthetic control and actual values for placebo tests

(a) All ethnic groups



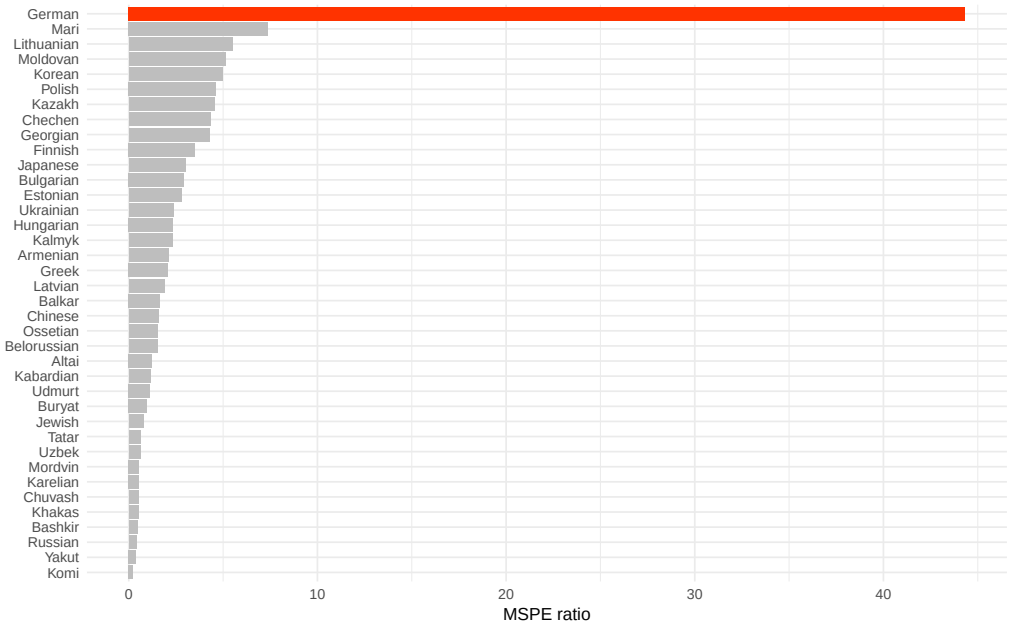
(b) Ethnic groups with pre-treatment MSPE higher than 20 times the MSPE of Germans excluded



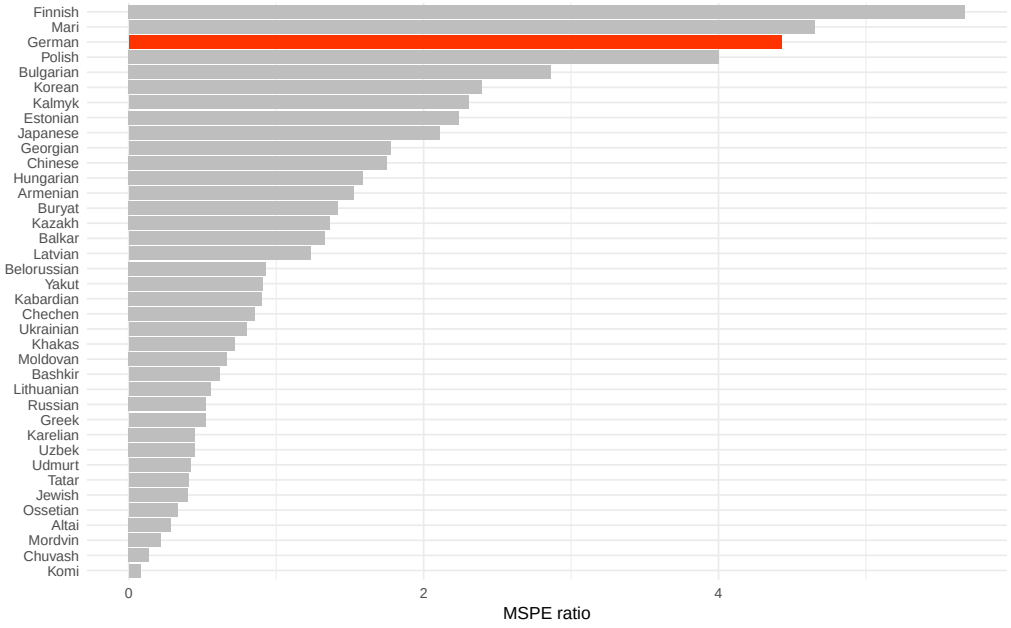
Notes: The predictors are the mean of log (1 + arrests) in the pre-treatment period, total population of the ethnic group in the USSR and its urbanization rate (both taken from the 1926 Soviet census), and linguistic similarity to Russian. Ethnicity and date of arrest were imputed. Full matrix adjustment was applied on ethnic group imputations. All 38 ethnic groups are included.

Figure 15: Ratios of post-treatment MSPE to pre-treatment MSPE

(a) The whole post-treatment period in the numerator (1933-1960)



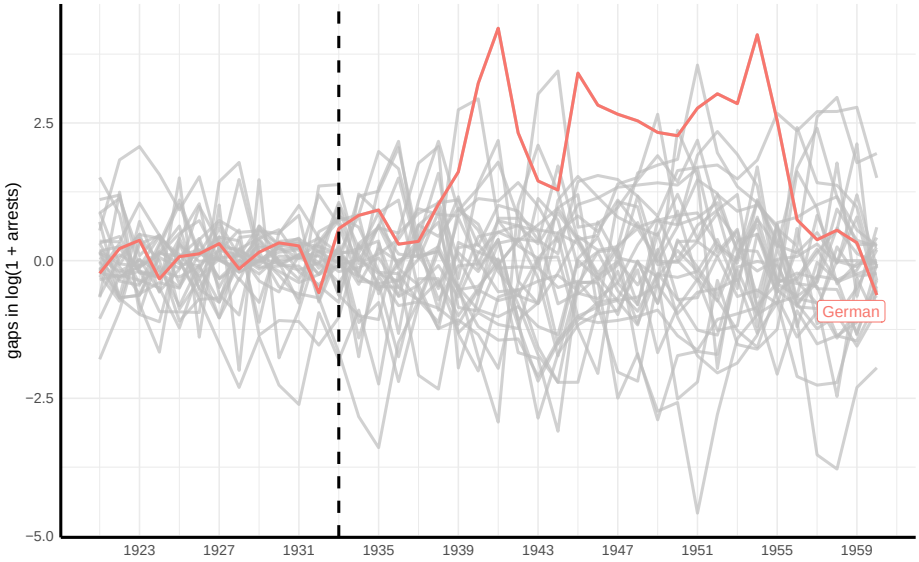
(b) Only the period from 1933 to 1939 in the numerator



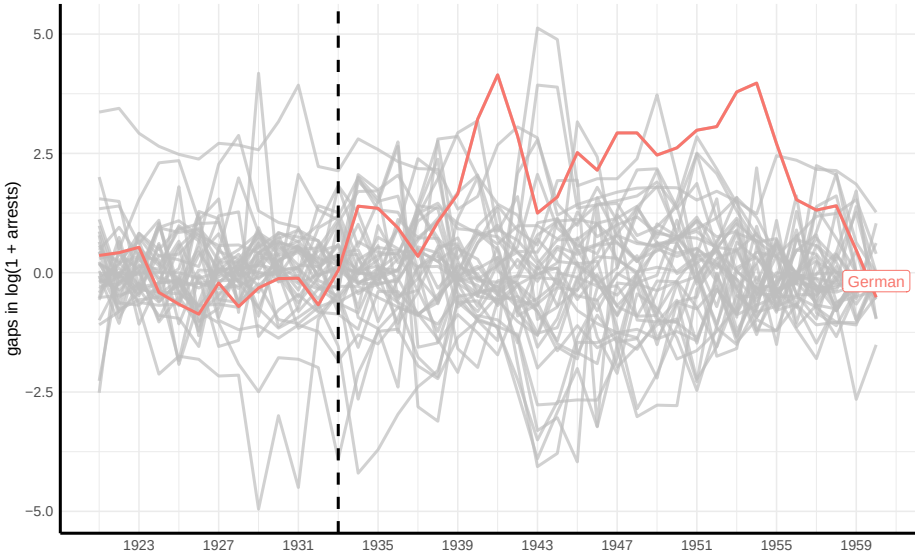
Notes: The same as for the figure 14.

Figure 16: Gaps between synthetic control and actual values for placebo tests

(a) Only ethnicities without independent state



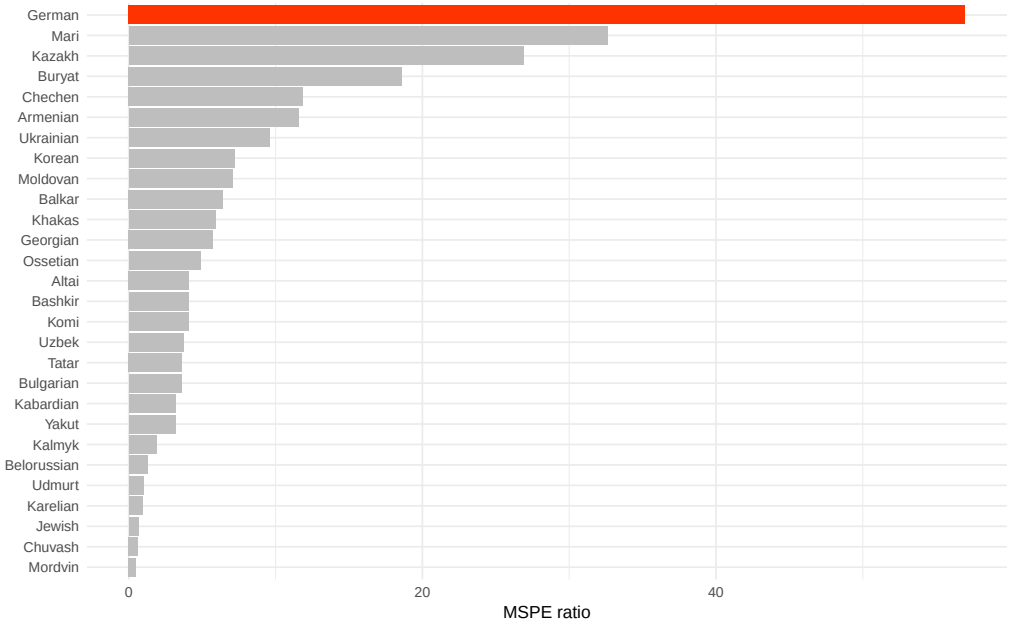
(b) Only rehabilitated individuals



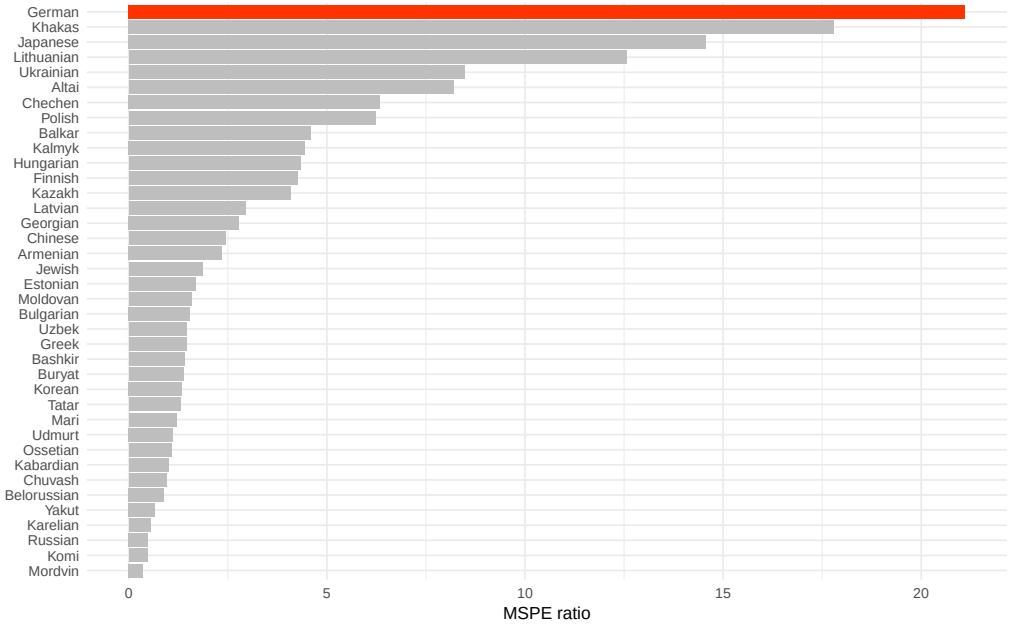
Notes: All pre-treatment outcomes were used as predictors. Ethnicity and date of arrest were imputed. Full matrix adjustment was applied on ethnic group imputations.

Figure 17: Ratios of post-treatment MSPE to pre-treatment MSPE

(a) Only ethnicities without independent state



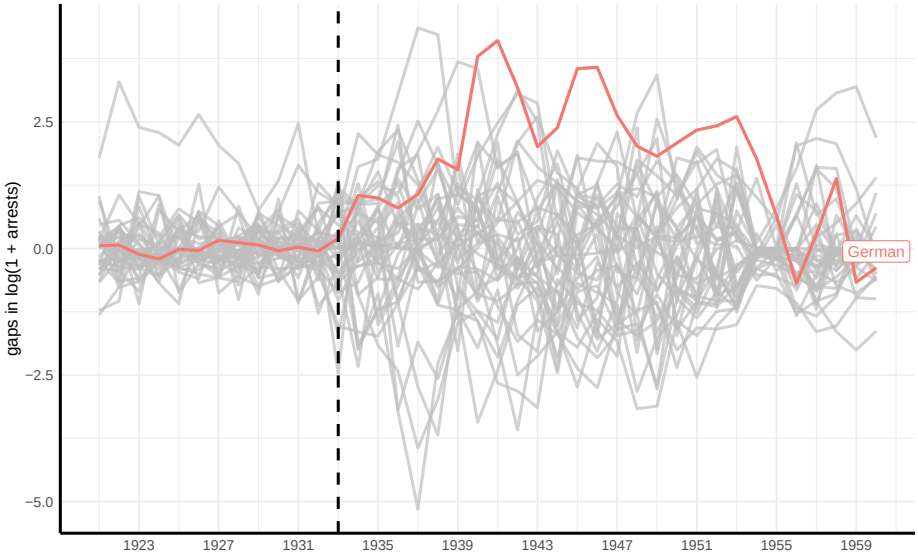
(b) Only rehabilitated individuals



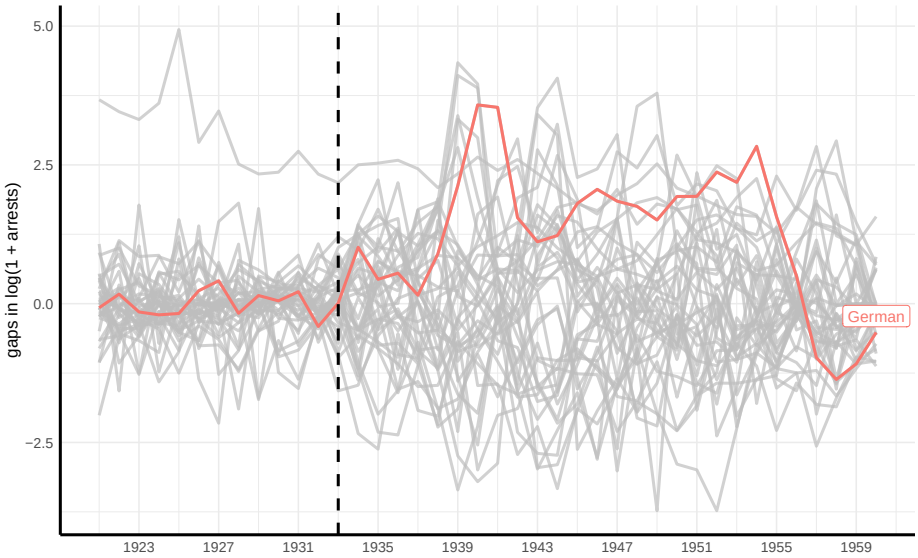
Notes: The whole post-treatment period in the numerator (1933-1960) for both figures. Otherwise same as for the figure 16.

Figure 18: Gaps between synthetic control and actual values for placebo tests

(a) Only Border Regions



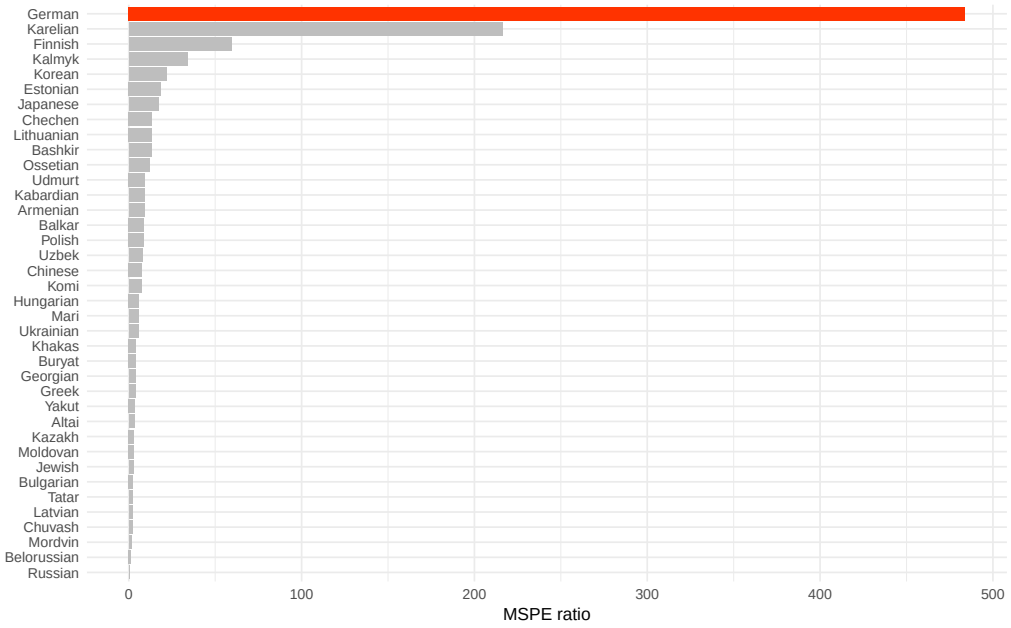
(b) Border Regions Excluded



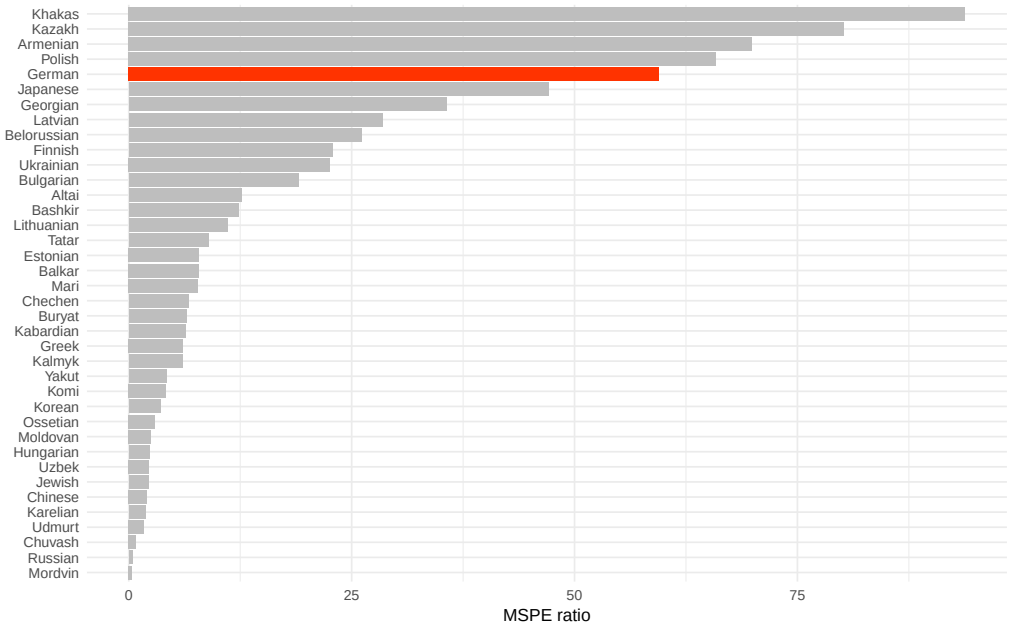
Notes: All pre-treatment outcomes were used as predictors. Ethnicity and date of arrest were imputed. Full matrix adjustment was applied on ethnic group imputations.

Figure 19: Ratios of post-treatment MSPE to pre-treatment MSPE

(a) Only Border Regions



(b) Border Regions Excluded



Notes: The whole post-treatment period in the numerator (1933-1960) for both figures. Otherwise same as for the figure 18.