Contents

Introduction

What determines the attitude of a state toward ethnic minorities within its borders? Why are some minorities accommodated or assimilated and others are politically excluded and repressed? Furthermore, why does the position of a state toward its minorities change in time? For example Soviet Union largely accommodated its minorities by in 1920s but it heavily repressed them in the campaigns of mass terror 10 years later.

Mylonas (2013) argues that the geopolitical concerns play an important role. Specifically, a state is likely to choose repression and exclusion if the ethnic minority's country of origin is seen as an geopolitical enemy. The minority is then viewed by the state as unreliable and as a potential fifth column of the foreign country.

We test this hypothesis on the case of German minority in Soviet union. In 1933, Hitlers rise to power changed Germany from a neutral actor to ideological and geopolitical enemy in the perspective of the Soviet Union. We can then see how the repression changed before and after 1933 and compare it with other minorities. In particular, we use the individual arrests by the Soviet secret police as a dependent variable and employ the difference in difference strategy and the synthetic control method.

The thesis has the following structure. First, we summarize the existing literature on the topic in the section 1. Next, the section 2 provides necessary historical context on German-Soviet relations, political repressions, and position of ethnic minorities in the USSR. This is followed by the section 3 where we describe the sources of the data and provide summary statistics of our main dataset. In the section 4, we present the methods that we use to impute missing information on ethnicity and the date of arrest. In the section 5, we describe the two methods that we use to empirically estimate the effect. The results of our analysis are provided in the section 6. Additional robustness checks are performed in the section 7 in order to asses the sensitivity of results to different specifications. Finally, we discuss implications of the results in the conclusion.

1 Literature Review

Existing literature on repressions has focused mostly on their consequences and legacies (Rozenas et al., 2017; Lupu and Peisakhin, 2017; Zhukov and Talibova, 2018). As far as the strategic use of repressions by the state is studied, it is usually in relation to domestic factors such as institutions and economic shocks (Davenport, 2007; Greitens, 2016; Blaydes, 2018) with less attention being given to external forces.

Davenport (2007) finds that democracy is correlated with lower levels of repression. However, it is mainly free electoral competition rather than constraints on power of the executive that accounts for this negative effect. Greitens (2016) links the severity of repression to the threat from dictator's inner circle. A dictator who fears that he would be deposed in a coup rather than in popular uprising will fragment their coercive apparatus in order to weaken the power of a potential challenger from within. The weakened secret police will be, according to Greitens, more likely to use violence since it fails to identify the transgressors and cannot effectively deter dissent. Other scholars see repression as a substitute for co-option (Wintrobe, 1998; Svolik, 2012). Instead relying on the threat of persecution, an authoritarian ruler might buy the loyalty of the population by distributing rents to the supporters of the regime usually through the party apparatus. Negative economic shocks can then increase repression since the rents are no longer available. Blaydes (2018) illustrates this on the case of Iraq under Hussein where lower oil prices

Furthermore, Blaydes (2018) presents a theory attempting to explain different levels of repression across ethnic minorities within a country. She argues that the nature of repression depends on the legibility of the ethnic group to the state.[1] Since the state coercive institutions cannot reliably identify transgressors in less legible population (because of, for example, greater cultural and linguistic distance), they will more tend to resort to collective punishment. The logic behind this is that the members of the

[1] The term legibility in this context means ability of a state to identify individuals in a given population and gather information on them.

group will police its members to avoid collective punishment.

Our research also contributes to the literature studying factors that influence position of a state towards its ethnic minorities and under what conditions conflict is likely to occur. Size and and distribution of ethnic groups have been emphasized. Several scholars pointed out that states with large number of ethnic groups are more likely to violently repress calls for autonomy or secession to discourage other ethnic minorities from making similar demands in the future (Evera, 1994; Toft, 2005; Walter, 2009). Furthermore, Toft (2005) argues that geographically concentrated groups tend to view their ethnic homeland as indivisible and non-negotiable issue which increases the likelihood of violent conflict. However, these approaches fail to explain changes in state's attitudes to minorities over short periods of time when the size and distribution of ethnic groups remains roughly constant.

More recently, the role of international factors have received greater attention. Butt (2017) argues that response of a state to a secessionist movement depends on external security environment and outside actors. Specifically, a state located in a war-prone region is more likely to suppress demands for secession because loss of territory and population would make it vulnerable to a potential future attack. Furthermore, a state responds with more violence if a separatist movement receives a support from an external power since the outside assistance makes the secessionists stronger. In addition to these strategic reasons, Butt emphasizes that receiving external support incites a strong feeling of betrayal to the central state.

Mylonas (2013) puts forward a theory explaining how geopolitical relations influence the attitude of a state towards its minorities. The model features an ethnic minority living within a host state and an external power. Moreover, Mylonas distinguishes between hosts states with revisionist foreign policy which want change the international status quo (e.g. because they gained power or lost territory in the past) and host states that prefer the current international order.

The predictions of the theory are summarized in the table 1. First, if a

External Power Support

		Yes		**No**
		Interstate Relations		Assimilation
		Ally	**Enemy**	
Host's State	**Revisionist**	Accommodation	Exclusion	
Foreign Policy	**Status Quo**	Accommodation	Assimilation	

Table 1: Theoretical predictions of Mylonas (2013)

minority group is not supported by any external power, the theory predicts that the host state will pursue policy of assimilation towards the group to "immunize" it from possible future agitation of external powers. Second, if an ethnic minority is supported by geopolitical ally then accommodation is likely since more repressive policies towards the minority could jeopardize the alliance. Third, theory predicts assimilation if a minority is supported by an geopolitical enemy and a host state pursues non-revisionist foreign policy because exclusionary policies could trigger new hostilities threatening the status quo. Finally, support of an external enemy combined with revisionist foreign policy will likely lead to exclusion of a given ethnic group since it is view as a potential "fifth column" of the external power.

We can apply the theory to our case. First, the Soviet Union was arguably a revisionist state. The Bolsheviks had to accept large losses of territory under the Treaty of Brest-Litovsk in 1918 so that they could focus on fighting in Russian Civil War. Thus, the USSR would certainly prefer to change the international status quo. Second, the German-Soviet relations were neutral of even moderately friendly prior to 1933 but turned hostile after the rise of Hitler (described in greater detail in subsection 2.1). Therefore the theory predicts that the policy of the Soviet state towards the German minority should change from accommodation to exclusion. Mylonas (2013, p. 22) defines exclusion as "policies that aim at the physical removal of a non-core group from the host state." The Soviet political repression which usually featured either outright execution or a term in a labor camp in the Far East of the country fits this description well. Thus, the theoretical expectation is that repressions of Soviet Germans relative to other minorities should increase after 1933.

Our main contribution is empirical test of this theory using a credible identification strategy. Most studies presented in this review test their hypotheses only by qualitative comparison of selected cases. Quantitative research usually involve only cross-sectional regressions based on categorical dependent variables.

For example, Mylonas (2013) tests his theory with data on the post-World War I Balkans where the nation-building policies (categorized into 3 groups: accommodation, assimilation and exclusion) toward 90 ethnic groups are a dependent variable and information on their support by external powers is an explanatory variables (together with other control variables). However, the results of the cross-sectional regression, used in the study, might easily be biased due to omitted variables or reverse causality and we believe that our approach offers cleaner identification.

McNamee and Zhang (2019) is methodologically and thematically closet study to ours. They analyze how the 1958 split in Soviet-China relations affected the demographic composition of the population in the Soviet-Chinese border regions. Using difference-indifference strategy, they find that, after the split, both states supported expulsions of the minority group and sponsored immigration of the majority group but only in border regions without significant natural boundary (e.g. high mountains). They conclude that the states use demographic engineering as a way to protect their vulnerable border against a hostile power. Nonetheless, as McNamee and Zhang only measure the ethnic composition of the regions they cannot unambiguously identify expulsions as the main culprit since other factors could plausibly affect voluntary migration.

2 Historical Background

In this section, we provide brief historical context for selected topics. Specifically, we first describe changing geopolitical relations of Germany and the Soviet from 1920 to 1941. In the next subsection, we provide brief overview of the most important aspects of Soviet political repression. Finally, evolution of the Soviet policy towards its ethnic minorities is summarized.

2.1 German–Soviet Relations in the Interwar Period

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. 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. Moreover, Hitler soon after his rise to power banned the German Communist Party and started to persecute its members (Haslam, 1984).

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 to increasing tensions, Japan and Germany signed the Anti-Comintern Pact in 1936 in which they committed to co-operate for defense against communistic disintegration.

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. Hitler thus tried to negotiate neutrality of the Soviet Union in war to avoid simultaneously facing Western powers, Poland and the Soviet Union. Soviet neutrality was potentially beneficial for Stalin too. A long and costly war would weaken the both the capitalist and 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 1938 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.

Hitler enjoyed major success in the first years of the war. By summer 1940, German forces defeated French army and annexed Denmark and Norway. However, German industry was severely lacking raw materials needed in war effort against Britain 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.

2.2 Soviet Political Repressions

The Soviet Union had massive coercive apparatus. The Soviet secret police (which was throughout the years named the Cheka, OGPU, NKVD, MVD and the KGB)[2] 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 counterrevolutionary activities which included treason, espionage, counterrevolutionary propaganda or agitation and failure to report any of these crimes.

[2] 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

In practice, this broad definition meant that anyone regarded as politically inconvenient could be arrested and prosecuted.

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).

2.3 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.[3] 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, 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 (this was know as the "Piedmont Principle").

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.

[3] The census data were obtained Institute of Demography of the National Research University Higher School of Economics

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).

3 Data

Our data on Soviet repressions come from the Victims of Political Terror in the USSR database by the Russian human rights organization Memorial (2017).[4] The Memorial data has already been used in empirical research by Zhukov and Talibova (2018) to estimate the effects of the Soviet repressions on the current political participation. The main sources of the Memorial lists are declassified Russian Interior Ministry documents, prosecutor's offices and the Commission for the Rehabilitation of Victims of Political Repression, and "Books of Memory". Vast majority records in the database are individual arrests under Article 58. This means that the victims of other repressive activities of the Soviet state such mass forced migration, counter-insurgency operations of the Red Army during Russian Civil War or famines are mostly excluded. Therefore our research focuses on one particular type of repression (individual arrests by the NKVD).

Nevertheless, even if we restrict ourselves to the individual arrests, the Memorial database is still not complete. At the time of our access to the database, the databes contained 2.7 million records which is approximately 70% of estimated 3.8 million people convicted under Article 58 between 1921 and 1953 (Zhukov and Talibova, 2018).

Missing data presents another major challenge. Table 2 shows how many values are missing for the variables of our interest. We can see that information on ethnicity is not available for more than half of all observations. In contrast, a surname is recorded for every arrest in the dataset and first name is missing only for negligible fraction of observations. The availability of information on names enables us to use them to infer missing ethnicity of an individual. In particular, we will train a Naive Bayes classifier on the 1 197 373 with known ethnicity and use the model's predictions to impute the ethnicity for the remaining 1 507 177 observations (the details are described in the subsection 4.1).

[4] The database can be accessed and searched from http://base.memo.ru/ (new version) or at http://lists.memo.ru/ (older version). However, we downloaded the file with every record from the database from https://github.com/MemorialInternational/memorial_data_FULL_DB/blob/master/data/lists.memo.ru-disk/lists.memo.ru-disk.zip

Table 2: Missing Data by Variable

Variable	Number of Missing Obs.	Percent of Missing Obs.
Ethnicity	1 507 177	55.73
Date of Arrest	1 650 912	61.04
Date of Trial	943 108	34.87
First Name	14 006	0.52

Table 3: Missing Dates of Arrest and Trial

	Date of Arrest	
Date of Trial	Missing	Present
Missing	747 419	195 689
Present	903 493	857 949

Date of arrest, which is also necessary for our analysis, has even higher rate of missingness than ethnicity.[5] One solution, albeit only partial, might be to use the date of trial to extrapolate the missing date of arrest where the trial date is available. As is shown in the table 3, we could impute this way the missing date of arrests for 903 493 observations for which we have the date of trial. Total number of arrests used for our analysis would therefore increase from 1 053 638 to 1 957 131. However, there remains 747 419 observations for which neither the date of arrest nor the date of trail are known.

There are more than 100 different ethnicities in the Memorial data. However, in many cases there are no more than hundred of people in the whole database that belong to a given ethnic groups. Therefore we decided to limit ourselves only to ethnic groups with at least 900 individuals in the Memorial database. This leaves us with 38 ethnic groups.

After imputations of ethnicity and the date of arrest had been applied, we created our main dataset by counting number of arrest for each ethnicity by year. With 38 ethnic groups and 40 time periods (from 1921 to 1960), we have 1520 observations in total. Basic descriptive statistics for each ethnicity are provided in the table A4. The plot of arrest by ethnicity and year (after

[5]We consider a date missing only if not even year of arrest is available. Many observations have information on year of arrest even though the exact month or day are missing. We do not categorize these observation as having missing date of arrest since our fundamental unit of analysis is a year and thus month and day are not of high interest to us. The same applies to the date of trial.

applying the transformation $\log(1 + y_{it}))$ is shown in figure A2.

In addition to data on repressions, we also obtained some information on some characteristics of the ethnic groups in our dataset that we will use as covariates in the synthetic control method. In particular, we acquired total population of the ethnic groups and their urbanization rate from 1926 Soviet Census from the Demoscope website.[6] For each ethnic group, we also calculated the cladisitc similarity of its language to Russian based from Glottolog language trees (Hammarström et al., 2018). Cladistic measure of linguistic similarity counts the number of shared branching points between the two nodes on a language tree. It has been used by Fearon (2003) and Dickens (2018) among others. The full data are provided in the table A7 in the appendix.

[6]It is available online at http://www.demoscope.ru/weekly/ssp/ussr_nac_26.php

4 Imputation of Missing Data

4.1 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).

4.1.1 Naive Bayes Classifier

Let $\boldsymbol{x} = (x_1, x_2, x_3)$ be features used for predicting ethnicity, that is a 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

$$p(E_k \mid \mathbf{x}) = \frac{p(E_k)\ p(\mathbf{x} \mid E_k)}{p(\mathbf{x})}, \tag{4.1}$$

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

$$p(E_k \mid \mathbf{x}) = \frac{p(E_k)\ \prod_{i=1}^{3} p(x_i \mid E_k)}{p(\mathbf{x})}. \tag{4.2}$$

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

$$\hat{y} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \; p(E_k) \prod_{i=1}^{3} p(x_i \mid E_k). \tag{4.3}$$

However, one potential issue is that whenever a likelihood of a certain feature is estimated to be 0 then the posterior probability is always 0 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

$$\hat{\theta}_j = \frac{c_j + \alpha}{N + \alpha d} \qquad j = 1, \dots, d \tag{4.4}$$

where parameter $\alpha > 0$ is a smoothing parameter. This ensures that for any finite value of N, $\hat{\theta}_j$ will never be 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).

4.1.2 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 78.7%. However, we are also interested how it varies by ethnicity. For this reason we calculate sensitivity and specificity for each ethnic group.[7] The results, provided in the table A3 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 95% 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.

Therefore 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 actual the 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

$$P_{it} = \alpha_i R_{it} + (N_t - R_{it}) \cdot (1 - \beta_i)\,. \tag{4.5}$$

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

$$R_{it} = \frac{P_{it} - N_t\,(1 - \beta_i)}{\alpha_i + \beta_i - 1}. \tag{4.6}$$

[7] Sensitivity measures the proportion of observations in the class 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).

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 higher proportion of Armenians.

Fortunately, we can address this potential bias by more building complex modeling. 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

$$P_{it} = \sum_{j=1}^{K} b_{ij} R_{jt} \qquad i = 1, \ldots, K. \tag{4.7}$$

This equation can be expressed in matrix form as

$$\mathbf{P}_t = \mathbf{B} \cdot \mathbf{R}_t, \tag{4.8}$$

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

$$\mathbf{R}_t = \mathbf{B}^{-1} \cdot \mathbf{P}_t. \tag{4.9}$$

We will call this method the full matrix adjustment. Compared to the parsimonious adjustment (in equation 4.6), 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 significantly different.

4.2 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 to 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 the figure 1. First, we can see that there is fairly large variance in the variable with number of days ranging from 0 to more than 1000.[8] 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.

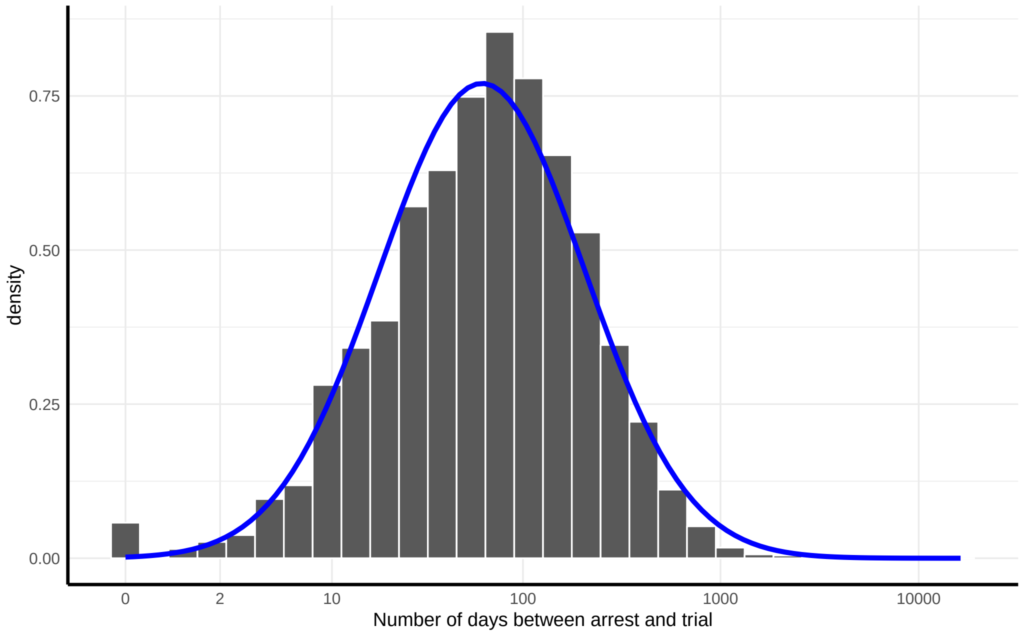
To address this problem, we model the zero and positive values separately in a two-stage process using a method described in Gelman and Hill (2006, p. 537-538). Let y be the number of days between arrest of an individual and his or her trial and X be set of dummy variables indicating the year of trial. 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

$$\Pr\left(I_i^y = 1\right) = \operatorname{logit}^{-1}\left(X_i \alpha\right). \tag{4.10}$$

In the second stage, simple log-linear regression is applied to predict only

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

Figure 1: Histogram for Number of Days between Arrest and Trial



Notes: The x-axis is shown on $\log(1+x)$ scale.

the positive values y^{pos}

$$\log\left(y_i^{\text{pos}}\right) \sim \mathrm{N}\left(X_i\beta, \sigma\right). \tag{4.11}$$

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 the table A6 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 76% 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 $\operatorname{logit}^{-1}\left(X_i\hat{\alpha}\right)$ 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 the figure A3 in the appendix. The resulting distribution highly resembles the distribution in the figure 1 including the fraction of zero values indicating our model captures the actual data fairly well.

Nevertheless, one issue 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 between the years 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.[9] 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 the figure A4 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.

[9]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.

5 Methodology

5.1 Difference-in-differences

We employ difference-in-differences design to estimate the effect of the changing geopolitical relations on repressions of Germans in the USSR. Our main specification is the dynamic difference-in-differences model:

$$\log\left(1+y_{it}\right)=\sum_{k=1922}^{1960}\beta_k\,\mathrm{German}_i\cdot\mathrm{Year}_t^k+\lambda_t+a_i+\mathrm{Relations}_{it}+u_{it},\quad(5.1)$$

where y_{it} is number of arrests of people with ethnicity i in year t (from 1921 to 1960), λ is year fixed effect, a is ethnicity fixed effect (both captured by respective dummy variables) and Year_t^k are dummy variables that equals 1 if its year k equals to t and 0 otherwise. Prior to 1933 they capture the lead (anticipatory) effects used to test if pre-treatment trends are parallel. After 1933 they capture the dynamic lagged effects. The time-varying vector $\mathrm{Relations}_{it}$ is set a of dummy variables that capture other major changes in relations of a state with core group i with the USSR during this period. The list of these changes is provided in appendix in the table 6.

We apply logarithmic transformation on y_{it} since it better fits the data. We use $\log\left(1+y_{it}\right)$ because some observations (although not many) have $y=0$. As discussed in Wooldridge (2015, p. 193), the percentage change interpretation is usually closely preserved (except for changes beginning at 0 which are not of great interest to us).

In additional specifications used to test sensitivity of results, we also include the terms $E_i\cdot t$ and $E_i\cdot t^2$ that can capture ethnicity-specific time trends. This allows for a certain deviation from the parallel trends if it can be described by the specified (quadratic or linear) function (Angrist and Pischke, 2009, chapter 5). However, a potential issue with the ethnicity-specific time trends is that they could absorb part of the treatment effect if it increases over time (Meer and West, 2016) which is why we do not include them in our default specification.

The identifying assumption for the specifications 5.1 is that the number of arrest of Germans after 1933 would go in parallel to arrests of other minorities in the absence shock to German-Soviet relations conditional on our control variables.[10] Although we cannot test this assumption, we can test whether the trends were parallel prior to 1933 (pre-treatment) which could increase our confidence that they were parallel after 1933 too. This can be testing if the coefficients β_k on the lead effects are significantly different from zero.

As Bertrand et al. (2004) show, the usual standard errors are downward-biased for most DiD regressions since they do not account for the serial correlation within the units of interests (states, countries etc.). A common solution to this problem is to estimate standard errors using robust covariance matrix that allows for clustering (i.e. cluster-robust standard errors). However for small number of groups (generally less than 40), the cluster-robust standard errors are downward-biased and not reliable. Angrist and Pischke (2009, chapter 8) suggest taking the maximum of cluster-robust as a simple rule of thumb to avoid gross misjudgements in precision. More rigorous solutions are cluster bootstrapping (Cameron et al., 2008; Cameron and Miller, 2015) and using t-distribution with $G - K$ degrees of freedom (where G is number of clusters and K number of parameters) rather than the standard Normal distribution (McCaffrey and Bell, 2002; Imbens and Kolesár, 2016). Since we have small number of groups we use the generalization of McCaffrey and Bell (2002) correction to models with arbitrary sets of fixed effects by Pustejovsky and Tipton (2018).

5.2 Synthetic Control Method

However, difference-in-differences makes a fairly strong assumption regarding the absence of time-varying individual-specific heterogeneity which might be unrealistic in some empirical settings. The synthetic control method has been increasingly applied in the economic literature to overcome this issue (Abadie and Gardeazabal, 2003; Abadie et al., 2010; Billmeier and Nan-

[10] In the presence of linear ethnicity-specific time trends, the identifying assumption is the parallel growth of the oucome variable of control and treatment group. This is equivalent to the parallel trends for the first differences of the outcome (Mora and Reggio, 2017).

nicini, 2013; Cavallo et al., 2013). The method works by constructing a synthetic version of the treated unit from the control units based on matching of pre-treatment variables. The outcomes of synthetic control are then compared to the actual outcomes to estimate the treatment effect. Unlike difference-in-differences, the synthetic control method allows for the presence of certain time-varying unit-specific confounders although only if they can be captured by the factor model in the equation 5.3.

Let Y_{it} be the outcome of a unit i at time t with $i = 1$ being the treated group and $i \in \{2, \ldots, J+1\}$ untreated units (we will also call them donor units). We denote D_{1t} as the treatment dummy, i.e. variable that equals 1 if $i = 1$ and $t > T_0$ and 0 otherwise (with T_0 being the start of the treatment). Let be Y_{1t}^N be a counterfactual outcome for the treated unit at time $t > T_0$ in the absence of treatment. The observed outcome Y_{it} is then assumed to be sum of the counterfactual outcome Y_{1t}^N and the effect of treatment at time t, α_{1t}, i.e.:

$$Y_{1t} = Y_{1t}^N + \alpha_{1t} D_{1t}. \tag{5.2}$$

The synthetic control method also assumes that Y_{1t}^N can be expressed by the following factor model:

$$Y_{1t}^N = \delta_t + \boldsymbol{\theta}_t \boldsymbol{Z}_i + \boldsymbol{\lambda}_t \boldsymbol{\mu}_i + \epsilon_{it}, \tag{5.3}$$

where is δ_t an unknown common factor with constant factor loadings across units, $\boldsymbol{Z}_i$ is a $(1 \times r)$ vector of observed time-invariant covariates (unaffected by the treatment), $\boldsymbol{\theta}_t$ is a $(1 \times r)$ vector of unknown parameters, $\boldsymbol{\lambda}_t$ is a $(1 \times F)$ vector of unobserved time-varying factors, $\boldsymbol{\mu}_i$ is an $(F \times 1)$ vector of unknown factor loadings and ϵ_{it} is the error term with zero mean.

Notice that if $\boldsymbol{\lambda}_t$ is constant for all t, we get the traditional difference-in-differences with time and unit-specific fixed effects. The synthetic control method thus offers more general model that, in contrast to difference-in-differences, allows the unobserved confounders vary with time.

We then try to estimate the counterfactual outcome Y_{1t}^N by constructing a synthetic control group as a convex combination of available comparison units (in our case other ethnic groups in the USSR) that most closely resembles the pre-treatment characteristics of the treated group (or more precisely, for which the average of its factor loadings $\boldsymbol{\mu}_i$ match the factor loadings of the treated unit $\boldsymbol{\mu}_1$).

More formally, we choose a vector of weights $W = (w_2, \ldots, w_J, w_{J+1})$ subject to $w_j \geq 0$ for $j = 2, \ldots, J, J+1$ and $w_2 + \cdots + w_J + w_{J+1} = 1$ that minimize $\|X_1 - X_0 W\|_V = \sqrt{(X_1 - X_0 W)^T V (X_1 - X_0 W)}$ where $X_1 = (Z_1, Y_1^{K_1}, \ldots, Y_1^{K_L})$ is a $(k \times 1)$ vector of pre-treatment characteristics of the treated unit, Y^{K_l} are linear combinations of pre-treatment outcomes, and X_0 is a $(k \times J)$ matrix with pre-treatment characteristics of untreated units analogous to X_1, V is a $(k \times k)$ matrix that weights the importance of different pre-treatment predictors in the minimization problem. V is usually chosen among symmetric and positive semidefinite matrices to minimize the mean squared prediction error (MSPE) in the outcome in the pre-treatment period so that the V-weights would reflect the predictive power of the covariates. The effect of the treatment α_{1t} at time $t > T_0$ is then estimated as a difference between the outcome for the synthetic control and the treated unit, that is:

$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}, \tag{5.4}$$

where w_j^* are the estimated optimal weights.

The synthetic control method, however, does not provide us with any standard errors that could measure the uncertainty in the estimates. Abadie et al. (2010) propose assessing significance using placebo tests and randomization inference. Synthetic control method is applied iteratively to every unit in the donor pool as if they were treated and the results of placebo test are then compared to the treated unit. If the estimated effect for the treated unit is much larger than the placebo effects, it implies that the effect is likely significant since such a result would not be likely under the null hypothesis of zero treatment effect.

However, Abadie et al. (2010) point out that placebo synthetic controls with poor pre-treatment fit do not provide good comparison for estimating rareness of large effect for a treated unit with a good pre-treatment fit. They thus recommend excluding excluding placebo units with substantially higher pre-treatment MSPE relatively to the treated unit.

Nevertheless, the choice of any level of the cutoff of pre-treatment MSPE is somewhat arbitrary. Better way to asses significance of results might be to compare the ratios of post/pre-treatment MSPE take into account different values of pre-treatment fit between the treated unit and the placebo tests. First, the MSPE ratio for unit j is defined as

$$\text{MSPE ratio}_j = \frac{\sum_{t=T_0+1}^{T}\left(Y_{jt} - \hat{Y}_{jt}^{N}\right)^2}{\sum_{t=1}^{T_0}\left(Y_{jt} - \hat{Y}_{jt}^{N}\right)^2} \tag{5.5}$$

where $\hat{Y}_{jt}^{N}$ is estimated synthetic control for unit j at time t. The p-values can then be calculated as

$$p_1 = \frac{\sum_{j=2}^{J+1} \mathbb{1}\left(\text{MSPE ratio}_j \geq \text{MSPE ratio}_1\right)}{J+1}. \tag{5.6}$$

where $\mathbb{1}$ is an indicator function that equals 1 if the equation in its argument is true and 0 otherwise.

6 Results

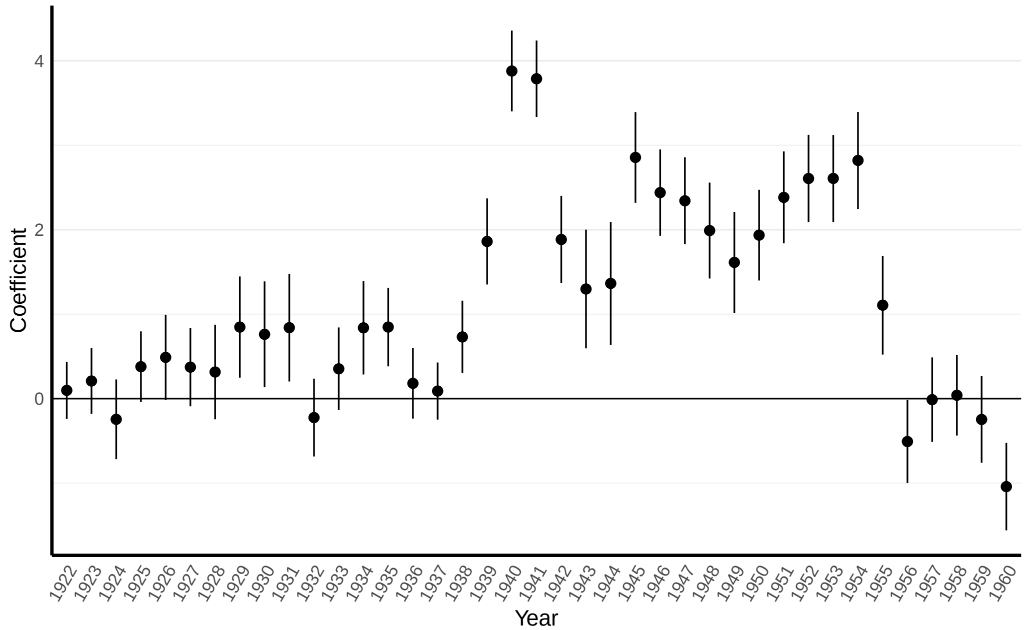
6.1 Difference-in-differences

We first present results from our preferred specification 5.1 with geopolitical controls and no ethnicity-specific time trends. We imputed missing ethnicity and date of arrest as described in the section 4. Full matrix adjustment from equation 4.9) was applied to the ethnicity imputations.

The estimated coefficients β_k from the dynamic difference-in-differences model (5.1) are plotted in the figure 2. Recall that our dependent variable is $\log(1 + \text{arrests}_{it})$ and therefore we can interpret the coefficients β_k as percentage change in arrests. We can draw several conclusions from the results. First, the coefficients from 1933 to 1938 are all smaller than 1 and most of them are not significantly different from zero which implies that hostility between Germany and the Soviet Union in this period does not seem to impact the repressions of Germans contrary to the theoretical predictions. Second, the political arrests of Germans start to rise in 1939 which is surprising given that Molotov-Ribbentrop pact guaranteeing neutrality Germany and the USSR was signed that year. The repressions then peak in 1940 and 1941 at 4% followed by sharp drop in years 1942-1944 to approximately 1.5%. However, we have to be careful when interpreting those coefficients. Since the Soviet Union was at that period at war and initially lost large amounts of territory, it is plausible that the arrests of Germans declined simply because there were fewer Germans on the territory controlled by the Soviets. In any case, the impact of war on the repressions appear to be highly persistent as the estimated effect stays high at around 2 to 3% even after 1945. Finally, starting from 1955, the coefficients rapidly fall to zero. This period coincides with the partial relaxation of repressions and censorship following the death of Stalin in 1953 and the subsequent rise of power of Khrushchev. This is reflected in our data where the number of arrested Germans in a given year after 1954 does not exceed 50 compared to hundreds of arrests in the preceding years.

However, the pre-1933 coefficients in the figure 2 give us some reason to

Figure 2: Estimates of β_k from the Specification 5.1



Notes: Ethnicity and date of arrest were imputed. Full matrix adjustment was applied on ethnic group imputations. All 38 ethnic groups are included. There are no ethnicity-specific time trends. 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.

doubt the validity of our model. Even though they are small in size relative to the post-1939 coefficient, they are all significantly different from 0 at 5% level. This provide some evidence that the pre-treatment trends for German minority were not parallel with trends for other ethnic groups. We can thus suspect that the post-treatment trends were not parallel either which would violate the basic identifying assumption of difference-in-differences. We address this problem by applying the synthetic control method in the next section.

Another potential issue is that some ethnic groups changed their treatment status in this period in various complicated ways. We tried to control for this by including a set of dummy variables that capture the most important changes in geopolitical relations (more on their definition in appendix). However, these dummy variables might miss more subtle changes in relations with the USSR. As a robustness check, we therefore exclude every ethnicity which constituted a core group in some independent state in the interwar period from the dataset (except for Germans, of course) and reestimate the model. This criterion removes 10 ethnic groups from the dataset. The full list of them is provided in the table A7. The results of the dynamic difference-in-differences are plotted in the figure A5 in the appendix. The estimated coefficients change very little compared to the results with all ethnic groups included, only confidence intervals appear slightly wider since we have fewer observations. Therefore our previous results are maintained.

In the section 7, we perform several additional robustness checks to further asses the sensitivity of our findings. In particular, we refit the models for different ethnicity-specific time-trends and imputation adjustments.

6.2 Synthetic Control Method

We implemented the synthetic control method in R using the MSCMT package (Becker and Klößner, 2018). Our outcome variable is again $\log(1 + \text{arrests})$. As in the difference-in-difference, we include all 38 ethnic groups and we impute missing date of arrest and ethnicity (which we adjust using the full-matrix correction). In our baseline model, the outcomes for all pre-treatment years (1921-1932) were included as predictors. This approach has been widely used in the literature (Billmeier and Nannicini, 2013; Cavallo et al., 2013; Bohn et al., 2014) and in contrast to other methods (such as using only mean of pre-treatment outcomes) it has the advantage of reducing opportunities of specification search (Ferman et al., 2018). We also include three time-invariant covariates that might potentially be predictive of post-1933 repressions: total population of the ethnic group in the USSR, its urbanization rate (both taken from the 1926 Soviet census), and linguistic similarity to Russian. However, including all pre-treatment outcomes as predictors renders other covariates unimportant in the optimization procedure (Kaul et al., 2018). On the other hand, Botosaru and Ferman (2019) argue that if there is a long set of pre-treatment outcomes (which is our case) then a perfect balance on covariates should not be required. Since there is no clear consensus in the literature, we apply both methods and show the synthetic control method with mean of the pre-treatment outcome as predictor

Table 4: Synthetic German minority weights

Ethnic group	W-Weight
Russian	0.36
Greek	0.28
Finnish	0.17
Lithuanian	0.08
Khakas	0.05
Yakut	0.05
Bulgarian	0.01

in section 7 as a robustness check in addition to our baseline specification with all pre-treatment outcomes as predictors that we present below.

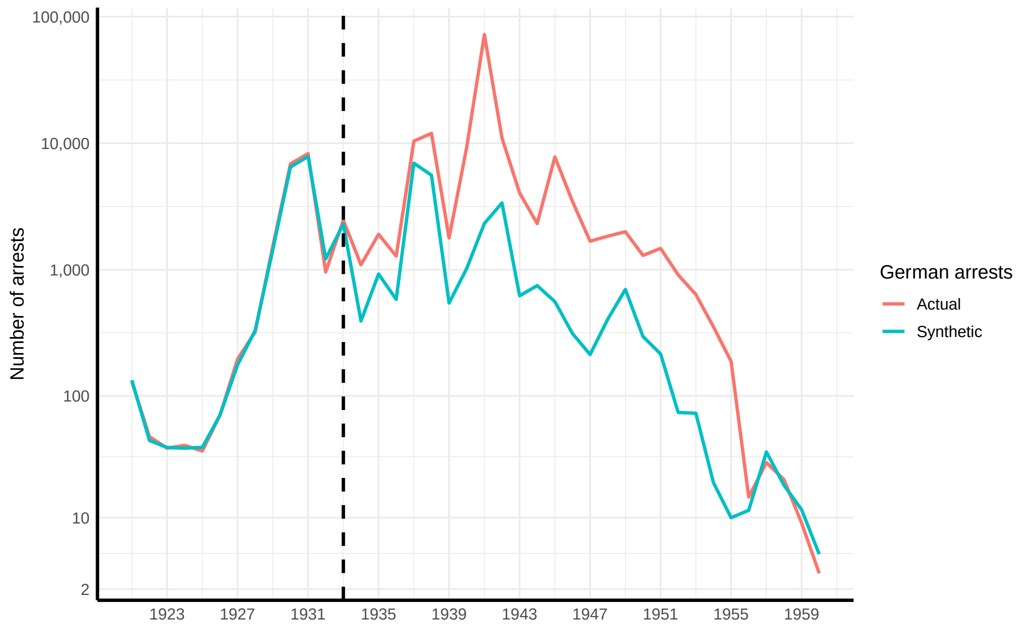
The calculated optimal weights W of ethnic groups in the synthetic German minority are provided in the table 4 (ethnic groups with zero weight are not shown). The highest contribution in the synthetic German minority have Russians, Greeks and Finns with weights 0.36, 0.28, and 0.17 respectively. Lithuanians, the Khakas, Yakuts, and Bulgarians are also represented in the synthetic control although only with weights smaller than 0.08.

Figure 3 shows the arrests of the German minority and its synthetic control. The synthetic control fits the actual pre-treatment values reasonably well. In the post-treatment period, both time series follow similar general trends (rise in 1937-1938 and decline after 1945). However up until 1955, the actual arrests of Germans are consistently higher than the predictions of synthetic control. We can infer the estimated effect size in the post-1933 period from the figure 4a which shows the difference between the actual arrest and their synthetic counterparts for each year (on $\log(1+x)$ scale).

From 1933 to 1939, this gap is close to 1. For the period from 1941 to 1955, the estimated effect is higher but also more volatile oscillating between 1 and 3 corresponding to 1 to 3% increase in arrests of Germans in that period. After 1955, the gap shrinks to zero. The results are similar to difference-in-differences in the overall trends. Nevertheless, the estimated effects by the the synthetic control are slightly smaller than the their difference-in-differences equivalents.

We estimate the significance of our results with placebo tests by applying

Figure 3: Comparison plot



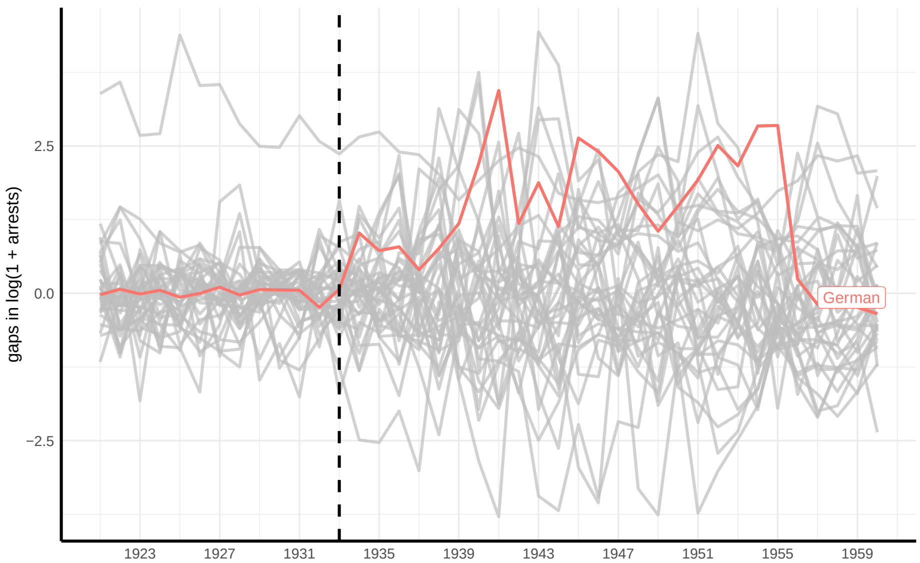
Notes: The values on y-axis are shown on $\log(1 + y)$ scale. Ethnicity and date of arrest were imputed. Full matrix adjustment was applied on ethnic group imputations. All 38 ethnic groups are included.

the same procedure to every ethnic group in the dataset. The differences between the actual values and the respective synthetic control for every placebo ethnicity are plotted in grey in the figure 4a. Even in comparison with the placebo tests, the estimated effect for German ethnicity still appears relatively large for the period from 1939 to 1955 although for years 1933-1938 the effect for Germans seems within the standard range of the placebo gaps.

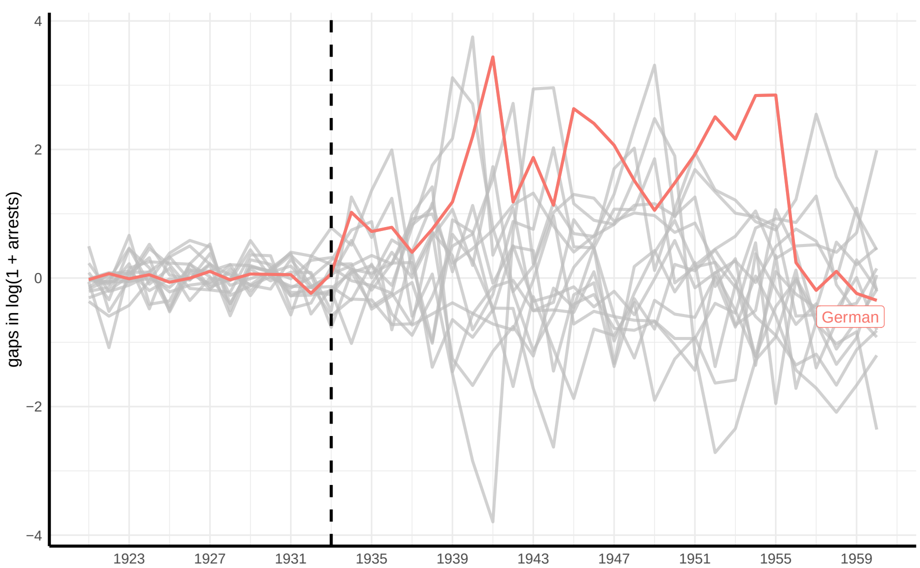
The figure 4a also shows that for some ethnic groups the pre-treatment gaps are large. For example, the gap for Russians stays above 2.5 for the whole pre-1933 period. This indicates that synthetic control of these ethnic groups does not capture the actual pre-treatment trends well. As we explained in the subsection 5.2, the placebo synthetic controls with with substantially worse pre-treatment then the treated unit should not be used in estimating significance of the treatment effect. We thus exclude the placebo ethnicities whose pre-treatment mean squared prediction error (MSPE) is 20 times higher than the same measure for German minority. Even though this is relatively lenient cutoff, it removes 11 ethnic groups that does not meet the criterion. The resulting plot is shown in the figure 4b. The post-1933 gaps in German arrests now stand out more clearly and even the gap for the years 1933-1938 now appears relatively significant.

Figure 4: Gaps between synthetic control and actual values for placebo tests

(a) All ethnic groups



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



Nevertheless, our preferred approach for assessing uncertainty in results that avoids choice of any arbitrary level of the pre-treatment MSPE cutoff for the exclusion of poorly fit placebos is to compare the ratios of post/pre-treatment MSPE. The values of these ratios for all ethnic groups are displayed in the figure 5a for the whole post-treatment period. The MSPE ratio for the German minority is by far the highest. The probability of German minority having the highest ratio of all under the null hypothesis of zero treatment effect is $1/38$ (≈ 0.026). In the figure 5b, we also provided the MSPE ratios with the post-treatment MSPE calculated only for the years 1933-1939 to estimate the significance just for this period. Although the gap between MSPE ratio for German minority and other ethnic groups shrunk somewhat, the German MSPE ratios stays the highest and therefore we again get the p-value of $1/38$. These results contradict our inferences from difference-in-differences where we could not reject the hypothesis of zero treatment effect for the period from 1933 to 1939.

7 Robustness Checks

7.1 Difference-in-differences

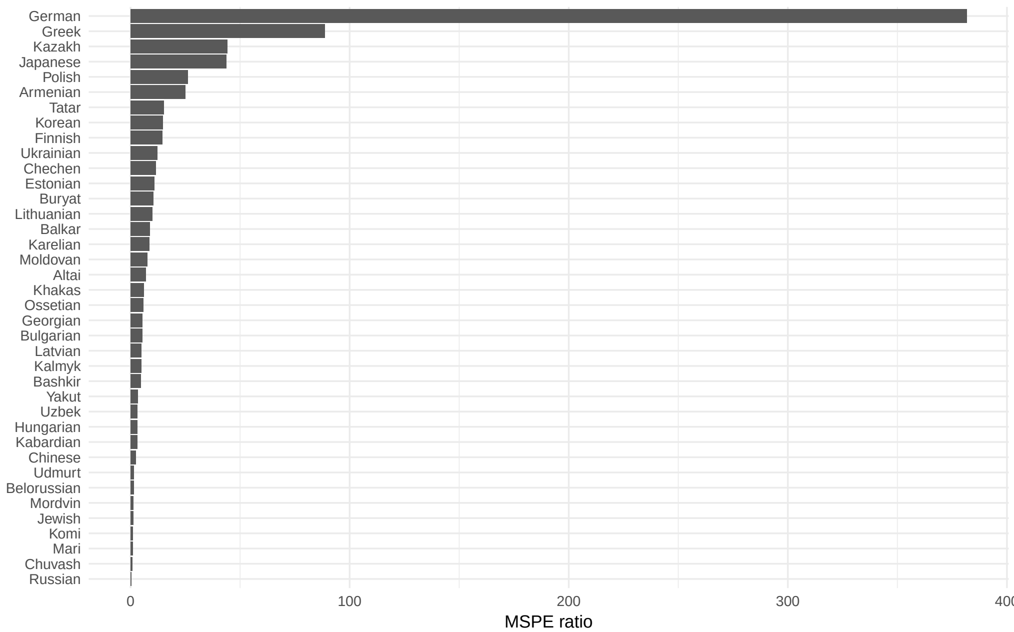
Our results have shown rise of arrests of Germans during and after the war. However, the question is if these people were arrested for aiding the invading German army or only based on their ethnicity. Since the Russia's 1991 Law "On Rehabilitation of Victims of Political Repression" specified that individuals who joined the German army[11] were not eligible for rehabilitation, we can filter out the cases of direct cooperation by restricting our analysis to only rehabilitated victims (Frierson, 2014). In the Memorial database, there are 1 257 796 individuals classified as rehabilitated (for rest of the data we either do not have the information on rehabilitation or they were not eligible for rehabilitation). We thus re-estimate our baseline specification using only those observations. The results are plotted in the figure A6. We see that the coefficients after 1933 change only little. Nonetheless, the pre-treatment estimates deviate from zero more.

Second, we consider if the inclusion of ethnicity-specific time trends affects the results. The estimates regressions with quadratic , linear, and no time (default) trends are plotted in the figure A7 in the appendix. The coefficients

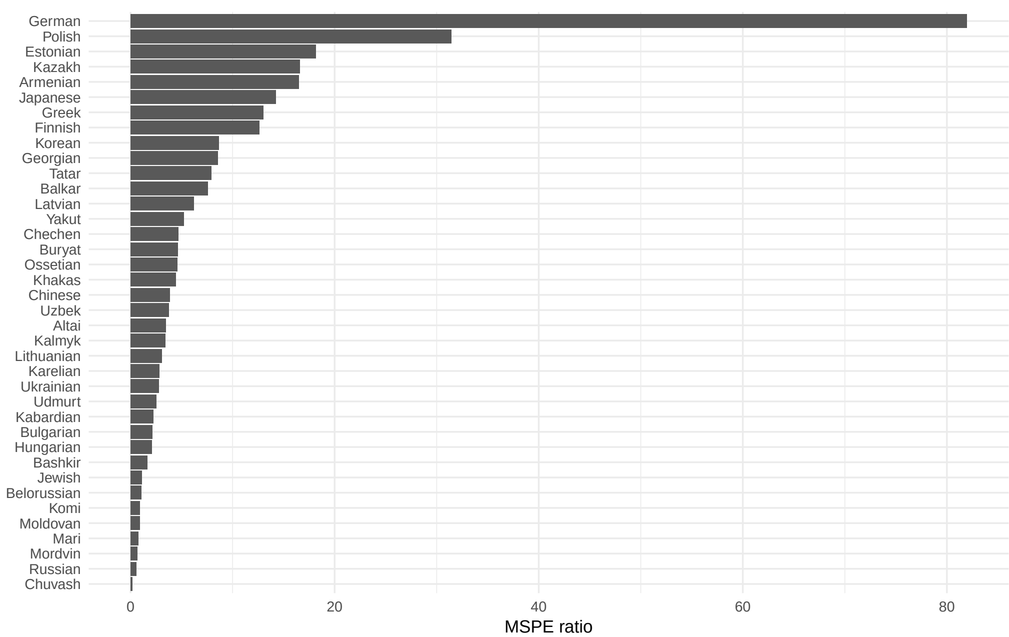
[11] Subsequent amendments expanded the restriction to anyone who aided the German army.

Figure 5: 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



for the model with no ethnicity-specific time trends are somewhat lower for the baseline model and the estimates for linear time trends are lower still. Nonetheless, the main general pattern is preserved in all specifications.

Third, we test the sensitivity of results to different adjustments of ethnicity imputation. As we explain in the subsection 4.1.2, the adjustments were applied in order to correct for the unbalanced accuracy in prediction of our Naive Bayes classifier across ethnic groups. The results from fitting our default specification to data with different adjustments are shown in the figure A8. We see that the estimates for the full matrix (used by default) and the parsimonious adjustment are virtually the same. When no adjustment is applied, the coefficients even get slightly larger.

7.2 Synthetic Control Method

We construct a synthetic control using mean of the outcome in the pre-1933 period instead of including outcome for every year as we did in our baseline model. This increases the importance of the time-invariant covariates (population, urbanization rate, and linguistic similarity) that are also included as predictors.

Table 5: Pre-treatment Predictor Means

	German minority			
Variable	Actual	Synthetic	Mean of all ethnicities	V weights
Log(1 + arrests)	5.59	5.48	3.649	0.15
Total population	1 238 549.00	1 747 879.23	3 681 250.421	0.84
Urbanization rate	14.92	21.56	17.844	0.00
Ling. similarity to Russian	1.00	1.16	0.763	0.00

Note:
Log(1 + arrests) is averaged over the pre-treatment period (1921-1932). All other predictor are time-invariant. Total population and urbanization rate are taken from 1926 Soviet census.

Recall that the predictors weights V are chosen to minimize pre-treatment MSPE. We can thus infer from the calculated weights V (provided in the table 5) that neither urbanization rate nor linguistic similarity to Russian are good predictors of pre-1933 repressions. Therefore the W-weights of ethnic groups in the synthetic control are chosen to mainly match the German

Table 6: Synthetic German minority weights

Ethnic group	W-Weight
Tatar	0.49
Polish	0.39
Korean	0.12

minority on population and the mean of $\log(1 + \text{arrests})$. The calculated W-weights are shown in the table 6. We see that Tatar and Polish minorities are now contributing with the largest share to the synthetic control.

When we compare the gaps between synthetic control and actual value (figure A9) with our baseline model, we notice three main differences. First, pre-treatment fit for Germany is substantially worse which is the the cost assigning greater importance to other covariates by using only the pre-treatment mean of the outcome. Second, the gap for the years 1933-1939 is slightly smaller. Finally, there is large drop in the estimated effect in the year 1944. This most likely a consequence of the sharp increase in repressions of Tatars in 1944 following their mass deportations from Crimea that year. Nonetheless, both our baseline model and its robustness check otherwise follow very similar trends. The p-value of the effect for the whole period 1933-1960 is again $1/38$ (≈ 0.026). However if we consider only the time window of 1933-1939, the p-value increases to $1/19$ (≈ 0.105). The MSPE ratios for all ethnic groups based on which the p-values were calculated are provided in the figure A10.

Furthermore, we apply our baseline synthetic control specification (with all pre-treatment outcomes as predictors) but only limiting ourselves to ethnic groups without independent state (e.g. Armenians, Kazakhs, etc.). The resulting synthetic control is weighted average of repressions of Tatars, Koreans , Ukrainians, and Jews as shown in the table A8. The estimated effects shown in the figure A11a are similar to the baseline specification. Nonetheless, the pre-treatment fit is slightly worse. As a consequence, German minority has only the second highest MSPE ratio (taken for he whole post-1933 period) as shown in the figure A12a and thus the corresponding p-value is $2/28$ (≈ 0.074).

Finally, we apply again the same baseline synthetic control method but only to data with rehabilitated individuals. The estimated W-weights are provided in the table A9. The trends in the gaps between the actual and synthetic repressions (plotted in the figure A11b) are broadly consistent with our previous estimates. The German MSPE for the whole post-treatment period is the highest of all (provided in the figure A12b) and therefore the implied p-value is again $1/38$ (≈ 0.026).

Conclusion

We used difference-in-differences and the synthetic control method to test how changińg geopolitical relations between Soviet Union and Germany affected repressions of Germans by the NKVD. Both methods provide evidence that the war significantly increased the arrests of Germans. Specifically, the models estimate 2% to 3% rise in the average of arrests during the war even though are some year-to-year deviation.

Furthermore, we find that the increased repression persist almost undiminished for at least 10 year after the end of the war when the security concerns are no longer present. This suggests that use of violence by the state might be largely driven by out-group hostility rather than the strategic considerations emphasized in the literature. The strong and long persistence of the hostile attitudes after the war could potentially help explain the phenomenon of conflict trap (i.e. why violence tends to reoccur in the same places). However, our methods do not enable us to determine the underlining mechanism. It could be the bias of rank and file officers of the secret police or some directives from the top.

For the period of hostilities (but not war) from 1933 to 1939, we get conflicting results. Whereas in the difference-in-differences we cannot reject the null hypothesis, our baseline synthetic control model implies positive and statistically significant effect. Nevertheless, even for the synthetic control, the estimated effect is fairly small corresponding to about 1% rise in the repressions of Germans for the years 1933-1939.

However, we also have to be aware of the limitations of this study. First, the standard errors in our results might be slightly underestimated since they do not take into account the uncertainty in the imputed values. Second, to correctly estimate the treatment effect, we have to assume that the treatment has no spillovers on the control units. Yet, it is fairly plausible that circumstances of war with Germany could increase repressions of other minorities as well. Nonetheless, since we would probably expect these spillovers to be positive, our estimates would in that case be biased downward.

References

Abadie, A., Diamond, A. and Hainmueller, J. (2010) "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program", *Journal of the American statistical Association* 105 (490), pp. 493–505.

Abadie, A. and Gardeazabal, J. (2003) "The Economic Costs of Conflict: A Case Study of the Basque Country", *American Economic Review* 93 (1), pp. 113–132, DOI: 10.1257/000282803321455188.

Angrist, J.D. and Pischke, J.-S. (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press.

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

Becker, M. and Klößner, S. (2018) "Fast and reliable computation of generalized synthetic controls", *Econometrics and Statistics* 5, pp. 1–19.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004) "How much should we trust differences-in-differences estimates?", *The Quarterly Journal of Economics* 119 (1), pp. 249–275.

Billmeier, A. and Nannicini, T. (2013) "Assessing Economic Liberalization Episodes: A Synthetic Control Approach", *The Review of Economics and Statistics* 95 (3), pp. 983–1001, DOI: 10.1162/REST_a_00324.

Blaydes, L. (2018) *State of Repression: Iraq under Saddam Hussein*, Princeton: Princeton University Press.

Bohn, S., Lofstrom, M. and Raphael, S. (2014) "Did the 2007 Legal Arizona Workers Act Reduce the State's Unauthorized Immigrant Population?", *Review of Economics and Statistics* 96 (2), pp. 258–269, DOI: 10.1162/REST_a_00429.

Botosaru, I. and Ferman, B. (2019) "On the Role of Covariates in the Synthetic Control Method", *The Econometrics Journal* (Forthcoming), DOI: 10.1093/ectj/utz001.

Butt, A.I. (2017) *Secession and Security: Explaining State Strategy against Separatists*, 1 edition, Ithaca: Cornell University Press.

Cameron, A.C., Gelbach, J.B. and Miller, D.L. (2008) "Bootstrap-based improvements for inference with clustered errors", *The Review of Economics and Statistics* 90 (3), pp. 414–427.

Cameron, A.C. and Miller, D.L. (2015) "A practitioner's guide to cluster-robust inference", *Journal of Human Resources* 50 (2), pp. 317–372.

Cavallo, E., Galiani, S., Noy, I. and Pantano, J. (2013) "Catastrophic Natural Disasters and Economic Growth", *The Review of Economics and Statistics* 95 (5), pp. 1549–1561, DOI: 10.1162/REST_a_00413.

Davenport, C. (2007) *State Repression and the Domestic Democratic Peace*, New York: Cambridge University Press.

Dickens, A. (2018) "Ethnolinguistic favoritism in african politics", *American Economic Journal: Applied Economics* 10 (3), pp. 370–402.

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

Domingos, P. and Pazzani, M. (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.

Evera, S.V. (1994) "Hypotheses on Nationalism and War", *International Security* 18 (4), pp. 5–39.

Fearon, J.D. (2003) "Ethnic and cultural diversity by country", *Journal of economic growth* 8 (2), pp. 195–222.

Ferman, B., Pinto, C. and Possebom, V. (2018) *Cherry picking with synthetic controls*, Working Paper, [Online]. Available at: \url{example.com} (Accessed: 10 January 2013).

Frierson, C.A. (2014) *Russia's Law 'On Rehabilitation of Victims of Political Repression': 1991-2011, An Enduring Artifact of the Dismantling of the Soviet Regime, Transitional Justice, and the Aspiration for a Rule of Law State*, tech. rep., National Council for Eurasian and East European Research.

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

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

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

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

Greitens, S.C. (2016) *Dictators and their Secret Police: Coercive Institutions and State Violence*, Cambridge: Cambridge University Press.

Hammarström, H., R. Forkel and M. Haspelmath, eds. (2018) *Glottolog 3.3*, Jena: Max Planck Institute for the Science of Human History.

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

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

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

Hofstra, B., Corten, R., Tubergen, F. van and Ellison, N.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, B. and Schipper, N.C. de (2018) "Predicting ethnicity with first names in online social media networks", *Big Data & Society* 5 (1), DOI: 10.1177/2053951718761141.

Imbens, G.W. and Kolesár, M. (2016) "Robust Standard Errors in Small Samples: Some Practical Advice", *The Review of Economics and Statistics* 98 (4), pp. 701–712, DOI: 10.1162/REST_a_00552.

Kaul, A., Klößsner, S., Pfeifer, G. and Schieler, M. (2018) *Synthetic control methods: Never use all pre-intervention outcomes together with covariates*, Working paper.

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

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

Lupu, N. and Peisakhin, L. (2017) "The legacy of political violence across generations", *American Journal of Political Science* 61 (4), pp. 836–851.

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

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

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

McCaffrey, D.F. and Bell, R.M. (2002) "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples", *Survey Methodology* 28 (2), pp. 169–181.

McNamee, L. and Zhang, A. (2019) “Demographic Engineering and International Conflict: Evidence from China and the Former USSR”, *International Organization* (forthcoming).

Meer, J. and West, J. (2016) “Effects of the Minimum Wage on Employment Dynamics”, *Journal of Human Resources* 51 (2), pp. 500–522.

Memorial (2017) “Zhertvy politicheskogo terrora v SSSR [victims of political terror in the USSR]”.

Mora, R. and Reggio, I. (2017) “Alternative diff-in-diffs estimators with several pretreatment periods”, *Econometric Reviews* 0 (0), pp. 1–22, DOI: 10.1080/07474938.2017.1348683.

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

Mylonas, H. (2013) *The Politics of Nation-Building: Making Co-Nationals, Refugees, and Minorities*, Cambridge: Cambridge University Press.

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

Pustejovsky, J.E. and Tipton, E. (2018) “Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models”, *Journal of Business & Economic Statistics* 36 (4), pp. 672–683, DOI: 10.1080/07350015.2016.1247004.

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

Rozenas, A., Schutte, S. and Zhukov, Y. (2017) “The political legacy of violence: The long-term impact of Stalin’s repression in Ukraine”, *The Journal of Politics* 79 (4), pp. 1147–1161.

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

Svolik, M.W. (2012) *The Politics of Authoritarian Rule*, Cambridge: Cambridge University Press.

Toft, M.D. (2005) *The Geography of Ethnic Violence: Identity, Interests, and the Indivisibility of Territory*, Princeton: Princeton University Press.

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

Walter, B.F. (2009) *Reputation and Civil War: Why Separatist Conflicts Are So Violent*, 1 edition, Cambridge: Cambridge University Press.

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

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

Wintrobe, R. (1998) *The Political Economy of Dictatorship*, Cambridge: Cambridge University Press.

Wooldridge, J.M. (2015) *Introductory econometrics: A modern approach*, Nelson Education.

Zhukov, Y.M. and Talibova, R. (2018) "Stalin's terror and the long-term political effects of mass repression", *Journal of Peace Research* 55 (2), pp. 267–283.

List of tables and figures

List of Tables

List of Figures

Appendix

Every table and graph in this thesis, 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`. If you have any troubles with replications or any other questions, please email me at martin.kosiik@gmail.com.

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 5.1) are provided in the table A1. 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 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 ethnicity with independent states from analysis.

Table A1: 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 A2: Total arrest by ethnicity, 1921-1960

	Reference		
Ethnicity	Only Labeled	Labeled + Unadj. Imputation	Labeled + Adj. Imputation
Russian	550 280	1 064 596	1 069 379
Belorussian	67 613	85 517	72 979
Polish	61 221	85 258	79 742
German	60 798	168 419	169 955
Ukrainian	54 403	91 814	97 042
Kazakh	37 125	46 540	43 541
Tatar	32 095	72 417	71 351
Jewish	31 050	43 710	42 613
Latvian	15 444	21 628	18 796
Chinese	9 693	11 506	10 466
Estonian	9 402	15 561	13 380
Chuvash	8 910	14 930	26 520
Bashkir	8 428	17 876	18 615
Finnish	8 337	14 594	13 550
Mordvin	6 011	12 682	20 642
Buryat	5 679	6 735	6 715
Mari	5 383	7 482	12 288
Lithuanian	4 651	5 474	5 522
Karelian	4 174	9 941	5 379
Korean	4 060	8 821	11 560
Komi	3 613	5 834	4 281
Ossetian	3 237	3 724	3 419
Udmurt	3 082	4 454	5 566
Armenian	2 937	4 850	4 674
Kabardian	2 733	4 438	4 021
Greek	2 246	24 500	25 514
Khakas	2 221	8 137	6 136
Altai	1 894	2 477	2 471
Georgian	1 621	3 049	1 993
Yakut	1 544	2 909	1 572
Moldovan	1 392	2 765	2 719
Kalmyk	1 293	2 168	2 059
Japanese	1 231	14 571	10 821
Uzbek	1 061	4 044	7 470
Hungarian	1 018	1 611	1 119
Bulgarian	1 015	2 479	1 904
Balkar	861	4 740	3 423
Chechen	696	8 508	11 548

Table A3: 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 A4: Descriptive Statistics of Arrests from 1921 to 1960 by Ethnicity, Part 2

	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	1 744
Armenian	55	112	0	524	2 210	63	127	0	614	2 514
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	794
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	3 234	6 483	183	697	0	3 411	7 316
Georgian	30	69	0	320	1 220	38	82	0	370	1 515
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	2 878	9 669
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	514	0	2 969	5 887
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	663	1 922
Komi	85	189	0	1 137	3 395	101	226	0	1 359	4 052
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	255	1 324
Mordvin	130	258	0	1 377	5 197	157	314	0	1 715	6 263
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	14 808	33 668	54	196 315	592 305
Tatar	688	1 406	0	6 275	27 539	764	1 562	0	7 098	30 560
Udmurt	71	134	0	779	2 857	77	145	0	848	3 067
Ukrainian	1 160	2 668	10	14 694	46 384	1 329	3 025	12	16 820	53 179
Uzbek	26	58	0	268	1 059	69	157	0	746	2 752
Yakut	36	60	0	280	1 437	46	73	0	340	1 860

Table A5: 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	160	0	950	2 477
Armenian	73	137	0	648	2 937	121	182	0	829	4 850
Balkar	22	63	0	374	861	118	225	0	916	4 740
Bashkir	211	509	0	2 102	8 428	447	1 150	0	6 108	17 876
Belorussian	1 690	3 434	5	19 491	67 613	2 138	4 026	9	22 380	85 517
Bulgarian	25	54	0	245	1 015	62	96	0	367	2 479
Buryat	142	431	0	2 202	5 679	168	450	0	2 243	6 735
Estonian	235	749	1	3 832	9 402	389	936	1	4 740	15 561
Finnish	208	712	0	3 456	8 337	365	859	0	3 751	14 594
Georgian	41	86	0	383	1 621	76	129	0	568	3 049
German	1 520	3 757	2	21 702	60 798	4 210	10 902	3	67 829	168 419
Greek	56	149	0	682	2 246	612	1 167	0	5 149	24 500
Hungarian	25	96	0	582	1 018	40	114	0	664	1 611
Chechen	17	33	0	151	696	213	428	0	1 975	8 508
Chinese	242	1 117	0	7 086	9 693	288	1 241	0	7 850	11 506
Chuvash	223	448	0	2 495	8 910	373	731	0	3 277	14 930
Japanese	31	95	0	550	1 231	364	751	0	3 373	14 571
Jewish	776	1 875	2	8 236	31 050	1 093	2 594	4	12 821	43 710
Kabardian	68	193	0	1 108	2 733	111	229	0	1 188	4 438
Kalmyk	32	120	0	754	1 293	54	159	0	934	2 168
Karelian	104	431	0	2 460	4 174	249	630	0	3 545	9 941
Kazakh	928	2 034	0	10 056	37 125	1 164	2 360	0	11 475	46 540
Khakas	56	136	0	551	2 221	203	498	0	2 479	8 137
Komi	90	195	0	1 170	3 613	146	281	0	1 526	5 834
Korean	102	382	0	2 267	4 060	221	492	0	2 404	8 821
Latvian	386	1 353	3	7 155	15 444	541	1 583	6	8 397	21 628
Lithuanian	116	281	0	1 542	4 651	137	306	1	1 680	5 474
Mari	135	268	0	1 374	5 383	187	360	0	1 510	7 482
Moldovan	35	55	0	221	1 392	69	99	0	400	2 765
Mordvin	150	289	0	1 530	6 011	317	631	0	2 635	12 682
Ossetian	81	181	0	1 091	3 237	93	197	0	1 168	3 724
Polish	1 531	3 276	0	15 059	61 221	2 131	4 086	0	17 689	85 258
Russian	13 757	30 122	61	172 363	550 280	26 615	52 282	74	233 893	1 064 596
Tatar	802	1 627	0	6 703	32 095	1 810	4 293	0	22 459	72 417
Udmurt	77	141	0	807	3 082	111	192	0	939	4 454
Ukrainian	1 360	2 966	16	16 312	54 403	2 295	4 245	30	21 027	91 814
Uzbek	27	58	0	268	1 061	101	194	0	899	4 044
Yakut	39	63	0	283	1 544	73	107	0	386	2 909

Table A6: Arrest Date Imputation - Model Results

	Dependent variable:	
	I^y	$log(y^{\mathrm{pos}})$
	logistic	*OLS*
	(1)	(2)
(Intercept)	-0.771^{***} (0.161)	0.888^{***} (0.025)
Year of Process - 1922	1.630^{***} (0.586)	1.038^{***} (0.033)
Year of Process - 1923	0.955^{*} (0.510)	0.976^{***} (0.039)
Year of Process - 1924	2.128^{**} (1.004)	1.122^{***} (0.043)
Year of Process - 1925	1.019^{*} (0.586)	1.112^{***} (0.043)
Year of Process - 1926	0.508^{*} (0.284)	0.972^{***} (0.027)
Year of Process - 1927	0.346 (0.234)	0.904^{***} (0.024)
Year of Process - 1928	-0.260^{**} (0.123)	0.422^{***} (0.016)
Year of Process - 1929	-0.209^{**} (0.101)	0.307^{***} (0.012)
Year of Process - 1930	0.012 (0.103)	0.754^{***} (0.012)
Year of Process - 1931	0.166 (0.111)	0.695^{***} (0.013)
Year of Process - 1932	0.299^{***} (0.106)	0.463^{***} (0.012)
Year of Process - 1933	0.193 (0.139)	0.457^{***} (0.016)
Year of Process - 1934	-0.198^{*} (0.116)	0.602^{***} (0.014)
Year of Process - 1935	-0.206^{*} (0.112)	0.874^{***} (0.014)
Year of Process - 1936	0.858^{***} (0.099)	-0.298^{***} (0.011)
Year of Process - 1937	1.116^{***} (0.101)	0.486^{***} (0.012)
Year of Process - 1938	1.845^{***} (0.151)	2.010^{***} (0.013)
Year of Process - 1939	1.560^{***} (0.162)	1.579^{***} (0.013)
Year of Process - 1940	0.463^{***} (0.113)	0.705^{***} (0.013)
Year of Process - 1941	0.282^{***} (0.109)	0.641^{***} (0.013)
Year of Process - 1942	0.234^{**} (0.114)	0.833^{***} (0.013)
Year of Process - 1943	-0.422^{***} (0.118)	0.804^{***} (0.015)
Year of Process - 1944	0.175 (0.127)	0.936^{***} (0.015)
Year of Process - 1945	0.264^{*} (0.142)	1.182^{***} (0.016)
Year of Process - 1946	0.164 (0.161)	0.987^{***} (0.018)
Year of Process - 1947	0.231 (0.179)	0.860^{***} (0.020)
Year of Process - 1948	0.810^{***} (0.192)	0.735^{***} (0.017)
Year of Process - 1949	0.512^{***} (0.186)	0.953^{***} (0.019)
Year of Process - 1950	0.532^{***} (0.188)	0.908^{***} (0.019)
Year of Process - 1951	-0.080 (0.197)	0.844^{***} (0.024)
Year of Process - 1952	0.077 (0.269)	0.619^{***} (0.031)
Year of Process - 1953	0.003 (0.589)	1.680^{***} (0.070)
Year of Process - 1954	-0.526 (0.462)	2.253^{***} (0.071)
Year of Process - 1955	-0.713^{*} (0.425)	1.324^{***} (0.071)
Year of Process - 1956	0.950^{***} (0.367)	0.683^{***} (0.029)
Year of Process - 1957	0.595 (0.367)	0.813^{***} (0.033)
Year of Process - 1958	-0.232 (0.333)	1.036^{***} (0.044)
Year of Process - 1959	-0.716 (0.516)	1.042^{***} (0.087)
Year of Process - 1960	-0.796^{**} (0.395)	0.864^{***} (0.069)
Year of Process - 1961	1.722^{*} (1.005)	0.909^{***} (0.052)
Year of Process - 1962	-0.328 (0.423)	0.910^{***} (0.059)
Year of Process - 1963	-1.182^{***} (0.398)	1.293^{***} (0.083)
Year of Process - 1964	4.292^{***} (0.094)	3.697^{***} (0.011)
Observations	813,726	806,913
R^2		0.235
Adjusted R^2		0.235
Log Likelihood	$-38{,}398.330$	
Akaike Inf. Crit.	76,884.660	
Note:		*p<0.1; **p<0.05; ***p<0.01

Table A7: Pre-treatment characteristics of ethnic groups in the USSR

Ethnic group	Total population	Ling. similarity to Russian	Urbanization rate	Ind. state
Armenian	1 567 568	1	35.45	0
Belorussian	4 738 923	4	10.32	0
Estonian	154 666	0	23.00	1
German	1 238 549	1	14.92	1
Greek	213 765	1	21.21	1
Chechen	318 522	0	0.98	0
Chinese	10 247	0	64.87	1
Jewish	2 599 973	1	82.43	0
Kabardian	139 925	0	1.27	0
Kalmyk	129 321	0	1.29	0
Korean	86 999	0	10.52	0
Latvian	141 703	2	42.31	1
Lithuanian	41 463	2	63.16	1
Ossetian	272 272	1	7.86	0
Polish	782 334	3	32.75	1
Tatar	2 916 536	0	15.48	0
Ukrainian	31 194 976	4	10.54	0
Altai	39 062	0	0.30	0
Balkar	33 307	0	1.23	0
Bashkir	713 693	0	2.12	0
Bulgarian	111 296	3	6.26	0
Buryat	237 501	0	1.05	0
Finnish	134 701	0	10.55	1
Georgian	1 821 184	0	16.93	0
Hungarian	5 476	0	63.33	1
Chuvash	1 117 419	0	1.60	0
Japanese	93	0	76.34	1
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
Mari	428 192	0	0.84	0
Moldovan	278 905	1	4.86	0
Mordvin	1 340 415	0	2.19	0
Russian	77 791 124	5	21.32	1
Udmurt	504 187	0	1.21	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 A8: Synthetic German minority weights, Only ethnicities without ind. state

Ethnic group	W-Weight
Tatar	0.42
Korean	0.21
Ukrainian	0.20
Jewish	0.16

Table A9: Synthetic German minority weights, Only rehabilitated individuals

Ethnic group	W-Weight
Korean	0.34
Polish	0.33
Tatar	0.32

Additional Figures

Figure A1: Number of Predicted Arrests by Ethnicity, Year, and Prediction Adjustment

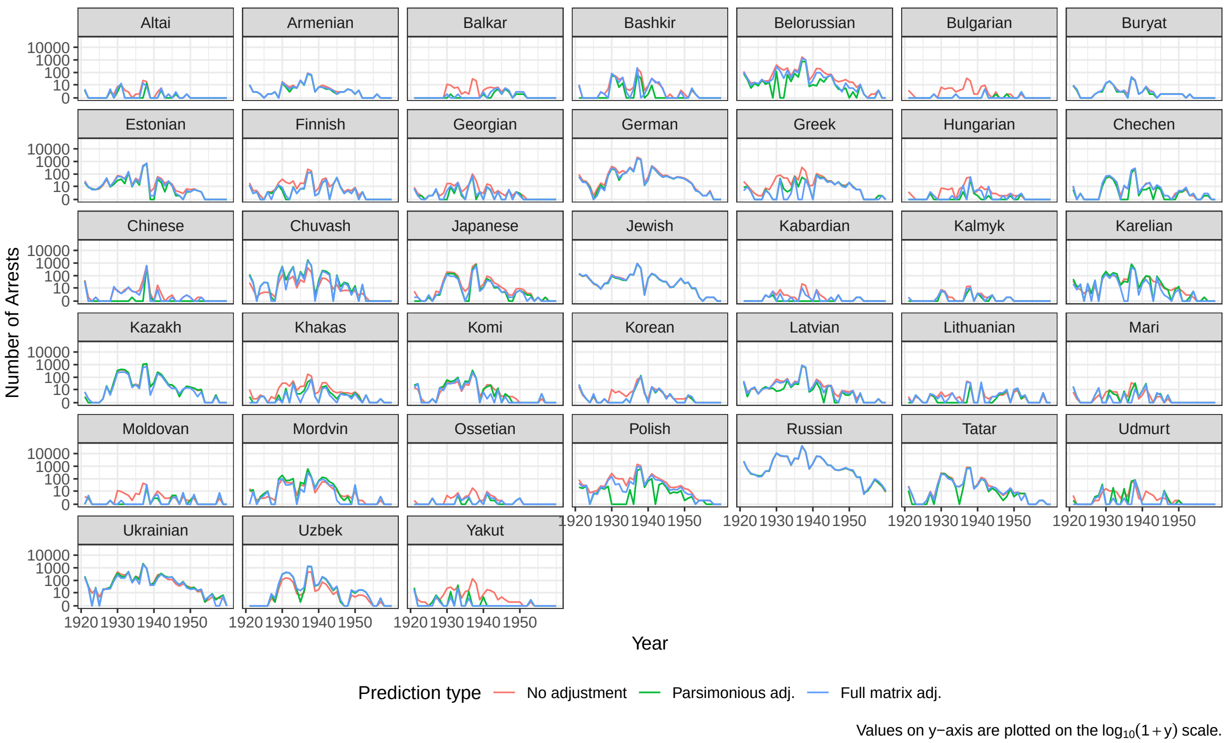


Figure A2: Histograms of Arrests by Ethnicity and Year

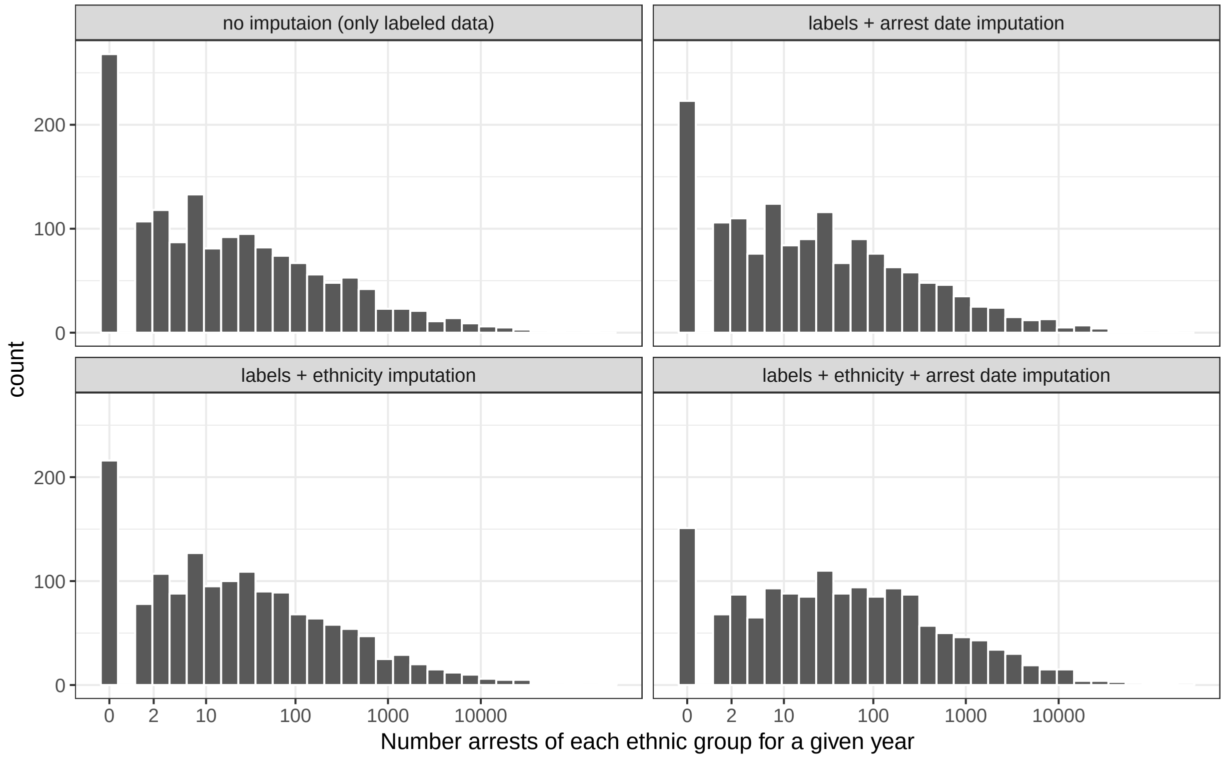


Figure A3: Histogram of Imputed Number of Days between Arrest and Process

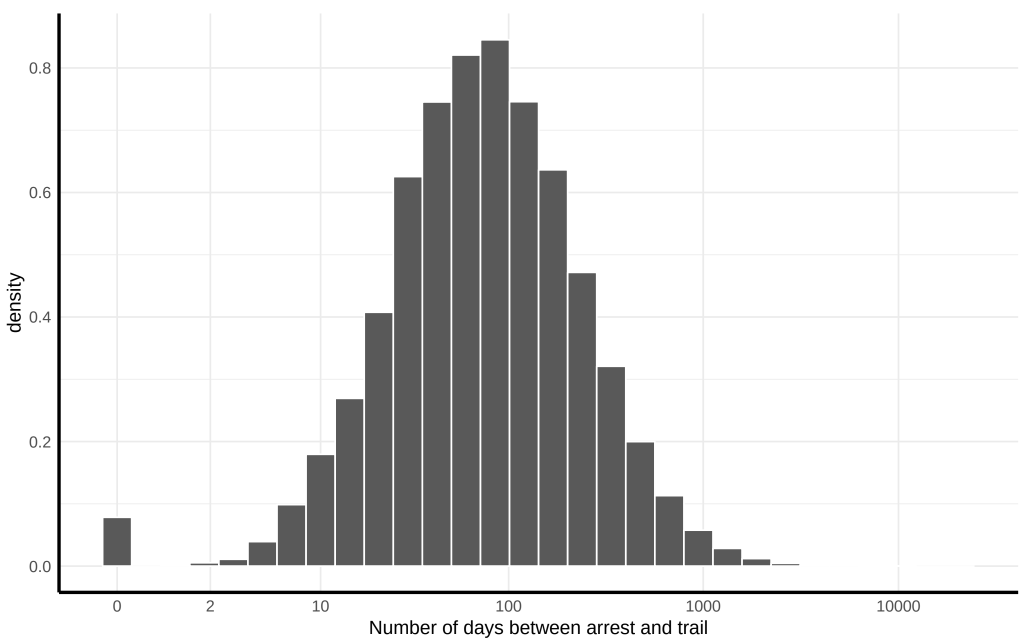
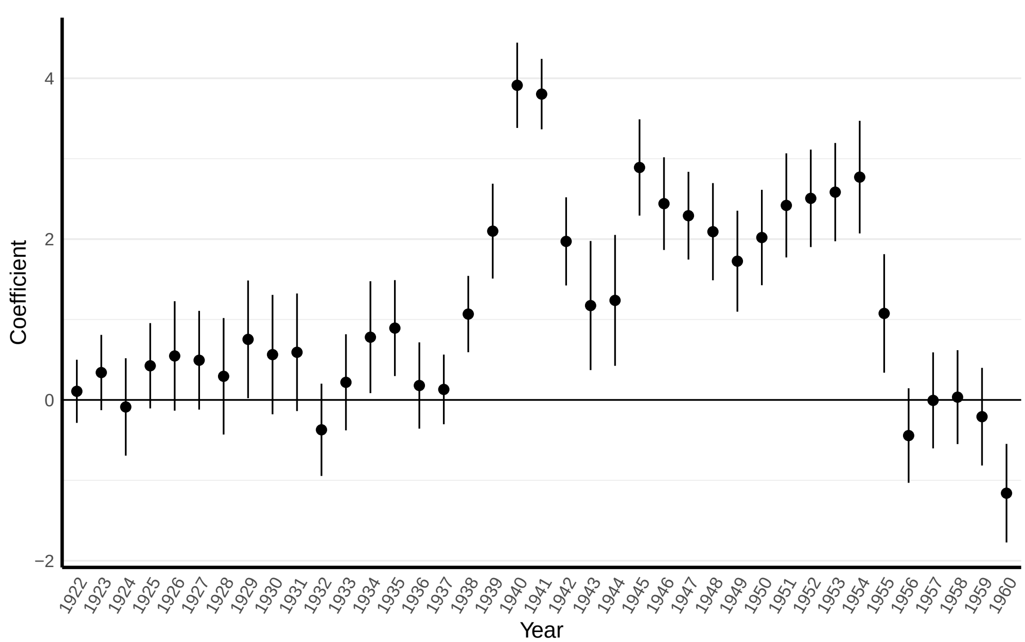


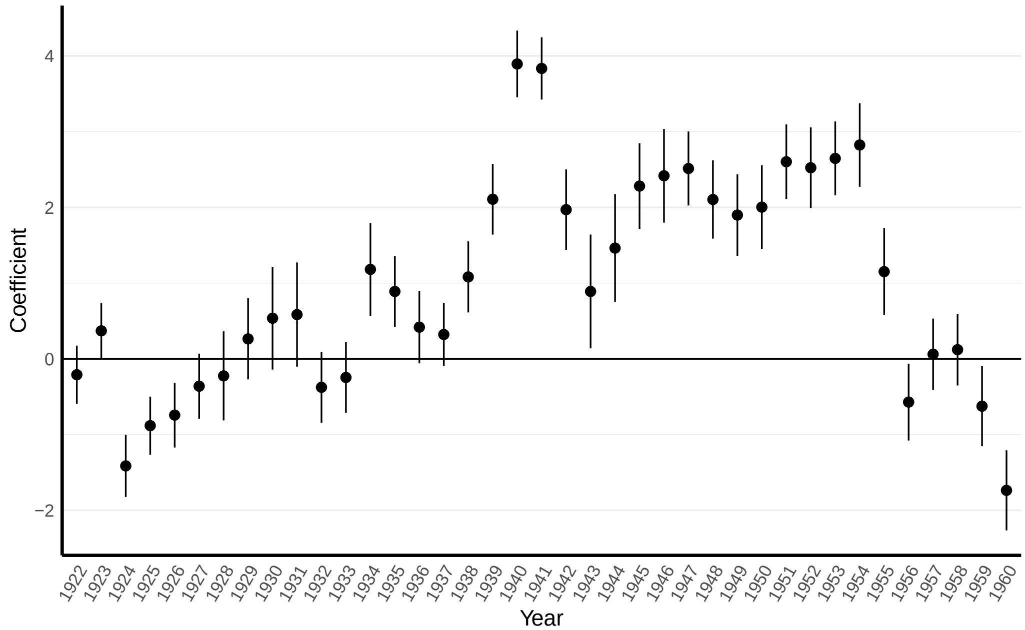
Figure A4: Time Series of Arrests

Figure A5: Dynamic DiD, Only Ethnicities without Independent State



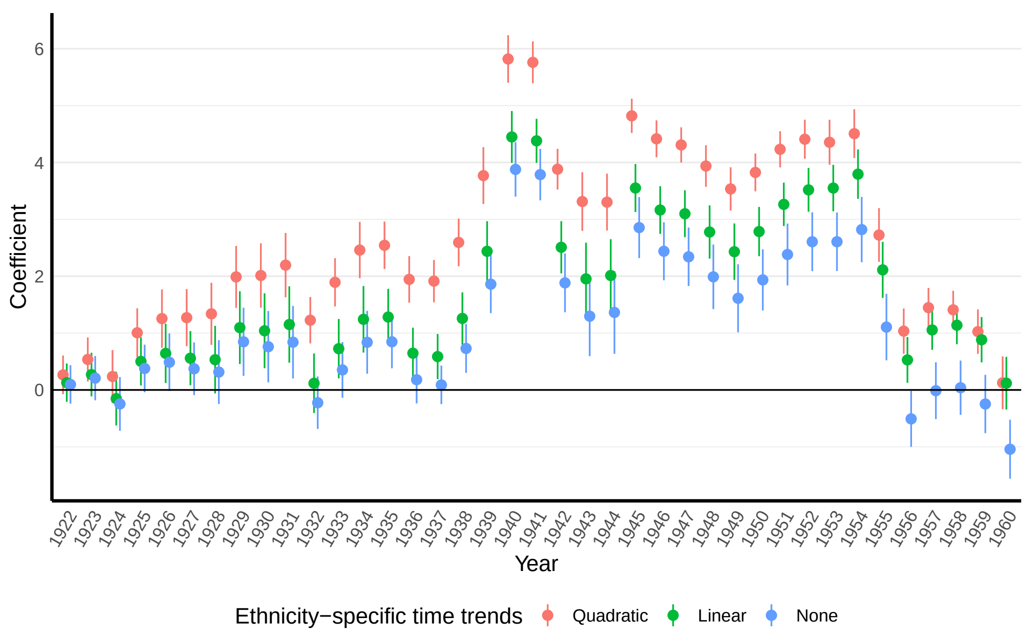
Notes: Only ethnic groups without independent state are included in the control group. Ethnicity and date of arrest were imputed. Full matrix adjustment was applied on ethnic group imputations. The quadratic ethnicity-specific time trends are included. Standard errors are clustered on the level of ethnicity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

Figure A6: Dynamic DiD, Only Rehabilitated Individuals



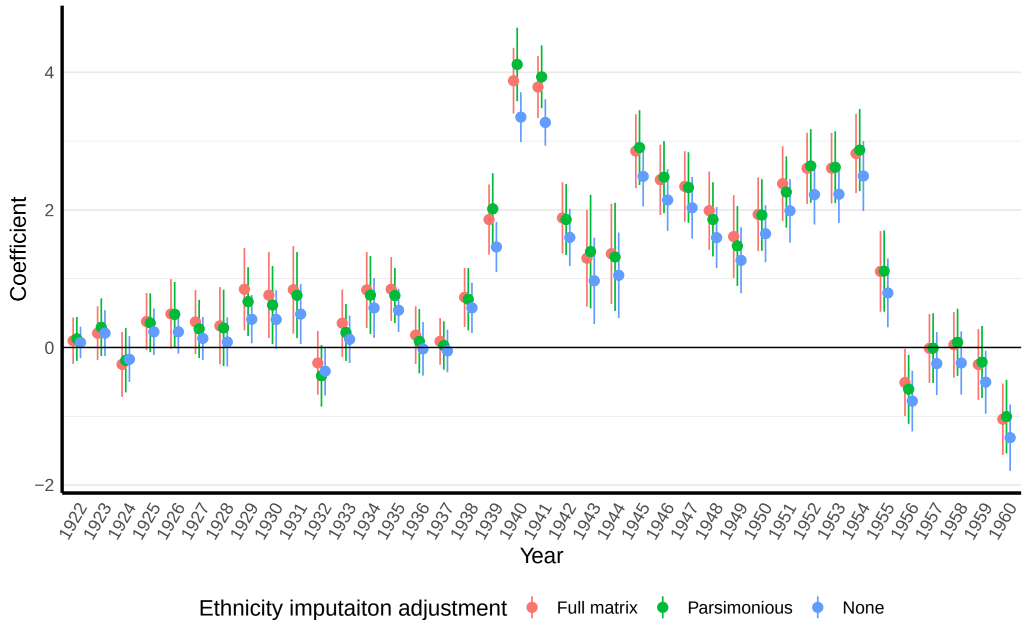
Notes: All 38 ethnic groups are included. Ethnicity and date of arrest were imputed. Full matrix adjustment was applied on ethnic group imputations. Controls for major changes in relations with the USSR are included. Standard errors are clustered on the level of ethnicity and are based on cluster robust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

Figure A7: 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.

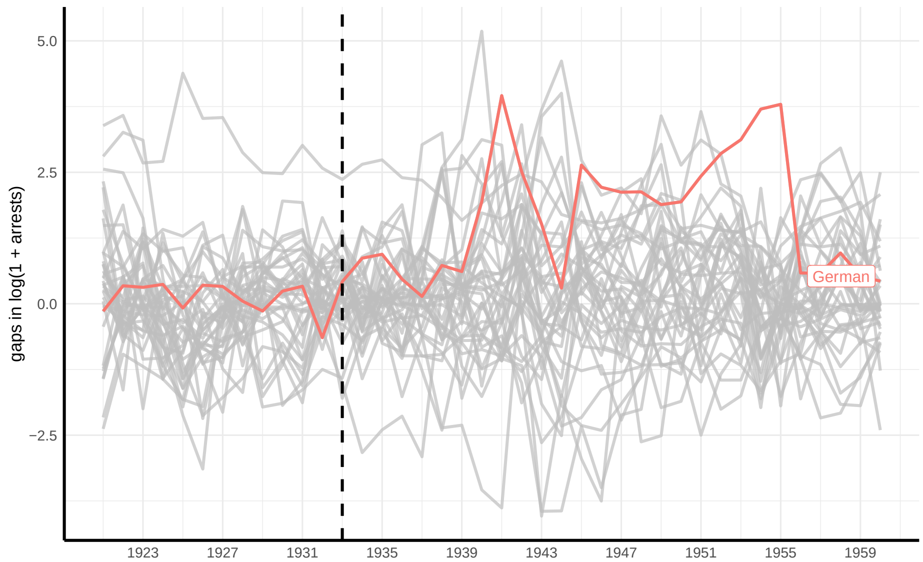
Figure A8: Comparison of Ethnicity Imputation Adjustments for DiD



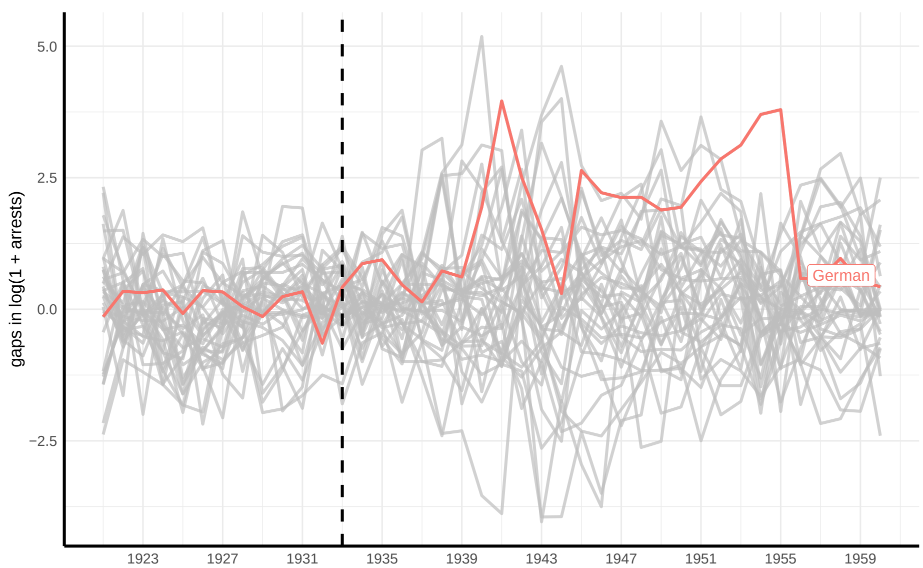
Notes: All 38 ethnic groups are included. Ethnicity and date of arrest were imputed. Standard errors are clustered on the level of ethnicity and are based on cluster robust estimator by Pustejovsky and Tipton (2018). Error bars show 95% confidence intervals.

Figure A9: Gaps between synthetic control and actual values for placebo tests

(a) All ethnic groups



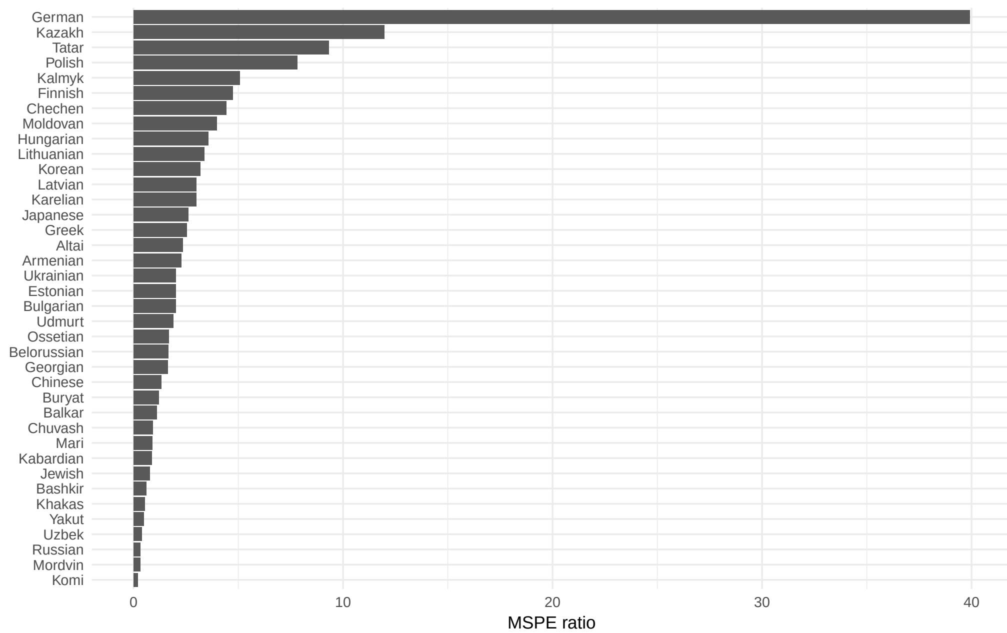
(b) Ethnic groups with pre-treatment MSPE higher than 20 times the MSPE of Germans excluded



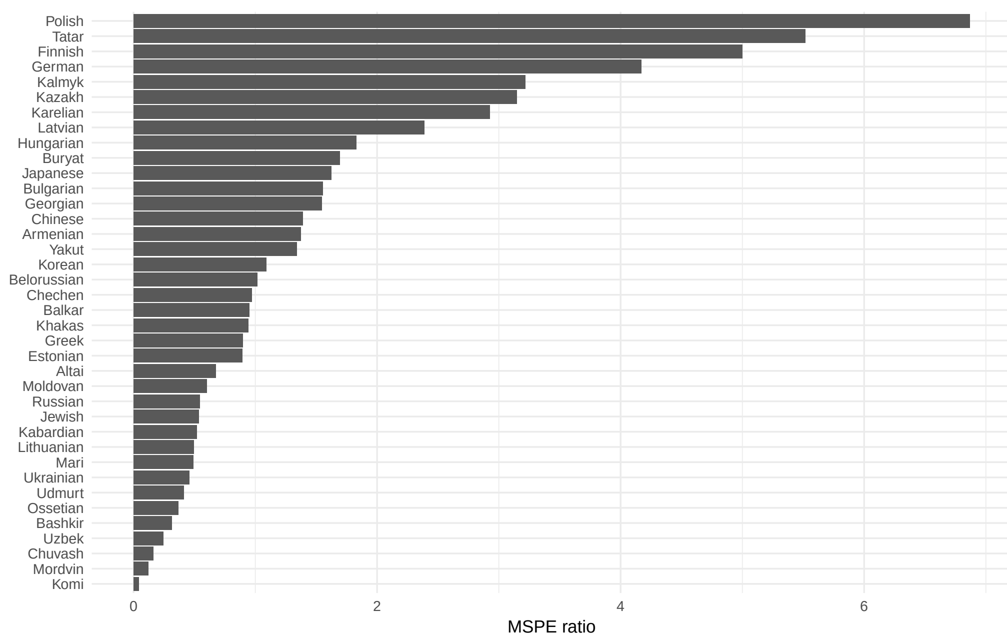
Notes: The predictors are the mean of log $(1 + \text{arrests})$ in the pre-treatment period, total population of the ethnic group in the USSR and its urbanization rate (both taken from the 1926 Soviet census), and linguistic similarity to Russian. Ethnicity and date of arrest were imputed. Full matrix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

Figure A10: Ratios of post-treatment MSPE to pre-treatment MSPE

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



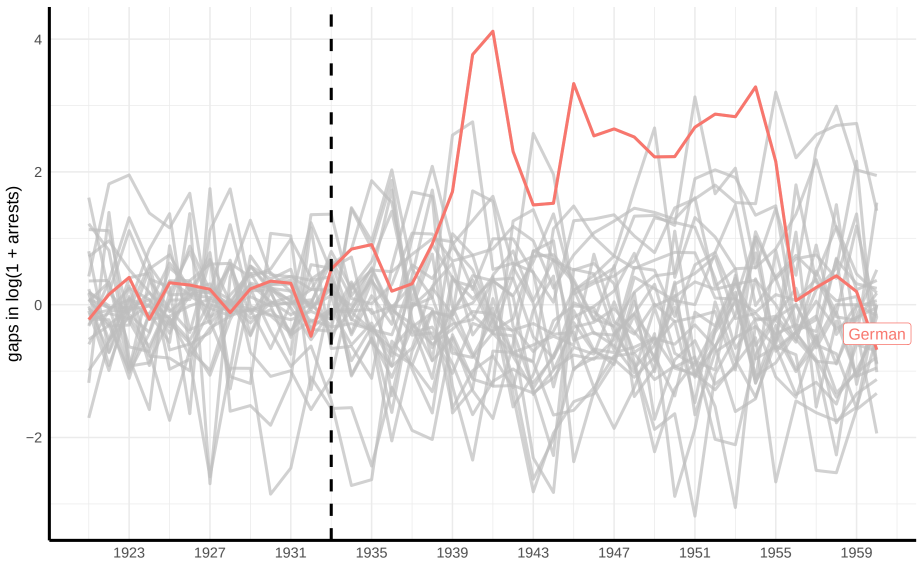
(b) Only the period from 1933 to 1939 in the numerator



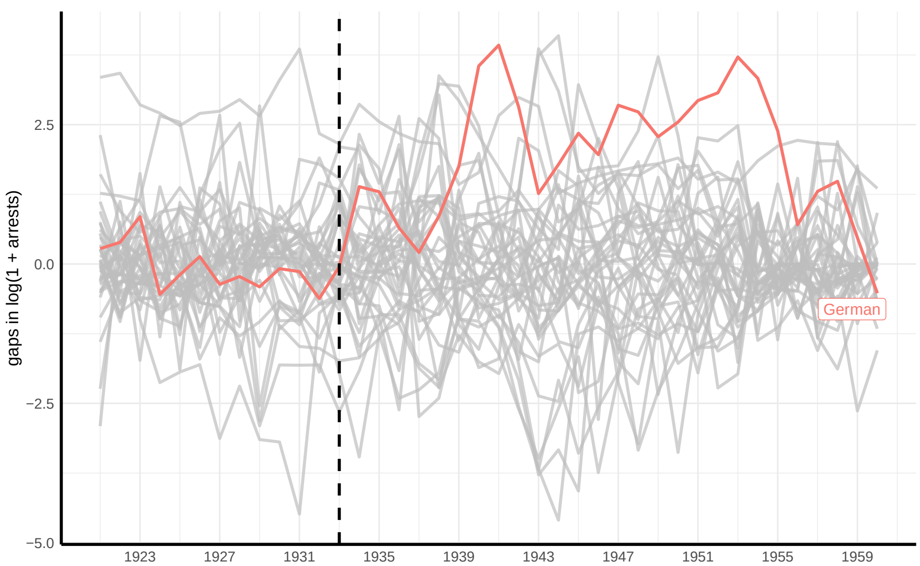
Notes: The same as for the figure A9.

Figure A11: Gaps between synthetic control and actual values for placebo tests

(a) Only ethnicities without independent state



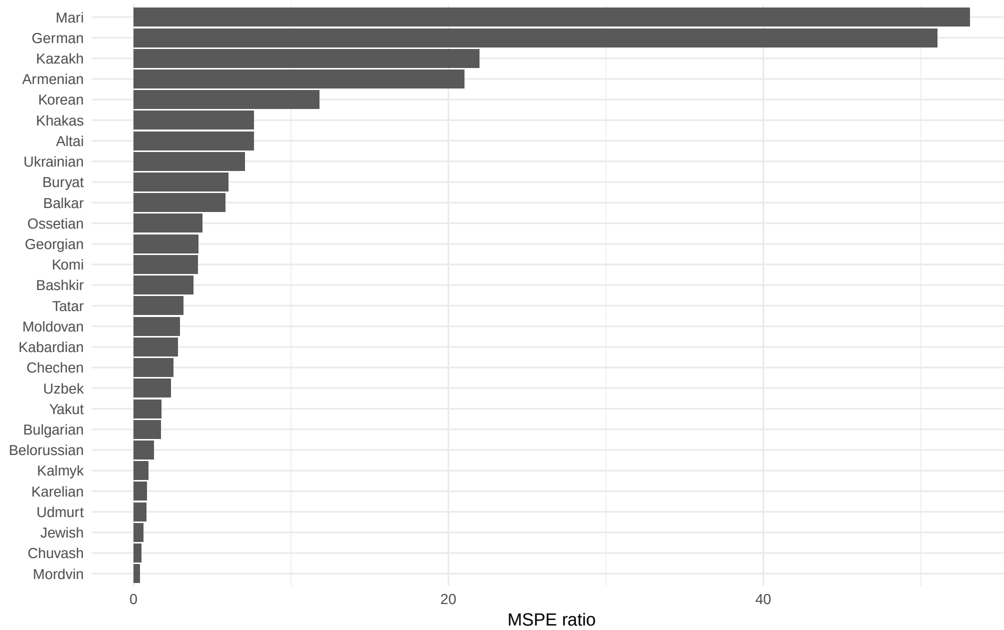
(b) Only rehabilitated individuals



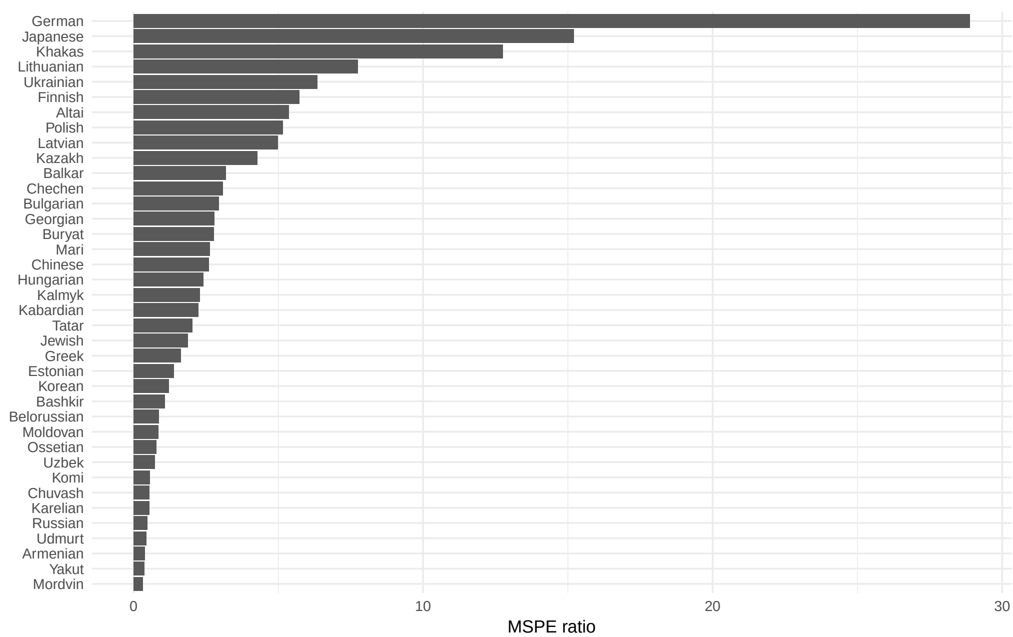
Notes: All pre-treatment outcomes were used as predictors. Ethnicity and date of arrest were imputed. Full matrix adjustment was applied on ethnic group imputations.

Figure A12: 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 A11.