Shaping States into Nations: The Effects of Ethnic Geography on State Borders

Carl Müller-Crepon* Guy Schvitz†

Lars-Erik Cederman‡

March 13, 2023

We thank the editor and three anonymous reviewers for their valuable comments and suggestsion and Nils-Christian Bormann, Michael Kenwick, Melissa Lee, Kan Li, Paul Poast, seminar and conference participants at APSA 2020, the University of Oxford, London School of Economics and Political Science, Harvard University, and the Perry World House Borders and Boundaries Conference, as well as members of the International Conflict Research group for their helpful feedback and Nicole Arnet, Camiel Boukhaf, Nicole Eggenberger, Benjamin Füglister, Sebastian Gmüer, Irina Siminichina, Tim Waldburger, Benjamin Wallin, and Roberto Valli for their invaluable research assistance. We are very grateful to Philipp Hunziker who provided crucial inputs to this work. All remaining errors are ours. We acknowledge generous financial support from the Advanced ERC Grant 787478 NASTAC, "Nationalist State Transformation and Conflict."

*Corresponding author: c.a.muller-crepon@lse.ac.uk. Department of Government, London School of Economics and Political Science. Houghton Street, London, WC2A 2AE, United Kingdom.

†guy.schvitz@ec.europa.eu. Center for Comparative and International Studies, ETH Zürich and European Commission, Joint Research Centre. Via Enrico Fermi, 2749, 21027 Ispra (VA), Italy.

‡cederman@icr.gess.ethz.ch. Center for Comparative and International Studies, ETH Zürich. Haldeneggsteig 4, 8092 Zurich, Switzerland.

Abstract

Borders define states, yet little systematic evidence explains where they are drawn. Putting current challenges to state borders into perspective and breaking new methodological ground, this paper analyzes how ethnic geography and nationalism have shaped European borders since the 19^{th} century. We argue that nationalism creates pressures to redraw political borders along ethnic lines, ultimately making states more congruent with ethnic groups. We introduce a Probabilistic Spatial Partition Model to test this argument, modeling state territories as partitions of a planar spatial graph. Using new data on Europe's ethnic geography since 1855, we find that ethnic boundaries increase the conditional probability that two locations they separate are, or will become, divided by a state border. Secession is an important mechanism driving this result. Similar dynamics characterize border change in Asia but not in Africa and the Americas. Our results highlight the endogenous formation of nation-states in Europe and beyond.

Keywords: Borders; Ethnicity; Europe; Spatial Partitioning; GIS; Computational Methods

Borders are constitutive features of the modern state system that define the size and shape of states and specify the limits of state sovereignty.[1] A growing literature documents borders' attributes (Simmons and Kenwick 2021) and consequences (Abramson and Carter 2016; Carter and Goemans 2011; Simmons 2005; Michalopoulos and Papaioannou 2016). Yet, their origins remain understudied with much research treating states and their borders as exogenous. Border formation has however gained renewed relevance as Russia invaded Ukraine, majorities support territorial revisionism in Hungary, Greece, Bulgaria, and Turkey (Fagan and Poushter 2020), and secessionist challenges in Scotland, Northern Ireland, and Catalonia. Ethnonationalist demands to redraw state borders along ethnic lines are central to all these cases.

Yet, despite their intuitive appeal, explanations that seek borders' origins in ethnicity are contested and not systematically tested. Addressing this gap, we ask whether, how, and to what extent ethnic geography has shaped Europe's partitioning into states since the 19th century. Following macrosociological theories, we argue that the historical rise of nationalism, "a political principle which holds that the political and national unit should be congruent" (Gellner 1983, p. 1), created demand for nation-states. As most nations are ethnically defined, nationalism prompted popular pressures to redraw borders along ethnic lines through secessionism, unification, and ir-

[1]See Sack (1986) on human territoriality more generally.

redentism (Weiner 1971; Hechter 2000; O'Leary 2001). Of these mechanisms, secessionism is most common and systematically studied. While the ethno-political roots of secessionist conflict are well evidenced (e.g. Cederman, Gleditsch and Buhaug 2013; Germann and Sambanis 2021), some studies of secessions discount ethnicity and nationalism in favor of pre-existing political units and power politics (Roeder 2012; Griffiths 2016; Coggins 2014). We contribute to this debate by integrating secessionist, unificationist, and irredentist border change into a common analytical framework and by overcoming previous studies' problematic reliance on geographically fixed units of analysis.[2]

We thus innovate the study of border determinants, which so far lacks a robust quantitative estimator to test theoretical arguments against potentially confounding alternative hypotheses. Realists argue that borders emerge along mountains and rivers, facilitating internal power-projection and effective defense (Morgenthau 1985, also Kitamura and Lagerlöf 2020). From an institutionalist perspective, borders are coordination devices based on states' preferences for territory and stability (Simmons 2005) and often follow local "focal" lines – rivers, watersheds, or historical precedents (Abramson and Carter 2016; Carter and Goemans 2011; Goemans 2006; Goemans and Schultz 2017). A third perspective highlights borders' origins in ethnic geography. Alesina and Spolaore (1997, 2005) theorize the trade-off

[2]See, e.g., (Griffiths 2016, ch. 2).

between economies of scale and costs of ethnic heterogeneity in large states (see also Friedman 1977; Desmet et al. 2011). We empirically test the effect of ethnic geography on state borders and provide comprehensive evidence that accounts for alternative explanations.

To do so, we overcome three challenges of assessing the determinants of borders and the spatial partitioning they produce. First, border formation is an intractable problem as infinitely many borders can partition space into an ex ante unknown number of units. Second, borders entail significant and complex spatial dependencies as they form contiguous, non-overlapping units. Third, unbiased estimation of ethnic geography's effect on borders requires consideration of confounding geographic features that affect both.

We address these challenges with a new *Probabilistic Spatial Partition Model* (PSPM) which allows us to estimate the conditional effect of spatial features (e.g., ethnic settlement patterns) on the partitioning of geographic space into non-overlapping units (e.g., states). The model discretizes geographic space as a planar network of points that encodes the main dependent and independent variables. It makes partitionings tractable, accounts for spatial dependencies, estimates effects conditional on covariates, and yields valid uncertainty estimates. Beyond our present use, the PSPM can be applied to model other partitionings, for example administrative or electoral units. We provide an accompanying open-source R package and code for handling spatial network data.[3].

[3]Available at github.com/carl-mc/pspm and github.com/

We use the PSPM to estimate the effect of ethnic geography on state borders. Our new, time-varying data on ethnic geography predate (changing) state borders in Europe since 1855. Digitized from 73 historical maps, the dataset enables us to analyze borders and border change based on pre-existing ethnic settlement areas. We address omitted variable and reverse causality bias by pairing a cross-sectional with a lagged dependent-variable model that captures the effect of ethnic geography on border change.

We find that an ethnic boundary between two locations increases the probability that they are or will become separated by an international border by 34 and 17 percentage points, respectively. This finding is robust to accounting for potentially endogenous changes in ethnic geography, alternative measures of ethnic differences, additional controls, and changes to the spatio-temporal data structure. Additional analyses highlight ethnic secession as a key mechanism: Since 1946, areas home to peripheral ethnic groups saw secessionist claims, civil wars, and border change 11, 21, and 50 times more often than other areas. Moving beyond Europe, we find that ethnic boundaries explain border change since the 1960s in Asia but not elsewhere.

`carl-mc/SpatialLattice`.

Nationalism and the shaping of states

We argue that the rise of nationalism created a demand for ethnically homogeneous nation-states, which caused an increasing alignment of Europe's borders with the underlying ethnic map. This development is part of a larger process of the "right-peopling" and "right-sizing" of states (O'Leary 2001). The former has received much attention in nationalism studies evidencing the formation of nations within states through assimilationist policies and ethnic violence (Weber 1976; White 2004; Bulutgil 2016; McNamee and Zhang 2019) or local dissimilation processes along state borders (Sahlins 1989). Yet, an exclusive focus on state-led identity formation that follows Hobsbawm's (1990, p. 10) claim that "[n]ations do not make states and nationalisms but the other way around" neglects parallel changes in state borders and risks underestimating the full impact of nationalism.[4] We therefore focus on the nationalist right-sizing of states along ethnic lines and address reverse processes as an empirical challenge.

How did nationalism transform Europe's borders? We start by defining ethnic groups as "those human groups that entertain a subjective belief in common descent" (Weber 1978, pp. 385-98), most frequently distinguished by their language and religion. Once groups' members desire to control a

[4]The two processes are linked as ethnic homogenization often focuses on contested territories (Bulutgil 2015, 2016; McNamee and Zhang 2019; Mylonas 2012).

state, they become ethnic nations, "a community of sentiment which would adequately manifest itself in a state of its own" (Weber 1978, p. 176). In consequence, ethno-nationalist ideology requires "that ethnic boundaries should not cut across political ones, and, in particular, that ethnic boundaries within a given state [...] should not separate the power-holders from the rest" (Gellner 1983, p. 1). Three constellations violate Gellner's congruence principle, each motivating a specific type of border change.

First and most common are ethnic minorities in a state dominated by a different group. Such "alien rule" (Hechter 2013) deprives groups of self-determination and state services that often favor ruling groups (De Luca et al. 2018). In response, stateless nations may try to attain statehood by secession. The break-up of empires and multi-ethnic states exemplifies this process (Beissinger 2002; Germann and Sambanis 2021). With many more potential ethnic nations than states,[5] secessionism is the most common type of border change (Gellner 1983; Griffiths 2016; Hechter 2000).

Second, ethno-nationalist grievances can also emerge if an ethnic group is divided by state borders, prompting nationalist calls for unification (Cederman, Rüegger and Schvitz 2022). The promise of benefits from governance over a larger and ethnically homogeneous territory and population can help their cause (Alesina and Spolaore 2005). Such efforts sometimes

[5]Particularly after the German and Italian unifications outside our empirical scope.

yield the merger of co-ethnic units, as illustrated by 19th-century Germany and Italy and the more recent reunifications of Vietnam, Yemen, and Germany. Concomitant to the decline of state death since 1945 (Fazal 2004, 2007), ethnic unification is exceedingly rare.

Third, a configuration in which an ethnic group dominates one state but forms a minority in another can pressure the homeland government to "liberate" their kin, thus resulting in irredentist nationalism (Weiner 1971; Siroky and Hale 2017). Named after Italian Veneto and Trento that remained "unredeemed" after the first wave of Italian unification, the stronger territorial integrity norm has reduced irredentist border change after World War II (Zacher 2001).

Nationalist ideology equips revisionist activists of all three situations with powerful arguments that legitimize their claims over ostensibly "indivisible" territory and mobilize elites and citizens for their projects (Hroch 1985; Murphy 2002; Goddard 2006). While collective action problems and resistance by the incumbent state can inhibit actual border change (Hardin 1995), nationalist grievances can lower the bar by making activists less risk averse (Petersen 2002; Nugent 2020; Germann and Sambanis 2021). Still, revisionist nationalism is unlikely to succeed without considerable material and organizational resources (Tilly 1978). Alternatively, geopolitical and economic crises create opportunities for change by weakening existing states, as illustrated by imperial collapse after the World Wars (Abramson

and Carter 2021; Skocpol 1979). In addition, nationalist "successes" can inspire nationalists elsewhere, further reinforcing the spatio-temporal clustering of border change. Nationalist ideas spread through 19th century Europe and globally thanks to the "Wilsonian moment" after World War I (Manela 2007).

Yet the diffusion of nationalism beyond Europe did not necessarily produce ethno-nationalist congruence. The disintegration of the massively multi-ethnic European colonial empires led to new borders that cut through ethnic groups and created ethnically diverse independent states (Englebert, Tarango and Carter 2002). While some activists supported pan-nationalism, the prevailing elites in the Global South generally subscribed to the legal norm of *uti possidetis*. This implied that new borders would follow colonial administrative borders regardless of their ethnic fit (Ratner 1996). Where ethnic groups were much smaller than states, as in sub-Saharan Africa, *uti possidetis* was particularly influential (Carter and Goemans 2011), a tendency that was further reinforced by a lack of interstate competition over sparsely populated areas (Herbst 2000) and international norms (Zacher 2001). Even under these conditions, Sub-Sahara Africa was far from immune to ethno-nationalist revisionism, as evidenced by Somali irredentism and Biafran separatism in Nigeria. In contrast and thanks to the presence of demographically dominant groups, ethno-nationalism had a larger influence on border drawing in post-colonial Asia.

Regardless of the specific historical context, those mobilizing for border change will base their territorial claims on their – often self-serving – understandings of ethnic geography. However, even where mobilization successfully achieves border change, "ethnically pure" borders tend to be elusive because of overlapping and non-contiguous ethnic settlement patterns (Sambanis and Schulhofer-Wohl 2009). As a result, ethnic geography determines the *approximate* location of new borders. In turn, sharp focal lines such as previous administrative borders, historical precedents, rivers, or watersheds inform their local settlement (Goemans 2006; Carter and Goemans 2011).

Analyzing the primacy of secession and the global generalizability of the argument in separate analyses, our main empirical focus is on the overall impact of ethnic settlement patterns on European state borders:

Hypothesis 1 *Ethnic settlement patterns shape state territories such that ethnic boundaries and state borders become increasingly congruent.*

Unit of analysis and data

We test our claims about the effect of ethnic boundaries on state borders using time-variant data on state borders and ethnic geography in Europe since 1886. This section explains how we go beyond previous studies of border determinants by modeling the European landmass as a spatial network of points. We use the network to encode our data and estimate our new *Prob-*

abilistic Spatial Partition Model (PSPM) presented subsequently.

Geographic space as a network of points

We model geographic space as a network of points, a move that addresses limitations of previous analyses of border locations. These have followed three approaches. First, Goemans (2006) and Carter and Goemans (2011) show that new borders are frequently drawn along focal lines such as natural frontiers, administrative borders, or historical precedent. This valuable description of border characteristics provides the ground for analyzing border precedents as influential causes of border stability (Carter and Goemans 2011; Abramson and Carter 2016). Yet, a focus on observed borders produces limited insights into their causes, since it neglects all potential but unrealized borders. In addition, a focus on locally aligned features risks missing factors such as ethnic geography that only determine borders' *approximate* location at a higher geographic level.

A second approach by Kitamura and Lagerlöf (2020) uses grid cells as seemingly independent units to examine the frequency with which they have been crossed by state borders. Doing so disregards nonmonotonic spatial dependencies inherent in the outcome of interest. Because borders partition space into contiguous territorial units, they interdependently emerge in grid cells. For example, a border will cross a string of pairs of neighboring grid cells, violating the assumption of unit-independence in standard

regression approaches as the outcome for any unit depends on its relation to the *ensemble* of neighboring cells (not) crossed by a border. Classic spatial error clustering (e.g. Conley 1999) and spatio(-temporal) diffusion models (Wucherpfennig et al. 2021) rely on a exogeneously imposed spatial connectivity matrix and are thus unable to recover such endogenous spatial dependency structures.

A third approach compares observed partitionings with simulated ones. Prominent in the gerrymandering literature (e.g., Fifield et al. 2020), such comparisons are based on aggregate statistics, as in our example the ethnic homogeneity of observed and simulated states. This approach yields information on the likelihood that an observed partitioning could have originated from the simulated process. Yet, because the observed partitioning is not modelled directly, such analyses do not produce inferences about the effects of a given spatial feature on the partitioning, in particular in the presence of confounders.

In response to these limitations, we introduce a simplified understanding of space as a planar network G of N points. Discretizing space makes tractable the problem of analyzing the partitioning of a continuous surface, which otherwise has infinitely many possible outcomes. Coupled with the partition model introduced below, the network structure of the data allows us to capture the spatial dependencies that characterize borders. Taking a network of points guarantees that G's vertices have unambiguous partition

memberships. G covers Europe[6] as a hexagonal lattice with 1096 nodes and 2905 edges. Its nodes j are connected to their up to six first-degree neighbors k at a distance of $\sim$100km (Figure 1a).[7]

Data on state borders

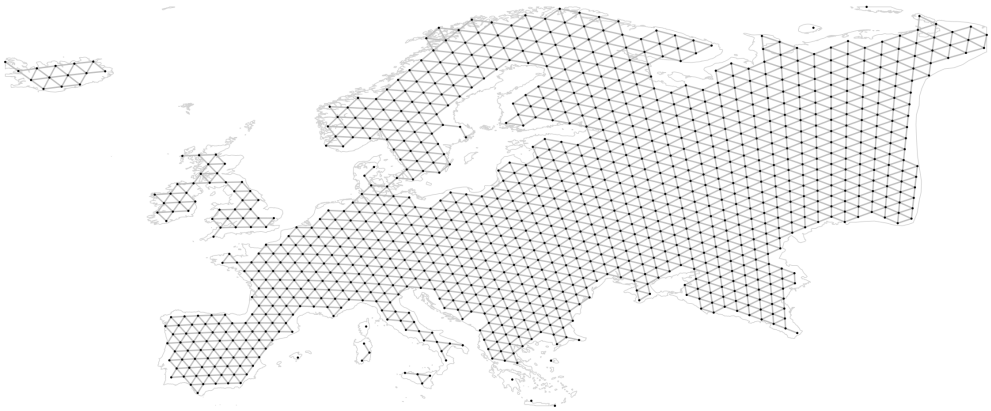
Our main outcome is the map of states: the partitioning P_t of the lattice G_t into states in year t. We measure P_t by retrieving the state each vertex belongs to between 1886 and 2019 from the CShapes 2.0 dataset (Schvitz et al. 2022). We analyze borders in every 25$^{\text{th}}$ year, i.e., in 1886, 1911,..., 2011.[8] The quarter-century intervals are long enough for cumulative border change to produce meaningful variation yet short enough to capture varying patterns of border change since 1886.

Figure 1b plots the outcome data in 1886. While we can distinguish "Spain" from "France," these labels are, for our purposes, completely interchangeable. Because we do not ex ante know the number or names of states, we are not interested in whether some vertices became part of "France." Instead, we study whether certain vertices together form a contiguous state –

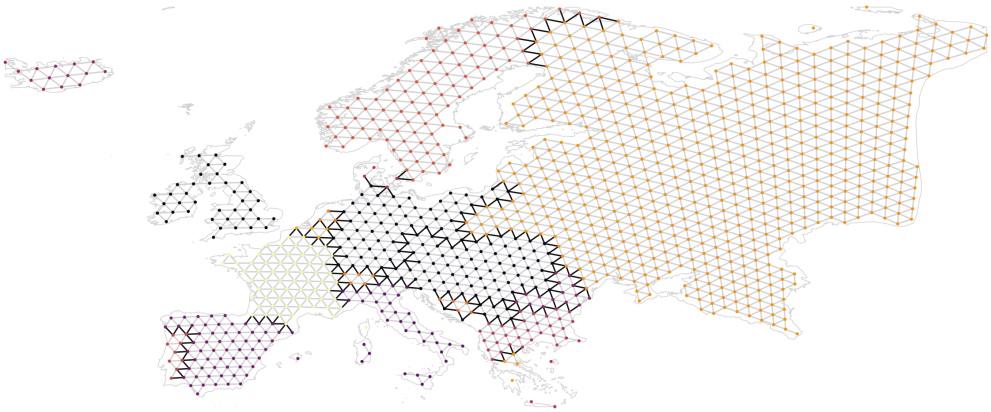
[6]We avoid state-based definitions and define Europe's eastern border from the Bosporus, via the Black Sea, the Carpathian mountain ridge, the Caspian Sea, and the Ural.

[7]This minimizes geographic distortion. Appendix D shows robustness to varying the graph's location, resolution, and structure.

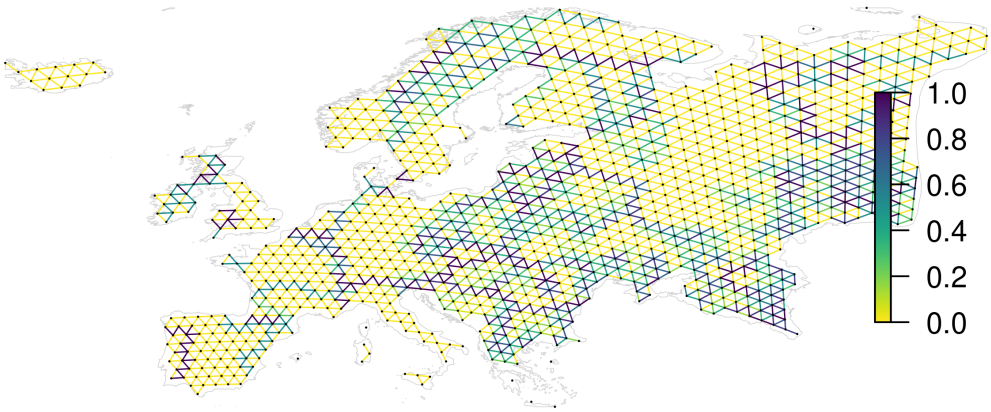
[8]Appendix D analyzes alternative temporal structures.



(a) Baseline lattice



(b) Partitioning into states in 1886. Border-crossing edges in black.



(c) Ethnic boundaries in 1836-1885. Color denotes fraction of maps in which an edge crosses an ethnic boundary.

Figure 1: Europe as a hexagonal spatial lattice

a partition. The set of all partitions defines the partitioning of Europe into states.

Data on historical ethnic settlement patterns

We collect new data on ethnic settlement areas in Europe since 1855. Our main independent variable measures whether an edge crosses and ethnic boundary or not. We construct this measure from 73 historical maps that together capture changes in ethnic settlement patterns over the past 165 years. Changes from genocides and population exchanges are well documented, while assimilation has more gradually altered ethnic geography. Our historical and time-varying data avoid reverse causality that may arise from backwards-projecting contemporary ethnic data.

Ethnic maps first emerged in the mid-19th century and proliferated due to at least two factors. First, innovations in statistics and cartography enabled the linguistic and religious categorization of local populations. Second, the rise of state-driven and peripheral nationalisms created a demand for maps of ethnic groups (Kertzer and Arel 2002; Hansen 2015). Initial efforts by German and Austrian geographers in the 1840s were followed by authors from Russia, the Balkans, and other parts of Europe, resulting in a scientific community dedicated to ethnic cartography.

For the most part, maps were drawn from census data on the town- or district-level, and defined ethnicity based on native languages (Cadiot 2005;

Hansen 2015). The production of ethnic maps was generally viewed as a scientific endeavor, motivated by enlightenment-era ideals of measuring and classifying the "natural" world (Livingstone and Withers 1999). Cartographers therefore sought to establish common standards and provided detailed justifications (Hansen 2015).

However, ethnic maps and census data were also used politically, employed by states and nationalist movements to shape perceptions of national homelands and support territorial claims (Herb 2002; Anderson 1991).[9] This was most evident at the Paris Peace conference of 1919, where all parties relied on their own maps to support their demands (Palsky 2002). Yet, the scope for manipulation was limited. Because cartographers largely relied on similar data and methods, they could not arbitrarily "invent" ethnic boundaries without jeopardizing their reputation (Hansen 2015; Herb 2002). Instead, most attempts to manipulate maps and census data involved the subtle use of politically convenient criteria such as the choice of sources, population thresholds (Hansen 2015), and the underlying list of ethnic groups (Hirsch 1997; Cadiot 2005).[10] While ethnic categorizations may have additionally affected ethnic consciousness (Kertzer and Arel 2002; Anderson 1991), such ethnic malleability was restricted too: while uni-

[9]See Branch (2013) for parallel consequences of mapping states.

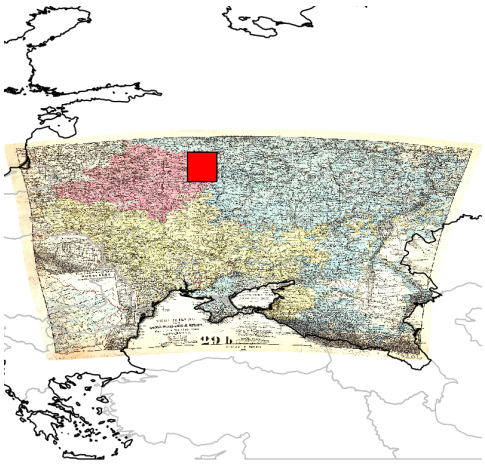
[10]For example, Kertzer and Arel (2002) note that Greek, Serbian and Bulgarian nationalists used alternative linguistic criteria to justify claims on parts of Macedonia.

fying German dialects into one self-conscious group was possible, more salient and sticky linguistic divides between mutually unintelligible languages were very difficult, if not impossible, to alter, invent, or make disappear.

As with all data on ethnic demographics, the political importance and potential manipulation of ethnic maps could bias our analysis. Our empirical strategy to test for and mitigate such biases consists of five components.

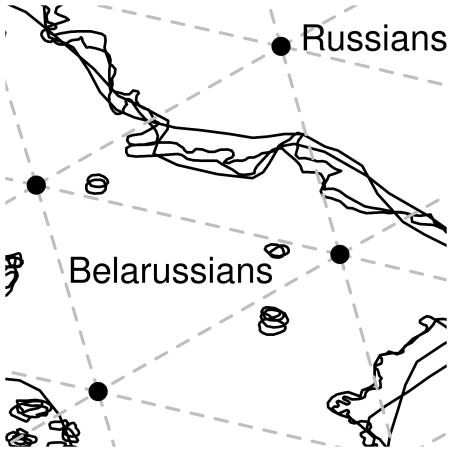
First, we carefully screened our map material. Starting with over 350 maps, we selected 73 maps based on high quality and spatial precision, and the absence of obvious political bias (Appendix C.1). Drawn by 64 authors from 18 nationalities, the maps cover various parts of Europe at different points in time using sometimes diverging categorizations of ethnic groups. Second, our spatial graph G is coarse with a resolution of 100km and up to 200km in a robustness check. Most differences between and likely manipulations of ethnic maps affect much smaller areas (see Figure 2).

Third, we average ethnic settlement patterns across all maps from a given period, reducing the impact of potential biases on any one map. Additionally, we find no "outlier maps" when re-estimating our main models for each map separately. Fourth, we show that our results are robust to exclusively using pre-1886 ethnic boundaries to explain changes of state borders between 1886 and 2011. This severely limits potential reverse causality, as well as strategic map manipulations during the World Wars. Fifth, we em-

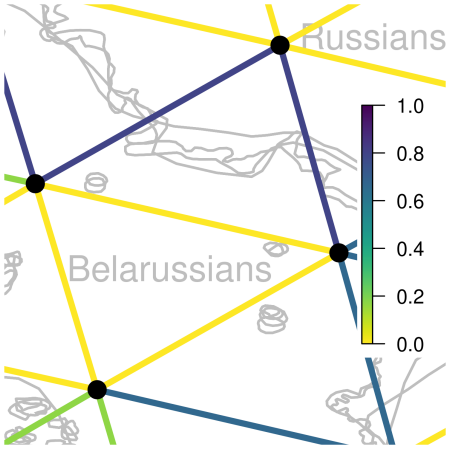


(a) 1878 map of Russians, Belarussians, and Ukrainians

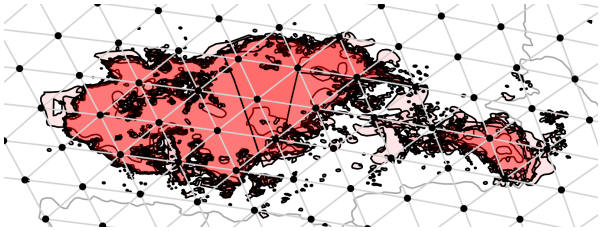
(b) Detail of the Belarussian-Russian ethnic boundary, red square in (a)



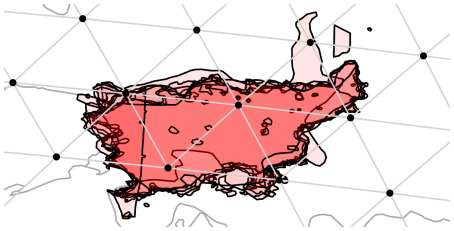
(c) Ethnic boundaries from (b) and other maps (1835-1885) overlaid with graph G



(d) Ethnic boundary$_{1886}$ measure



(e) Hungarian settlement area from 9 pre-1886 maps overlaid with G



(f) Slovenian settlement area from 8 pre-1886 maps overlaid with G

Figure 2: Constructing ethnic boundary from historical ethnic maps

Note: (a)-(d) show the transfer of ethnic settlement data onto graph G. (e) and (f) show Hungarian and Slovenian settlement areas from multiple maps.

ploy linguistic distances and ethno-demographic Austro-Hungarian census data as two alternative, continuous measures of "ethnic distance" to address remaining concerns of political biases. Discussed below, our results are robust across all tests.

We construct our main independent variable ethnic boundary as the proportion of maps from a given period in which an edge crosses an ethnic boundary:

$$\text{ethnic boundary}_{j,k,t} = \frac{1}{M_{j,k,t}} \sum_{m=1}^{M_{j,k,t}} \mathbb{1}_{g_{m,j} \neq g_{m,k}} \tag{1}$$

where j and k are an edge's constitutive nodes observed in year t. $M_{j,k,t}$ denotes the set of maps that cover the geographic location of j and k in one of the 50 years prior to t. The variable $\text{ethnic boundary}_{j,k,t}$ is the simple arithmetic mean of the map-level indicators that are 1 if a map m shows nodes j and k in different ethnic settlement areas and 0 otherwise.[11]

Modeling and estimation

We start from the intuition that the partitioning of space into states results from "attractive" and "repulsive" forces active between different locations. These forces correspond to factors that affect border formation, such as a river or an ethnic boundary separating two locations. If two points attract

[11]Where settlement areas overlap, we compute the share of non-mutual groups in j and k.

each other, they are likely part of the same state. If pushed apart by repulsive forces, they may become divided by a border. Each point is attracted to or repulsed by multiple neighboring points, but can only be part of one state. Directly capturing spatial dependence by only allowing for contiguous and non-overlapping state territories, a point's ultimate state "membership" is therefore the probabilistic result of the interplay of the attraction and repulsion exerted by and among all its neighbors and their state memberships.

Our Probabilistic Spatial Partition Model (PSPM) captures this logic by modeling the partitioning of a planar graph. The model allows us to estimate the attractive or repulsive forces resulting from attributes of the graph's edges. When estimating the effect of ethnic differences on state borders, we can thus account for covariates that influence ethnic settlement patterns and state borders. We next present and validate the PSPM and then introduce our empirical strategy to test our theoretical argument.

Probabilistic Spatial Partition Model

We model state territories as contiguous and mutually exclusive clusters of nodes (partitions) of graph G introduced above. Our modeling objective is to estimate the magnitude and uncertainty of the effects of edge-level attributes while accounting for spatial dependencies in the graph. We here present the models' fundamentals, explain our approach to estimation and

uncertainty, and validate the results with Monte Carlo experiments. Appendix A contains all further details.

The model: We model the distribution over all possible partitionings P of lattice G as a Boltzmann distribution:

$$Pr(P = p_i) = \frac{e^{-\epsilon_i}}{\sum_{i=1}^{|\mathbb{P}|} e^{-\epsilon_i}}, \tag{2}$$

where the realization probability of partitioning p_i decreases with its *energy* ϵ_i. The term energy reflects the origin of the Boltzmann distribution in modeling the condition of a system in statistical mechanics (e.g., Park and Newman 2004).[12] Because systems typically move towards a low energy, low-energy partitionings have higher probabilities. Applied to the partitioning of space into states, we can interpret the energy ϵ_i as the sum of inter- and intrastate tensions that result from a given partitioning.

Figure 3 illustrates this intuition for a simple graph of four vertices. The plot maps five (out of twelve possible) partitionings, with "countries" shown as nodes' color and number. Solid edges run within country borders and dashed ones across them. The top and bottom edges span across the red boundary between two ethnic groups, while the top and left edges cross

[12]The PSPM can be reformulated as an Exponential Random Graph Model, where $P(Y = y_i)$ is the probability of the realization of subgraph y_i of lattice G where y_i exclusively connects members of the same partition.

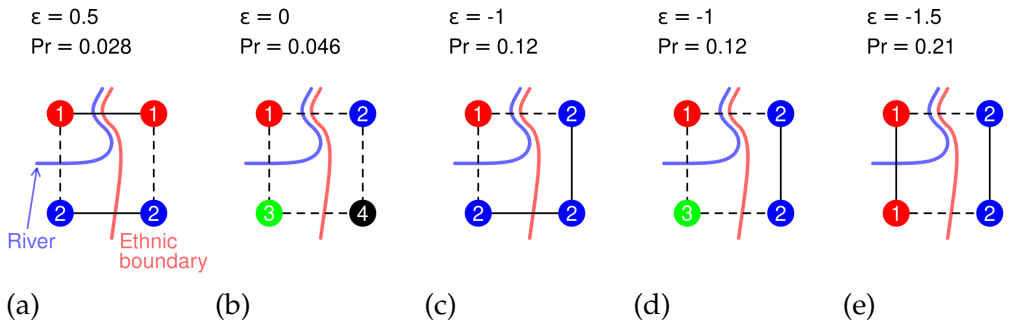


Figure 3: Illustration of the PSPM

Note: See main text for discussion. For illustrative purposes, we set parameters as $\beta_0 = -1$; $\beta_{\text{ethnic boundary}} = 1$, $\beta_{\text{river}} = 0.5$. The potential energy of each edge (from top, clockwise) is therefore .5, -1, 0, and -.5 (Eq. 5).

the blue river. For illustrative purposes, we assume that political tensions ϵ result when states are too small (b, d), multi-ethnic (a, c), or divided by the river (a, e). Intuitively, Eq. 2 holds that partitionings with ubiquitous tensions on the left have a lower probability than those with less tension to the right. Note also the spatial consistency of the graph. We cannot, for example, switch the left edge in (a) from dashed to solid since this would make the partitioning intransitive.

We assume that a partitioning's total energy ϵ_i is determined by the sum of realized energies of the edges that connect all first-degree neighbor node pairs L on the lattice:[13]

$$\epsilon_i = \sum_{j,k \in L} \epsilon_{j,k} * s_{j,k}, \tag{3}$$

[13]More complex total energy functions could account for higher-level predictors working, for example, at the level of emerging partitions (e.g., their size) or the partitioning as a whole (e.g., number of partitions or their size distribution).

whereby the potential energy $\epsilon_{j,k}$ of the edge between nodes j and k is realized if j and k are part of the same partition ($s_{j,k} = 1$, solid lines in Figure 3) and is not realized if they are part of different partition ($s_{j,k} = 0$, dotted lines in Figure 3). Our empirical interest focuses on the determinants of each edges' potential energy:

$$\epsilon_{j,k} = \beta_0 + \beta \, \mathbf{x}_{j,k}, \tag{4}$$

which defines the potential energy ϵ of the edge between nodes j and k as the sum of a constant β_0 that captures the baseline repulsion between nodes and edge-level characteristics $\mathbf{x}_{j,k}$ weighted by the parameter vector β. In our case and as discussed in the next section, $\mathbf{x}_{j,k}$ includes the indicator $\text{ethnic boundary}_{j,k}$ and additional edge-level covariates. While we have manually set the β parameters in Figure 3 for illustrative purposes, our empirical goal is to estimate them from the observed partitioning of Europe.

Because the realization probability of a partitioning decreases with its total energy (Eq. 2), coefficient estimates can be interpreted as follows: Variables associated with a positive estimate exert a *repulsive* force on nodes and increase the probability of them ending up in different partitions. Those with a negative estimate exert an *attractive* force, decreasing the chance that a border separates two points.

Applied to Figure 3 where we have manually set $\beta_{\text{ethnic boundary}} > \beta_{\text{river}}$, this means that ethnically aligned state territories have the highest probability (Panels d and e). Borders along the river in Panel (c) have a reduced

probability. Finally, because of a baseline attraction between nodes ($\beta_0 < 0$), partitionings with many small countries have a low likelihood (Panels b and d).

Because edges' values of $s_{j,k}$ are interdependent, it is difficult to interpret coefficients directly. This holds except for *bridge edges* that connect two otherwise disjoint network parts (i.e., a peninsula with the continent) and can therefore independently switch $s_{j,k}$ without violating transitivity. For these edges, we can interpret coefficient estimates as in a logistic regression model, computing odds ratios, predicted probabilities, and marginal effects (see also Cranmer and Desmarais 2011, p. 73).

Estimation and uncertainty: We estimate the β-parameters in Eq. (4) using a maximum composite likelihood approach (Lindsay 1988). Here, the likelihood function is the product over the conditional probabilities of vertices' observed partition memberships, defined based on their neighbors' memberships. We implement a Gibbs sampler that follows this logic to sample from the set of possible partitionings $|\mathbb{P}_G|$ of graph G, given edge-level predictors $\mathbf{x}_{i,j}$ and known parameters β. The sampler allows us to derive standard errors from a parametric bootstrap.[14]

Validation: We test the validity of inferences drawn from our model in an extensive series of Monte Carlo experiments presented in detail in Ap-

[14]See Appendix A.2.

pendix B. Our estimator is asymptotically unbiased in the size and number of independent networks across varying β parameter combinations, and parametric bootstrapping produces consistent frequentist uncertainty estimates. Appendix D.8 compares the PSPM with a benchmark that disregards spatial dependence, showing that the latter produces upwards-biased and overconfident estimates.

Empirical strategy

To test our main Hypothesis, we estimate the effect of ethnic geography on the partitioning of our spatial lattice G_t into states specifying the edge-level energy function as:

$$\epsilon_{j,k,t} = \beta_0 + \beta_1 \,\text{ethnic boundary}_{j,k,t} + \gamma \,\mathbf{X}_{j,k}, \tag{5}$$

where β_0 is the baseline repulsion between nodes and $\text{ethnic boundary}_{j,k,t}$ captures whether the nodes of an edge are located in different ethnic settlement areas (Eq. 1 above). To avoid bias from omitted spatial features, $\mathbf{X}_{j,k}$ must capture factors that cause ethnic as well as state borders. We therefore include time-invariant indicators for the length of each edge in kilometers,

the size of the largest river[15] and watershed[16] crossed by an edge, and the mean elevation (Hastings et al. 1999) along it. Taken together, these covariates capture important geographic causes of ethnic geography and state borders (e.g., Kitamura and Lagerlöf 2020). We scale all variables to range between 0 and 1 to ensure coefficients' comparability.

Our second analysis uses a lagged dependent variable (LDV) model to test whether ethnic boundaries affect border *change* such that both become increasingly congruent and address reverse causality as the main inferential threat affecting the baseline model. If ethnic settlement patterns results from identity formation within state borders (e.g., Hobsbawm 1990) the estimate of β_1 in Eq. 5 could be systematically biased. We therefore account for past borders leaving ethnic boundary to affect only border change:

$$\begin{aligned} \epsilon_{j,k,t} = \beta_0 + \beta_1 \,\text{ethnic boundary}_{j,k,t-1} + \beta_2 \,\text{state border}_{j,k,t-1} + \\ \beta_3 \,\text{deep lag}_{j,k} + \gamma \,\mathbf{X}_{j,k}, \end{aligned} \tag{6}$$

where we model edges' potential energy in period t as depending on ethnic and state borders 25 years earlier in $t-1$. In other words, to explain

[15]Based on a river scale in the Natural Earth data: https://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-rivers-lake-centerlines/. Appendix D.3 shows robustness to non-linear river effects.

[16]We derive an ordinal variable from Pfaffstetter watershed codes (Lehner, Verdin and Jarvis 2008).

state borders in 1936, we control for state borders in 1911 and construct ethnic boundary$_{j,k,t-1}$ from ethnic maps drawn between 1860 and 1910. Because ethnic boundaries are measured in data from the 50 years preceding the lagged dependent variable (Eq. 1), border change between $t-1$ and t cannot impact ethnic boundary$_{j,k,t-1}$. This avoids bias from reverse causality. Appendix D.1 shows robustness to interacting controls with state border$_{j,k,t-1}$ to differentiate between border emergence and persistence and to modeling duration dynamics.

Furthermore, borders in the deep historical past may have caused ethnic boundaries and may form precedents for "new" borders (Abramson and Carter 2016; Simmons 2005). To avoid such omitted variable bias, we add a "deep lag" of state borders, the share of years in which an edge crosses a border in AD 1100, 1200, ..., 1600, and 1790.[17] Because we lack early-19$^{\text{th}}$ century ethnic maps, we cannot estimate the LDV model for the 1886 outcome data.

We first estimate our baseline and LDV models on the pooled sample of all periods. In a second step, we estimate separate models for each period to gauge temporal variation ethnic geography's effects. Throughout, we use a parametric bootstrap to derive confidence intervals.[18]

[17]Data is from Abramson (2017) and stops in 1790.

[18]Appendix D.7 shows robustness to varying burn-in rates of the underlying Gibbs sampler.

Results

Overall, we find consistent support for our theoretical argument with a strong correlation of ethnic boundaries with state borders in the baseline model. Moreover, we find similarly sized effects in our LDV mode: even when accounting for current and past political borders, ethnic boundaries are strongly and positively related to the formation of new borders over the next 25 years.

Main results: Table 1 presents the main results obtained from estimating the baseline and LDV models on the pooled data. The findings support our theoretical argument and corroborate further predictions from the broader literature. The negative constant shows that the nodes in our lattice are generally *attracted* to each other when we set all covariates to zero. This attraction is mitigated by our independent variables.

First, the coefficient of (lagged) ethnic boundaries is positive: nodes separated by an ethnic boundary repulse each other and likely become separated by state borders. The respective effect is only slightly larger in the baseline than in the LDV model, which accounts for past borders and their determinants. The baseline estimates are thus not simply driven by reverse effects of state borders on ethnic geographies and omitted variables that affect both. Importantly, the effects of ethnic boundaries are sizeable. They are associated with almost two thirds of the energy attributed to a lagged state

Table 1: Determinants of state borders in Europe, 1886–2011

	1: Baseline	2: Lagged Dep. Var.
Constant	-2.50^*	-3.01^*
	$[-3.04; -1.91]$	$[-3.98; -2.47]$
Ethnic boundary$_t$	1.22^*	
	$[1.06; 1.40]$	
Ethnic boundary$_{t-1}$		1.02^*
		$[0.79; 1.24]$
State border$_{t-1}$		1.65^*
		$[1.46; 1.96]$
Deep lag		0.74^*
		$[0.36; 1.15]$
No. of periods	6	5
No. of vertices	6769	5412
No. of edges	17923	14243
No. of states	189	177
Controls	yes	yes

Notes: Each period t has a length of 25 years. 95% confidence intervals from parametric bootstrap in parenthesis. * Statistically significant at the 95% level.

border. Consistent with the prevalence of secessionist border change since 1886, we find that ethnic boundaries affect the emergence of new borders more than the stability of old ones (Appendix D.1).

Consistent with the findings by Abramson and Carter (2016), the LDV model shows that state borders from between the 10th and 18th century continue to separate nodes after 1886 conditional on ethnic geography. Shown in Appendix D.4, estimated effects of natural border determinants support previous arguments. Large watersheds and rivers, but not high altitudes are likely to divide locations into different states, in particular at a high spatial resolution and without conditioning on post-treatment ethnic boundaries and historical state borders.

Interpretation of effect sizes: Table 1 says little about the estimated absolute effect of ethnic boundaries on state borders. As discussed above, we can interpret the coefficients in parallel to those of a logistic regression for edges that bridge otherwise disjoint parts of the lattice and are therefore independent. For these bridge edges, the coefficient of ethnic boundary implies an odds ratio of 3.4 [2.9, 4.0][19] for the baseline model. Holding all covariates at their median values, an ethnic boundary thus leads to an increase in the probability of crossing a state border from 11.2 [9.7, 12.4] to 29.9 [27.6, 33.0] percent. The LDV model yields an odds ratio of 2.8 [2.2, 3.4] and a change in the border probability from 6.1 [4.6, 7.9] to 15.3 [11.0, 19.4] percent.[20] These substantial effects constitute a lower bound to the effects of ethnic boundaries which increase as they cross multiple interdependent edges.

For the more common case of *interdependent* edges, we use our estimates to sample 120 partitionings of the type plotted in Figure 4a and compute predicted border probabilities as the fraction of partitionings in which an edge crosses a border. The joint effect of all ethnic boundaries can be assessed by sampling two types of partitionings. The first type is sampled from the observed data in 2011 (Figure 4b). The second, counterfactual type is sampled assuming that all of Europe belongs to the same ethnic group[21]

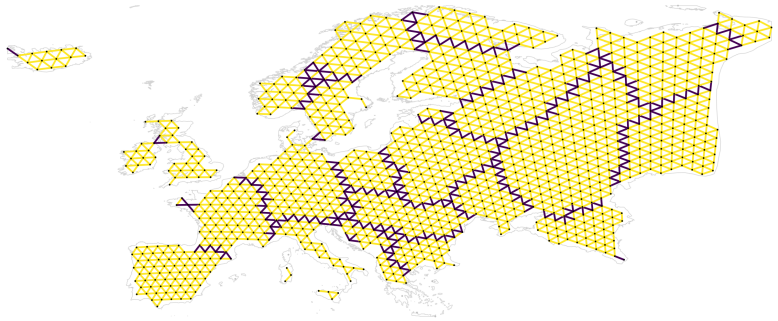
[19] 95% CI in parentheses.

[20] This change is conditional on no border in $t-1$, hence the lower probability.

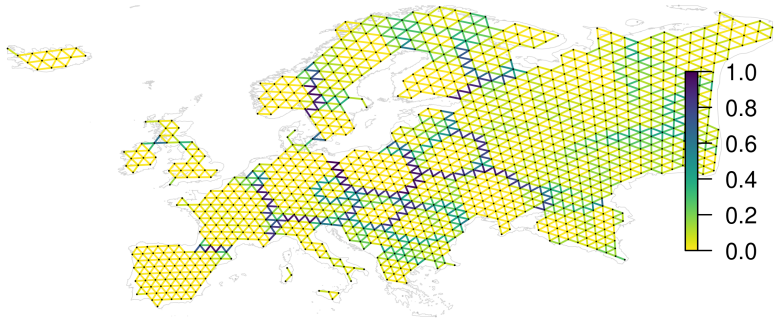
[21] I.e., setting all ethnic boundaries to zero.

but holding all other covariates at their observed values (4c).

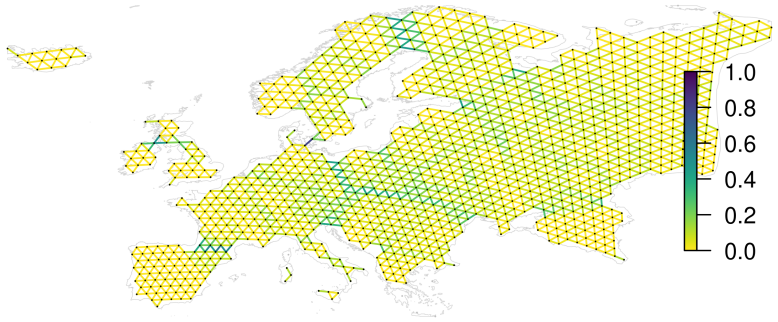
Predicted probabilities based on observed data in 4b overall closely resemble Europe's political map. Portugal is a prominent false negative, likely due its small size, narrowness, and rivers and watersheds that cross it. In the Balkans, diffuse border probabilities reflect overlapping ethnic settlement areas. Lastly, false positives cross Switzerland, a state that defies ethnically aligned borders. Yet, a comparison to Panel 4c shows that incorporating ethnic boundaries greatly improves our prediction, increasing the area under the ROC curve from 63 to 88 percent.



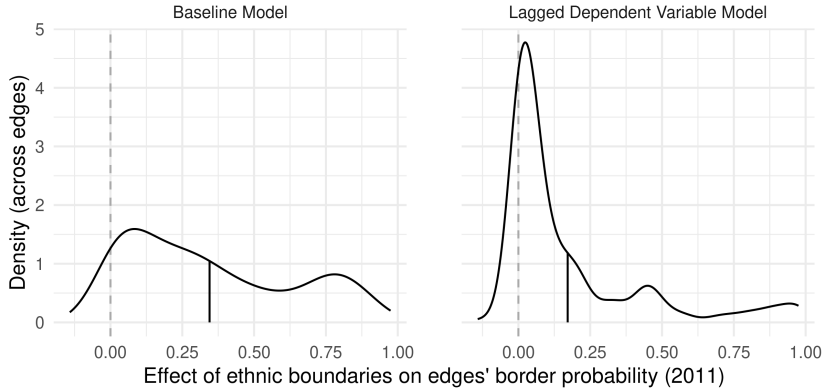
(a) One partitioning sampled from observed data (2011), baseline model



(b) Predicted border probabilities based on 120 partitionings sampled from observed data (2011), baseline model



(c) Border probabilities predicted without ethnic boundaries, baseline model



(d) Distribution of effect of ethnic boundaries on edge-level border probability

Figure 4: Effect of ethnic boundaries on edges' predicted border probability.

The difference between border probabilities in Panels 4b and 4c constitutes the joint effect of all observed ethnic boundaries, shown in Panel 4d. Being larger than the bridge-edge effects discussed above, ethnic boundaries increase border probabilities by 34 percentage points in the baseline model. In the LDV model, border probabilities increase by 17 percentage points over a relatively small baseline probability of border change. In sum, these results confirm a substantial effect of ethnic boundaries on the location of (newly drawn) state borders.

Variation over time: Figure 5 sheds light on temporal dynamics by showing separate estimates for each 25th year since 1886. Consistent with our argument, the baseline association between state borders and ethnic boundaries increases over time. The temporally disaggregated LDV models show

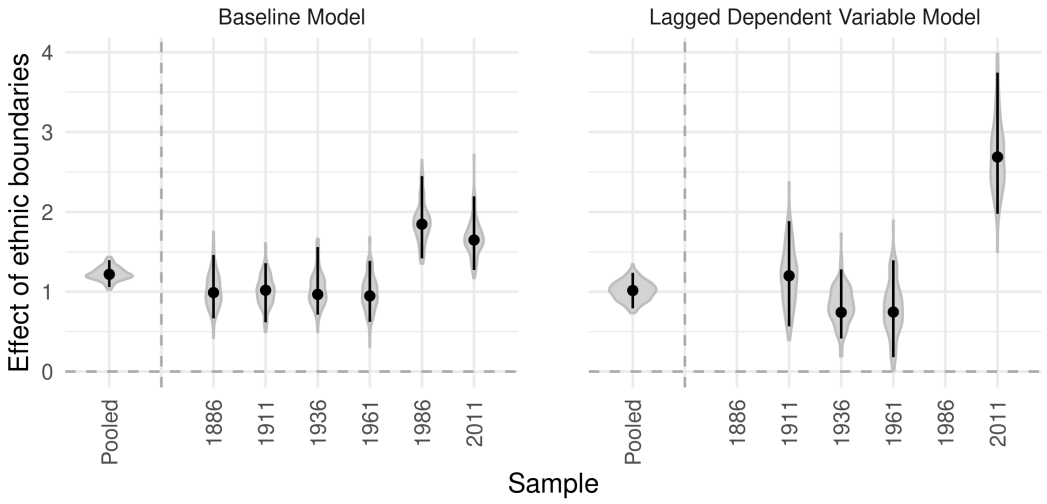


Figure 5: Effect of ethnic boundaries on the partitioning of Europe into states

Note: 95% CIs and grey areas show the distribution of bootstrapped estimates.

that ethnic geography affected *changes* in state borders particularly around the turn of the 19th century, World War I, and between 1986 and 2011 when the Soviet Union and Yugoslavia collapsed.[22] World War II brought slightly lesser ethnic alignment of state borders, and borders were stable from 1961 to 1986. In sum, systemic instability comes with nationalist border change (cf., Skocpol 1979; Abramson and Carter 2021).

Robustness checks

Our robustness checks assess whether the main findings are driven by potentially endogenous changes in ethnic geography, the choice of data on ethnicity and control variables, or the spatio-temporal data structure. Appendix D presents all details.

Pre-1886 ethnic boundaries: Political biases may affect in particular ethnic data produced during the World Wars. In addition, our results could be biased by omitted factors that first changed ethnic settlement patterns and, temporarily lagged, correlated border change. As a remedy, we use ethnic boundaries observed in the 50 years prior to 1886 as time-invariant predictor. The estimates in Figure 6 show effects of historical ethnic boundaries that are only marginally smaller than our baseline estimates. We also find an

[22]Post-Soviet and -Yugoslav borders mostly followed administrative borders often drawn based on ethnic geography (e.g., Hirsch 2000).

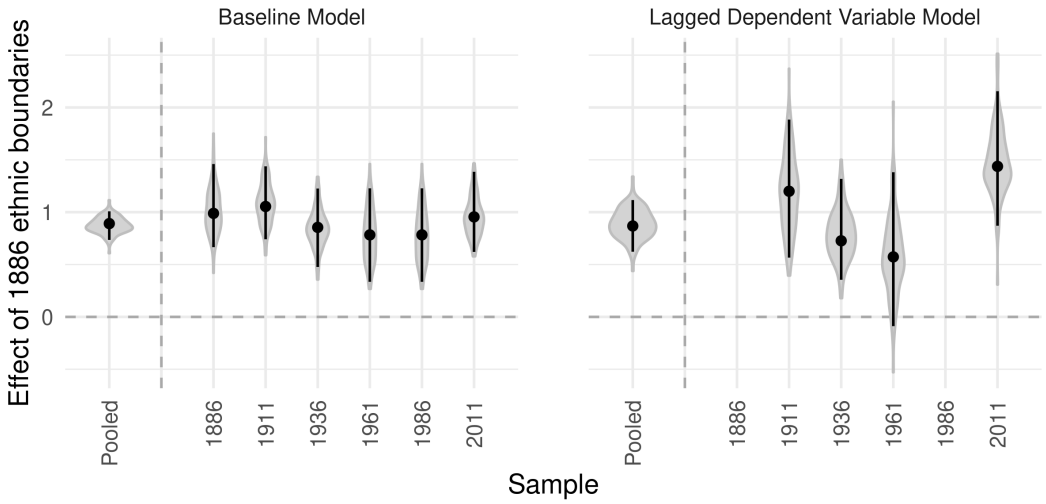


Figure 6: Effect of pre-1886 ethnic boundaries on the partitioning of Europe into states

Note: 95% CIs and grey areas show the distribution of bootstrapped estimates.

increasing alignment of state borders to ethnic boundaries in the LDV models. Reaffirming the absence of reverse causality and providing evidence against political bias, the LDV results show that pre-1886 ethnic boundaries affected border change even a century later. These results hold when we account for subnational regional borders in 1800 and 1900 (Appendix D.3).

Alternative measures of ethnic difference: We further test robustness regarding three alternative measures of edge-level ethnic differences (Appendix D.2). First, estimating our main specification for each ethnic map yields a smooth estimate distribution without "outlier maps" and evidences no undue influence of any one map. Second, we inquire whether effects of ethnic maps on ethnic identities may have caused our results. Such effects would most likely arise between linguistically close groups, yet our

estimates increase with the linguistic distance between groups. Third, politically biased ethnic maps may result from manipulated population thresholds. Using continuous census data on compositional ethnic differences between districts in pre-WWI Austria-Hungary yields stronger and more precise results, likely due to more precise measurement. In sum, we find no evidence that political or other biases from our historical ethnic maps affect our results.

Control variables: Our main results are insensitive to re-estimating models without controls or extending them to account for ruggedness, population density, the edges' geographic orientation, and non-linear river effects, as well as administrative borders in 1800 and 1900.

Variation of the data structure: We find that our results are robust to varying the length of periods t between 5 and 65 years.[23] We also vary the spatial data structure regarding (1) the graph's exact location, (2) its spatial resolution, and (3) its connectivity structure. For each variation, estimates remain statistically and substantially significant and similar to the baseline results. As additional evidence against potential bias from ethnic maps that are erroneous or manipulated, effects *increase* with coarser networks in which spatial error becomes less relevant and manipulation less likely.

[23] 65 years is the maximum period length that produces at least two periods.

In sum, our robustness checks show that the main results are not due to either endogenous changes in ethnic boundaries over time or potentially arbitrary modeling decisions of ours. The consistency of the results with early and alternative ethnic data as well as coarse spatial networks suggests the absence of substantive bias from political manipulation of ethnic data. In the next section, we provide evidence on secessionist claims and conflicts as an important mechanism through which ethnic geography shapes state borders in the age of nationalism.

Mechanism: Secessionist claims and conflict

Because there are more potential ethnic nations than realized states, secessionism likely drives much of the border-changing effects of nationalism. In an auxiliary analysis in Appendix E, we find that ethnically distinct peripheral areas were more likely experience ethnic secessionism since 1946. For this analysis, we recur to the vertices of our spatial network as units of analysis. For each year, we code whether a point became part of a secessonist claim (Germann and Schvitz 2023), was settled by a politically relevant ethnic group associated with an onset of secessionist civil war (Vogt et al. 2015), and became part of a newly independent state (Schvitz et al. 2022). We model the effect of co-ethnicity of the point with its state's capital on these outcomes using a Cox Proportional Hazard model, which mitigates the problem of successful secession leading to selection out of the treatment

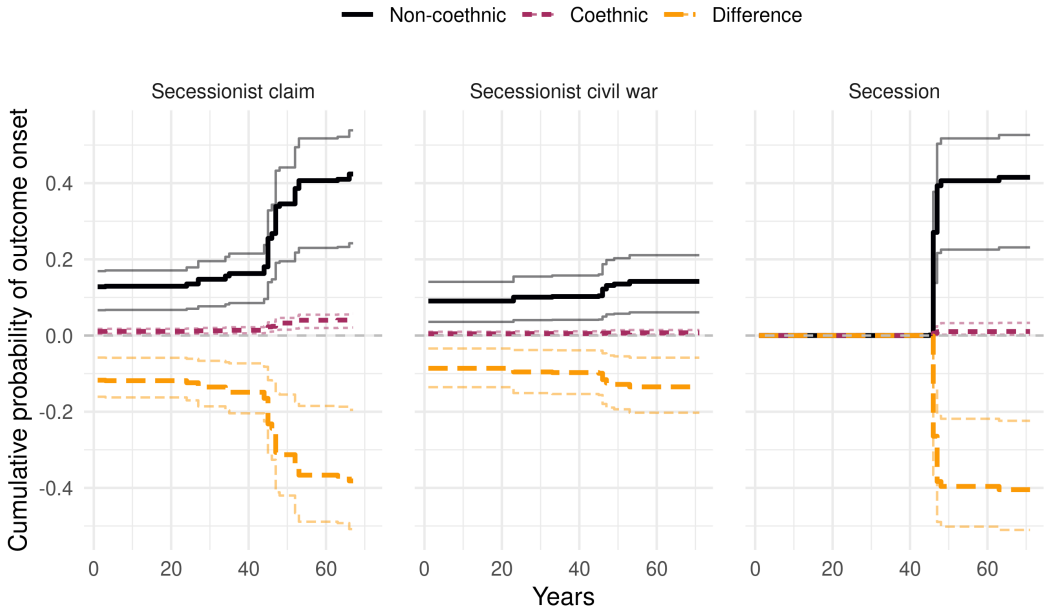


Figure 7: Effect of ethnic boundaries on secessionism.

Note: Predictions with 95% CIs based on Models 1, 3, and 5 in Appendix Table A6, setting covariates to median values.

group.

We find large and statistically significant effects of being ruled from a non-coethnic capital on demands for and realizations of secession. Over 50 years and holding covariates at their median value, Figure 7 shows that ethnically distinct regions have a probability of 35 percent to be part of a claimed, violently pursued (14 percent), or realized border change (41 percent). The respective probabilities for co-ethnic areas are close to zero. While the break-up of the USSR and Yugoslavia dominate the temporal pattern of secessions, our results hold when we stratify by country-year. In sum, they show that ethnic secessions drive the alignment of state borders with the ethnic map.

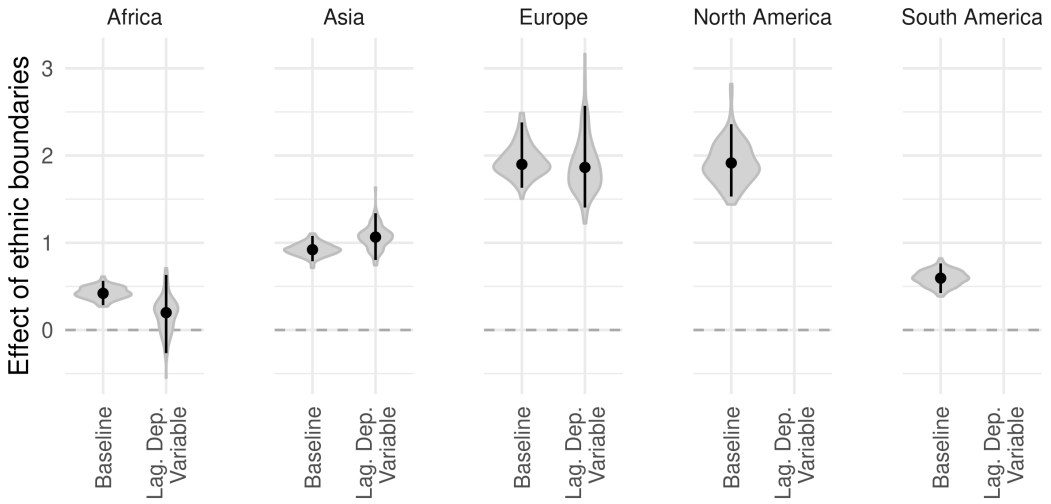


Figure 8: Effect of ethnic boundaries in 1964 on state borders across continents

Note: 95% CIs and grey areas show the distribution of bootstrapped estimates.

Global comparison

Our findings have so far been limited to 19th and 20th century Europe. We here analyze their generalizability by comparing the effects of ethnic geography on recent borders and border change across continents.

To do so, we create spatial lattices for each continent and use our main PSPM specifications to estimate the effect of ethnic boundaries on state borders in 2017. We use the earliest global data on ethnic geography from the 1963 Soviet *Atlas Narodov Mira* (Weidmann, Rød and Cederman 2010) and control for 1964 state borders in the LDV model.[24]

Starting with Africa, the results in Figure 8 support the conventional wisdom that decolonization and the *uti possedetis* norm preserved colonial

[24]Lacking global data, we omit the "deep lag."

borders drawn with little reference to ethnic geography (Griffiths 2015; Michalopoulos and Papaioannou 2016). The baseline coefficient is relatively small (yet statistically significant, see also Paine, Qiu and Ricart-Huguet 2021) and the LDV result shows no significant effect on border changes since 1964. Ethnic boundaries have had a more substantive effect on Asian borders. Though "only" half the size compared to Europe, ethnic boundaries significantly correlate with borders in 2017 and with post-1964 border change, mostly driven by Soviet Republics' independence. Lastly, we observe a stronger cross-sectional correlation between ethnic and state boundaries in North than in South America. The absence of recent border change prohibits estimating LDV models. In an auxiliary test in Appendix D.9, we find that ethnic boundaries have a larger effects on border change in densely populated regions in Europe and globally, suggesting that the nationalist reshaping of states occurs mostly where territory is of high value and competed over (cf. Herbst 2000).

In sum, these results yield two insights. First, state borders are cross-sectionally aligned with ethnic boundaries at a global scale, with states in Africa showing the least alignment. Second, ethnic boundaries seem to affect border change in Asia and Europe but not elsewhere. Ongoing ethnonationalist conflicts from secessionist Kurdistan to border disputes between India and Pakistan suggest an ongoing risk of ethnic reshaping of Asian states. In contrast, outright secessionist conflict is rare in Africa where

the territorial integrity norm is generally upheld (Englebert and Hummel 2005; Zacher 2001), low population densities decrease territorial competition (Herbst 2000), but ethnic conflict fragments some states internally.

Conclusion

Assessing nationalism's impact empirically, this study has analyzed whether, by how much, and how the nationalist principle reshaped European states along ethnic boundaries since 1886. Bringing systematic evidence to bear, we contribute to the literature on state and border formation which has so far been relatively fragmented as regards the ethnic origins of the partitioning of geographic space into states.

Theoretically, we have drawn on a rich yet mostly qualitative literature that highlights the impact of nationalism on international borders through secession and, less frequently, unification and irredentism. Over time, these processes gradually aligned state borders with the ethnic map. We have tested this proposition with new spatial data on ethnic settlement patterns since 1855 and a new Probabilistic Spatial Partition Model that allows us to estimate the effect of ethnic geography on the partitioning of Europe into states.

While developed for this study, the PSPM can be adapted to study other partitionings such as administrative units or electoral districts. To improve its flexibility, future developments could focus on supra-edge predictors,

different samplers, compositional membership outcomes, computational efficiency, and statistical properties. Lastly, innovative modellers may want to jointly assess the reciprocal relationship between state borders and ethnic geography, thus moving beyond the partial effects estimated here.

Our empirical results show that ethnic boundaries substantively affected borders and border change since 1886. We estimate that an ethnic boundary between two locations increases the likelihood of an interstate border between them by 34 percentage points. Conditional on past state borders, ethnic boundaries increase border probabilities by 17 percentage points. Supporting the claim that secessionist border change drives the ethnic reshaping of states, we find that peripheral ethnic minorities are at high risk to be subject to secessionist claims, conflict, and final break away. Our results also suggest the ethnic alignment of state borders to be ongoing macro-historical process. The Russian invasion of Ukraine and secessionist demands across the continent underscore the continuing centrality of nationalist revisionism in European politics. Looking beyond Europe, we have found similar dynamics of ethno-nationalist border change in Asia but less so elsewhere.

In sum, our findings suggest that ethnic geography has an important and continuing impact on the shape of European states. In consequence, the common treatment of states (and other political units) as fixed and exogenous entities comes at the risk of selection and reverse causality biases. Selection bias might, for example, deflate estimated effects of ethno-political

exclusion on conflict (Cederman, Gleditsch and Buhaug 2013) if previous secessions caused lower levels of ethnic exclusion and conflict. Reverse causality might inflate estimated effects of ethnic diversity on economic performance (Alesina and Ferrara 2005) if economic development sparked centripetal *and* centrifugal nationalism (Gellner 1983, ch. 7), secessions, and thus lower ethnic diversity. Knowing about units' origins is therefore an important prerequisite to inferring the consequences of at least some of their attributes.

Our analysis of post-1886 Europe being primarily structuralist, we caution against deterministic extrapolations. While the *potential* of ethnic centrifugal forces merits full recognition, previous research offers perspectives on how to contain them through ethnic power-sharing and regional accommodation (Cederman, Gleditsch and Buhaug 2013). More radical if perhaps utopian, dissociating states from nations altogether may succeed in depoliticizing ethnic divides (Mamdani 2020). Internationally, territorial integrity norms could rein in nationalist excesses (Zacher 2001), even though the recent revival of nationalist forces could endanger such progress.

References

Abramson, Scott and David B Carter. 2021. "Systemic Instability and the Emergence of Border Disputes." *International Organization* 75(1):103–146.

Abramson, Scott F. 2017. "The Economic Origins of the Territorial State." *International Organization* 71(1):97–130.

Abramson, Scott F and David B Carter. 2016. "The historical origins of territorial disputes." *The American Political Science Review* 110(4):675.

Alesina, Alberto and Eliana La Ferrara. 2005. "Ethnic diversity and economic performance." *Journal of economic literature* 43(3):762–800.

Alesina, Alberto and Enrico Spolaore. 1997. "On the number and size of nations." *The Quarterly Journal of Economics* 112(4):1027–1056.

Alesina, Alberto and Enrico Spolaore. 2005. *The size of nations*. Cambridge, MA: MIT Press.

Anderson, Benedict. 1991. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. 2nd ed. London: Verso.

Beissinger, Mark R. 2002. *Nationalist Mobilization and the Collapse of the Soviet Union*. Cambridge: Cambridge University Press.

Branch, Jordan. 2013. *The cartographic state: Maps, territory, and the origins of sovereignty*. Vol. 127 Cambridge University Press.

Bulutgil, H Zeynep. 2015. "Social cleavages, wartime experience, and ethnic cleansing in Europe." *Journal of Peace Research* 52(5):577–590.

Bulutgil, H Zeynep. 2016. *The Roots of ethnic cleansing in Europe*. Cambridge: Cambridge University Press.

Cadiot, Juliette. 2005. "Searching for nationality: statistics and national categories at the end of the Russian Empire (1897-1917)." *The Russian Review* 64(3):440–455.

Carter, David B and Hein E Goemans. 2011. "The making of the territorial order: New borders and the emergence of interstate conflict." *International Organization* 65(2):275–309.

Cederman, Lars-Erik, Kristian Skrede Gleditsch and Halvard Buhaug. 2013. *Inequality, Grievances, and Civil War*. Cambridge: Cambridge University Press.

Cederman, Lars-Erik, Seraina Rüegger and Guy Schvitz. 2022. "Redemption through Rebellion: Border Change, Lost Unity and Nationalist Conflict." *American Journal of Political Science* 66:24–42.

Coggins, Bridget. 2014. *Power Politics and State Formation in the Twentieth Century: The Dynamics of Recognition*. Cambridge: Cambridge University Press.

Conley, Timothy G. 1999. "GMM estimation with cross sectional dependence." *Journal of econometrics* 92(1):1–45.

Cranmer, Skyler J and Bruce A Desmarais. 2011. "Inferential network analysis with exponential random graph models." *Political analysis* 19(1):66–86.

De Luca, Giacomo, Roland Hodler, Paul A Raschky and Michele Valsecchi. 2018. "Ethnic favoritism: An axiom of politics?" *Journal of Development Economics* 132:115–129.

Desmet, Klaus, Michel Le Breton, Ignacio Ortuño-Ortín and Shlomo Weber. 2011. "The stability and breakup of nations: a quantitative analysis." *Journal of Economic Growth* 16(3):183.

Englebert, Pierre and Rebecca Hummel. 2005. "Let's stick together: Understanding Africa's secessionist deficit." *African Affairs* 104(416):399–427.

Englebert, Pierre, Stacy Tarango and Matthew Carter. 2002. "Dismemberment and suffocation: A contribution to the debate on African boundaries." *Comparative Political Studies* 35(10):1093–1118.

Fagan, Moira and Jacob Poushter. 2020. NATO Seen Favorably Across Member States. Report Pew Research Center.

Fazal, Tanisha M. 2004. "State death in the international system." *International Organization* 58(2):311–344.

Fazal, Tanisha M. 2007. *State Death: The Politics and Geography of Conquest, Occupation, and Annexation*. Princeton: Princeton University Press.

Fifield, Benjamin, Michael Higgins, Kosuke Imai and Alexander Tarr. 2020. "Automated redistricting simulation using Markov chain Monte Carlo." *Journal of Computational and Graphical Statistics* 29(4):715–728.

Friedman, David. 1977. "A Theory of the Size and Shape of Nations." *Journal of Political Economy* 85(1):59–77.

Gellner, Ernest. 1983. *Nations and Nationalism*. Ithaca: Cornell University Press.

Germann, Micha and Guy Schvitz. 2023. "Representing Self-Determination Claims in Space: The GeoSDM Dataset." Mimeo, University of Bath.

Germann, Micha and Nicholas Sambanis. 2021. "Political Exclusion, Lost Autonomy, and Escalating Conflict over Self-Determination." *International Organization* 75(1):178–203.

Goddard, Stacie E. 2006. "Uncommon ground: Indivisible territory and the politics of legitimacy." *International Organization* 60(1):35–68.

Goemans, Hein E. 2006. "Bounded communities: territoriality, territorial attachment, and conflict." *Territoriality and conflict in an era of globalization* pp. 25–61.

Goemans, Hein E and Kenneth A Schultz. 2017. "The politics of territorial claims: A geospatial approach applied to Africa." *International Organization* 71(1):31–64.

Griffiths, Ryan D. 2015. "Between Dissolution and Blood: How Administrative Lines and Categories Shape Secessionist Outcomes." *International Organization* 69(3):731–751.

Griffiths, Ryan D. 2016. *The Age of Secession: The International and Domestic Determinants of State Birth*. Cambridge: Cambridge University Press.

Hansen, Jason D. 2015. *Mapping the Germans: Statistical Science, Cartography, and the Visualization of the German Nation, 1848-1914*. Oxford Studies in Modern Europe.

Hardin, Russell. 1995. *One For All: the Logic of Group Conflict*. Princeton: Princeton University Press.

Hastings, David A, Paula K Dunbar, Gerald M Elphingstone, Mark Bootz, Hiroshi Murakami, Hiroshi Maruyama, Hiroshi Masaharu, Peter Holland, John Payne, Nevin A. Bryant, Thomas L. Logan, J.-P. Muller, Gunter Schreier and John S. MacDonald. 1999. "The global land one-kilometer base elevation (GLOBE) digital elevation model, version 1.0." *National Oceanic and Atmospheric Administration, National Geophysical Data Center* 325:80305–3328.

Hechter, Michael. 2000. *Containing Nationalism*. Oxford: Oxford University Press.

Hechter, Michael. 2013. *Alien rule*. Cambridge: Cambridge University Press.

Herb, Guntram Henrik. 2002. *Under the Map of Germany: Nationalism and propaganda 1918-1945*. Routledge.

Herbst, Jeffrey. 2000. *States and Power in Africa*. Princeton: Princeton University Press.

Hirsch, Francine. 1997. "The Soviet Union as a work-in-progress: ethnographers and the category nationality in the 1926, 1937, and 1939 censuses." *Slavic Review* 56(2):251–278.

Hirsch, Francine. 2000. "Toward an empire of nations: border-making and the formation of Soviet national identities." *The Russian Review* 59(2):201–226.

Hobsbawm, Eric J. 1990. *Nations and Nationalism Since 1780*. Cambridge: Cambridge University Press.

Hroch, Miroslav. 1985. *Social Preconditions of National Revival in Europe: A Comparative Analysis of the Social Composition of Patriotic Groups among the Smaller European Nations*. Cambridge: Cambridge University Press.

Kertzer, David and Dominique Arel. 2002. Census and identity. In *The Politics of Race, Ethnicity, and Language in National Censuses*, ed. Jack Caldwell, Andrew Cherlin, Tom Fricke, Frances Goldscheider et al. Cambridge: Cambridge University Press.

Kitamura, Shuhei and Nils-Petter Lagerlöf. 2020. "Geography and state fragmentation." *Journal of the European Economic Association* 18(4):1726–1769.

Lehner, Bernhard, Kristine Verdin and Andy Jarvis. 2008. "New global hydrography derived from spaceborne elevation data." *Eos, Transactions American Geophysical Union* 89(10):93–94.

Lindsay, Bruce G. 1988. "Composite likelihood methods." *Contemporary mathematics* 80(1):221–239.

Livingstone, David N and Charles WJ Withers, eds. 1999. *Geography and enlightenment*. Chicago: University of Chicago Press.

Mamdani, Mahmood. 2020. *Neither Settler nor Native*. Cambridge, MA: Harvard University Press.

Manela, Erez. 2007. *The Wilsonian Moment: Self-Determinatino and the International Origins of Anticolonial Nationalism*. Oxford: Oxford University Press.

McNamee, Lachlan and Anna Zhang. 2019. "Demographic Engineering and International Conflict: Evidence from China and the Former USSR." *International Organization* 73(2).

Michalopoulos, Stelios and Elias Papaioannou. 2016. "The long-run effects of the scramble for Africa." *American Economic Review* 106(7):1802–48.

Morgenthau, Hans. 1985. *Politics among nations: The struggle for power and peace.* New York: Knopf.

Murphy, Alexander. 2002. "National claims to territory in the modern state system: Geographical considerations." *Geopolitics* 7(2):193–214.

Mylonas, Harris. 2012. *The politics of nation-building: Making co-nationals, refugees, and minorities*. Cambridge University Press.

Nugent, Elizabeth R. 2020. "The Psychology of Repression and Polarization." *World Politics* 72(2):291–334.

O'Leary, Brendan. 2001. The elements of right-sizing and right-peopling the state. In *Right-sizing the state: The politics of moving borders*, ed. Brendan O'Leary, Ian Lustick, Thomas Callaghy, Thomas M Callaghy et al. Oxford University Press.

Paine, Jack, Xiaoyan Qiu and Joan Ricart-Huguet. 2021. "Endogenous Colonial Borders: Precolonial States and Geography in the Partition of Africa." *Available at SSRN 3934110* .

Palsky, Gilles. 2002. "Emmanuel de Martonne and the ethnographical cartography of central Europe (1917–1920)." *Imago Mundi* 54(1):111–119.

Park, Juyong and Mark EJ Newman. 2004. "Statistical mechanics of networks." *Physical Review E* 70(6):066117.

Petersen, Roger D. 2002. *Understanding Ethnic Violence: Fear, Hatred, and Resentment in Twentieth-Century Eastern Europe*. Cambridge: Cambridge University Press.

Ratner, Steven R. 1996. "Drawing a Better Line: Uti Possidetis and the Borders of New States." *American Journal of International Law* 90(4):590–624.

Roeder, Philip G. 2012. *Where nation-states come from*. Princeton University Press.

Sack, Robert David. 1986. *Human territoriality: its theory and history*. Cambridge University Press.

Sahlins, Peter. 1989. *Boundaries: the making of France and Spain in the Pyrenees*. Univ of California Press.

Sambanis, Nicholas and Jonah Schulhofer-Wohl. 2009. "What's in a line? Is partition a solution to civil war?" *International Security* 34(2):82–118.

Schvitz, G, S Rüegger, L Girardin, L-E Cederman, N Weidmann and KS Gleditsch. 2022. "Mapping The International System, 1886-2017: The Cshapes 2.0 Dataset." *Journal of Conflict Resolution* 66(1):144–161.

Simmons, Beth A. 2005. "Rules over real estate: trade, territorial conflict, and international borders as institution." *Journal of Conflict Resolution* 49(6):823–848.

Simmons, Beth A and Michael R Kenwick. 2021. "Border Orientation in a Globalizing World." *American Journal of Political Science, Early View* .

Siroky, David S and Christopher W Hale. 2017. "Inside irredentism: A global empirical analysis." *American Journal of Political Science* 61(1):117–128.

Skocpol, Theda. 1979. *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge: Cambridge University Press.

Tilly, Charles. 1978. *From Mobilization to Revolution*. New York: McGraw-Hill.

Vogt, Manuel, Nils-Christian Bormann, Seraina Rüegger, Lars-Erik Cederman, Philipp M Hunziker and Luc Girardin. 2015. "Integrating Data on Ethnicity, Geography, and Conflict: The Ethnic Power Relations Dataset Family." *Journal of Conflict Resolution* 59(7):1327–1342.

Weber, Eugen. 1976. *Peasants into Frenchmen: The Modernization of Rural France 1870-1914*. Stanford: Stanford University Press.

Weber, Max. 1978. *Economy and Society*. New York: Bedminster.

Weidmann, Nils B., Jan Ketil Rød and Lars-Erik Cederman. 2010. "Representing ethnic groups in space: A new dataset." *Journal of Peace Research* 47(4):491–499.

Weiner, Myron. 1971. "The Macedonian syndrome an historical model of international relations and political development." *World Politics* 23(4):665–683.

White, George W. 2004. *Nation, State and Territory. Origins, Evolutions and Relationships*. Lanham: Rowman & Littlefield.

Wucherpfennig, Julian, Aya Kachi, Nils-Christian Bormann and Philipp Hunziker. 2021. "A fast estimator for binary choice models with spatial, temporal, and spatio-temporal interdependence." *Political Analysis* 29(4):570–576.

Zacher, Mark W. 2001. "The Territorial Integrity Norm: International Boundaries and the Use of Force." *International Organization* 55(2):215–250.

Supplementary Material

Shaping States into Nations: The Effects of Ethnic Geography on State Borders

for online publication only

Table of Contents

A Probabilistic Spatial Partition Model

A.1 A distribution over partitionings

Our model operates on a lattice graph G, typically a planar graph with grid-like structure that is superimposed over the area of interest. G consists of N nodes and M edges, where edges connect neighboring nodes.

Our model is based on a probability distribution defined over all contiguous partitionings of G. A contiguous partitioning is an assignment of G's nodes into $K \leq N$ groups, called partitions, such that any two member nodes of a partition k are connected on G through a path that only passes through other member nodes of k. To give an example, consider a simple lattice with four nodes, arranged in a square, each connected to their two orthogonally adjacent neighbors. There are 12 contiguous partitionings possible on this baseline lattice: One where all nodes are isolated, 2 partitionings of 2+2, 4 partitionings of 3+1, 4 partitionings of 2+1+1, and one partitioning where all nodes are in the same partition.

We model the probability distribution over partitionings as

$$Pr(P = p_i) = Z^{-1} e^{-\epsilon_i}, \tag{A1}$$

a Boltzman distribution where P is a random variable denoting the partitioning of G, p_i is some realized partitioning with index i, and ϵ_i is the 'energy' associated with partitioning i. The term 'energy' for ϵ is owed to the Boltzman distribution's origin in statistical mechanics (Park and Newman 2004). Besides the usefulness of having a name for ϵ, ϵ can be intuitively interpreted as 'political tension' in the system when applying the model to political partitionings. Finally, Z is a normalizing sum,

$$Z = \sum_{i=1}^{|\mathbb{P}|} e^{-\epsilon_i}, \tag{A2}$$

with $\mathbb{P}$ being the set of possible contiguous partitionings.

In our model, the partitionings' total energy ϵ_i is the sum of all realized edge-level energies. Let $\epsilon_{j,k}$ represent the energy value of the edge that connects nodes j and k. Further, let $s_{j,k}$ be a realization variable that takes a value of 1 if nodes j and k are part of the same partition, and zero otherwise. Then we define

$$\epsilon_i = \sum_{j,k \in L} \epsilon_{j,k} * s_{j,k}, \tag{A3}$$

where L are the node pairs that are connected by G's edges.

Distribution (A1) assigns higher probability to partitionings where partition borders coincide with high-energy edges. This relationship allows us to formulate a model where the probability of observing any given partitioning is a function of edge-level covariates (like observed natural obstacles). We specify a linear relationship,

$$\epsilon_{j,k} = \beta \, \mathbf{x}_{j,k}, \tag{A4}$$

where $\mathbf{x}_{j,k}$ is a vector of edge-level covariates and a unit constant, and β is a parameter vector of corresponding length.

To illustrate how the edge-level covariates and parameters determine the probability of different partitionings, let us discuss a simple example. Say we have a covariate measuring whether an edge crosses a river. If the respective β parameter is positive, then the presence of rivers will increase the energy of all edges crossing rivers. As a result, ceteris paribus, partitionings where partition borders run along rivers are now more probable than other partitionings. Naturally, the same applies to any covariate measuring any type of distance. For these, positive β parameters imply that larger distances increase the likelihood of partition boundaries between nodes, and vice-versa for negative β parameters.

A.2 Sampling from the model

Before we discuss the estimation of our model, it is useful to discuss our approach to sampling. Note that sampling from the distribution over partitionings directly is infeasible for non-trivial sizes of G as the number of possible partitionings to iterate over grows exponentially.To our best knowledge, the exact function that maps lattices onto the number of possible contiguous partitionings is unknown. For instance, the number of possible contiguous partitionings of a 3x3 quadratic lattice is 1434; for a 10x10 quadratic lattice it is approximately 10^{45} (see Sloane et al. 2003, A145835).

A more practical approach is Gibbs sampling. Specifically, we sample the partition membership of each node in G, conditioned on the partition membership of all other nodes. A single Gibbs sample is completed once we have iterated over all nodes in the baseline lattice.

To illustrate our Gibbs sampling approach, it is useful to think of partition membership not as a node attribute, but as a relational attribute between any two nodes. To this end, let us slightly rewrite our probabilistic model over partitionings. Let H be a graph between all N nodes in G. H will have $N(N-1)/2$ edges. Each edge of H is associated with a binary random variable $S_{j,k}$ that captures whether nodes j and k are in the same partition ($s_{j,k} = 1$) or in distinct partitions ($s_{j,k} = 0$). Distribution (A1) can then be rewritten as

$$Pr(\mathbf{S} = \mathbf{s}) = \begin{cases} Z^{-1}\, exp\left(-\sum_{j,k \in L} \epsilon_{j,k} * s_{j,k}\right) & \text{if } \mathbf{s} \in \mathbb{P} \\ 0 & \text{otherwise,} \end{cases} \tag{A5}$$

where $\mathbb{P}$ is the set of valid contiguous partitionings on G, and $\mathbf{S}$ is a random vector of all $N(N-1)/2$ edge-wise S variables. Assigning a non-zero probability only if the realized state vector $\mathbf{s}$ is in $\mathbb{P}$ is necessary because there are many permutations of $\mathbf{s}$ that do not yield valid contiguous partitionings. For one, there are many permutations of $\mathbf{s}$ where transitivity is violated, e.g. where node pairs (j,k) and (k,l) are each assigned to the same partition ($s_{j,k} = 1$ and $s_{k,l} = 1$), but node pair (j,l) is not ($s_{j,l} = 0$). Moreover, there are many permutations of $\mathbf{s}$ where transitivity holds, but the partitioning is not contiguous. We assign these permutations a zero probability weight because they are not part of the sampling space of (A1).

We can sample from (A5) using block-wise Gibbs sampling. Specifically, we sample from the conditional distribution $Pr(\mathbf{S}_j|\mathbf{S}_{-j})$, where $\mathbf{S}_j$ is a vector of all S for those edges adjacent to node j, and $\mathbf{S}_{-j}$ is a vector of all remaining $\mathbf{S}$. In other

words, we sample the partition membership of node j conditioned on the partition memberships between all other nodes. The conditional distribution is given by

$$Pr(\mathbf{S}_j = \mathbf{s}_j | \mathbf{S}_{-j} = \mathbf{s}_{-j}) = \frac{Pr(\mathbf{S} = \mathbf{s})}{\sum_{\mathbf{s}'_j \in \mathbb{S}_j} Pr(\mathbf{S}_j = \mathbf{s}'_j | \mathbf{S}_{-j} = \mathbf{s}_{-j})}$$
$$= \begin{cases} \frac{exp\left(-\sum_{j,k \in N_j} \epsilon_{j,k} * s_{j,k}\right)}{\sum_{\mathbf{s}'_j \in \mathbb{S}_j} exp\left(-\sum_{j,k \in N_j} \epsilon_{j,k} * s'_{j,k}\right)} & \text{if } \mathbf{s} \in \mathbb{P} \\ 0 & \text{otherwise,} \end{cases} \tag{A6}$$

where $\mathbf{S}_j$ is the set of all possible permutations of $\mathbf{s}_j$ and N_j is the set of edges adjacent to node j in G. At first sight, expression (A6) seems difficult to sample from, as it requires us to sum over all 2^{N-1} permutations of $\mathbf{s}_j$. In practice, however, we only care about permutations that yield a valid contiguous partitioning, of which there are few. In fact, there are only two types: One where $\mathbf{s}_j$ is a zero-vector and node j forms its own partition, and one where node j is part of a partition in its neighborhood in G. These relevant permutations of $\mathbf{s}_j$ are very easily identified, and thus (A6) can be computed rapidly.

A.3 Estimation by Composite Likelihood

We are interested in obtaining an estimate for the parameter vector β. Ideally we would do so by exact maximum likelihood, i.e. by solving

$$\widehat{\beta} = \arg\max_{\beta} ln\ \widehat{\mathcal{L}}(\beta\ ;\ p, \mathbf{X}), \tag{A7}$$

where

$$ln\ \widehat{\mathcal{L}} = ln\ Pr(P = p \mid \beta, \mathbf{X})$$
$$= -(\sum_{j,k \in L} \mathbf{x}_{j,k}\beta * s_{j,k}) - ln(Z). \tag{A8}$$

p denotes the observed partitioning, and $s_{j,k}$ is a binary scalar indicating whether nodes j and k are observed to be in the same partition. Unfortunately, computing (A8) exactly is impossible for non-trivially sized Gs, as we would have to compute the normalizing sum Z.

Instead, we pursue a maximum composite likelihood approach, where we approximate the full likelihood using a product over conditionals (Lindsay 1988; Varin, Reid and Firth 2011). Specifically, we use expression (A6) and estimate β by maximizing the following log composite likelihood,

$$ln\ \widehat{\mathcal{L}}_C = \sum_{j=1}^{N} \ln\ Pr(\mathbf{S}_j = \mathbf{s}_j | \mathbf{S}_{-j} = \mathbf{s}_{-j}). \tag{A9}$$

This is similar in structure to the pseudolikelihood proposed by Besag (1974), with the key difference that Besag's model estimates vertex-level outcomes on a lattice, whereas we are interested in partition memberships. Though inefficient, maximum composite likelihood generally yields consistent estimates (Lindsay 1988). How-

ever, it is important to note that asymptotic theory only ensures consistency as the number of independent samples approaches infinity, not the number of random variables in the joint distribution that is approximated. In our case, this means that consistency is only ensured in the number of independent graphs G, not in the graph size N (Varin, Reid and Firth 2011). Hence, whether consistency also holds in N is an empirical question, which we address in Appendix B below.

In order to obtain stable estimates where the likelihood is relatively flat, we augment (A9) with a penalization parameter σ that nudges our estimate towards 0,Throughout this paper, we set $\sigma = 10$. thus obtaining our parameter estimates from

$$\widehat{\beta} = \arg\max_{\beta} ln\ \widehat{\mathcal{L}_C}(\beta\ ;\ p, \mathbf{X}) - \frac{\beta^2}{2\sigma} \tag{A10}$$

A.4 Standard errors

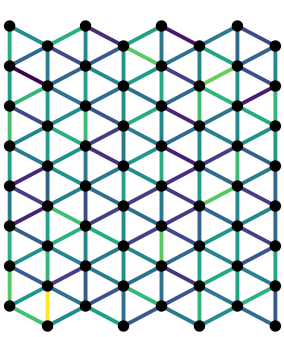
Because we estimate β by maximizing the (intentionally misspecified) composite likelihood (A9), we cannot use the observed Fisher information to estimate $var(\widehat{\beta})$. One common approach for computing appropriate standard errors for composite likelihood estimates is to substitute the Fisher information matrix with the Godambe information matrix (Godambe 1960). However, obtaining unbiased estimates of the Godambe matrix is difficult without many independent samples (Varin, Reid and Firth 2011, pp. 29ff). For this reason, we adopt a resampling approach, relying on a parametric bootstrap algorithm to estimate standard errors and confidence intervals (e.g., James et al. 2013, pp. 187-190).

Our algorithm consists of three steps. First, we obtain B partitioning samples from the fitted model using the Gibbs sampling (Section A.2, each with a separate Gibbs chain. To achieve good mixing, we initialize each chain by assigning each vertex its own partition and discard the first 100 'burn-in' samples.See Section B.2 for an evaluation of effects of the burn-in rate on parameter estimates. Second, we refit the model to each of the B partitioning samples, obtaining B parameter vectors $\widehat{\beta}^B$. Third, we obtain confidence interval estimates for parameter β_k by computing the empirical quantiles over the B β_k^B samples. See Section B.2 for simulation results showing unbiased coverage of the resulting confidence intervals.

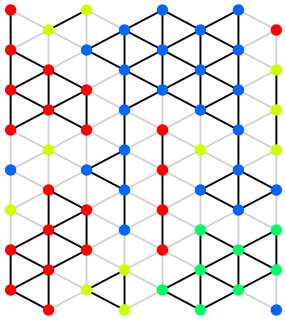
B Model Evaluation: Monte Carlo Simulations

We conduct Monte Carlo experiments to test the performance of our model and the Maximum Composite Likelihood estimator estimator. The main experiments explore potential biases in estimates recovered by the estimator and investigate the precision of uncertainty estimates while varying the (1) burn in rate of our sampler, (2) the size of networks, and (3) the number of independent instances. Biases stabilize after a relatively short burn in period and decrease with the size and number of networks. Biases are mainly concentrated in areas with separation issues. Standard errors derived from the Hessian of the Maximum Composite Likelihood estimator are consistent in most cases. Parametric bootstrapping offers an alternative method to derive uncertainty estimates.

B.1 Simulation setup



(a) Predictor.



(b) Sampled partitioning: $\beta_0 = -1$; $\beta_1 = 1$; burn-in rate of 100

Figure A1: Monte Carlo simulation setup

Our simulation setup is visualized in Figure A1. For every simulation, we construct a set of I instances of graphs G, each consisting of N vertices. Each lattice covers a quadratic area and exhibits a hexagonal network structure. Each edge is associated with a value of a single predictor. As shown in Figure A1a, the predictor x – the experimental equivalent to an ethnic boundary, river, or mountain ridge – is drawn from a normal distribution with mean 1 ($x \sim N(1,1)$) for the first, third, fifth, ..., column of edges, and from the normal distribution with mean 0 ($x \sim N(0,1)$) for all other columns as well as vertical edges. x are drawn once and stable across instances of experiments if equally-sized lattices. The differing means combined with random local variation introduce a 'typical' geographic structure similar to, e.g., mountain ranges.We use our Gibbs sampler to sample the partitioning of G based on the following edge-level energy function:

$$\epsilon_{j,k} = \beta_0 + \beta_1 \, x, \tag{A11}$$

where we experimentally set β_0 and β_1 to 'realistic' values. We let vertices have a baseline attraction (β_0) ranging between -2 and 0, and let the predictor's repulsion (β_1) range between 0 and 2.

In a last step, we use the sampled partition of G to estimate $\widehat{\beta}_0$ and $\widehat{\beta}_1$. For each experiment, we vary one set of parameters. For each parameter combination, we analyze 100 independently sampled networks. Table A1 summarizes the parameters governing each experiment, run on a high-performance server with 40 CPUs and 1.5TB RAM.

Table A1: Monte Carlo Experiment Parameters

	Parameter values:						
Experiment	Iterations	Beta 0	Beta 1	Network size	Instances	Burn-in rate	Std. error
1. Burn-in rate	100	[-2, -1, 0]	[0, 1, 2]	1024	1	[1, 5, 10, .., 1000]	–
2. Network size	100	[-2, -1, 0]	[0, 1, 2]	[16, 64, .., 4096]	1	100	–
3. Instances	100	[-2, -1, 0]	[0, 1, 2]	256	[1, 2, 4, 8, 16]	100	–
4. Para. bootstrap	100	[-2, -1, 0]	[0, 1, 2]	1024	1	100	Bootstrap

B.2 Simulation results

Following Table A1, we start by examining the upward or downward bias in the results of our experiments. The bias of an estimated $\widehat{\beta}_k$ parameter is defined in a straightforward manner as $\widehat{\beta}_k - \beta_k$. We examine this bias as a function of the burn-in rate, the size of graphs, and the number of independent graphs. Lastly, we examine the quality of confidence intervals derived from a parametric bootstrap. In sum, the results show that parameter estimate are asymptotically consistent and that estimate uncertainty is well reflected in the bootstrapped confidence intervals.

1. Burn-in rate: Figure A2 plots the results of experiment 1, examining the relationship between the burn-in rate of our Gibbs sampler and the bias in parameter estimates. The graph shows that the bias decreases quickly, approaching 0 only after 10–50 burn-in periods. In a set of experiments with a high baseline attraction between nodes ($\beta_0 = -2$) and no effect of our predictor ($\beta_1 = 0$), we see that the decrease in the bias in $\widehat{\beta}_0$ is matched by an *increase* in the bias in $\widehat{\beta}_1$. This is due to separation issues in the networks, which cause the two biases being negatively correlated. Based on these results, we choose as baseline burn-in rate of 100 for all following experiments and examine the behavior of estimate biases as we vary the size and number of networks.

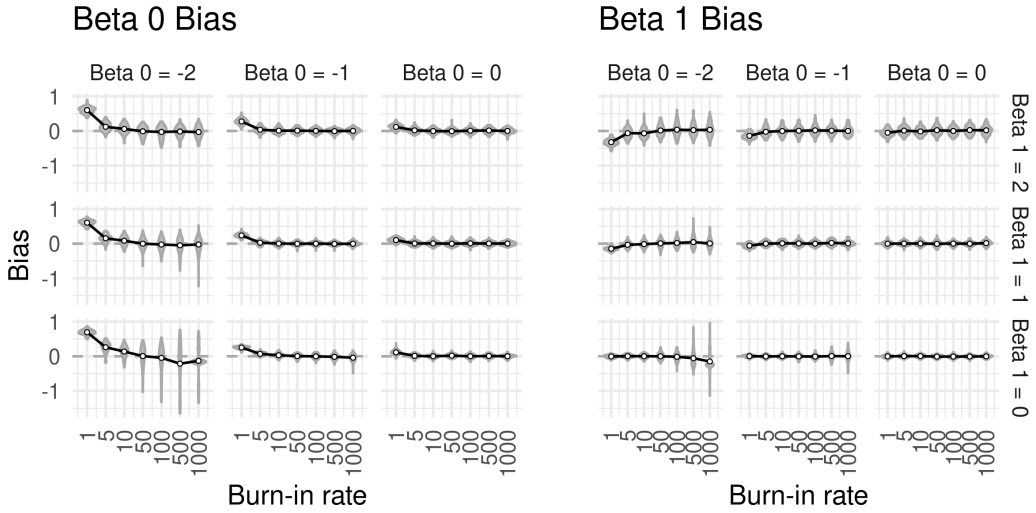


Figure A2: Bias in parameter estimates and the burn-in rate.

Note: Resulting from Monte Carlo simulations with the following parameters: 100 iterations; 1024 nodes on a hexagonal lattice; 1 instance; burn-in rate, β_0, and β_1 as shown in graph.

2. Network size: Next, we examine whether biases in our estimates decrease as we increase the size of networks. This is a necessary test as the consistency of the Maximum Composite Likelihood estimator is only ensured in the number of independent graphs G, not in the graph size N (see Section A.3 above; Varin, Reid and Firth 2011). Increasing the size of our experimental graphs in exponential steps from $N = 16$ to $N = 4096$ shows that the estimator is asymptotically consistent. As plotted in Figure A3 the estimator bias and variance decrease sharply in N and

approaches 0 for all combinations of $beta$ parameters. This decrease is slowest in areas where our data is vulnerable to separation problems, i.e. for $\beta_0 = -2$. With such high baseline attraction, only very large networks yield unbiased estimates.

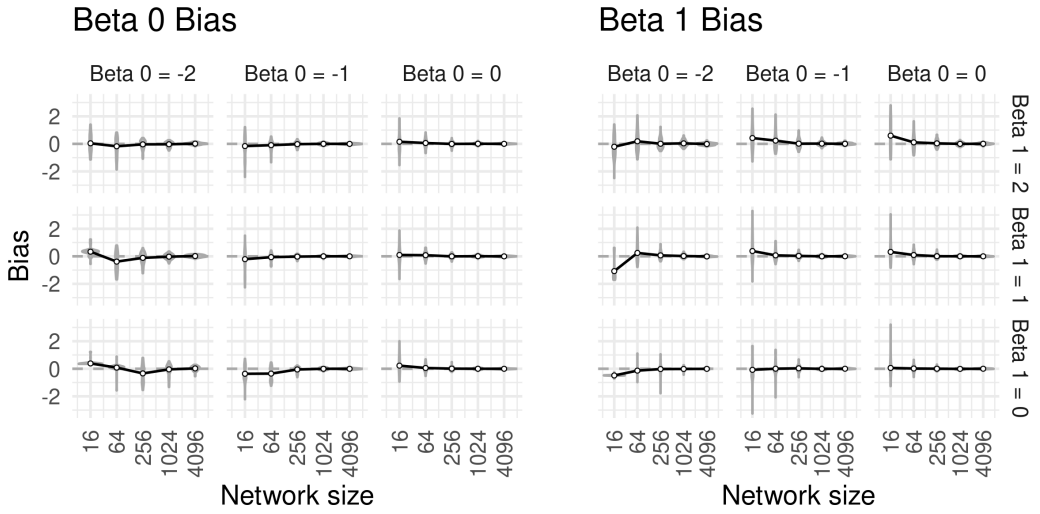


Figure A3: Bias in parameter estimates and the size of spatial lattices.

Note: Resulting from Monte Carlo simulations with the following parameters: 100 iterations; 1 instance each; burn-in rate of 100; network size (hexagonal structure), β_0, and β_1 as shown in graph.

3. Number of instances: In the next step, we test whether our estimator is asymptotically consistent in the number of independent instances of graphs G. For that purpose, we increase the number of instances in exponential steps from 1 to 16. Figure A4 shows that the resulting biases and variance in $\widehat{\beta}_0$ and $\widehat{\beta}_1$ decrease as our estimator draws on more independent data. We again note that this decrease is slowest in areas where our data is vulnerable to separation problems, i.e. for $\beta_0 = -2$. With this high baseline attraction between nodes, we need many (or large, or both) networks to obtain unbiased estimates.

4. Parametrically bootstrapped confidence intervals: Lastly, we test he consistency of our procedure for obtaining standard error described above in Section A.4. To that intent, we first compute bootstrapped 95% confidence intervals for the $beta$ estimates of 100 Monte Carlo experiments for each combination of β parameters. For each set of 100 experiments, we then compute the 'coverage' of confidence intervals, i.e. the fraction of confidence intervals that contain the real β value. If our bootstrapped confidence intervals are consistent, this fraction is close to and statistically indistinguishable from .95.

Figure A5 shows that for most β parameter combinations, around 95% of our bootstrapped confidence intervals contain the real value of β. Confidence intervals are slightly overconfident (i.e. too small) for very small values of $beta_0$. This result is directly related to the (small) biases that affect our estimates in this corner of the parameter space where separation problems occur. Statistically, it is not surprising that parametrically bootstrapped confidence intervals for biased estimates are not consistent. However, even for those biased cases, the resulting coverage gap is

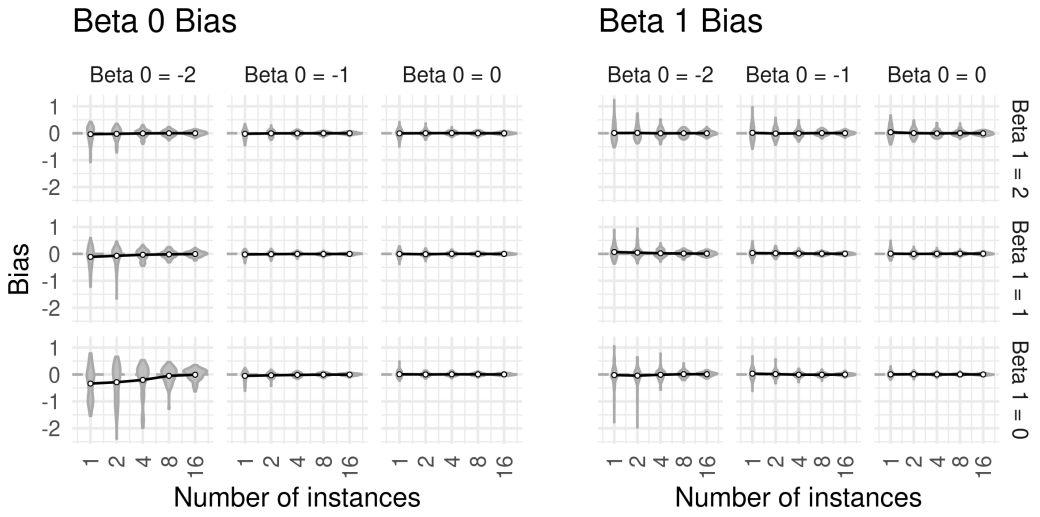


Figure A4: Bias in parameter estimates and the number of independent of spatial lattice instances.

Note: Resulting from Monte Carlo simulations with the following parameters: 100 iterations; network size $N = 256$; burn-in rate of 100; number of instances, β_0, and β_1 as plotted.

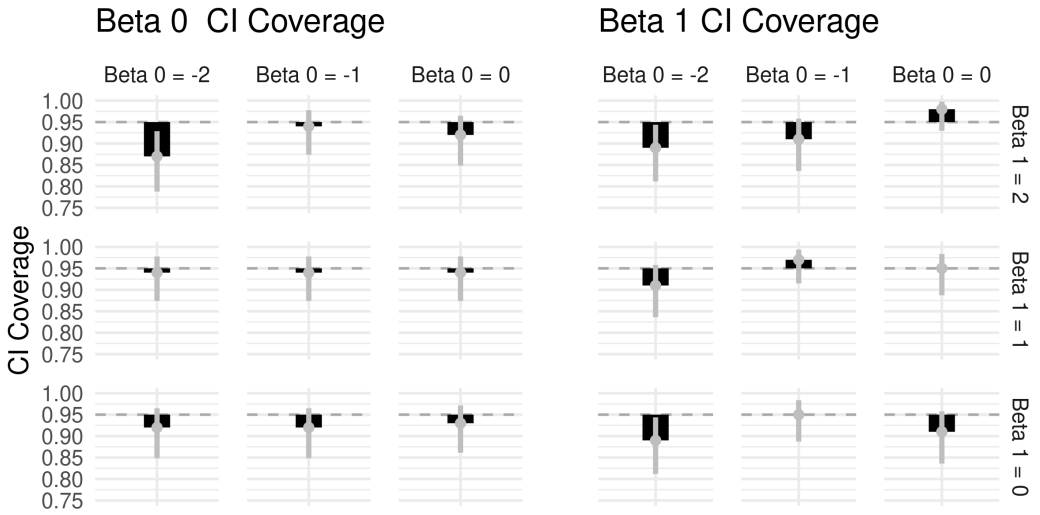


Figure A5: Precision of confidence interval coverage: Standard errors and 95% confidence intervals derived from parametric bootstraps (Section A.4.

Note: Grey bars denote the 95% confidence interval of the CI coverage estimates. Monte Carlo simulations with 100 iterations; 1 instance each; 1024 nodes; burn-in rate of 100; β_0 and β_1 as plotted.

relatively small (ca. 90% instead of 95%). Adding the above insight that our estimator is asymptotically consistent, these results show that the parametric bootstrap is able to derive consistent confidence intervals.

C Data

C.1 Historical ethnic map collection

We worked with a team of research assistants to gather ethnographic maps of Europe and the Levant from the 19th century to the present, relying on 25 different online and archival resources. This yielded a total of ca. 350 digital map scans,This count is approximate since we digitized many maps on the basis of library catalogue entries which ended up not being maps of ethnic groups in the first place. from which we selected 73 maps that we considered the most suitable. Five criteria determined maps' suitability. (1) Maps must depict ethnic settlement areas (as opposed to general maps of race or religion, or maps of groups' population share). (2) Maps should depict a snapshot in time close to the year they were published (as opposed to ex-post maps of historical ethnic geography). (3) Maps must have sufficient level of detail and precision. (4) They should not exhibit obvious signs of political bias. This led to the exclusion of the map in Figure A6 published by the Lithuanian National Committee in 1918. Not only is it published by a nationalist organization but also pictures an obviously inflated settlement area of Lithuanians if compared to other maps from the same time and area. We also exclude maps from German nationalist and national socialist Paul Langhans whose geographic journal was boycotted by geographers of the time for its political biases. We have not identified any other cases of egregious political biases in our maps. (5) Maps cannot be duplicates of other maps (some maps were just slightly altered, republished versions of earlier ones).

Table A2 lists all 73 maps that we use as source material, along with the relevant metadata and Figure A7 summarizes the resulting historical data on ethnicity. Panel (a) shows the (unstandardized) boundaries across all maps from all years (see (c) for the temporal distribution), showing a relatively 'clear' ethnic geography in Western Europe and substantial local ethnic diversity across Central and Eastern Europe.

In Panel (b), we present a systematic analysis of the overlaps between the same groups' depicted on different maps. The upper panel compares all maps with each other, while the lower panel compares all maps with the currently earliest digital ethnic data of Europe derived from the Soviet Atlas Narodov Mira (ANM; Bruk and Apenchenko 1964; Weidmann, Rød and Cederman 2010) only. For both comparisons, we "standardize" ethnic labels on the maps by linking them to the tree of languages (Lewis 2009) and only compare groups associated with the same language. In both analyses, we find high overlap with maps from the same decade overlapping to 90% with the ANM and to 80% across all maps. Maps diverge partially where they disagree on settlement patterns, in particular where populations are ethnically diverse. They also diverge because of differing definitions of ethnic groups that lead to imperfect linguistic standardization – i.e. some maps simply include Bretons in their definition of French and some do not. These uncertainties about the definition and geography of ethnic groups are captured by our edge-level ethnic boundary measure which averages across all maps from a given period. Additional robustness checks employing data on linguistic distances and compositional ethnic census data (D.2), estimating effects separately by map (D.2), and varying in spatial resolution (D.6) further address potential problems arising

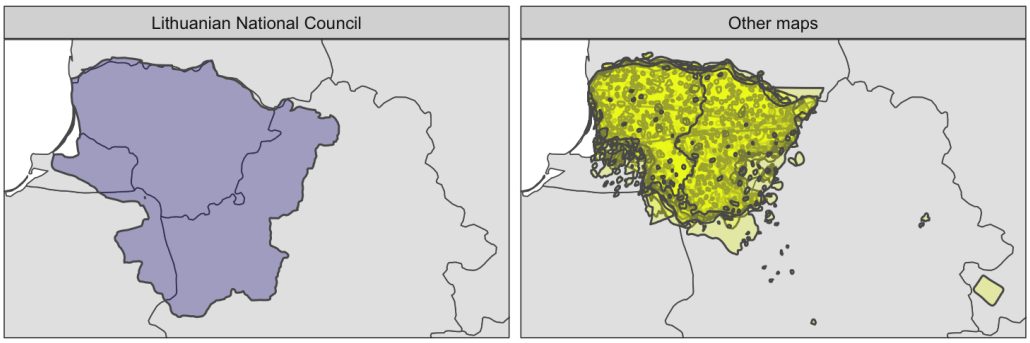


Figure A6: Lithuanian settlements as defined by the LNC compared to 24 other maps (1863-1963)

from the ethnic maps.

Lastly, ethnic change leads to increasingly large difference as the temporal distance between two maps increases. This is visible in Panel (d) which plots the intertemporal correlation of our edge-level ethnic boundary variable and shows clear breaks in Europe's ethnic geography after WWI and WWII.

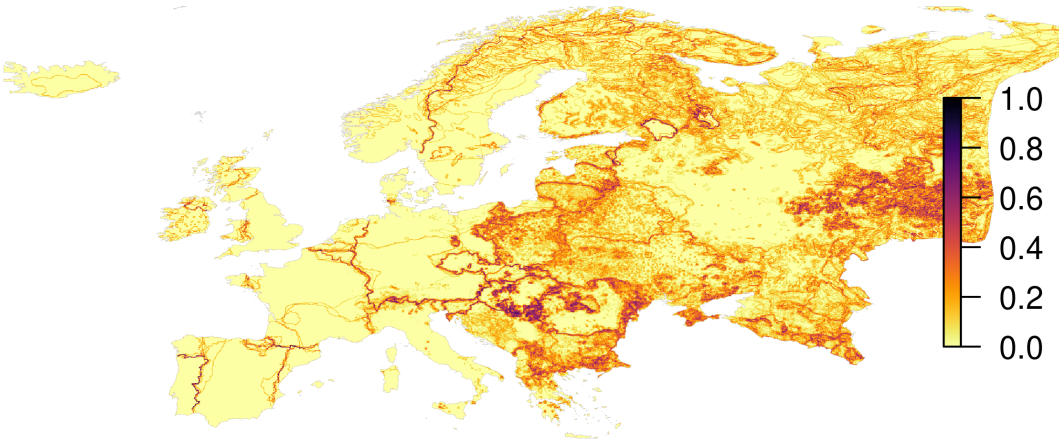
Figure A7 summarize maps temporal distribution (c), as well as the correlation of the final edge-level measure of ethnic boundaries over the main time periods in our analysis (d). We include the metadata and images of all digitized maps as well as examples of discarded ones in the replication files.

Title	Year	Author	Nationality
Ethnographische Karte der Östereichischen Monarchie	1855	Czoernig, Karl Freiherr von	Austrian
Carte Ethnographique de la Turquie d'Europe et des États Vassaux Autonomes	1861	Lejean, Guillaume Marie	French
Tableau Ethnographique	1863	Erckert, Roderich von	German
Völker und Sprachenkarte von Deutschland und den Nachbarländern	1867	D. Reimer	German
Völker- und Sprachen-Karte von Österreich und den Unter-Donau-Ländern	1869	Kiepert, Heinrich von	German
Europe Ethnographic	1870	Unknown (Russian author)	Russian
Specialkarte der deutsch-französischen Grenzländer mit Angabe der Sprachgrenze (neue berichtigte Ausgabe)	1870	Kiepert, Heinrich von	German
Ethnic Map of European Russia	1875	Rittikh, Aleksandr Fedorovich	Russian
Die Neueste Eintheilung, die Türkischen Gebiete & die Confessionen in der Türkei	1876	Petermann, August, Habenicht, Hermann	German
Ethnographische Übersicht des Europäischen Orients	1876	Kiepert, Heinrich von	German
Deutsche & Romanen in Süd-Tirol & Venetien	1877	Petermann, August	German
Ethnographische Karte der Europäischen Türkei	1877	Carl Sax	Austrian
Ethnographische Karte von Russland (Nördliches Blatt)	1878	Rittikh, Aleksandr Fedorovich	Russian
Ethnographische Karte von Russland (Südliches Blatt)	1878	Rittikh, Aleksandr Fedorovich	Russian
Etnograficheskaia Karta Kavkazskago Kraia	1878	Rittikh, Aleksandr Fedorovich	Russian
Vertheilung der Gross-, Weiss- & Klein-Russen	1878	Petermann, August	German
Europa um 1880	1880	Berghaus, Heinrich	German
Sprachen-Karte der westlichen Kronländer von Oesterreich	1880	Held, F.	Austrian
Sprachen-Karte von Österreich-Ungarn	1880	Franz Ritter v. Le Monnier	Austrian
Sprachenkarte, Religionskarte Schweiz	1881	Andree, Richard	German
Völkerkarte von Russland.	1881	Andree, Richard	German
Die Polen in Deutschland: Nordöstliches Deutschland nebst Polen. Ethnographische Karte	1885	Geographisches Institut Weimar	German
Politisch-Ethnographische Übersichtskarte von Bulgarien, Ost-Rumelien	1885	Geographisches Institut Weimar	German
Ethnographic map of Austria-Hungary and Romania	1892	Kiepert, Heinrich von	German

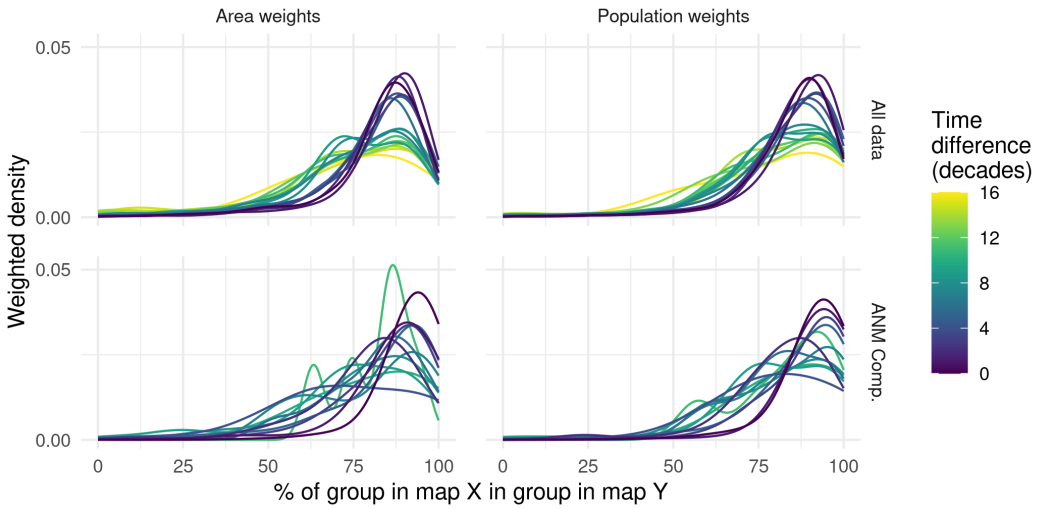
Title	Year	Author	Nationality
Völker- und Sprachenkarte von Mitteleuropa	1893	Karl Peucher	Austrian
Deutsches Reich. Religionskarte. Völkerkarte	1896	Diercke, Carl	German
Ethnographic map of Austria-Hungary	1900	Andree, Richard	German
Ethnographic map of the Balkan Peninsula.	1900	Andree, Richard	German
Völker u. Sprachenkarten. Europa. Konfessionskarten.	1902	Wagner, Hermann	German
Völkerkarte des rumänischen Sprachgebietes	1909	Weigand, Gustav	German
Die Sprachgebiete der Schweiz unter besonderer Berücksichtigung der Hohenregionen, nach Walser	1910	Deutsches Ausland-Institut, Isbert, O.A., Strotha, M.K.v	German
Map of Eastern Turkey in Asia, Syria and Western Persia (Ethnographical)	1910	Royal Geographical Society	British
Sprach- und Schulkarte Mähren und Schlesien	1910	Perko, Franz, Perko, Otto	Austrian
Das Bulgarentum auf der Balkanhalbinsel im Jahre 1912	1912	Ishirkov, A.	Bulgarian
Ethnographische Übersichtskarte von Osteuropa	1916	Freytag, G.	Austrian
Carte Ethnographique de l'Europe Centrale et des États Balkaniques.	1918	Bolzé, R., Chesneau, M.	French
Carte Ethnographique de la Péninsule des Balkans	1918	Cvijić, Jovan	Serbian
Ethnographic map of the Balkan Peninsula	1918	Cvijić, Jovan	Serbian
G. Freytags Völker und Sprachenkarte von Mittleuropa nebst Italien und der Balkanhalbinsel	1918	Freytag, G.	Austrian
Germany. Ethnographical map, Poland. Ethnographical map, Northern Italy. Ethnographical map, South East Europe. Ethnographical map	1918	Great Britain. General Staff. Geographical Section	British
The Daily Telegraph. Language map of Eastern Europe	1918	Gross, Alexander	Hungarian
Völker- und Sprachenkarte Österreich-Ungarn	1918	Mayer, Rudolf	German
Carte Ethnographique des Régions Habitées par les Roumains et des Colonies Étrangeres Qui s'y Trouvent	1919	Demetresco, Atanasiu, Borcea	Romanian
The Question of Thrace. Greeks, Bulgars and Turks	1919	Mills, J.S., Chrussachi, M.G.	British, Greek
Carte Ethnographique de l'Albanie	1920	Délegation de la Colonie Albaise de Turquie	Albanian
Völker und Staaten in Mitteleuropa	1924	Winkler, Wilhelm	Austrian
Carte ethnographique de l'Empire Ottoman. Faute de données statistiques exactes, depuis la Guerre balkanique [...].	1925	Unknown (French author)	French
Volksbodenkarte der Slowakei	1930	Isbert, O.A.	German
Carte ethnographique et linguistique de l'Europe nouvelle	1933	Wehrli, Max	Swiss
Völkerkarte der Sowjet-Union	1938	Klante, M. (Reichsamt für Landesaufnahme)	German
Die Völker des Donauraumes und der Balkanhalbinsel	1940	Generalstab des Heeres, Abteilung für Kriegskarten u. Vermessungswesen	German
Rumänien. Volksgruppen	1940	Generalstab des Heeres, Abteilung für Kriegskarten u. Vermessungswesen	German
Albanian Minority in Yugoslavia	1941	Great Britain. Foreign Office. Research Department.	British
Völkerkarte des Kaukasus. Aufgrund der vom Bataillon der Waffen-SS z. b. v. sichergestellten 'Ethnographischen Karte des Kaukasus.'	1942	Kommission für das Studium der Völker der UdSSR und ihrer Nachbarländer, Reichsamt für Landesaufnahme.	German
Poland language map	1945	United States. Office of Strategic Services. Research and Analysis Branch	USA
Karta Narodov SSSR. Uchebia dlia Spednei Shkoly.	1955	Unknown (Russian author)	Russian
Ethnic Map of the Soviet Union	1959	Main Directorate of Geodesy and Cartography, Ministry of Geology and Mineral Resources of the USSR	Russian
Atlas Narodov Mira / Geo-referencing of Ethnic Groups	1964	Bruk, S.I., Apenchenko, V.S., Digitized by Weidmann et al. (2010)	NA
Map of People of the USSR	1972	Main Directorate of Geodesy and Cartography, Ministry of Geology and Mineral Resources of the USSR	Russian

Title	Year	Author	Nationality
Cyprus, Ethnic Distribution	1973	U.S. Central Intelligence Agency	USA
Völker und Sprachen Europas unter besonderer Berücksichtigung der Volksgruppen	1978	Straka, Manfred	Austrian
Ethnic Groups in Southern Soviet Union and Neighboring Middle Eastern Countries	1986	U.S. Central Intelligence Agency	USA
Map of Slovenian Dialects	1986	Logar, Tine, Rigler, Jakob	Slovenian
Ethnic map of the Soviet Union	1988	Main Directorate of Geodesy and Cartography, Ministry of Geology and Mineral Resources of the USSR	Russian
Herrien Europa. Europa de Los Pueblos. L'Europe de Peuple. Europe of the People	1992	Herreros Agüi, Sebastián, Durán Rodriguez, Adolfo	Spanish
Ethnolinguistic Groups in the Caucasus Region	1995	U.S. Central Intelligence Agency	USA
The Levant: Ethnic Composition	2009	Izady, M.	Belgian
Languages of North Africa	2013	Izady, M.	Belgian
Ethnolinguistic Groups in the Caucasus and Vicinity	2014	Izady, M.	Belgian
Ethnologue / World Language Mapping System. Language Maps. Version 17	2014	SIL International	USA
Ethnic Ukrainians and Russians in the Caspian-Black Sea Basin	2017	Izady, M.	Belgian
Languages of Europe	2017	Unknown	
Middle East: Ethnic Groups	2020	Izady, M.	Belgian

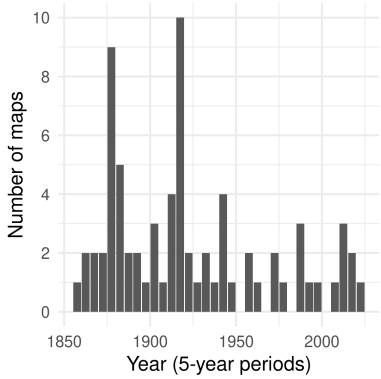
Table A2: List of 73 ethnographic maps used as source material



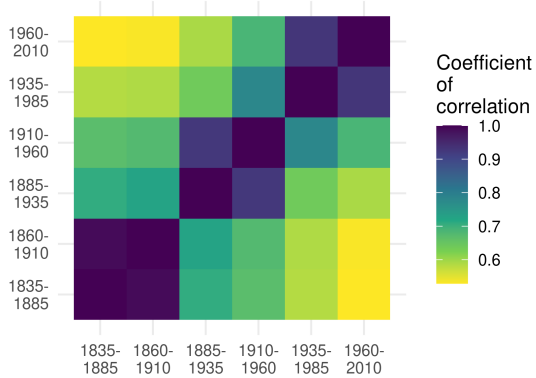
(a) Ethnic boundaries as fraction of maps covering an area, 1836–2010.



(b) Systematic comparison of ethnic settlement patterns across maps.



(c) Number of maps over time



(d) Correlation of ethnic boundary across periods t

Figure A7: Historical ethnic data: Summary

C.2 Data on self-determination claims: GeoSDM

To capture secessionist claims, we draw on new spatial data from GeoSDM (**?**). This dataset maps territorial claims made by 466 self-determination movements worldwide since 1945, as identified by the Self Dermination Movements (SDM) dataset (Sambanis, Germann and Schädel 2018). Our analysis is limited to secessionist claims in Europe, a subset of the full GeoSDM data.

GeoSDM codes the "dominant" territorial claim as expressed by representatives of each SDM. In addition, the dataset accounts for changes in territorial claims over time that may result from changes in international borders or changes in a group's stated objectives. Territorial claims are coded based on the detailed background information on each movement provided by the SDM dataset's supplementary information, as well as multiple primary and secondary sources describing the territories claimed by separatist movements (e.g. Minahan 1996, 2002; Roth 2015). Where possible, GeoSDM relies on existing spatial datasets to geocode territorial claims (e.g GADM 2019; Weidmann, Rød and Cederman 2010). Where available GIS data was insufficient, claim polygons were based on digitized maps, mostly taken from Roth (2015).

D Robustness checks: Probabilistic Spatial Partition Model

This section presents the design and results of robustness checks of the paper's main analysis.

D.1 Border emergence vs. persistence and duration dynamics

Drawing on studies of temporal dynamics in binary time-series-cross-sectional data, we here assess (1) differential effects of ethnic boundaries on border emergence and persistence and (2) potential bias from unmodelled temporal dynamics in the lagged dependent variable specification.

Because processes of border emergence differ from those of border persistence, we test the assumption implicit in the lagged dependent variable model that ethnic boundaries have similar effects on both. Discussed by Beck et al. (2001) in their treatment of restricted transition Probit models, we can do so by interacting all variables with the lagged dependent variable. Beyond testing whether ethnic boundaries similarly affect border emergence and persistence, the difference between their effect on both processes tests whether ethnic secession (newly emerging borders) or unification (disappearance of old borders) drive the results. Irredentism features both types of border change in parallel. Table A3 present the results, first introducing only interactions of with the control variables in Model 1, then only interacting ethnic boundaries with lagged state borders, and finally showing the fully specified model. The results in 1 show that our main estimate of ethnic boundaries is if at all downward biased by the exclusion of controls interacted with the state border lag. Consistent with Model 2, Model 3 shows that ethnic boundaries' effects on the emergence of new borders (1.35 [1.07, 1.71]) is more than twice as large as that on the persistence of old borders (0.53 [-0.07, 1.10], $p < .1$), with the difference amounting to -0.82 [-1.45, -0.18]. The results thus show that secessions and the

creation of new borders substantively but not exclusively drive the effect of ethnic boundaries in the lagged dependent variable model.

Table A3: Border emergence, stability, and duration

	(1)	(2)	(3)	(4)	(5)
Constant	-2.87^*	-3.13^*	-2.96^*	-1.80^*	-1.67^*
	$[-3.80; -2.13]$	$[-3.94; -2.39]$	$[-3.88; -2.11]$	$[-2.64; -0.86]$	$[-2.72; -0.82]$
State border$_{t-1}$ (SB)	0.79	2.12^*	1.09^*	0.62^*	-0.03
	$[-0.22; 2.12]$	$[1.71; 2.53]$	$[0.09; 2.62]$	$[0.08; 1.21]$	$[-1.05; 1.60]$
Ethnic boundary$_{t-1}$ (EB)	1.11^*	1.34^*	1.35^*	0.95^*	1.27^*
	$[0.89; 1.32]$	$[0.97; 1.60]$	$[1.07; 1.71]$	$[0.69; 1.28]$	$[0.99; 1.65]$
EB$_{t-1}$ × SB$_{t-1}$		-0.98^*	-0.82^*		-0.82^*
		$[-1.54; -0.28]$	$[-1.45; -0.18]$		$[-1.53; -0.17]$
No. of periods	5	5	5	5	5
No. of vertices	5412	5412	5412	5412	5412
No. of edges	14243	14243	14243	14243	14243
No. of states	177	177	177	177	177
Controls	yes	yes	yes	yes	yes
SB$_{t-1}$ × controls	yes	no	yes	no	yes
Cubic duration	no	no	no	yes	yes
SB$_{t-1}$ × cub. dur.	no	no	no	yes	yes

Notes: Each period t has a length of 25 years. 95% confidence intervals from parametric bootstrap in parenthesis. * Statistically significant at the 95% level.

A second addition to the main lagged dependent variable specification addresses concerns that unmodelled *duration dynamics* might bias the results (Beck, Katz and Tucker 1998). First, since borders' stability increases and their chance of reemergence decreases with time, our outcome partitioning in t is likely dependent on the outcome in all previous periods. Second, the duration of the presence or absence of a border might directly affect ethnic geography: places that have been located in the same state for a long time are more likely to be the same ethnic settlement area than places that until recently have been separated by a state border. This may bias our estimates upwards. In order to control for such temporal dynamics, we follow Carter and Signorino (2010) and construct an edge-level cubic polynomial of the time since an edge has attained its border-crossing status (0/1) observed in $t-1$.To that intent, we combine our 25-yearly edge-level cshapes-based variable with the pre-1790 variables that are coded based on Abramson (2017). To account for differing duration dependence of borders and non-borders, we then interact the resulting three polynomial terms with edges' border-crossing status in $t-1$. As Model 4 in Table A3 shows, doing so does not substantively affect the estimated effect of ethnic boundaries. Model 5 lastly adds the duration controls to Model 3, showing again stable estimates. In sum, this suggests that the effects of ethnic boundaries are duration independent.

D.2 Varying measures of geospatial ethnic difference

Effects by ethnic map:

To gauge whether the results are driven by a (small) set of potentially biased maps, we re-estimate the main specifications separately for each map, using only data on ethnic geography from that map. We only include periods t observed after the creation date of a map. Naturally, as maps are of different size and cover different areas, not all of them yield positive and statistically significant estimates. Yet, the

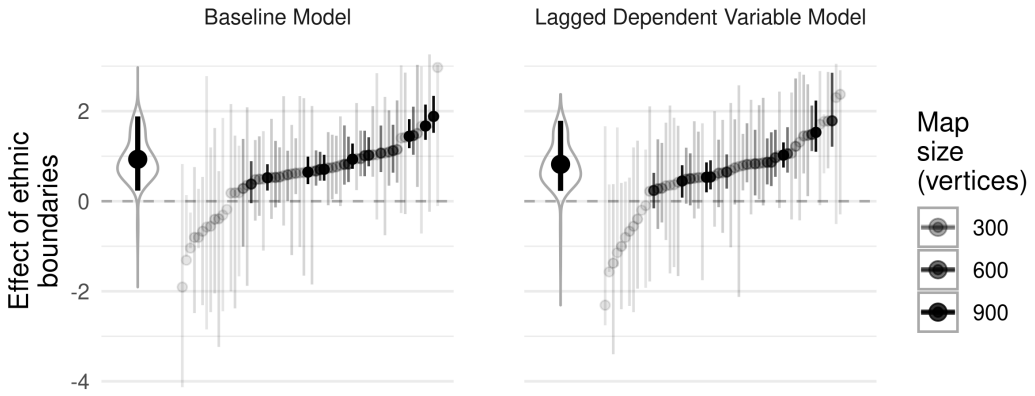


Figure A8: Effect of ethnic boundaries on the partitioning of Europe into states, estimated separately for each ethnic map

Note: Pooled sample. Darker coefficients are based on maps that cover larger areas of our spatial graph. 95% CIs show the distribution of bootstrapped estimates. Large coefficients to the left of each plot show the size-weighted average estimate across all maps with 95% CI.

distributions plotted in Figure A8 show that only a quarter of all maps yields estimates below 0,These estimates originate from small maps and are all statistically insignificant. that all large maps feature positive and significant estimates and that the median map yields estimates that is similar to our main estimates. We find no sign of undue influence exerted by a few, potentially biased, "outlier maps."

Linguistic distance:

We test whether linguistic distances as a continuous measure of ethnic difference affect border probabilities in our spatial graph. To that intent, we manually encode the language(s) associated with the ethnic groups depicted on each map in our map collection. On the basis and following Fearon (2003), we use this information to compute, for every edge in the graph, the linguistic distance it traverses on the philogenetic tree of all known languages (Lewis 2009). We compute the distance between two languages L_1 and L_2 as the fraction of their paths to the linguistic tree root they share:

$$D_{L_1,L_2} = 1 - \left(\frac{2d(w(L_1,..,O) \cap w(L_2,..,O))}{d(w(L_1,..,O)) + d(w(L_2,..,O))} \right)^{\delta},$$

where $d(w(L_1,..,O))$ is the length of L_1's path to the origin and $d(w(L_1,..,O) \cap w(L_2,..,O))$ is the length of the paths' intersection. δ is an exponent to discount distances further away from the root of the tree and is set to .5 (Fearon 2003).Where vertices are located in overlapping settlement areas, we take, for each group of a vertex, the minimum distance to the other vertex' groups and average across groups. This yields an edge-level linguistic distance measure that is bounded between 0 (same groups) and 1 (e.g., Magyar-German). For further analysis of potentially non-linear effects, we cut this measure into four bins of roughly equal size: [0], (0,.25], (.25,.5], (.5,1], containing (in 1886) 1640, 482, 464, and 461 edges, respectively.

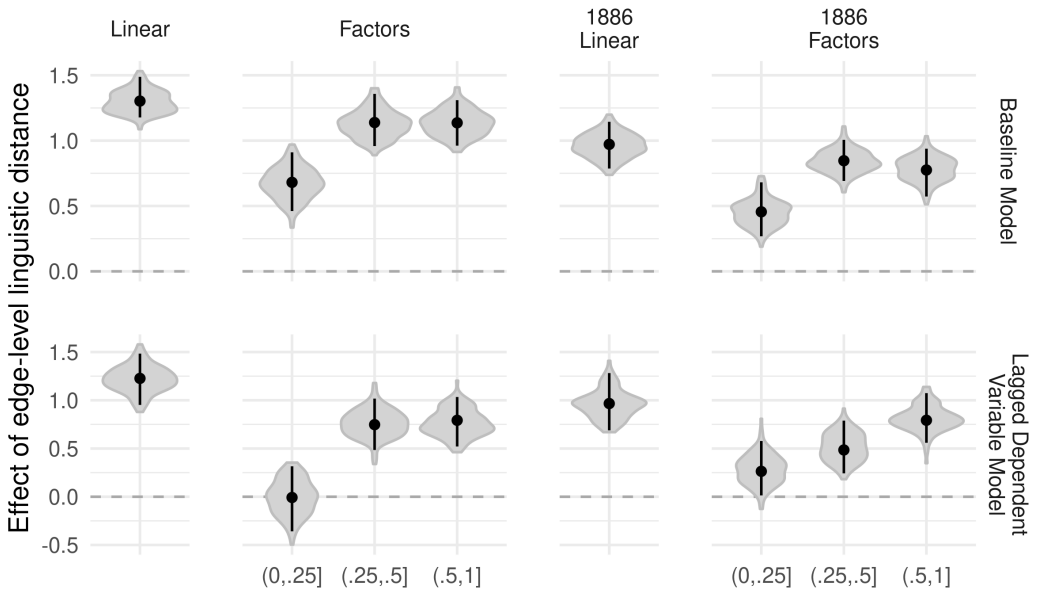


Figure A9: Effect of linguistic distances on the partitioning of Europe into states

Note: Pooled sample. 95% CIs and grey areas show the distribution of bootstrapped estimates.

With these variables, we estimate partition models that substitute our main ethnic boundary (1) with the edge-level linguistic distance and (2) the three positive distance factors (0 being the reference category). We estimate the baseline and lagged dependent variable models using both, time-varying ethno-linguistic data, as well as data observed only before 1886. The results in Figure A9 show that linguistic distances exert a substantive effect on edges' border-crossing probability, which is similar in size to the main estimates. Importantly, effects are stronger for larger linguistic distances. This is consistent with states' inability to overcome ethnic divisions through assimilation in such cases. This finding also provides strong evidence that our findings are not biased by effects of ethnic maps as instigators of common ethnicity. Such bias could arise where ethnic consciousness and subsequent political strife for nationhood originated primarily from one or multiple ethnic maps. While maps were potentially influential in such ways, they will not have produced large linguistic differences such as those between Hungarians and Rumanians (1), Germans and French (.75), or Poles and Czechs (.57) from scratch. The findings thus show that such effects of ethnic maps do not drive our results.

Using Austro-Hungarian census data from 1900 and 1910:

In order to assess the importance of (1) measurement error and (2) potential political biases in the historical maps of ethnic groups, we replicate our analysis using district-level data from the pre-WWI 1900 and 1910 censuses in Austria-Hungary.Available at: https://alex.onb.ac.at. District shapefiles from: MPIDR and CGG (2012) The census data records the shares of the main ethnic groupsGermans, Czechs (Bohemians & Moravians), Polish, Ruthenian, Slovenes, Serbo-Croats, Italians, Romanians, and Hungarians in 508 districts. We first construct a spatial graph by

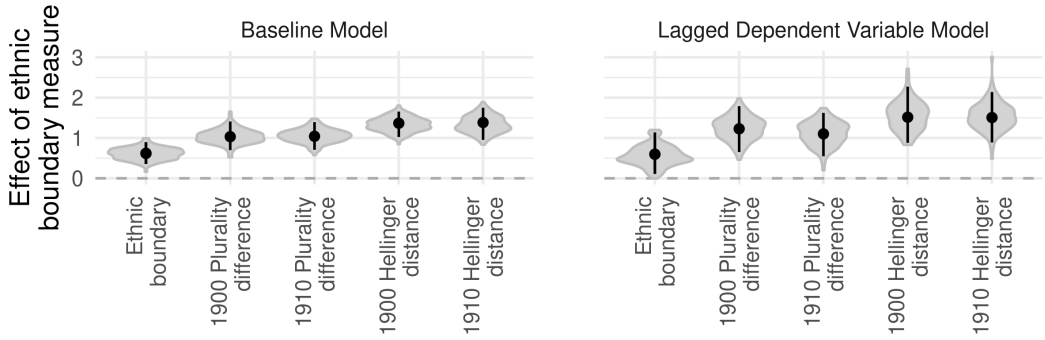


Figure A10: Point estimates of the effect of ethnic boundary measures from the 1900 and 1910 census data in Austria-Hungary

Note: 95% CIs and estimate distributions result from a parametric bootstrap with 120 iterations.

connecting districts' centroids via a Delaunay triangulation. We then compute, for each edge, three measures of ethnic difference between two districts A and B: (1) our baseline ethnic boundary measure, computed from maps produced between 1861 and 1910; (2) the average difference between the census shares of each district's plurality group and its share in the other district (plurality difference) and (3) the Hellinger distance between the districts' ethnic composition.Computed as $H(A,B) = \frac{1}{\sqrt{2}} \sqrt{\sum_{g=1}^{G} (\sqrt{p_{g,A}} - \sqrt{p_{g,B}})^2}$, where $p_{g,A}$ and $p_{g,B}$ denote the shares of group g in districts A and B, respectively. The latter is a metric bounded between 0 and 1 that makes use of the full information while not being influenced by irrelevant and absent groups (Hellinger 1909). We use these data to re-estimate our baseline and lagged dependent variable models for the years 1911–2011 and 1936–2011, respectively. In order to control for potential effects of regional borders on the census and future state borders, we add a dummy for whether two districts are separated by a "Landesgrenze" – borders at the first administrative level – within the Austro-Hungarian empire.

The results in Figure A10 indicate that our main results are, if at all, *downward* biased. All models indicate significant and sizeable effects of ethnic differences on edges' border probability with no detectable difference between variables derived from the 1900 and 1910 census data. Yet, the measure of differing plurality group shares as well as the Hellinger distance yield significantly larger point estimates while operating on the same 0 to 1 scale as the ethnic boundary indicator. That suggests that measurement error in our maps – most likely from the categorical depiction of continuous ethnic geography – exerts attenuation bias. With that, the result also provides strong evidence against the argument that our findings are caused by ethnic maps that have been manipulated for political reasons.

D.3 Varying control variables:

We first assess the sensitivity of the results to the choice of control variables. We (1) drop all controls from our model except the state border lags in the lagged dependent variable model and (2) add the following variables:

- River size^2 and Any river: To account for potentially non-linear effects of rivers on ethnic geography and state borders, we include a squared term as well as a dummy variable capturing whether an edge crosses any river.Strictly speaking, most if not all edges cross minor water bodies such as creeks, which are however not included in the Natural Earth Data used here.
- Δ Longitude, Δ Latitude: (Laitin, Moortgat and Robinson 2012) show that countries tend to have an east-west orientation due to low latitudinal environmental variation . If ethnic geographies follow the same pattern, the direction of edges may present an omitted variable.We therefore include the distance an edge traverses in each direction in decimal degrees.
- Population density in 1880 (estimate): High population-density regions may feature higher levels of ethnic diversity and smaller countries, which may bias our estimates. We therefore add the average population density in 1880 estimated for the two vertices an edge connects. Population density estimates are retrieved from Goldewijk, Beusen and Janssen (2010) who base their projection on all available (historical) sub-national census data combined with higher-level population projections and environmental variables. Though currently the best available data source, we note that their estimation procedure may add post-treatment bias to our model.
- Cumulative altitude change: While our main analysis controls for the *average* altitude along an edge, an edge's *ruggedness* may explain the structure of ethnic and state geographies. Rugged terrain may pose a natural barrier and thus separate ethnic groups and cause country borders. We therefore add the cumulative altitude change along an edge. This is computed by sampling first a set of points at every 1km on each edge and then taking the sum of absolute difference between each pair of neighboring points.
- Standard deviation of altitude: Following the same logic we construct an alternative (and more widely used) measure of an edge's ruggedness as the simple standard deviation of the altitude of the points along an edge.
- Administrative borders: Lastly, we address concerns that our results are driven by administrative borders. Lacking time-variant regional borders over the past 150 years, we do so by controlling for regional-level borders in 1800 and 1900 retrieved from Nuessli (2010). To avoid bias in the estimated effect of ethnic boundaries from changes in administrative borders after 1900, we restrict our ethnicity data to maps observed before 1886. Note that this is a conservative test since administrative borders in 1900 have likely been affected at least to some extent by ethnic geography.

Following the main analysis, we standardize all additional variables to fall between 0 and 1 to compare coefficient magnitudes directly with the estimate of ethnic boundary. Table A4 presents the results of the main specification, from dropping the main covariates, and from adding the additional ones. We first note that the size of the coefficient of interest, ethnic boundary, barely changes from the value estimated in the main analysis. Hence, observed covariates do not bias our results. If these are, ex ante, the most likely biasing spatial features, the result furthermore suggests a very small magnitude of omitted variable bias.

Table A4: Determinants of state borders in Europe, 1886–2011: Varying control variables

	Main results		No controls		Add. controls		Admin. borders	
	1: Baseline	2: LDV	3: Baseline	4: LDV	5: Baseline	6: LDV	7: Baseline	8: LDV
Constant	-2.50^*	-3.01^*	-2.03^*	-2.69^*	-2.71^*	-0.87	-5.20^*	-5.63^*
	$[-3.04; -1.91]$	$[-3.98; -2.47]$	$[-2.15; -1.92]$	$[-2.93; -2.45]$	$[-3.57; -1.54]$	$[-3.08; 0.24]$	$[-5.81; -3.45]$	$[-6.12; -3.33]$
Ethnic boundary$_t$	1.22^*		1.31^*		1.23^*			
	$[1.06; 1.40]$		$[1.18; 1.51]$		$[1.08; 1.41]$			
Ethnic boundary$_{t-1}$		1.02^*		1.07^*		0.98^*		
		$[0.79; 1.24]$		$[0.81; 1.28]$		$[0.73; 1.19]$		
State border$_{pre-1886}$							0.78^*	0.81^*
							$[0.57; 0.96]$	$[0.49; 1.16]$
State border$_{t-1}$		1.65^*		1.66^*		1.66^*		1.75^*
		$[1.46; 1.96]$		$[1.44; 1.90]$		$[1.46; 2.03]$		$[1.56; 2.08]$
Deep lag		0.74^*		0.75^*		0.82^*		0.41
		$[0.36; 1.15]$		$[0.37; 1.13]$		$[0.42; 1.27]$		$[-0.03; 0.79]$
Edge length	0.16	-0.34			-0.38	-1.79	5.29^*	5.37
	$[-1.06; 1.06]$	$[-1.79; 1.28]$			$[-1.86; 0.82]$	$[-3.28; 0.77]$	$[1.17; 6.83]$	$[-0.25; 6.18]$
Elevation mean	0.24	0.15			0.27	-0.19	0.43	0.01
	$[-0.42; 0.80]$	$[-0.82; 0.85]$			$[-0.99; 1.33]$	$[-2.01; 1.88]$	$[-0.55; 1.07]$	$[-1.57; 0.73]$
Largest watershed	0.62^*	0.76^*			0.72^*	0.84^*	0.40^*	0.56^*
	$[0.41; 0.81]$	$[0.33; 1.10]$			$[0.49; 0.95]$	$[0.48; 1.19]$	$[0.09; 0.63]$	$[0.09; 1.03]$
Largest river	0.28^*	0.26			0.61	1.44^*	0.18	0.14
	$[0.11; 0.48]$	$[-0.03; 0.64]$			$[-0.08; 1.50]$	$[0.05; 2.33]$	$[-0.04; 0.40]$	$[-0.32; 0.52]$
Largest river^2					-0.93	-2.27		
					$[-2.37; 0.14]$	$[-3.85; 0.18]$		
Any river					0.51^*	0.51		
					$[0.04; 0.93]$	$[-0.22; 1.08]$		
Δ Longitude					-0.15	-2.25^*		
					$[-1.08; 0.61]$	$[-3.66; -0.65]$		
Δ Latitude					0.51	-0.82		
					$[-0.41; 1.36]$	$[-2.25; 0.89]$		
Pop. dens. 1880					1.34^*	-1.29		
					$[0.50; 1.89]$	$[-3.00; 0.10]$		
Cum. Δ altitude					-1.03	0.15		
					$[-2.27; 0.24]$	$[-1.71; 2.18]$		
Std. dev. altitude					1.45^*	0.01		
					$[0.45; 2.37]$	$[-1.36; 1.43]$		
Adm. bord.$_{1800}$							0.54^*	0.33
							$[0.30; 0.83]$	$[-0.05; 0.81]$
Adm. bord.$_{1900}$							0.63^*	0.53
							$[0.38; 0.90]$	$[-0.06; 1.03]$
No. of periods	6	5	6	5	6	5	6	5
No. of vertices	6769	5412	6769	5412	6769	5412	4680	3870
No. of edges	17923	14243	17923	14243	17923	14243	12294	10170
No. of states	189	177	189	177	189	177	176	158

Notes: Each period t has a length of 25 years. 95% confidence intervals from parametric bootstrap in parenthesis. * Statistically significant at the 95% level.

D.4 Unconditional effects of natural border determinants

We here assess the effects of natural border determinants, which have hitherto served only as control variables, which makes direct interpretation of the respective coefficients difficult due to post-treatment bias from the inclusion of our ethnic boundary and historical border measures. We therefore drop these variables for this analysis. Because natural determinants such as rivers might have effects at a lower spatial level than ethnic geography, we re-estimate our main specifications for spatial graphs at spatial resolutions varying between 50 and 200km (see also below in Subsection D.6).

While not providing a comprehensive and definitive analysis, results in Figure A11 indicate consistent, positive effects of rivers and watersheds. These are sizeable and precisely estimated at high resolutions (50km) and more unstable at lower spatial resolutions, both in the baseline and lagged dependent variable models. We only find consistent positive effects for high-altitude edges at high spatial resolutions in the baseline model, suggesting that mountainous terrain did not affect border change since 1886.

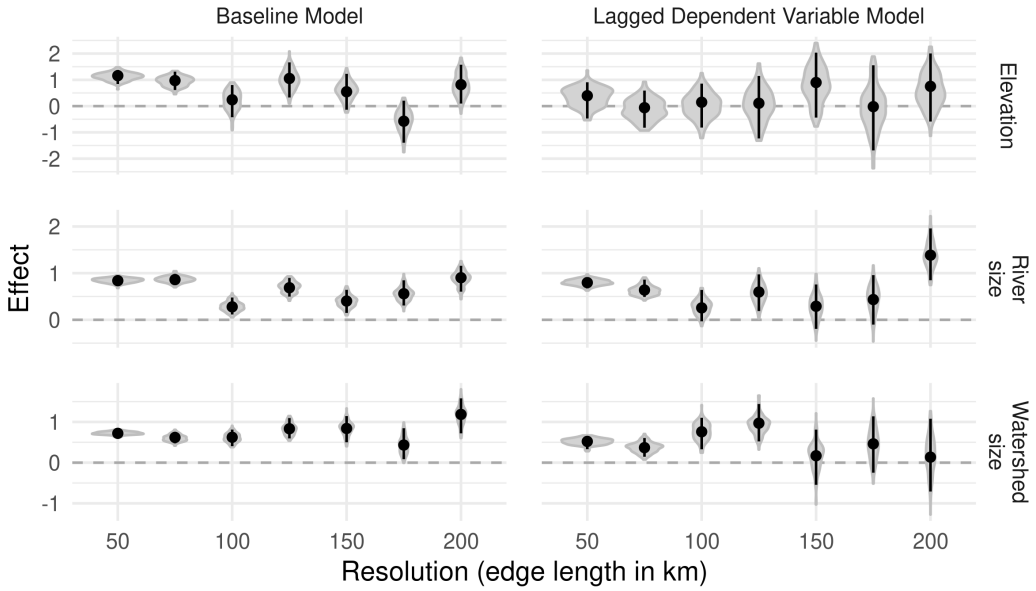


Figure A11: Point estimates of the effect of natural features on the partitioning of Europe into states at varying spatial resolutions

Note: Models without ethnic boundary and deep lag of state borders. 95% CIs and estimate distributions result from a parametric bootstrap with 120 iterations.

D.5 Varying the temporal structure of the data:

One important design choice at the outset of our main analysis is the choice of the length of periods that structure the temporal dimension of our data. For our main analysis, we measure state borders and ethnic geographies every 25 years, starting in 1886 and ending in 2011. While representing a middle ground between

very short and long periods, the period length of 25 years is arbitrarily set and our results may differ substantially for differing period lengths.

This robustness check tests whether this is the case by varying the period in 10-year steps length between 5 and 65 years.65 years is the longest period length for which we can split the available data since 1886 into two periods: 1886–1951 and 1951–2016. As in the baseline analysis, each dataset starts in 1886 and thus exhibits the following temporal structure: $t \in 1886 + 0\,p, 1886 + 1\,p, ..., 1886 + I\,p$, such that $1886 + I\,p <= 2019$. This setup entails that our data for $p = 35$ and $p = 45$ end in 1991 and 1976, respectively, thus omitting part of the breakdown of the USSR and former Yugoslavia.

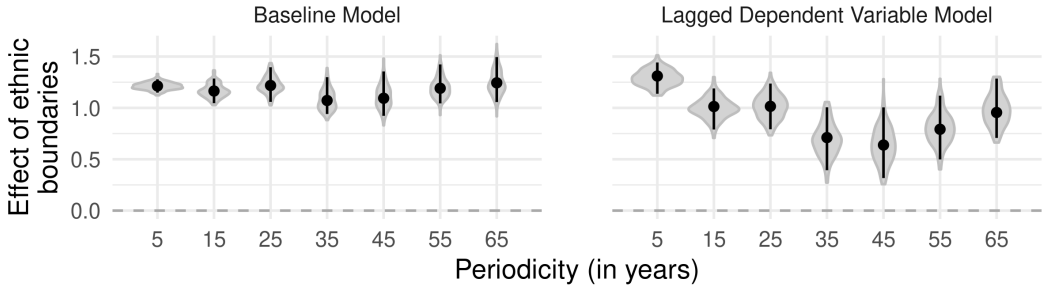


Figure A12: Point estimates of the effect of ethnic boundaries on the partitioning of Europe into states: Varying the period of the the length of periods t

Note: 95% CIs and estimate distributions result from a parametric bootstrap with 120 iterations.

Re-estimating our main specifications for each newly generated dataset yields results that broadly conform with our main results. Summarized in Figure A12, the estimates for the baseline (cross-sectional) model show coefficients that remain stable with the length of periods. The estimate for the 25 year period data is close to the average of all estimates.

The results for the lagged dependent variable model are somewhat more varied but consistently yield substantive and statistically significant estimates for the effect of ethnic boundaries. Upon closer inspection, we note that the downward deviations from our main result stem from the two datasets with a period of 35 and 45 year that omit the 1990s, an important period of ethnic secession in the former Soviet Union and on the Balkans. The results therefore leave us confident that the temporal structure of our main dataset does not substantially bias our results.

D.6 Varying the spatial lattice:

Similar to the temporal structure of our data, the making of the spatial lattice we analyze is based on three potentially influential parameters. The first parameter is the geographic location of the "anchor" of the lattice that determines the location of all vertices. The second parameter is the spatial resolution of the network. The third parameter is the spatial structure of the lattice.

Shifting the lattice anchor: The first parameter that determines the spatial make-up of our baseline lattice consists in the location of the "anchoring" point (in our

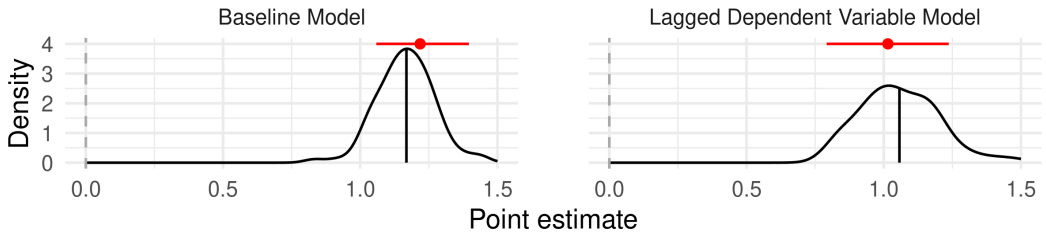


Figure A13: Point estimates of the effect of ethnic boundaries on the partitioning of Europe into states: Shifting the spatial lattice

Note: Main estimates from Table 1 in red. Distributions result from re-estimating the main models 100 times, with data from a randomly shifted hexagonal lattice.

case in the utmost south-west of the sampling area) from which the remainder of the lattice is constructed. We test whether shifting that point – and thereby the rest of the lattice – slightlyWe shift the lattice by displacing the anchoring point with random draws from a uniform distribution between 1 and 10 decimal degrees in each direction. along the north-south and east-west axes affects the results.

Following this procedure, we construct 100 lattices and recreate the entire dataset for each. Re-estimating the baseline models for each resulting network gives rise to a distribution of estimates for the baseline and lagged dependent variable specifications. Figure A13 shows that our main estimates are well centered at the 67^{th} and 38^{st} percentiles of the respective distributions. This shows that our main results are not sensitive to the exact location of the anchoring point of our spatial lattice.

Varying lattice resolution: The second parameter that governs the spatial dimension of our data consists in the length of edges on our lattice. We here present results from alternative specifications that let this spatial resolution vary between 50 and 200 km, in steps of 25km. Networks with a lower resolution (200km) feature less vertices and edges but may be able to capture more diffuse spatial patterns, i.e. capturing effects of ethnic geographies even if they are not precisely marked on a map or are in fact more gradual than our categorical maps suggest. Graphs with a higher resolution (25km) are more informative and have more statistical power but may miss more diffuse spatial effects due to their high level of detail. We therefore create alternative datasets with the alternative spatial resolutions that use the same spatial raw data to encode the very same variables as our main lattice.

Figure A14 presents the estimates for the effect of ethnic boundaries derived from the baseline and lagged dependent variable model estimated with the alternative lattices. The results show that our estimates slightly *increase* as we decrease the resolution of our data beyond an edge length of 100km. This suggest that ethnic geographies can have more diffuse effects that are not always captured by high-resolution data. Reassuringly, the effects estimated at resolutions lower than 100km are very similar and statistically indistinguishable from our baseline results.

Varying lattice structure The third parameter that determines the spatial makeup of our data consists in the structure of the spatial lattice. In particular, the vertices of the main lattice are the centroids of the tiles of a hexagonal tiling. There are two

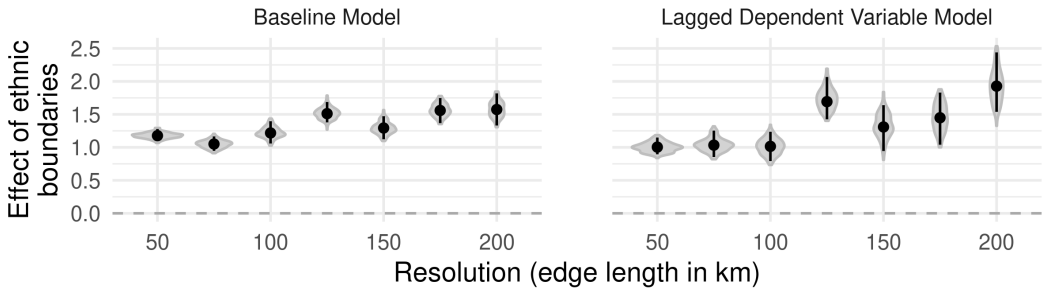


Figure A14: Effect of ethnic boundaries on the partitioning of Europe into states at varying resolutions of the spatial lattice

Note: 95% CIs and estimate distributions result from a parametric bootstrap with 120 iterations.

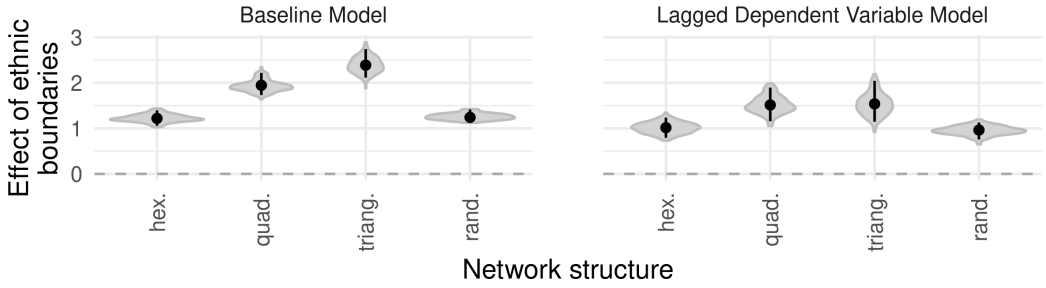


Figure A15: Effect of ethnic boundaries on the partitioning of Europe into states using a hexagonal, quadratic, triangular, and random lattice structure

Note: 95% CIs and estimate distributions result from a parametric bootstrap with 120 iterations.

other regular tilings, the quadratic and triangular tiling from which we can generate regular lattices.As in the hexagonal case, a tiling is transformed into a lattice by connecting the centroid (vertex) of each tile with the centroids of neighboring tiles. Together with the hexagonal tiling, the resulting lattices feature a constant edge length which is only slightly disturbed by the earth's surface curvature. However, quadratic and the triangular lattices feature fewer edges per vertex. Given a constant edge-length, they are therefore, theoretically, less able to capture spatial dependencies. A fourth possible lattice structure consists of a set of randomly located vertices connected by edges from a simple Delaunay triangulation. While the degree of vertices and edge-length in the random lattice is not constant, it is on average similar to the hexagonal structure.

In order to test whether our results are robust to these alternative networks structures, we construct additional lattices with a quadratic, triangular, and random structure. For each lattice, we again construct the same set of variables as in our main analysis and re-estimate our baseline and lagged dependent variable specification. Figure A15 summarizes the resulting estimates for the effect of ethnic boundaries. We note that the effect is *increasing* in the quadratic and triangular structure, yielding a similar effect as obtained when we decrease the spatial resolution of the lattice (see above). The random lattice structure yields estimates that are indistinguishable from those estimated from the hexagonal structure. In sum, these

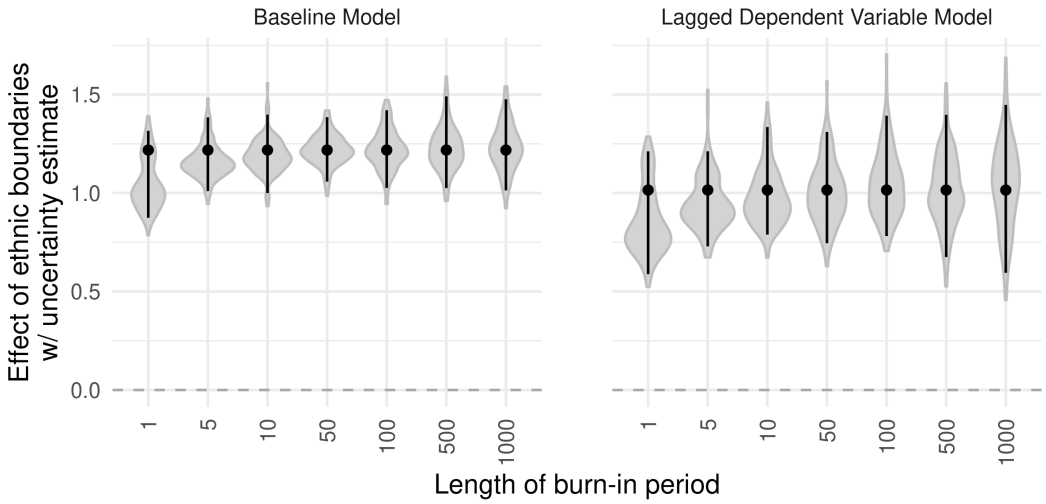


Figure A16: Uncertainty estimates with varying burn-in rates

Note: 95% confidence intervals result from a parametric bootstrap with 120 iterations an a burn-in rate as indicated on the x-axis. Shaded grey areas show distribution of bootstrapped estimates.

results suggests that the hexagonal lattice structure yields if at all conservative estimates doe to its increased ability of capturing spatial interdependence.

D.7 Burn-in rate in parametric bootstrap

We also assess whether the choice of the burn-in period (100 iterations) substantively affects the uncertainty estimates produced by our parametric bootstrap (see also Appendix Section A.4). Figure A16 plot the confidence intervals and parameter distribution retrieved from parametric bootstraps with a burn-in rate varying between 1 and 1000 iterations. The results show that the choice of the burn-in rate does not substantively affect the results above a very low burn-in rate of 10 iterations. This result coincides with the stability of the results in most areas of the parameter space assessed in our Monte Carlo experiments in Appendix Section B.2.

D.8 Logistic regression with edge-level data

To demonstrate the advantage of the PSPM in accounting for spatial dependence, we can alternatively estimate logistic regression that model the probability that an edge in our network crosses a state borders in a given year. We here do so with a cross-sectional baseline and lagged dependent variable specification that directly mirror the main PSPM specification with the important exception that we treat edges as fully independent.

Table A5 shows the logistic regression results and illuminates the effects of the invalid independence assumption. We can directly compare these results to the PSPM estimates for truly independent "bridge edges" (see discussion of the model in main text). Doing so shows that the estimates for ethnic boundaries from the logistic regression are approximately 2.5 times larger than in the corresponding

Table A5: Edge level modeling: Logit results

	Dependent variable:	
	Baseline model	Lagged dependent variable
	(1)	(2)
Constant	-5.14^{***}	-5.24^{***}
	(0.14)	(0.22)
Ethnic boundary$_t$	3.28^{***}	
	(0.07)	
Ethnic boundary$_{t-1}$		2.51^{***}
		(0.12)
State border$_{t-1}$		4.49^{***}
		(0.11)
Deep lag		1.67^{***}
		(0.14)
Controls	*yes*	*yes*
Observations	17,923	14,243
Log Likelihood	-4,322.71	-1,924.55
Akaike Inf. Crit.	8,657.42	3,865.10

Notes: Each period t has a length of 25 years. Robust standard errors in parenthesis..
Significance codes: $*$p<0.1; $**$p<0.05; $***$p<0.01

PSPM models, implying odds ratios that are 12 times too large. In addition, we see that standard errors are significantly smaller than in the PSPM. This overconfident upward bias is explained by edges' spatial dependencies, which are captured by the PSPM.

D.9 Assessing regional variation: The role of population density

Why did ethnic geography shape border change since 1964 in Europe, to some extent in Asia, but not elsewhere around the globe? We here present suggestive evidence that supports Herbst's (2000) claims about the centrality of population densities as a driving force behind states' competition over territory. Taking dense population as a proxy for the value of local territory, we expect ethnic boundaries to affect border change most in densely settled regions. In turn, low population densities suggest territory of low value in which competition over people will have precedence over competition over territory. We test this argument by recurring to our data on all continents and estimating the LDV model with an interaction term between logged population densityMeasured in 1880, from Goldewijk, Beusen and Janssen (2010). and the ethnic boundary measure from GREG as well as the lagged dependent variable measured in 1964.

The results in Figure A17 show that the effect of ethnic boundaries strongly increases in population density globally and within Europe. The interaction effects within Asia and Africa are much weaker and not statistically significant, suggesting that the global result is driven by variation within Europe and variation between Europe and Asia and the historically less densely settled Africa and the Americas. While only suggestive, these results are in line with valuable territory driving

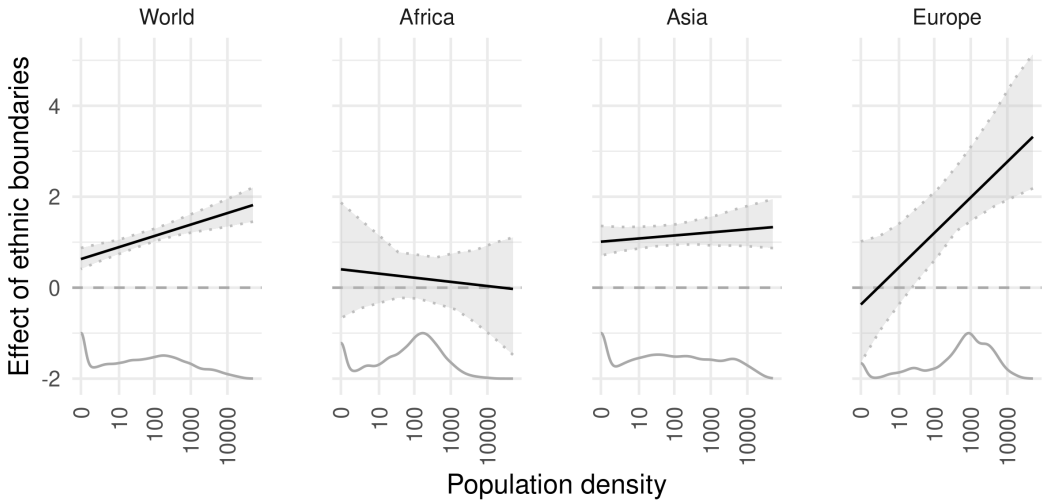


Figure A17: Marginal effect of ethnic boundary in the LDV model by population density (inhabitants/km^2): Global and by continent with observed border changes.

territorial competition and the alignment of states with ethnic geography.

E Analysis of secessionist claims and conflict

This section presents the analysis of secessionist claims and conflict. The type of the additional analysis partially mirrors the additional analyses conducted for the analysis of the partitioning of Europe into states.

Data

The vertices of our baseline lattice G constitute the units of analysis,Appendix E.1 shows robustness to different spatial data structures. avoiding units that are either spatially misaligned with our (in)dependent variables or defined based on state borders. We code whether points are (1) claimed by a self-determination movement, (2) fought over in a secessionist ethnic civil war, and (3) affected by a successful secession. Yearly data on secessionist self-determination claims between 1946 and 2012 come from the GeoSDM dataset (Germann and Schvitz 2023, and Appendix C.2). The Ethnic Power Relations data (Vogt et al. 2015) enlists the settlement regions of ethnic groups associated with secessionist civil wars between 1946-2016. Lastly, we code secession when a point becomes part of a newly independent state in the CShapes 2.0 data (Schvitz et al. 2022).

We expect that areas that are ethnically distinct from states' core groups are most likely to experience secessionism. We capture this logic by using our historical ethnic maps to measure whether a point is 'non-coethnic' to their state's capital.We construct this variable in parallel to the network-based variable ethnic boundary (Eq. (1), main text). Appendix E.1 shows robustness with pre-1886 ethnic data.

Empirical strategy

We model the onset of secessionist claims, conflicts, and successful secession using a Cox Proportional Hazard Model, which mitigates the problem of successful secession leading to selection out of the treatment group:Accounting for further potential endogeneity by analyzing only point-years unaffected by post-1946 border change increases effect sizes (Appendix E.1).

$$h(\tau)_{j,t} = h_0(\tau)\, exp(\beta_1 \text{ non-coethnic capital}_{j,t} + \gamma\, \mathbf{X}_{j,t} + \epsilon_{j,\tau}) \tag{A12}$$

where $h(\tau)_{j,t}$ is the expected onset risk of one of the three outcomes in point j in calendar year t and relative time τ – the years since j became a member of its current state.This counter starts with our data in 1946. The end of World War II as a critical juncture arguably restarted the survival 'clock' in much of Europe. Next to our variable of interest non-coethnic capital$_{j,t}$, we add controls $\mathbf{X}_{j,t}$ that account for the most important joint structural causes of peripheral minority status and secessionist conflict (e.g., Carter, Shaver and Wright 2019). These follow two logics. The first mirrors the dyadic controls from our main analysis, capturing the distance (logged), size of largest river and watershed, as well as the mean elevation between point j and its capital $C_{j,t}$, and the fraction of centuries (1000-1790) in which the two were part of the same state. The second logic focuses on points j only, with controls for the local population density (logged; Goldewijk, Beusen and Janssen 2010), the altitude and terrain slope (FAO 2015), as well as each points' distance to the closest border (logged).

We additionally estimate stratified models where the baseline hazard $h_0(t)$ varies by country-year. Similar to country-year fixed effects, this accounts for time-varying confounders within states (e.g., the breakup of the USSR). We cluster standard errors on 'stable state segments,' sets of points that were always jointly members of the same states.

E.1 Results

Main results: Table A6 presents the main results discussed in the paper.

Within borders from 1946 only: One important caveat of the main analysis is that border changes observed during the temporal coverage of the panel, i.e. after 1946, are endogenous to secessionism which is the main object of interest here. Because secessionism reduces mismatches between ethnic boundaries and state borders leaving only the "hard" cases with low secession probability in the sample, we may underestimate the effect of ethnic boundaries on the occurrence of secessionist dynamics. We test this conjecture by analyzing points only as long as they are situated in the state they were member of in 1946 and drop all other point-years. Table A7 presents the respective results. All coefficient increase substantially in size (on average around 50 percent). This suggests that selection bias in the original analysis leads us to underestimate the effect of mismatches between state and ethnic geographies on secessionism.

Table A6: Ethnic boundaries and the onset of self-determination claims, conflict, and border change

	Cox Proportional Hazard Model					
	Secessionist Claim		Secessionist Civil War		Secession	
	(1)	(2)	(3)	(4)	(5)	(6)
Non-coethnic capital	2.546***	1.684***	3.048***	2.211***	3.918***	2.924***
	(0.386)	(0.460)	(0.445)	(0.480)	(0.601)	(0.771)
Events:	211	211	116	116	153	153
Country-year strata:	no	yes	no	yes	no	yes
Controls:	yes	yes	yes	yes	yes	yes
Observations	64,810	64,810	71,057	71,057	75,387	75,387
R^2	0.007	0.004	0.005	0.003	0.007	0.005
Max. Possible R^2	0.044	0.030	0.023	0.017	0.027	0.021
Log Likelihood	-1,248.955	-833.687	-650.577	-514.381	-782.777	-620.219

Notes: Cox Proportional Hazard models. The unit of analysis is the point-year between 1946 and 2012.. Standard errors clustered on state-segments. Full results with control variables are reported in Table **??**. Significance codes: *p<0.1; **p<0.05; ***p<0.01

Table A7: Ethnic boundaries and self-determination: Within 1946 borders only

	Cox Proportional Hazard Model					
	Secessionist Claim		Secessionist Civil War		Secession	
	(1)	(2)	(3)	(4)	(5)	(6)
Non-coethnic capital	2.320***	1.744***	3.227***	2.444***	3.893***	2.924***
	(0.323)	(0.451)	(0.500)	(0.496)	(0.602)	(0.771)
Events:	200	200	104	104	152	152
Country-year strata:	no	yes	no	yes	no	yes
Controls:	yes	yes	yes	yes	yes	yes
Observations	58,805	58,805	64,217	64,217	68,403	68,403
R^2	0.007	0.005	0.005	0.004	0.008	0.006
Max. Possible R^2	0.045	0.032	0.022	0.018	0.030	0.024
Log Likelihood	-1,154.087	-811.985	-548.017	-476.614	-775.799	-620.219

Notes: Cox Proportional Hazard models. The unit of analysis is the point-year between 1946 and 2012. Standard errors clustered on state-segments. Significance codes: *p<0.1; **p<0.05; ***p<0.01

Using pre-1886 data on ethnic geography: Our analysis of secessionism may be biased if changes in ethnic boundaries are caused by causes of subsequent state border change. We therefore recur to ethnic settlement patterns mapped at the earliest point, in the 50 years prior to 1886. Estimating their effect on post-1946 secessionim yields estimates of non-coethnic capital that are marginally smaller than the baseline estimates but nevertheless of substantive size (Table A8). Given the reduced precision of the data, standard errors slightly increase. Together with the overall stability of ethnic geographies, this suggests that endogenous changes of ethnic geographies are unlikely to cause the results.

Varying the spatial sampling of points: As in the PSPM analysis (see Section D.6 above), we vary the spatial sampling of points by (1) randomly shifting points 100 times, (2) varying the spatial resolution (50 to 200km), and (3) retrieving points quadratic and triangular tiles, as well as from a spatial random draw. Our main estimates are well centered in the distribution of estimates yielded from (1) (Figure

Table A8: Ethnic boundaries and self-determination: Pre-1886 ethnic geography

	Cox Proportional Hazard Model					
	Secessionist Claim		Secessionist Civil War		Secession	
	(1)	(2)	(3)	(4)	(5)	(6)
Non-coethnic$_{1886}$ capital$_t$	1.359**	0.971	2.733***	2.193**	2.941***	1.921***
	(0.636)	(0.595)	(0.668)	(0.855)	(0.604)	(0.727)
Events:	211	211	116	116	153	153
Country-year strata:	no	yes	no	yes	no	yes
Controls:	yes	yes	yes	yes	yes	yes
Observations	64,972	64,972	71,207	71,207	75,537	75,537
R^2	0.005	0.004	0.004	0.003	0.005	0.005
Max. Possible R^2	0.044	0.030	0.023	0.018	0.027	0.021
Log Likelihood	-1,309.532	-851.996	-666.253	-520.320	-840.963	-647.417

Notes: Cox Proportional Hazard models. The unit of analysis is the point-year between 1946 and 2012.. Standard errors clustered on state-segments. Full results with control variables are reported in Table **??**. Significance codes: *p<0.1; **p<0.05; ***p<0.01

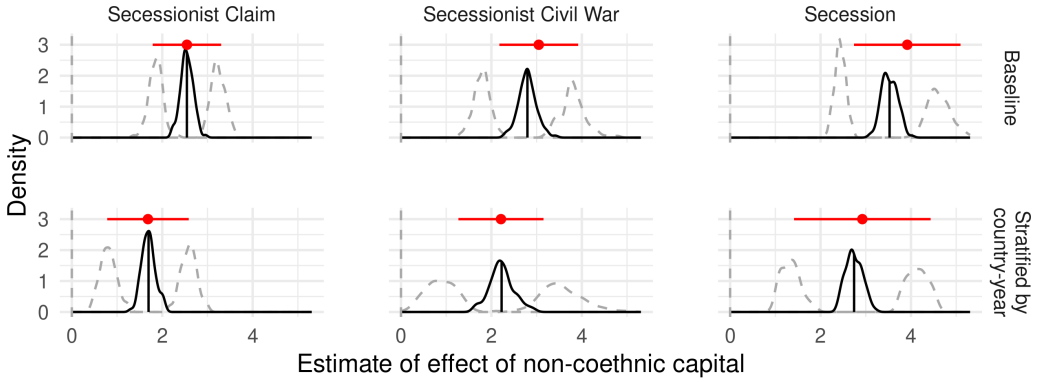


Figure A18: Secessionism robustness check: Shifting points (unit of analysis)

Note: Results from Table A6 in red. Solid lines denote distribution of main estimates, dotted lines distributions of upper and lower bounds of 95% CIs. Distributions result from re-estimating the main models 100 times, with data from a randomly shifted hexagonal lattice.

A18). Figure A19 demonstrates robust results when varying the spatial resolution of our data. Lastly, Figure A20 shows that the sampling strategy used for constructing our point-level data has no substantial effect on our results. In all, these results suggest that our results are robust to the parameter choices behind the spatial data structure.

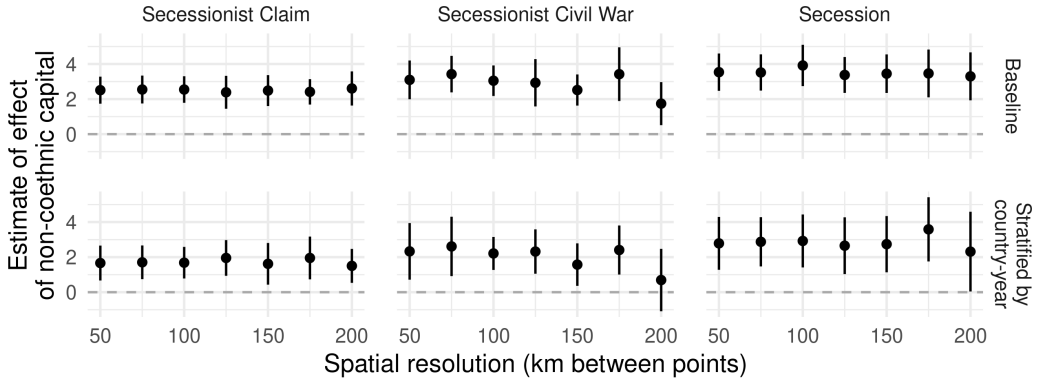


Figure A19: Estimates of the effect of non-coethnic capitals on secessionism at varying spatial resolutions lattice

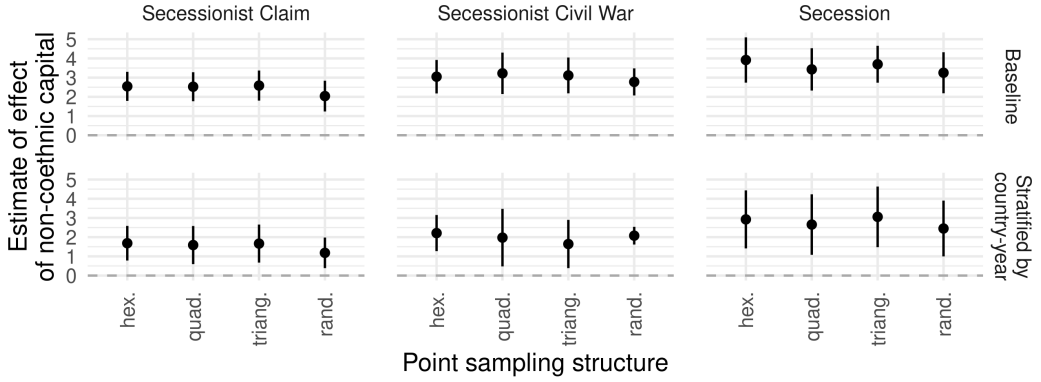


Figure A20: Estimates of the effect of non-coethnic capitals on secessionism with a hexagonal, quadratic, triangular, and random lattice structure

F References (Appendix)

Abramson, Scott F. 2017. "The Economic Origins of the Territorial State." *International Organization* 71(1):97–130.

Beck, Nathaniel, David L Epstein, Simon Jackman and Sharyn L O'Halloran. 2001. "Alternative models of dynamics in binary time-series-cross-section models: The example of state failure.".

Beck, Nathaniel, Jonathan N Katz and Richard Tucker. 1998. "Taking time seriously: Time-series-cross-section analysis with a binary dependent variable." *American Journal of Political Science* 42(4):1260–1288.

Besag, Julian. 1974. "Spatial interaction and the statistical analysis of lattice systems." *Journal of the Royal Statistical Society: Series B (Methodological)* 36(2):192–225.

Bruk, Solomon I. and V. S. Apenchenko. 1964. *Atlas Narodov Mira*. Moscow: Glavnoe upravlenie geodezii i kartografii gosudarstvennogo geologicheskogo

komiteta SSSR and Institut etnografii im. H. H. Miklukho-Maklaia, Akademiia nauk SSSR.

Carter, David B, Andrew C Shaver and Austin L Wright. 2019. "Places to Hide: Terrain, Ethnicity, and Civil Conflict." *The Journal of Politics* 81(4):1446–1465.

Carter, David B and Curtis S Signorino. 2010. "Back to the future: Modeling time dependence in binary data." *Political Analysis* 18(3):271–292.

FAO. 2015. "Global Agro-Ecological Zones: Crop Suitability Index." *Dataset, available online at: http://gaez.fao.org* .

Fearon, James D. 2003. "Ethnic and Cultural Diversity by Country." *Journal of Economic Growth* 8(2):195–222.

GADM. 2019. "GADM database of Global Administrative Boundaries Version 3.6. 2019." *https://gadm.org/* .

Germann, Micha and Guy Schvitz. 2023. "Representing Self-Determination Claims in Space: The GeoSDM Dataset." Mimeo, University of Bath.

Godambe, Vidyadhar P. 1960. "An optimum property of regular maximum likelihood estimation." *The Annals of Mathematical Statistics* 31(4):1208–1211.

Goldewijk, Kees Klein, Arthur Beusen and Peter Janssen. 2010. "Long-term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1." *The Holocene* 2010(1):1–9.

Hellinger, Ernst. 1909. "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen." *Journal für die reine und angewandte Mathematik* 1909(136):210–271.

Herbst, Jeffrey. 2000. *States and Power in Africa*. Princeton: Princeton University Press.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. *An introduction to statistical learning*. Vol. 112 Springer.

Laitin, David D, Joachim Moortgat and Amanda Lea Robinson. 2012. "Geographic axes and the persistence of cultural diversity." *Proceedings of the National Academy of Sciences* 109(26):10263–10268.

Lewis, M. Paul, ed. 2009. *Ethnologue: Languages of the world*. Vol. 16 SIL International Dallas, TX.

Lindsay, Bruce G. 1988. "Composite likelihood methods." *Contemporary mathematics* 80(1):221–239.

Minahan, James. 1996. *Nations without states: A historical dictionary of contemporary national movements*. Greenwood.

Minahan, James. 2002. *Encyclopedia of the stateless nations: DK*. Vol. 2 Greenwood Publishing Group.

MPIDR and CGG. 2012. "MPIDR Population History GIS Collection." *Available at* https://censusmosaic.demog.berkeley.edu/data/historical-gis-files.

Nuessli, Christos. 2010. "Euratlas Historical Atlas and Gazetteer of Europe.". **URL:** *https://www.euratlas.com/*

Park, Juyong and Mark EJ Newman. 2004. "Statistical mechanics of networks." *Physical Review E* 70(6):066117.

Roth, Christopher Fritz. 2015. *Let's Split!: A Complete Guide to Separatist Movements and Aspirant Nations, from Abkhazia to Zanzibar*. Litwin Books.

Sambanis, Nicholas, Micha Germann and Andreas Schädel. 2018. "SDM: A new data set on self-determination movements with an application to the reputational theory of conflict." *Journal of Conflict Resolution* 62(3):656–686.

Schvitz, G, S Rüegger, L Girardin, L-E Cederman, N Weidmann and KS Gleditsch. 2022. "Mapping The International System, 1886-2017: The Cshapes 2.0 Dataset." *Journal of Conflict Resolution* 66(1):144–161.

Sloane, Neil JA et al. 2003. "The on-line encyclopedia of integer sequences.".

Varin, Cristiano, Nancy Reid and David Firth. 2011. "An overview of composite likelihood methods." *Statistica Sinica* 21(2011):5–42.

Vogt, Manuel, Nils-Christian Bormann, Seraina Rüegger, Lars-Erik Cederman, Philipp M Hunziker and Luc Girardin. 2015. "Integrating Data on Ethnicity, Geography, and Conflict: The Ethnic Power Relations Dataset Family." *Journal of Conflict Resolution* 59(7):1327–1342.

Weidmann, Nils B., Jan Ketil Rød and Lars-Erik Cederman. 2010. "Representing ethnic groups in space: A new dataset." *Journal of Peace Research* 47(4):491–499.