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

Carl Müller-Crepon[*] Guy Schvitz[†] Lars-Erik Cederman[†]

February 3, 2022

Abstract

Borders define states, yet little systematic evidence explains where they are drawn. Putting recent 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 19th 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; Computational Methods

We thank Nils-Christian Bormann, Michael Kenwick, Melissa Lee, Kan Li, Paul Poast, seminar and conference participants at APSA 2020, the University of Oxford, 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.

[*] Department of Politics and International Relations, University of Oxford. Corresponding author: carl.muller-crepon@politics.ox.ac.uk

[†] Center for Comparative and International Studies, ETH Zürich

Borders are constitutive features of the modern state system that define the size and shape of states and specify the limits of state sovereignty. While 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), their origins remain understudied. Instead, most quantitative research treats borders as exogenous and sidesteps their formation. This question, however, has gained relevance as existing borders have come under increasing pressure. Most notably, Russia's annexation of Crimea in 2014 and its support of secessionist civil war in Eastern Ukraine have signaled a revival of revisionism. Majorities in Hungary, Greece, Bulgaria, and Turkey still view parts of neighboring countries as rightfully theirs (Fagan and Poushter 2020), the Catalan impasse persists, and Brexit has fueled Scottish secessionism and renewed tensions in Northern Ireland. Nationalist demands to redraw state borders along ethnic lines are at the core of all these cases.

Despite its obvious importance, we currently lack systematic evidence on the ethnic roots of border formation. We address this gap and ask whether, how, and to what extent ethnic geography has shaped Europe's state borders since the 19th century. Following macro-sociological 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 ethnically homogeneous nation-states. As most nations are ethnically defined, nationalism prompted popular pressure to redraw borders along ethnic lines, mostly through secessionism in multi-ethnic states, but occasionally also through unification and irredentism (Weiner 1971; Hechter 2000; O'Leary 2001).[1] While previous studies have highlighted the causes of secessionism and other nationalist demands for border change (see e.g., Coggins 2014; Griffiths 2016; Germann and Sambanis 2021), there is much less systematic evidence about how nationalism ultimately reshaped states along ethnic lines.

To fill this gap, it is necessary to overcome three empirical challenges. The first concerns the unit of analysis. Previous analyses focus exclusively on existing bor-

[1] In a similar vein and adopting a cultural and mostly individualist perspective, Alesina and Spolaore (1997, 2005) argue that states' size and shape results from trading off economies of scale and costs of ethnic heterogeneity (see also Friedman 1977; Desmet et al. 2011).

ders (Carter and Goemans 2011) or study border formation within individual grid cells (Kitamura and Lagerlöf 2020). They thereby either select on the dependent variable or ignore the spatial dependencies that characterize borders. Second, unbiased estimation of ethnic geography's effect on state borders requires us to consider confounding geographic features that affect both. Third, we need data on ethnic geography that predate (changing) state borders to avoid reverse causality from state-led ethnic assimilation and cleansing.

We address the first two challenges with a novel *Probabilistic Spatial Partition Model* (PSPM) that 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 treats geographic space as a planar network of points that encodes the main dependent and independent variables. It accounts for the spatial interdependencies that characterize partitionings, estimates effects conditional on covariates, and yields valid uncertainty estimates. Beyond our present use to estimate the effect of ethnic geography on state borders, the PSPM's approach can be used to model other types of spatial partitioning, for example administrative units or electoral districts.[2]

To solve the third problem, we collect new, time-varying spatial data on ethnic settlement areas in Europe since 1855 digitized from 73 historical ethnographic maps. After mitigating potential political biases, the dataset enables us to analyze borders and border change based on pre-existing ethnic settlement areas thus avoiding reverse causality. We minimize omitted variable bias by pairing a static baseline with a lagged dependent-variable model that captures the effect of ethnic geography on border change.

We find that the presence of an ethnic boundary between two locations increases the probability that they are or will become separated by an international border by 35 and 17 percentage points, respectively. This finding is robust to accounting for potentially endogenous changes in ethnic geography, additional controls, and changes to the spatio-temporal data structure. Additional analyses of the post-World War II period highlights ethnic secession as a key mechanism: Areas home to peripheral ethnic groups have an approximately 16, 22, and 50 times

[2] We distribute the PSPM as an R package upon publication.

greater risk of experiencing secessionist claims, civil wars, and border change, respectively. Finally, we explore the generalizability of our findings beyond Europe and find that ethnic boundaries explain border change since the 1960s only in Europe and Asia, while post-colonial Africa and the Americas have thus far avoided extensive ethno-nationalist border change.

Nationalism and the shaping of states

Our core argument holds that the rise of nationalism created a growing demand for ethnically homogeneous nation-states, which caused an increasing realignment of Europe's borders with the underlying ethnic map. This development is one part of a larger process that O'Leary (2001) labels the "right-peopling" and "right-sizing" of states. The former dynamic has received much attention in nationalism studies stressing how states shape ethnic affiliations, rather than the reverse. For instance, Hobsbawm (1990, p. 10) posits that "[n]ations do not make states and nationalisms but the other way around." This constructivist argument reflects evidence on states' efforts to mold their populations into nations through assimilationist policies and ethnic violence (Weber 1976; White 2004; Darden 2013; McNamee and Zhang 2019). While we do not dispute this evidence, we argue that an exclusive focus on state-led identity formation neglects changes in state borders and risks underestimating the full impact of nationalism.[3] We therefore focus on the nationalist 'right-sizing' of states along ethnic lines.

How did nationalism transform Europe's borders? To answer this question, we start by considering the link between ethnic and national identities. Following Weber (1978, pp. 385-98), we define ethnic groups as "those human groups that entertain a subjective belief in common descent," with language and religion being the most frequent markers that distinguish ethnic groups. Once the members of ethnic groups desire to control a state, they become ethnic nations. Again following Weber, a nation is "a community of sentiment which would adequately manifest itself in a state of its own" and hence "tends to produce a state of its own" (p. 176).

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

Gellner's congruence principle at the core of 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 this principle, each motivating a specific type of border change.

First and most common are ethnic minorities in a state dominated by a different ethnic group. Such a "state-to-nation deficit" Miller (2007) or "alien rule" (Hechter 2013) deprives ethnic groups of self-determination and state services often provided in favor of the ruling group (De Luca et al. 2018). In response, stateless nations may try to attain statehood by secession. The break-up of European empires represents the most important example of this process (Kumar 2017; Beissinger 2002).

Second, ethno-nationalist grievances can emerge if an ethnic group is divided by state borders prompting nationalist activists to call for unification of their kin (Cederman, Rüegger and Schvitz 2022). The promise of benefits from governance over a larger, yet ethnically homogeneous population can help their cause (Alesina and Spolaore 2005). Their efforts can result in the merger of co-ethnic units, as illustrated by 19th-century Germany and Italy and the more recent (re)unifications of Vietnam, Yemen, and Germany. While usually less contentious than secession, unification may trigger resistance in smaller units, or power competition in the unified nation. Concomitant to the decline of state death since 1945 (Fazal 2004, 2007), ethnic unification is exceedingly rare.

Third, mixed incongruence exists where an ethnic group dominates one state but forms a minority in another. This configuration creates a pressure for the homeland government to "liberate" the group in question, resulting in irredentist nationalism (Weiner 1971). 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). Russia's annexation of Crimea in 2014, however, illustrates that irredentism has not disappeared.

Whether striving for secession, unification, or irredentist border change, nationalist ideology equips political activists with powerful normative arguments to justify their claims over seemingly 'indivisible' territory and mobilize elites and citizens for their revisionist projects (Hroch 1985; Goddard 2006). While actual border change is difficult to achieve due to collective action problems (Hardin 1995) and resistance by the incumbent state, 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 (Abramson and Carter 2021; Skocpol 1979), as illustrated by the collapse of the European empires after the two world wars (Roshwald 2001). In addition, nationalist 'successes' can inspire nationalists elsewhere, further reinforcing the spatio-temporal clustering of border change. Such diffusion of ideas was well advanced in 19th century Europe and spread globally thanks to the "Wilsonian moment" after World War I (Manela 2007).

Our discussion has highlighted the impact of ethnic geography on border change. Because there have been many more ethnic groups that may strive for nationhood than there have been states since the late 19th century,[4] we expect secession to be the most important type of border change in this process (Gellner 1983; Griffiths 2016; Hechter 2000). Testing the primacy of secession in a separate mechanism analysis, our empirical work mostly focuses on the overall impact of ethnic settlement patterns on 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 historical, time-variant data on state borders and ethnic geography in Europe since 1886. This section explains how we go beyond previous approaches to analyzing the determinants of borders by modeling the European landmass as a spatial

[4] Even more so after the German and Italian unifications which fall outside our present empirical scope.

network of points on which we encode our data. We use the resulting dataset to test our hypothesis with the newly developed *Probabilistic Spatial Partition Model* (PSPM), which we introduce in the subsequent section.

Geographic space as a network of points

We model geographic space as a network of points, a move that addresses limitations of previous quantitative analyses of the determinants of state borders. These have followed two approaches. First, Carter and Goemans (2011) assess characteristics of newly drawn borders, focusing on whether they follow previous subnational administrative divisions. While a valuable description of border characteristics, this approach exclusively selects on the dependent variable (new borders), neglecting all potential but unrealized international borders, in particular the remaining set of administrative boundaries.

A second approach by Kitamura and Lagerlöf (2020) examines whether arbitrary grid cells are crossed by a border or not. While featuring no selection issues, doing so disregards nonmonotonic spatial dependencies inherent to the outcome of interest. Because borders partition space into contiguous territorial units, they are interdependently assigned to 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. Classic spatial error clustering or lags are unable to recover this spatial dependency structure.

In response to the limitations of previous approaches, we start from 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, while also avoiding selection bias. Coupled with the partition model introduced below, the network structure of the data allows us to capture the spatial dependencies that characterize territorial borders. Taking a network of points instead of one of grid cells additionally guarantees that our units of analysis have unambiguous outcomes. While points can only be in one state at any time, grid cells likely straddle state borders. G covers Europe[5] as a hexagonal lattice with 1096 nodes and 2905 edges. Its nodes j are

[5] We define 'Europe' in physical geographic terms, its eastern border being the Bosporus, the

connected to their up to 6 first-degree neighbors k at a distance of $\sim$100km (Figure 1a).[6]

Data on state borders

Our main outcome is the map of states at a given time, 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 limit ourselves to analyzing borders in every 25th year, i.e., in 1886, 1911,..., 2011.[7] 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 the colored partitions on the map carry substantive meaning in that we can distinguish "Spain" from "France," these partition labels are, for the purpose of this study, completely interchangeable. Because we do not ex ante know the number or names of states, we are not interested in whether certain vertices become part of a state named 'France.' Rather, the outcome of interest is whether certain vertices together belong to a contiguous state territory – a partition. The set of all partitions defines the partitioning of Europe into states.

Data on historical ethnic settlement patterns

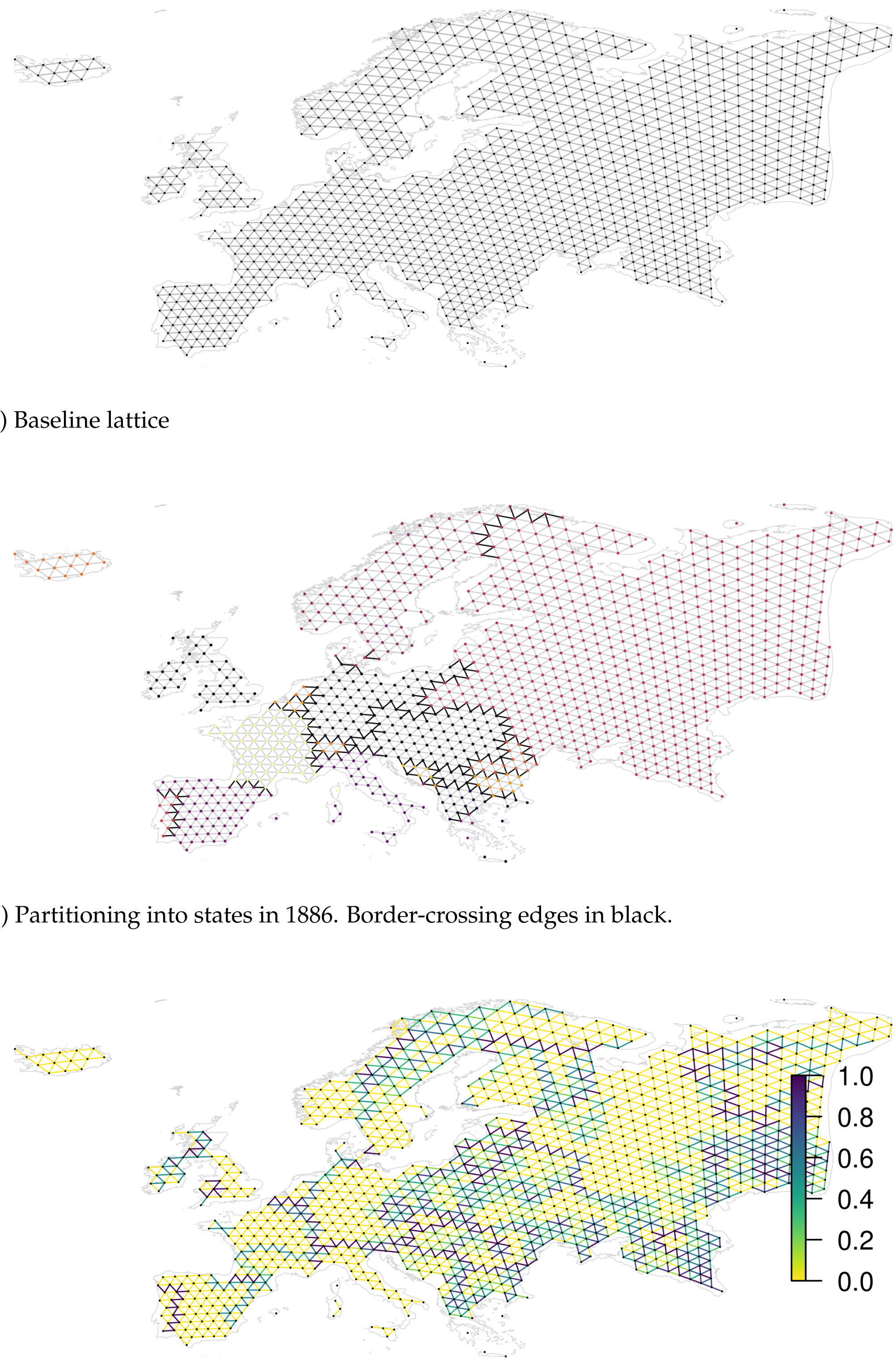
We collect new data on ethnic settlement areas in Europe since 1855. Our main independent variable is defined at the edge-level and measures whether its vertices j and k are located in the same ethnic group or not. We construct this measure from 73 historical maps that depict changes in ethnic settlement patterns over the past 165 years. Some of these changes are well known and documented – in particular genocides and population exchanges[8] – while assimilation has altered the

Black Sea, the Carpathian mountain ridge, the Caspian Sea, and the Ural. This avoids bias from definitions based on existing states.

[6] The hexagonal structure minimizes geographic distortion. Appendix D shows robustness to varying the graph's exact location, resolution, and structure.

[7] Appendix D analyzes alternative temporal structures.

[8] Such as the Armenian genocide (1915-1923) or the 1923 population exchange between Greece and Turkey.



(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

ethnic map more gradually. Accounting for these dynamics, our data avoid reverse causality that may arise when contemporary data on ethnic geography are projected into the past.

Ethnic maps first emerged in the middle of the 19th century and became increasingly widespread ever since. Their proliferation was driven by two developments: First, innovations in statistics and cartography enabled the categorization of local populations based on language and religion. Second, the rise of state-driven and peripheral nationalisms created a demand for maps of the various ethnic groups across Europe (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 classifying and mapping ethnic groups.

For the most part, maps were drawn based on census data on the town- or district-level,[9] and relied on native language as the defining ethnic marker (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, Withers et al. 1999). Mapmakers therefore sought to establish common scientific standards and provided detailed justifications (Dörflinger 1999; Hansen 2015).

At the same time, however, ethnic maps and census data were also used for political purposes. In particular, states and nationalist movements employed them to shape perceptions of national homelands and support territorial claims (Herb 2002; Anderson 1991). 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). But the scope for manipulation was limited. Because mapmakers largely relied on the same data and methods, they could not arbitrarily "invent" ethnic boundaries (Hansen 2015) without jeopardizing their reputation (Herb 2002).[10] 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

[9] Some maps were also based on philological research, travel reports, local ethnographic research, and previous maps (Dörflinger 1999; Hansen 2015).

[10] Blatant manipulation had consequences, as when geographers boycotted the journal *Petermann's Geographische Mitteilungen* due to its nationalist editor (Herb 2002).

2015), and the underlying list of ethnic groups to be counted and mapped (Hirsch 1997; Cadiot 2005).[11] At the same time, early ethnic categorizations may have affected ethnic identity formation itself where people identified with the groups they were assigned to (Kertzer and Arel 2002; Anderson 1991).

As with all data on ethnic demographics, the political importance and potential manipulation of ethnic maps could bias our analysis. Lacking ground-truth information on 19th century ethnic geography in Europe, our mitigation strategy consists of four parts:

First, we carefully screened our map material to exclude the most obvious cases of political bias. Starting with over 350 maps, we selected the 73 most suitable maps based on the absence of obvious bias and a high spatial precision.[12] These maps were drawn by 64 authors from 18 nationalities and cover various parts of Europe at different points in time, sometimes using different categorizations of ethnic groups.[13] Second, we average ethnic settlement patterns across all maps from a given period, reducing the impact of potential biases on any one map. Third, our spatial graph G is relatively coarse with a baseline spatial resolution of 100 km and up to 200 km in a robustness check. Most differences between and manipulations of ethnic maps will affect much smaller areas (see Figure 2). 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 rules out reverse causality, as well as strategic map manipulations during the World Wars.

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. The variable is formally defined as

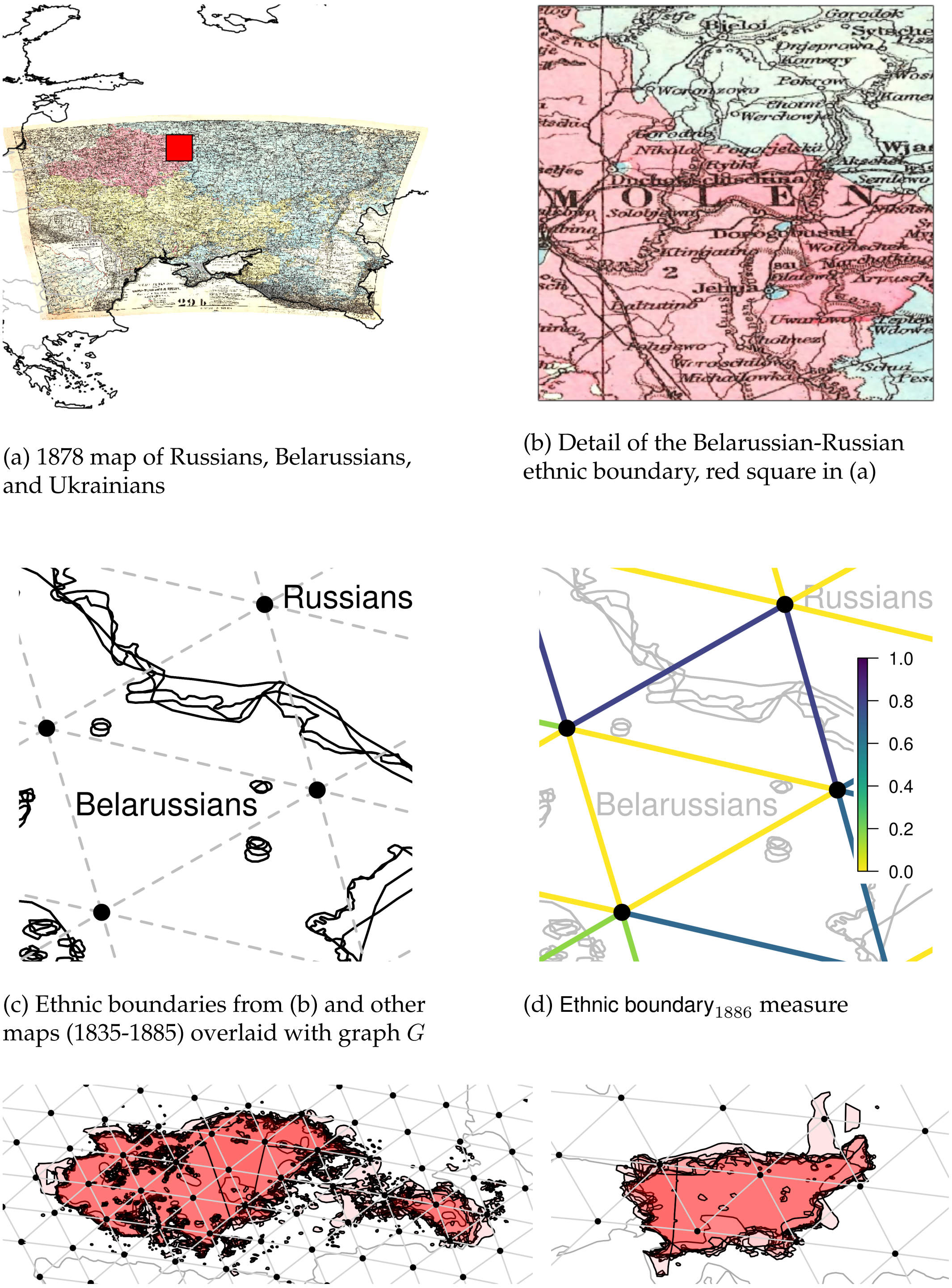
$$\textsf{ethnic boundary}_{j,k,t} = \frac{1}{M_{j,k,t}} \sum_{m=1}^{M_{j,k,t}} \mathbb{1}_{g_{m,j} \neq g_{m,k}} \tag{1}$$

where j and k are an edge's constitutive nodes observed in year t. The ensemble of maps $M_{j,k,t}$ consists of the set of maps that cover the geographic location of j

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

[12] Appendix C.1 details our selection criteria.

[13] See Posner (2004, 850-1) on the grouping problem of ethnic identities.



(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.

and k in one of the 50 years prior to t. The variable $\textsf{ethnic boundary}_{j,k,t}$ is the simple arithmetic mean of the map-level indicators that are 1 if a map m shows nodes j and k as being located in different ethnic settlement areas and 0 otherwise.[14]

Modeling and estimation

We start from the idea 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 each other, they are likely part of the same state. If pulled 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. A point's ultimate 'membership' is therefore the probabilistic result of the interplay of the attraction and repulsion exerted by and among all its neighbors.

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 multiple 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, as for example rivers. In the following, we first present and validate the PSPM. We 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 dependencies in the graph. We here present the fundamentals of the model, explain our approach to estimation and the quantification of uncertainty, and summarize the results of validating Monte Carlo experiments. We refer to Appendix A

[14] Where settlement areas overlap, we compute the share of groups for which $g_{m,j}$ differs from $g_{m,k}$.

for any 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).[15] Because systems typically move towards a low energy, low-energy partitionings are associated with comparatively high 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 four interconnected vertices separated by an ethnic boundary and a river. The plot maps five (out of twelve possible) partitionings, the color and numbering of each node indicating its 'country.' In the example, tensions result when states are too small (b, d), multi-ethnic (a, c), or divided by the river (a, e). Intuitively, partitionings with ubiquitous tensions (left) are less likely than those with less tension (right).

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:[16]

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

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). At the focus of our empirical interest are the determinants of each edges' potential

[15] 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.

[16] More complex PSPMs could account for higher-level predictors.

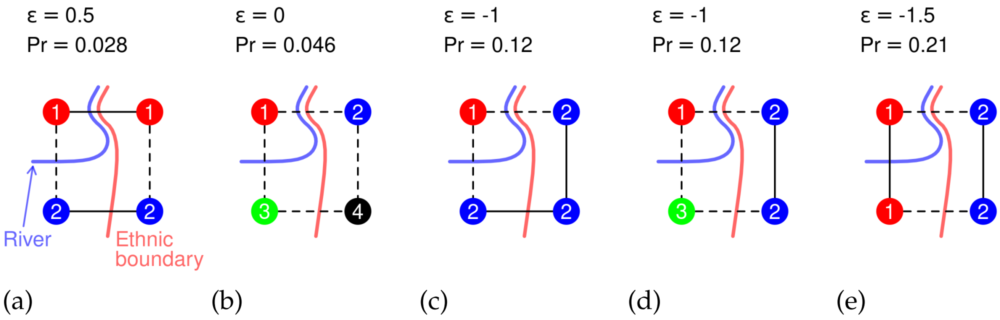


Figure 3: Illustration of the PSPM

Note: Spatial lattice with two border determinants, an ethnic boundary (red) and a river (blue). Depicts five possible partitionings of the lattice, each attributed with a total energy ϵ and a probability Pr. 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).

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 our illustration in 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 (d and e). Borders along the river (c) have a reduced probability. Finally, because of a baseline attraction between nodes (negative β_0), partitionings with many small countries have a low likelihood (b and d).

Because edge values of $s_{j,k}$ are strongly interdependent, a direct interpretation of coefficients is difficult for most edges in G. The one exception consists in *bridge edges*. Bridge edges connect two otherwise disjoint network parts (i.e. a peninsula with the continent) and can therefore independently switch $s_{j,k}$ without violating the transitivity requirement. 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; Varin, Reid and Firth 2011). 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 uses the same 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.[17]

Validation: We test the validity of inferences drawn from our model in an extensive series of Monte Carlo experiments presented in detail in Appendix B. Across varying β parameter combinations, our results demonstrate that our estimator is asymptotically unbiased in the size and number of independent networks, and that parametric bootstrapping produces consistent frequentist uncertainty estimates.

Empirical strategy

To test our main Hypothesis, we estimate the effect of ethnic geographies on the partitioning of our spatial lattice G_t into states with the following baseline specification of the edge-level energy function:[18]

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

[17] See Appendix A.2.

[18] Benchmark edge-level logit estimates are upwards biased and overconfident. See replication materials.

where β_0 is the baseline repulsion between nodes and $\textsf{ethnic boundary}_{j,k,t}$ 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, the size of the largest river[19] and watershed[20] 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., Morgenthau 1985; Kitamura and Lagerlöf 2020). We scale all variables to range between 0 and 1 to facilitate the comparison of our coefficients.

Our second analysis uses a lagged dependent variable 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:

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

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 state borders in 1936, we control for state borders in 1911 and construct $\textsf{ethnic boundary}_{j,k,t-1}$ 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 $\textsf{ethnic boundary}_{j,k,t-1}$. This avoids bias from reverse causality.

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 bor-

[19] Based on the ordinal Natural Earth data: https://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-rivers-lake-centerlines/

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

ders, the share of years in which an edge crosses a border in AD 1100, 1200, ..., 1600, and 1790.[21] Because we lack early-19th century ethnic maps, we cannot estimate the lagged dependent variable specification for the 1886 outcome data.

We first estimate our baseline and lagged dependent variable models on the pooled sample of all historical snapshots. In a second step, we estimate separate models for each period to gauge variation in the effects of ethnic geography over time. Throughout, we use a parametric bootstrap to derive confidence intervals.[22]

Results

Overall, we find consistent support for our theoretical argument. We do not only estimate a strong correlation of ethnic boundaries with state borders in the baseline model, but also find similarly sized effects in our lagged dependent variable models. In other words, even when accounting for current and past political borders, we find that ethnic boundaries are strongly and positively related to the formation of new borders. We discuss a series of robustness checks thereafter.

Main results: Table 1 presents the main results obtained from estimating the baseline the lagged dependent variable 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, showing that nodes located in differing ethnic settlement areas repulse each other and become increasingly separated by state borders. The respective effect is only slightly larger in the baseline model than in the lagged dependent variable model which accounts for past borders and their determinants. This result shows that the baseline estimates are not simply driven by reverse effects of state borders on ethnic geographies and omitted variables that have a simultaneous effect on both. Importantly, the effects of ethnic boundaries are sizeable. They are associated with almost two

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

[22] Appendix D.4 shows robustness to varying burn-in rates of the underlying Gibbs sampler.

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

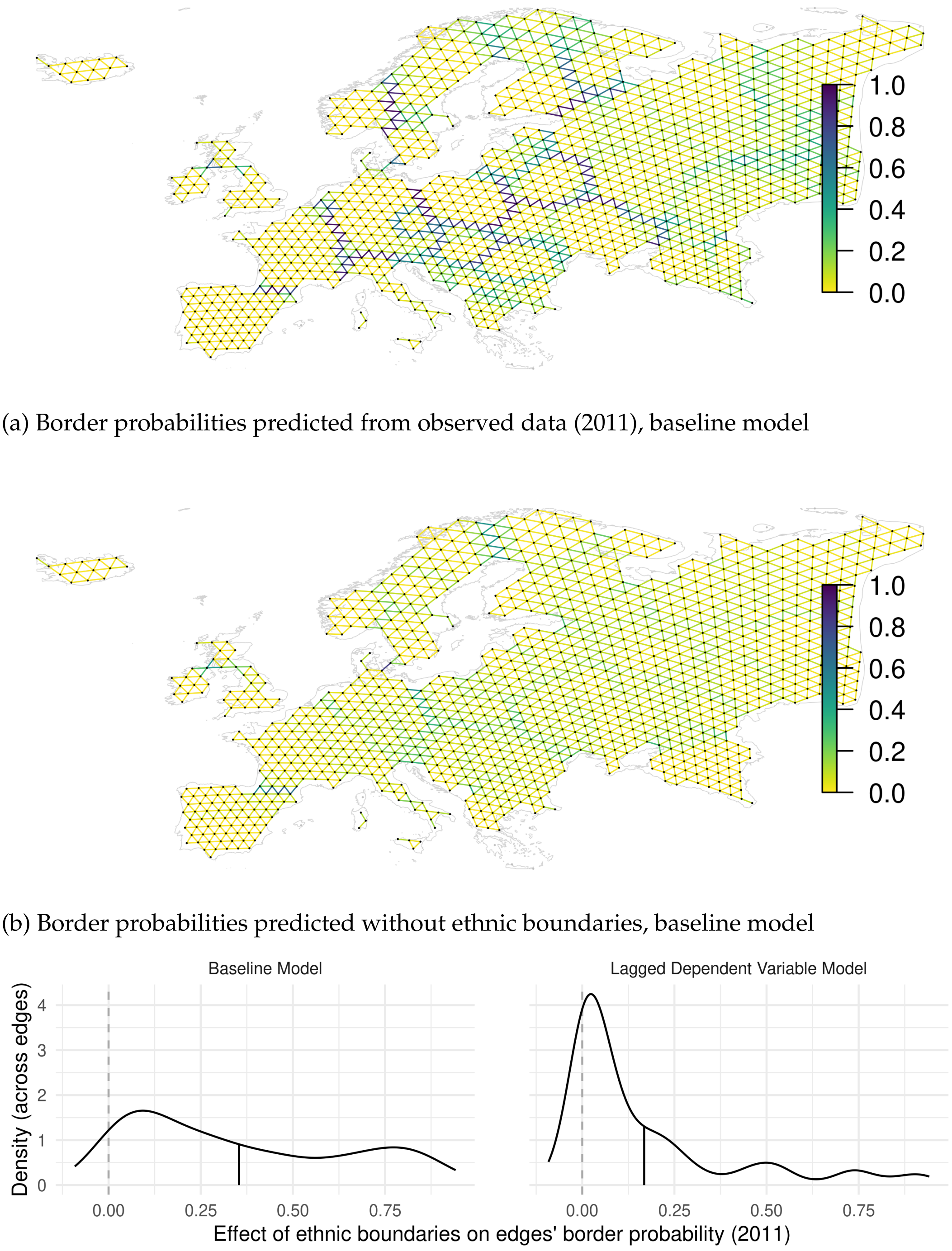
	Baseline	Lagged Dep. Var.
Constant	-2.25^{*}	-3.01^{*}
	$[-2.44; -1.98]$	$[-3.45; -2.55]$
Ethnic boundary$_t$	1.24^{*}	
	$[1.12; 1.45]$	
Ethnic boundary$_{t-1}$		1.03^{*}
		$[0.81; 1.26]$
State border$_{t-1}$		1.65^{*}
		$[1.44; 1.92]$
Deep lag		0.80^{*}
		$[0.42; 1.20]$
Edge length	-0.30^{*}	-0.32^{*}
	$[-0.49; -0.15]$	$[-0.61; -0.04]$
River	0.25^{*}	0.22
	$[0.05; 0.48]$	$[-0.18; 0.53]$
Watershed	0.64^{*}	0.76^{*}
	$[0.42; 0.82]$	$[0.47; 1.09]$
Elevation mean	0.26	0.31
	$[-0.48; 0.82]$	$[-0.90; 0.99]$
No. of periods	6	5
No. of vertices	6769	5412
No. of edges	17923	14243
No. of states	189	177

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

thirds of the energy attributed to a lagged state border and 4 to 5 times the energy attributed to the largest European river (the Danube).

Conditional on ethnic boundaries, the remaining estimates mostly support previous theoretical arguments. Large watersheds and rivers are likely to divide locations into different states. We find no robust evidence that high-altitude terrain supports border formation. Lastly, and consistent with the findings by Abramson and Carter (2016), the lagged dependent variable model shows that state borders from between the 10th and 18th century continue to separate nodes after 1886.

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 oth-



(a) Border probabilities predicted from observed data (2011), baseline model

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

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

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

Note: Sampled (a) based on observed data from 2011 and (b) based on counterfactual data without ethnic boundaries, using parameters from Model (1), Table 1. Panel (c) plots the distribution of the difference in the predicted probabilities for edges crossing an ethnic boundary. Straight lines at mean values.

erwise disjoint parts of the lattice and are therefore independent. For these bridge edges, the coefficient of ethnic boundary implies an odds ratio of 3.5 [3.1, 4.3][23] 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 10.6 [8.9, 12.0] to 29.0 [26.4, 32.2] percent. The lagged dependent variable model yields an odds ratio of 2.8 [2.3, 3.5] and a change in the border probability from 5.5 [4.1, 7.3] to 14.1 [10.9, 17.8] percent.[24] These substantial effects constitute a lower bound to the effects of ethnic boundaries which increase as they cross multiple interdependent edges.

Interpreting the results for the more common case of interdependent edges requires using our estimates to repeatedly sampling partitionings of graph G. With the resulting set of partitionings, we can compute predicted edge-level border probabilities as the fraction of partitionings in which an edge crosses a border. To assess the joint effect of all ethnic boundaries, we sample two types of partitionings. The first type is sampled from the observed data in 2011. The second, counterfactual type is sampled assuming that all of Europe belongs to the same ethnic group[25] but holding all other covariates at their observed values. The joint effect of all observed ethnic boundaries on an edge is then the difference between its probability of crossing a state border derived from the observed and that obtained from the counterfactual data.

Figure 4 plots the results of this procedure. Panel (a) and (b) map the predicted probabilities of each edge derived from the observed and counterfactual data for the year 2011 using the estimates from our baseline model. Comparing Panel (a) with (b), we see that incorporating information from ethnic boundaries in (a) greatly increases the fit of the predicted border probabilities with the contemporary map of Europe. Panel (c) plots the distribution of the difference between these two estimates for all edges that cross an ethnic boundary. The plot clearly shows that ethnic boundaries substantially increase border probabilities, with effects that are larger than the ones for bridge-edges discussed above. On average, border probabilities increase by 35 percentage points in the baseline model. In the

[23] 95% CI in parentheses.

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

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

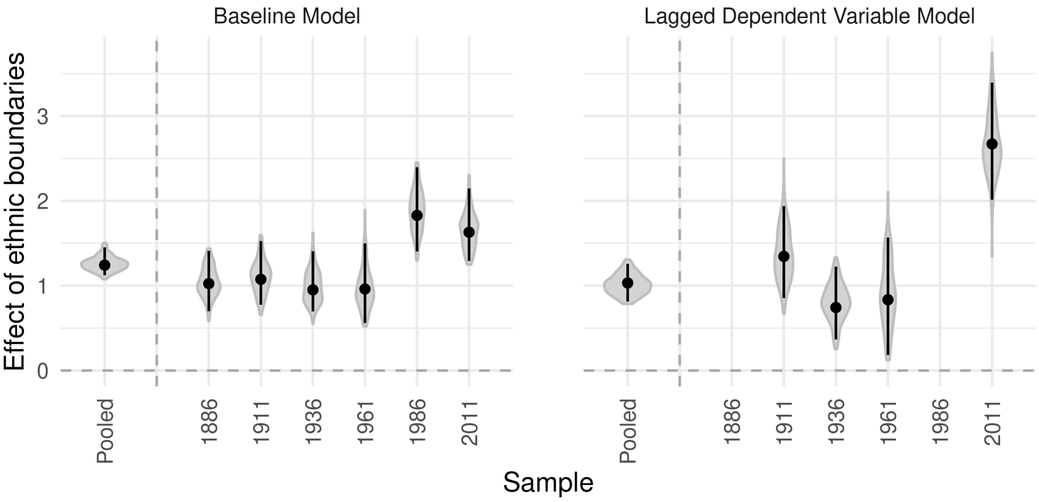


Figure 5: Effect of ethnic boundaries on the partitioning of Europe into states
Note: 95% CIs and grey areas show the distribution of bootstrapped estimates.

lagged dependent variable model, border probabilities increase by 17 percentage points. This lower effect results from the 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 disaggregates the results of the pooled models. To shed light on the temporal dynamics in the reshaping of states, we estimate a separate model for each 25th year in our data (1885, 1911, ..., 2011). We see that the association of state borders with ethnic boundaries estimated from the baseline specification increases over time. This is consistent with the main hypothesis and the lagged dependent variable estimates. The temporally disaggregated lagged dependent variable models show 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.[26] World War II did come with a slightly smaller ethnic alignment of state borders, and no changes occurred in Europe between 1961 and 1986. In line with Skocpol (1979) and Abramson and Carter (2021), these patterns suggest that systemic instability increases nationalist

[26] Our results are consistent with the fact that Post-Soviet and Post-Yugoslav borders mostly followed administrative boundaries. These were often created based on ethnic geography (e.g., Hirsch 2000) and only administrative borders that roughly coincided with ethnic divides were 'upgraded' to state borders.

border change.

Robustness checks

Our robustness checks assess whether the main findings are driven by potentially endogenous changes in ethnic geography, the choice of control variables, as well as the spatio-temporal structure of our data. Appendix D presents all details and results of the analyses summarized below.

Pre-1886 ethnic boundaries: Political biases may affect in particular ethnic data produced during the World Wars. In addition, our main 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 and re-estimate our models. The results in Figure 6 show that the effects of these stable historical ethnic boundaries are only marginally smaller than our baseline estimates. We also observe a similarly increasing alignment of state borders to ethnic boundaries as above. Reaffirming the absence of reverse and providing evidence against political bias in our analysis, the lagged dependent variable results show that pre-1886 ethnic boundaries continued to affect border changes even a century later.

Control variables: We assess whether our main results are sensitive to the specification of control variables. First, we re-estimate our main models without control variables. Second, we add control variables to the baseline specifications, controlling for terrain ruggedness, 1880 population density around vertices,[27] as well as the absolute longitude and latitude change covered by an edge.[28] These variations do not substantively change the estimated effects of ethnic boundaries.

Variation of the data structure: We also test the sensitivity of our results to our spatio-temporal data structure. Regarding the temporal dimension, our results are

[27] From Goldewijk, Beusen and Janssen (2010).

[28] See Laitin, Moortgat and Robinson (2012) who show that countries tend to be east-west oriented
.

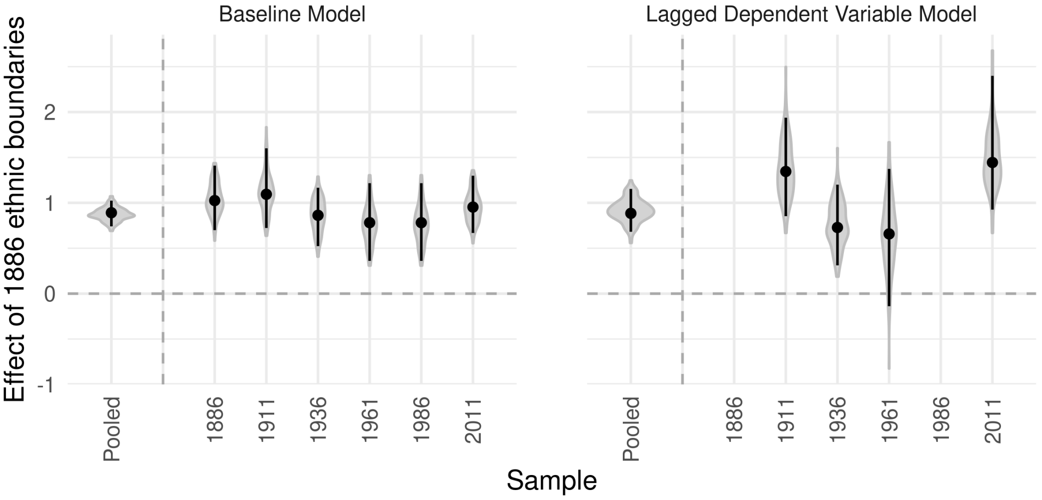


Figure 6: Effect of pre-1886 ethnic boundaries on the partitioning of Europe into states

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

robust to varying the length of periods t between 5 and 65 years.[29] We also implement robustness checks that vary the three parameters that determine the spatial data structure: the location of the 'anchor' vertex, the length of its edges, and its connectivity structure. First, we shift our network 100 times in the east-west and north-south direction. Second, we vary the length of edges between 50 and 200km. Third, we implement triangular, quadratic, and random lattice structures. For each resulting network, we regenerate the entire dataset and re-estimate our main specification. Our estimates remain statistically and substantially significant and similar to the baseline results across all network specifications. As additional evidence against potential bias from ethnic maps that are erroneous or manipulated, effects *increase* with coarser networks in which spatial measurement error becomes less relevant.

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 results with early ethnic data and coarse spatial networks also suggests the absence of substantive bias from political manipulation of ethnic data. In the next section, we provide evidence on secessionist

[29] 65 years is the maximum period length that produces at least two periods.

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, we posit that that secessionism drives the border-changing effects of nationalism. We here test this conjecture by analyzing whether ethnically distinct peripheral regions were more likely to experience secessionist claims, conflict, and ultimate secession since 1946.

Data

The vertices of our baseline lattice G constitute the units of analysis,[30] 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 (Schvitz, Germann and Sambanis 2021, 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.[31]

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

[30] Appendix E shows robustness to different spatial data structures.

[31] We construct this variable in parallel to the network-based variable ethnic boundary (Eq. 1). Appendix E shows robustness with pre-1886 ethnic data.

secession leading to selection out of the treatment group:[32]

$$h(\tau)_{j,t} = h_0(\tau)\, exp(\beta_1\, \textsf{non-coethnic capital}_{j,t} + \gamma\, \mathbf{X}_{j,t} + \epsilon_{j,\tau}) \tag{7}$$

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.[33] Next to our variable of interest $\textsf{non-coethnic capital}_{j,t}$, 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.

Results

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 about 39 percent to be part of a claimed, violently pursued (19 percent), or realized border change (40 percent). The respective probabilities for

[32] Accounting for further potential endogeneity by analyzing only point-years unaffected by post-1946 border change increases effect sizes (Appendix E).

[33] 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.

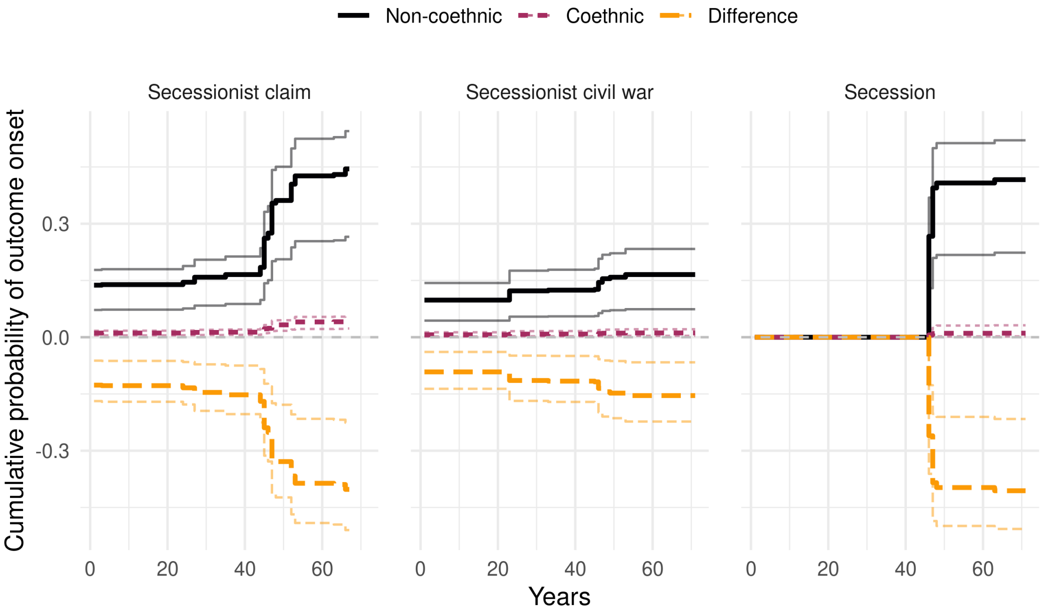


Figure 7: Effect of ethnic boundaries on secessionism.

Note: Predictions with 95% CIs based on Models 1, 3, and 5 in Appendix Table A3, setting covariates to median values.

co-ethnic areas are close to zero.[34] 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.

Global comparison

Our findings have so far been limited to 19th and 20th century Europe. We here analyze its generalizability by comparing the effects of ethnic geography on recent borders and border change in Africa, Asia, Europe, and the Americas.

To do so, we create one lattices of the same spatial structure as above for each continent. We then 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). Adapted to this data, the lagged dependent variable models control for state

[34] This corresponds to a 16, 22, and 50-times greater risk for non-coethnic regions to experience the three respective outcomes (see E).

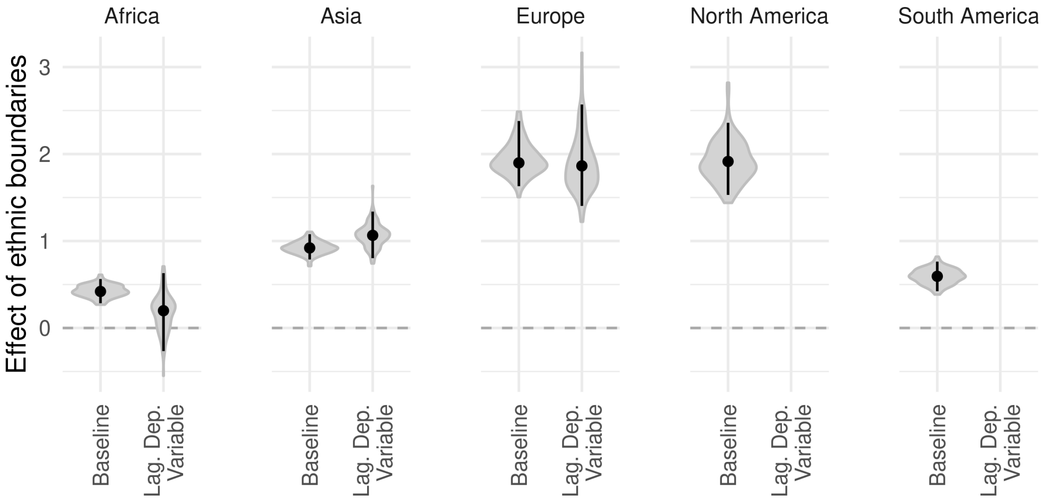


Figure 8: Effect of ethnic boundaries in 1964 on state borders across continents
Note: 95% CIs and grey areas show the distribution of bootstrapped estimates.

borders in 1964.[35]

Starting with Africa, the results in Figure 8 support the conventional wisdom that decolonization and the *uti possedetis* norm preserved colonial borders drawn with little reference to ethnic geography (Griffiths 2015; Michalopoulos and Papaioannou 2016). The baseline coefficient is relatively small (yet statistically significant) and the lagged dependent variable result shows no significant effect on border changes since 1964. Turning to Asia, the results suggest a more substantive effect of ethnic boundaries. Though 'only' half the size compared to Europe, ethnic boundaries significantly correlate with borders in 2017 and with post-1964 border change. This result is mostly driven by the independence of ethnically distinct Soviet Republics. 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 lagged dependent variable models.

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 ethno-nationalist conflicts from secessionist Kurdistan to border disputes between India and Pakistan suggest an ongoing

[35] Lacking global data on historical state borders, we omit the 'deep lag.' This does not affect results for Europe.

risk of ethnic reshaping of Asian states. In contrast, outright secessionist conflict is comparatively rare in Africa where the territorial integrity norm is generally upheld (Englebert and Hummel 2005; Zacher 2001) 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. In doing so, we contribute to the literature on international borders that has so far said little about their origins.

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

Our results show that ethnic boundaries have large and consistent effects on the location and change of state borders since 1886. We estimate that an ethnic boundary between two locations increases the likelihood of the presence of an interstate border between them by 35 percentage points. Ethnic boundaries have a similarly large effect on border change, increasing by 17 percentage points the probability of a state border conditional on past state borders. Supporting the claim that secessionist border change drives the ethnic reshaping of states, we find that peripheral ethnic minorities are at substantially higher risk to be subject to secessionist claims, conflict, and final break away from their multi-ethnic state.

In sum, our findings suggest that ethnic geography has had a substantial and continuing impact on the shape of European states. This has important implications for our understanding of state formation and its effects. State borders and the distribution of ethnic groups within them should not be treated as exogenously given. Quite to the contrary, the number, territorial shape, and ethnic makeup of states often resulted from nationalist struggles for ethnic self-determination. This should be kept in mind when comparing ethnically homogeneous European 'nation-states' with their mostly multi-ethnic counterparts elsewhere on the globe.

Moreover, our results indicate that the ethnic alignment of state borders is ongoing as we estimate the largest effect of ethnic geography on border change after 1986 when the USSR and Yugoslavia collapsed. Secessionist movements continue to challenge the borders of, for example, the Ukraine, Spain, and even France. The rising demands for Scottish independence and Irish unification only underscore the central role of nationalist struggles over borders in contemporary politics in Europe. Looking beyond Europe, we have found similar dynamics of ethnonationalist border change in Asia but not elsewhere. Mostly driven by former Soviet Republics, the available data have insufficient historical depth to draw firm conclusions on whether the region (and any other continent) follows a macrohistorical trajectory similar to Europe or not.

The answer to this question will shape the future of many multi-ethnic states. Since our analysis of post-1886 Europe is primarily structuralist, we caution against deterministic extrapolations. While it is important to recognize the *potential* of ethnic centrifugal forces, previous research points to possibilities for their peaceful containment. For example, ethnic power-sharing and regional accommodation may help defuse nationalist tensions (Cederman, Gleditsch and Buhaug 2013). More radically, dissociating states from nations (Mamdani 2020) may succeed in depoliticizing ethnic divides. Internationally, territorial integrity norms may have rained in nationalist excesses (Zacher 2001). Alarmingly, however, the recent revival of nationalist forces around the globe 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 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.

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, 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 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 Twenthieth Centuiry: The Dynamics of Recognition*. Cambridge: Cambridge University Press.

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

Darden, Keith. 2013. "Resisting occupation: Mass schooling and the creation of durable national loyalties." *Book manuscript* pp. 825–50.

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.

Dörflinger, Johannes. 1999. Zu den Sprachen- und Völkerkarten von Heinrich Kiepert. In *Antike Welten, Neue Regionen. Heinrich Kiepert - 1818-1899*, ed. Lothar Zögner. Berlin: Staatsbibliothek zu Berlin.

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

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

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

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.

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 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.

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.

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 et al. 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.

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.

Kumar, Krishan. 2017. *Visions of Empire: How Five Imperial Regimes Shaped the World*. Princeton: Princeton University Press.

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.

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, Charles WJ Withers et al. 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.

Miller, Benjamin. 2007. *States, Nations, and the Great Powers: The Sources of Regional War and Peace*. Cambridge: Cambridge University Press.

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

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.

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.

Posner, Daniel N. 2004. "Measuring ethnic fractionalization in Africa." *American Journal of Political Science* 48(4):849–863.

Roshwald, Aviel. 2001. *Ethnic Nationalism and the Fall of Empires: Central Europe, Russia and the Middle East, 1914–1923*. London: Routledge.

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.

Schvitz, Guy, Micha Germann and Nicholas Sambanis. 2021. "Mapping Self-Determination Claims 1946-2012: The GeoSDM Dataset." *Unpublished Working Paper.* .

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* .

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.

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.

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.

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

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 give the probability distribution over partitionings the form of a Boltzman distribution,

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

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 and as explained in the main paper, ϵ can be intuitively interpreted as total 'political tension' in the system when applying the model to the partitioning of space into political units. 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 partitioning energy ϵ_i is a function of 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 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 is the set of all node pairs that are connected by an edge in G. In other words, the energy of a partitioning is given by the sum of the energy of all edges that connect two nodes of the same partition.

It is worth noting an important implication of this setup: Distribution (A1) 'prefers' (i.e 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). In practice, 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.[36] 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 complete 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 s is in $\mathbb{P}$ is necessary because there are many permutations of s that do not yield valid contiguous partitionings. For one, there are many permutations of 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 s where transitivity

[36] To our best knowledge, the exact function that maps lattices onto the number of possible contiguous partitionings is unknown.

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

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

where $\mathbb{S}_j$ 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

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

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). However, 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,[37] thus obtaining our parameter estimates from

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

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.[38] 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

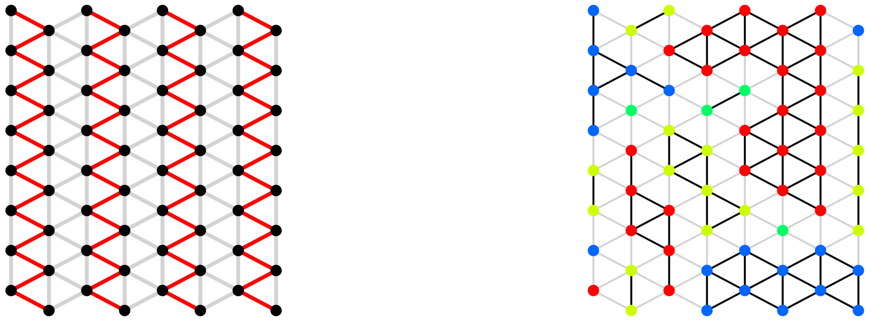
We conduct Monte Carlo experiments to test the performance of our model and the Maximum Composite Likelihood estimator estimator. The main experiments explore potential biases in estimates recovered by the estimator and investigate the precision of uncertainty estimates while varying the (1) burn in 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

[37] Throughout this paper, we set $\sigma = 10$.

[38] See Section B.2 for an evaluation of effects of the burn-in rate on parameter estimates.

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. Grey: $x \sim N(0,1)$; red: $x \sim N(1,1)$

(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. The differing means combined with random local variation introduce a 'typical' geographic structure similar to, e.g., mountain ranges.[39]

We use the Gibbs sampler described in A.2 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 control β_0 and β_1, setting them to 'realistic' values, i.e. letting vertices have a baseline attraction with β_0 ranging between -2 and 0, and making the predictor repulse vertices with a β_1 ranging 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 particular set of parameters and fix all others at a constant value. For each parameter combination, we analyze 100 independently sampled networks. We conduct one additional experiment to evaluate the consistency of uncertainty estimates derived from a parametric bootstrap. Table A1 summarizes

[39] Note that values of x are drawn only once and are stable across instances of our experiments where lattices are of the same size.

the parameters governing each experiment. We run the experiments 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 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.

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.

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

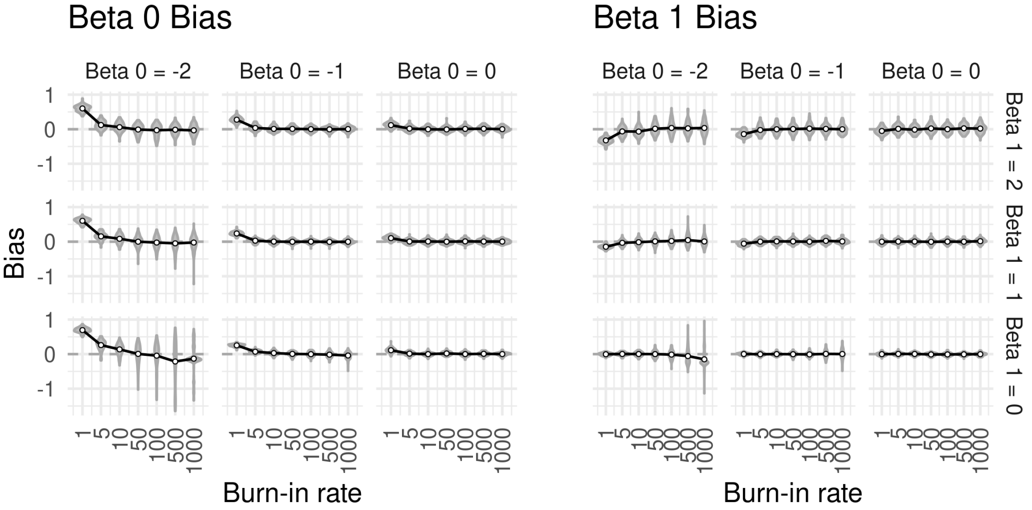


Figure A2: Bias in parameter estimates and the burn-in rate.

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.

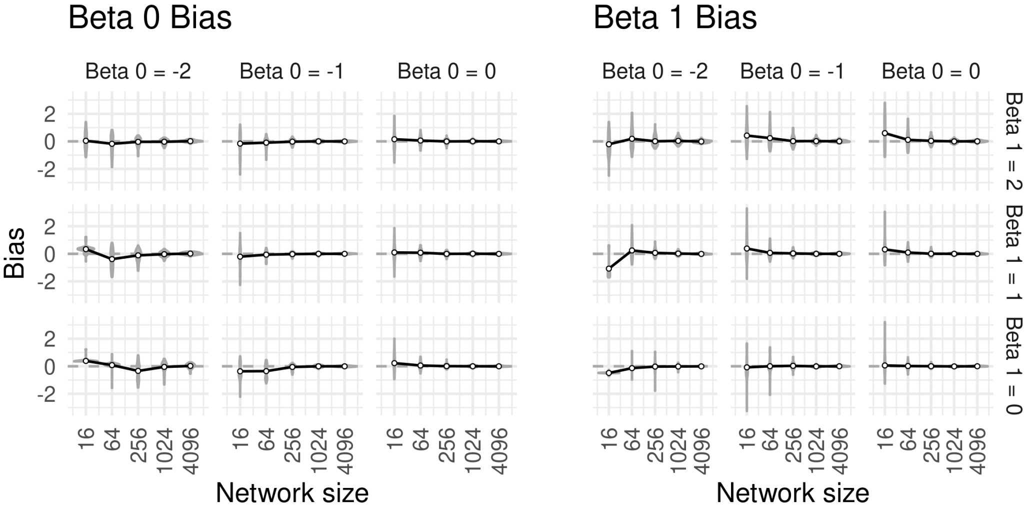


Figure A3: Bias in parameter estimates and the size of spatial lattices.

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.

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.

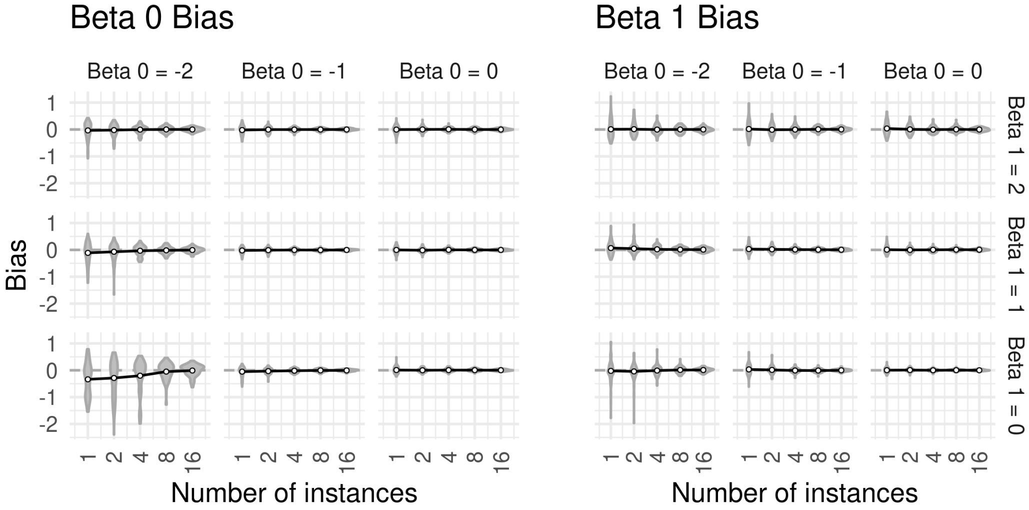


Figure A4: Bias in parameter estimates and the number of independent of spatial lattice instances.

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.

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 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.

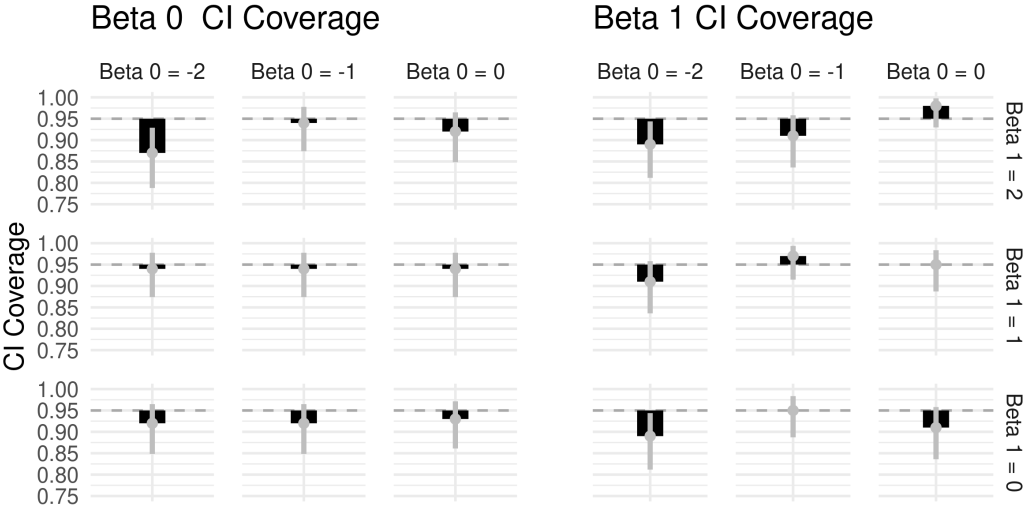


Figure A5: Precision of confidence interval coverage: Standard errors and 95% confidence intervals derived from parametric bootstraps (Section A.4.

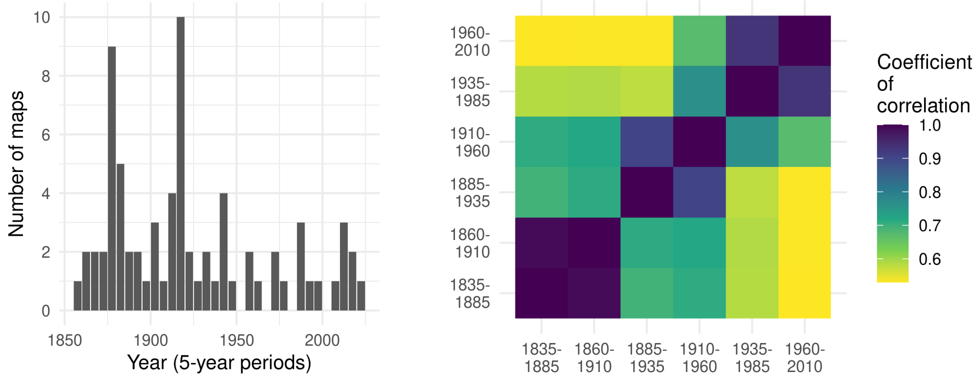
Note: Grey bars denote the 95% confidence interval of the CI coverage estimates. Monte Carlo simulations with 100 iterations; 1 instance each; 1024 nodes; burn-in rate of 100; β_0 and β_1 as plotted.

C Data

C.1 Historical ethnic map collection

We worked with a team of research assistants to gather ethnographic maps of Europe from the 19th century to the present, relying on 25 different online and archival resources. This yielded a total of ca. 350 digitized maps,[40] 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 a group's 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. (5) Maps cannot be duplicates of other maps (some maps were just slightly altered, republished versions of earlier ones). Figure A6 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.

[40] 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.



(a) Number of maps over time

(b) Correlation of ethnic boundary across periods t

Figure A6: Historical ethnic data: Summary

C.2 Data on self-determination claims: GeoSDM

To capture secessionist claims, we draw on new spatial data from GeoSDM (Schvitz, Germann and Sambanis 2021). 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 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:

- 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.

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 A2 presents the results of dropping the main and adding the additional covariates. 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.

In addition, the additional coefficients exhibit some interesting patterns. First, the estimated coefficient for Δ Longitude, suggest that edges with an east-west orientation are less likely to separate two states only in the lagged dependent variable model. The results also suggest that there are more border-crossing edges in densely populated areas. Lastly, in the baseline but not lagged-dependent variable specification, the edges' ruggedness affects their likelihood to push vertices into different states.

Table A2: Determinants of state borders in Europe, 1886–2011: Varying control variables

	Baseline	Lagged Dep. Var.	Baseline	Lagged Dep. Var.
Constant	-2.03^*	-2.69^*	-2.69^*	-1.59^*
	$[-2.15; -1.93]$	$[-2.94; -2.45]$	$[-3.50; -1.58]$	$[-3.12; -0.56]$
Ethnic boundary$_t$	1.31^*		1.24^*	
	$[1.19; 1.52]$		$[1.10; 1.44]$	
Ethnic boundary$_{t-1}$		1.07^*		1.01^*
		$[0.81; 1.29]$		$[0.77; 1.24]$
State border$_{t-1}$		1.66^*		1.65^*
		$[1.44; 1.90]$		$[1.44; 2.03]$
Deep lag		0.75^*		0.85^*
		$[0.37; 1.13]$		$[0.42; 1.26]$
Edge length			-0.33^*	-0.27^*
			$[-0.51; -0.16]$	$[-0.55; -0.02]$
Largest river			0.26^*	0.14
			$[0.04; 0.48]$	$[-0.24; 0.42]$
Largest watershed			0.72^*	0.82^*
			$[0.52; 0.92]$	$[0.51; 1.13]$
Elevation mean			0.57	0.19
			$[-0.78; 1.60]$	$[-1.35; 2.42]$
Δ Longitude			-0.09	-1.86^*
			$[-1.17; 0.74]$	$[-2.90; -0.38]$
Δ Latitude			0.42	-0.96
			$[-0.58; 1.25]$	$[-1.96; 0.56]$
Population density 1880			1.46^*	-1.00
			$[0.64; 1.96]$	$[-2.55; 0.39]$
Cumulative altitude change			-1.20	-0.03
			$[-2.40; 0.25]$	$[-1.58; 1.34]$
Std. dev. altitude			1.35^*	-0.03
			$[0.35; 2.12]$	$[-1.31; 1.34]$
No. of periods	6	5	6	5
No. of vertices	6769	5412	6769	5412
No. of edges	17923	14243	17923	14243
No. of states	189	177	189	177

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.2 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 (see Section and Figure 5 in the main paper). 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.[41] 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.

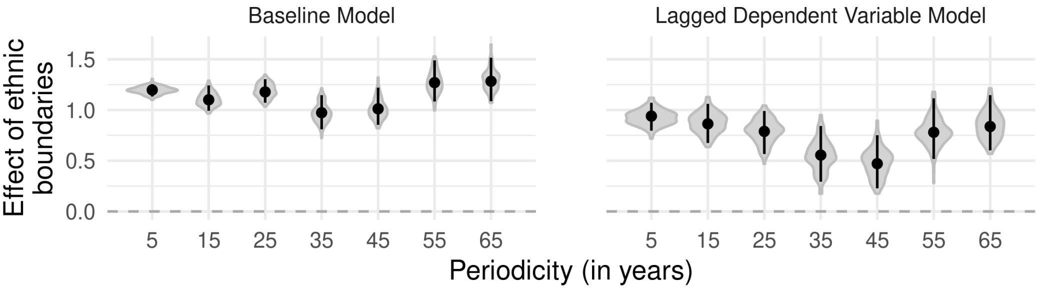


Figure A7: Point estimates of the effect of ethnic boundaries on the partitioning of Europe into states: Varying the period of the the length of periods t

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 A7, 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.3 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

[41] 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.

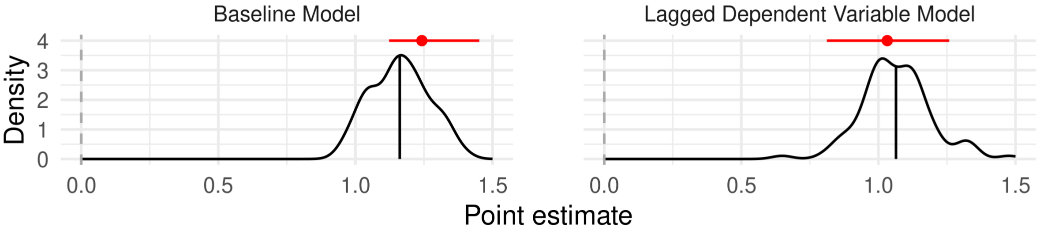


Figure A8: Point estimates of the effect of ethnic boundaries on the partitioning of Europe into states: Shifting the spatial lattice

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.

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 makeup of our baseline lattice consists in the location of the "anchoring" point (in our 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 – slightly[42] 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 A8 shows that our main estimates are well centered at the 77th and 45st 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 A9 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

[42] We shift the lattice by displacing the anchoring point with random draws from a uniform distribution between 1 and 10 decimal degrees in each direction.

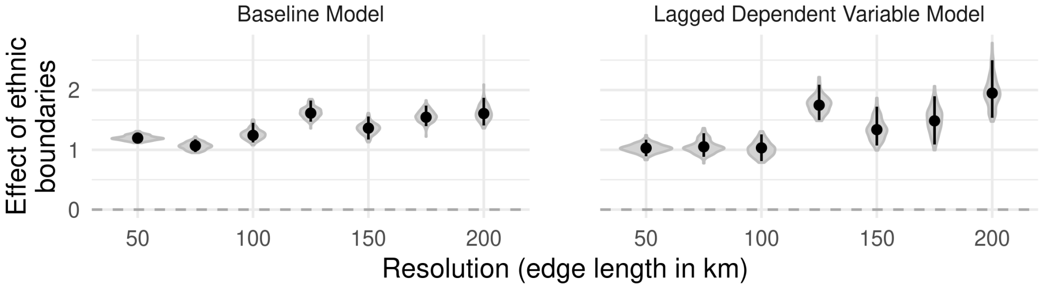


Figure A9: Effect of ethnic boundaries on the partitioning of Europe into states at varying resolutions of the spatial lattice

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

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 other regular tilings, the quadratic and triangular tiling from which we can generate regular lattices.[43] 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 lattice structures feature less edges per vertex. Given a constant edge-length, they therefore yield a thinner network structure and are, 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 A10 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 results suggests that the hexagonal lattice structure yields if at all conservative estimates doe to its increased ability of capturing spatial interdependence.

[43] 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.

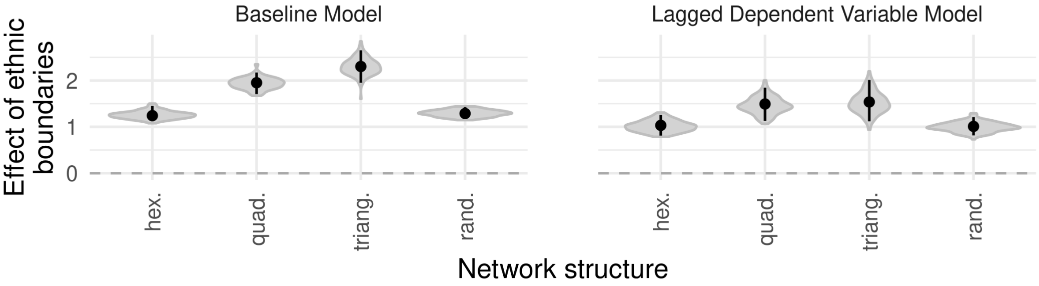


Figure A10: Effect of ethnic boundaries on the partitioning of Europe into states using a hexagonal, quadratic, triangular, and random lattice structure

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

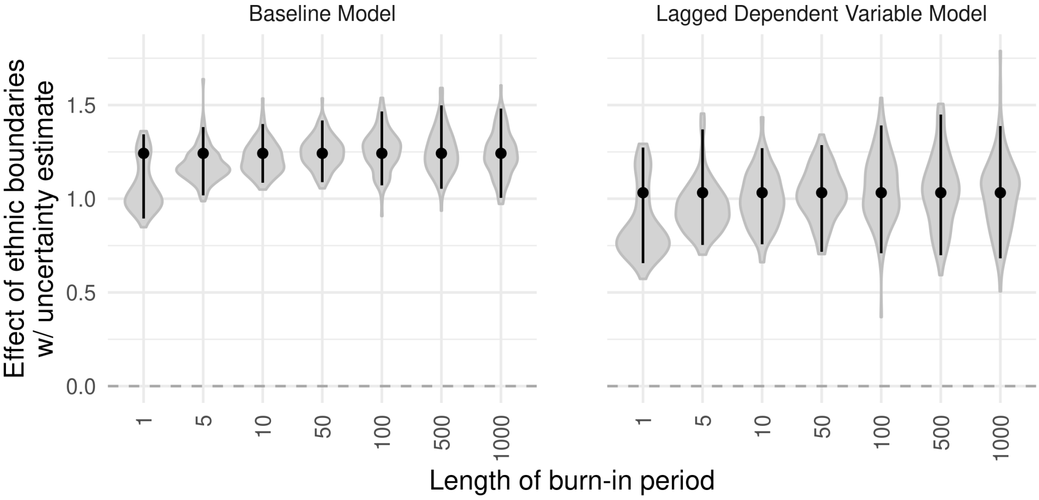


Figure A11: Uncertainty estimates with varying burn-in rates

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.

D.4 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 A11 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.

E Robustness checks: Analysis of secessionist claims and conflict

This section presents the robustness check for 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.

Main results: Table A3 presents the main results discussed in the paper.

Table A3: 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	2.602***	1.736***	2.766***	2.086***	3.918***	2.922***
	(0.337)	(0.381)	(0.471)	(0.369)	(0.609)	(0.694)
Events:	207	207	122	122	153	153
Country-year strata:	no	yes	no	yes	no	yes
Controls:	yes	yes	yes	yes	yes	yes
Observations	61,607	61,607	67,587	67,587	71,851	71,851
R^2	0.007	0.005	0.005	0.003	0.007	0.005
Max. Possible R^2	0.045	0.031	0.025	0.019	0.029	0.023
Log Likelihood	-1,217.990	-826.011	-697.294	-534.679	-781.121	-623.632

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

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 A4 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.

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 A5). Given the reduced precision of the data, standard errors slightly increase. Together with the

Table A4: 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	2.391***	1.801***	3.281***	2.459***	3.904***	2.922***
	(0.314)	(0.404)	(0.510)	(0.530)	(0.611)	(0.694)
Events:	197	197	102	102	153	153
Country-year strata:	no	yes	no	yes	no	yes
Controls:	yes	yes	yes	yes	yes	yes
Observations	55,640	55,640	60,807	60,807	64,905	64,905
R^2	0.007	0.005	0.005	0.004	0.008	0.006
Max. Possible R^2	0.047	0.033	0.023	0.019	0.032	0.025
Log Likelihood	-1,129.301	-804.624	-538.761	-468.544	-780.951	-623.632

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: $^*p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

overall stability of ethnic geographies, this suggests that endogenous changes of ethnic geographies are unlikely to cause the results.

Table A5: Ethnic boundaries and the onset of self-determination claims, conflict, and border change: Ethnicity data from before 1886

	Cox Proportional Hazard Model					
	Secessionist Claim		Secessionist Civil War		Secession	
	(1)	(2)	(3)	(4)	(5)	(6)
Non-coethnic$_{1886}$ capital$_t$	1.443**	0.989*	2.400***	1.726***	2.933***	1.693***
	(0.652)	(0.595)	(0.516)	(0.632)	(0.580)	(0.636)
Events:	207	207	122	122	153	153
Country-year strata:	no	yes	no	yes	no	yes
Controls:	yes	yes	yes	yes	yes	yes
Observations	61,709	61,709	67,677	67,677	71,941	71,941
R^2	0.005	0.004	0.004	0.003	0.006	0.005
Max. Possible R^2	0.045	0.031	0.025	0.019	0.029	0.023
Log Likelihood	-1,278.070	-845.274	-713.677	-542.935	-839.608	-652.928

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: $^*p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Varying the spatial sampling of points: As in the PSPM analysis (see Section D.3 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 A12). Figure A13 demonstrates robust results when varying the spatial resolution of our data. Lastly, Figure A14 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.

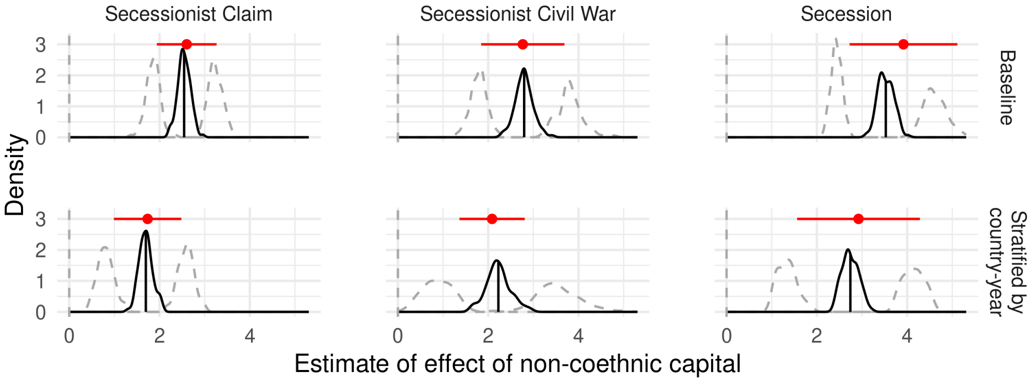


Figure A12: Secessionism robustness check: Shifting points (unit of analysis)

Note: Results from Table A3 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.

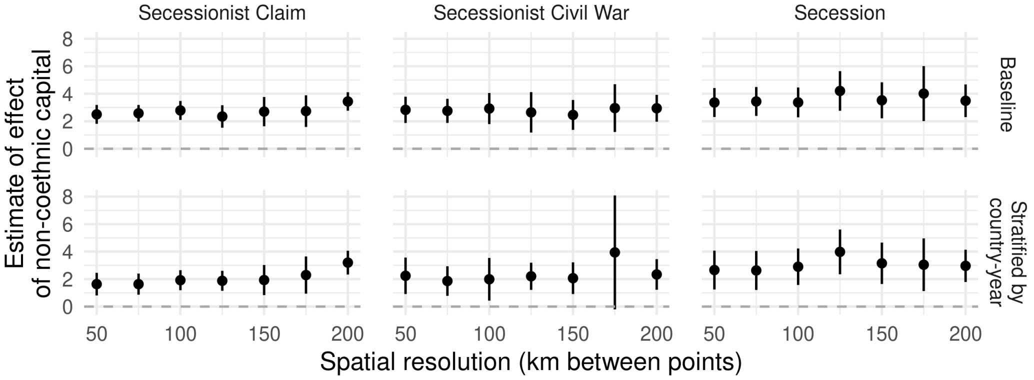


Figure A13: Estimates of the effect of non-coethnic capitals on secessionism at varying spatial resolutions lattice

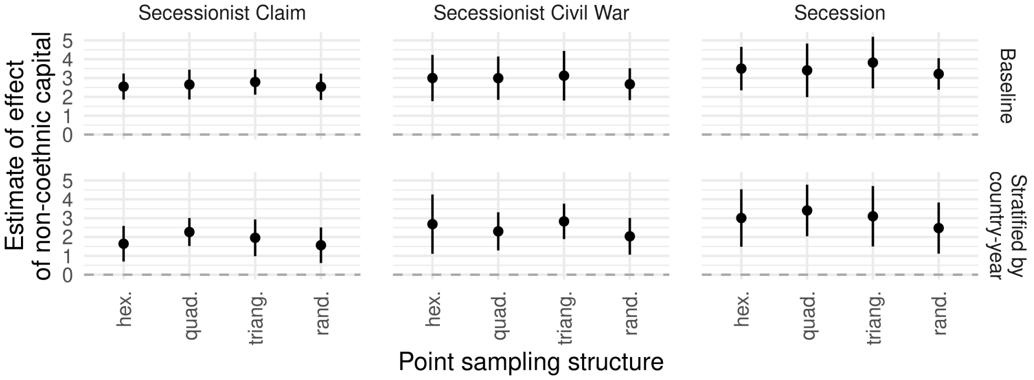


Figure A14: Estimates of the effect of non-coethnic capitals on secessionism with a hexagonal, quadratic, triangular, and random lattice structure

F References (Appendix)

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.

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

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.

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.

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.

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, Guy, Micha Germann and Nicholas Sambanis. 2021. "Mapping Self-Determination Claims 1946-2012: The GeoSDM Dataset." *Unpublished Working Paper.* .

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.

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.