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

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

May 28, 2021

Abstract

Borders define states, yet little systematic evidence explains how and where they are drawn. Putting recent challenges to state borders into perspective, 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 test this argument with a newly developed Probabilistic Spatial Partition Model that models state territories as partitions of a planar spatial graph. Introducing new data on Europe's ethnic geography since 1855, we consistently find that ethnic boundaries between two locations strongly increase the probability that they are, or will become, separated 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: State formation; borders; ethnicity; nationalism; Europe

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

Introduction

Borders are constitutive features of the modern state system that define the size and shape of states and specify the limits of their sovereignty. While a growing literature has focused on the political implications of borders (e.g., 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 the process of their formation. This question, however, has gained relevance as existing borders have come under pressure and risk being replaced by new ones. Not only Russia's annexation of Crimea in 2014 has 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 conflict in Northern Ireland.

Despite its theoretical and political importance, we lack systematic evidence on the drivers of border formation. In this paper, we examine the role of ethnic geography in shaping state boundaries, focusing on Europe since the late 19th century. We argue that the historical rise of nationalism created a demand for ethnically homogeneous nation-states and prompted efforts to redraw borders along ethnic lines. Nationalism mostly motivated secessionism in multi-ethnic states, but occasionally also powered unification and irredentism. As a result, borders became increasingly aligned with the underlying ethnic map. Although previous accounts support these claims (Beissinger 2002; Hechter 2000; Weiner 1971), the relationship between ethnicity and state borders remains contested with some arguing for the primacy of state-led efforts of creating ethnic nations (Hobsbawm 1990). Absent systematic evidence, it remains unclear whether, how, and by how much ethnic geography has shaped today's state borders.

Addressing this question requires solving three empirical challenges: The first concerns the right unit of analysis. Previous analyses have focused exclusively on existing borders (Carter and Goemans 2011) or used grid cells to study border formation (Kitamura and Lagerlöf 2020). Such approaches fail to account for counterfactual outcomes and ignore spatial dependencies that characterize borders. Sec-

ond, estimating the effect of ethnic geography on state borders requires considering geographic features that affect both. Third and lastly, we require data on ethnic geography that predate state borders and their changes as to avoid reverse causality from state-led ethnic assimilation and cleansing.

To address these issues, we develop a new *Probabilistic Spatial Partition Model* (PSPM) that allows us to estimate the effect of ethnic settlement patterns and other geographic features on the partitioning of Europe's landmass into states since 1886. By treating geographic space as a planar network of points that is partitioned into state territories, our model accounts for spatial dependencies and estimates effects conditional on covariates. Importantly and beyond its current application, the PSPM can be used to model any type of spatial partitioning, such as administrative units or electoral districts.[1]

Our analysis pairs the PSPM with new time-varying spatial data on ethnic settlement areas in Europe since 1855 digitized from 73 historical ethnographic maps. Combined with strategies that mitigate their potential political biases, these allow us to study borders and border changes based on pre-existing ethnic settlement areas, which prevents reverse causality. To minimize omitted variable bias, we pair a static baseline with a lagged dependent variable model of the effect of ethnic geography on border change.

We find that the presence of an ethnic boundary between two locations substantially 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 changing the spatio-temporal data structure. Additional analyses of post-World War II ethno-nationalist secession suggest it to be a key driver of the realignment of state borders: Areas home to peripheral ethnic groups have a 6 to 50 times greater risk of experiencing secessionist claims, civil wars, and border change. We finally explore whether our findings generalize beyond the European context. Although we find a static correlation between ethnicity and borders on all continents, ethnic boundaries explain border change only in Europe and Asia. Post-colonial Africa and the Americas have thus far avoided extensive ethno-nationalist

[1]The PSPM will be distributed as an R package released upon publication.

border change.

The origins of international borders

The partitioning of geographic space into states is one of the most consequential political processes that shapes national and international politics. It determines the number, size, and shape of states, as well as internal attributes such as their geography, demography, and economy. Understanding border formation can furthermore yield insights where new borders may emerge, and which borders may become unstable. However, systematic empirical evidence on arguments how and where borders are drawn is scarce (but see Carter and Goemans 2011; Kitamura and Lagerlöf 2020).[2] Below, we briefly outline realist and institutionalist arguments, before discussing a complementary perspective on ethnicity and nationalist border change. We then turn to the main empirical problems of assessing the origins of borders.

The realist perspective commonly starts from the assumption that power-maximizing states compete over territory and resources (e.g., Morgenthau 1985; Tilly 1990). In the struggle of all against all, borders emerge along natural obstacles such as mountains and waterbodies or man-made barriers that allow states to project power internally while keeping invaders at bay (Morgenthau 1985; Kitamura and Lagerlöf 2020). Some borders may also reflect geopolitical concerns over the balance of power (Møller 2014).

A second theoretical view holds that borders are institutions that help coordinate peaceful interstate relations. Correspondingly, they result from a "mixed-motive" game in which states compete over territory, but also desire border stability (Simmons 2005). To simplify bargaining and increase stability, states settle on borders that follow geographic features, cartographic lines, or historical precedents (Carter and Goemans 2011). The general expectation is that such borders help mitigate future conflict (Goemans and Schultz 2017; Abramson and Carter 2016).

A third, complimentary perspective highlights ethnic geography and national-

[2]This stands in contrast to evidence on the impact of borders on, for example, conflict (Abramson and Carter 2016; Carter and Goemans 2011; Michalopoulos and Papaioannou 2016), trade (Simmons 2005; Carter and Goemans 2018), and development (Alesina, Easterly and Matuszeski 2011).

ism as a main driver of border formation. Adopting a cultural and mostly individualist perspective, Alesina and Spolaore (1997, 2005) explain the size and shape of states based on a trade-off between economies of scale and and costs of ethnic heterogeneity in large states (Friedman 1977; Desmet et al. 2011). In turn, macro-sociological theories explain the origins of state borders in nationalism, defined by (Gellner 1983, p. 1) as "a political principle which holds that the political and national unit should be congruent." They hold that violations of the nationalist ideal caused by a mismatch between states and nations generate collective grievances and popular pressures to redraw the map. These frequently result in attempts to establish new borders along ethnic lines, either by secession, unification, or irredentism (Weiner 1971; Hechter 2000; O'Leary 2001).

Because there are more potential ethnic nations then states (Gellner 1983), secessionist conflict is the most frequent type of conflict emerging from structural state-nation incongruence Hechter (2000). The phenomenon has been extensively analyzed, partly because it is prone to violence, which is often fueled by ethnic inequality (Cederman, Gleditsch and Buhaug 2013). Historically, secessionist nationalism has contributed to the collapse of multi-ethnic empires (Hiers and Wimmer 2013; Beissinger 2002; Roshwald 2001) and continues to threaten multi-ethnic states (Germann and Sambanis 2020). In contrast, research on rarer processes of unification and irredentism is more limited (Breuilly and Speirs 2005). In one of the few systematic studies of national unification, Griffiths (2010) finds that it requires linguistic homogeneity. Studying irredentist conflict on the Balkans before World War I, Weiner (1971) argues that intra- and interstate dynamics can escalate nationalist claims over territory and fuel conflict. More recently, Siroky and Hale (2017) show that irredentism can be fueled by political grievances of minorities with an ethnic kin state abroad.

While a substantive literature documents secessionist, unifying, and irredentist nationalism and conflict, the claim that ethnic groups shape states remains contested. In fact, Hobsbawm's (1990, p. 10) claims that "[n]ations do not make states and nationalisms but the other way around." This argument coincides with many studies that show how states form nations and ethnic groups (Weber 1976; White 2004; Darden 2013; McNamee and Zhang 2019). While we do not dispute this evi-

dence, we argue that a one-sided focus on identity formation captures only half the impact of nationalism.

Despite the importance and prominence of this debate for understanding the origins of (nation-)states, the border-transforming effects of nationalism and ethnic geography have not been systematically assessed by the literature. We identify three challenges in addressing this gap.

The first challenge consists in modeling borders and their partitioning of space. Previous studies have been either limited to existing borders (Carter and Goemans 2011) or have examined border formation at the level of grid cells (Kitamura and Lagerlöf 2020). Restricting the analysis to existing borders ignores the infinite set of counterfactual borders that could have been drawn, which makes it difficult to explain border formation. Cell-based approaches, in turn, disregard non-monotonous spatial dependencies inherent to borders, which partition space into contiguous territorial units. To address these problems, we introduce a novel spatial partition model that accounts for counterfactual borders and spatial dependencies.

A second challenge in existing theoretical and empirical accounts consists of alternative explanations and potential omitted variables. To examine the implications of ethno-nationalism without bias, one needs to account for geographic features that may have shaped both state borders and ethnic geography. Our empirical approach allows us to do so by including these variables as controls, similar to standard regression models.

The third challenge relates directly to the objection that states made ethnic nations but not the other way around. So far, the lack of historical data on ethnic geography has made it impossible to untangle the reciprocal relationship between ethnicity and state borders. We here introduce data on ethnic settlement areas since 1855 and examine their impact on *changes* in international borders. Because our ethnic data temporally precedes the border change it explains, we can rule out reverse causality.

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. While the former refers to state-led assimilation and extreme outcomes such as population transfers or genocide, our focus here is on the latter.[3]

Historians have extensively documented both mechanisms across three distinct phases in European history (Schieder 1964; Alter 1989). In the first phase around the French Revolution, nationalism spread to West-European states and triggered policies to assimilate peripheral ethnic groups into national identities (e.g., Weber 1976). The second phase spread nationalism through the Napoleonic wars and the reaction of German and Italian nationalists, which ultimately led to the "right-sizing" of their states through national unification. In the third phase, modernization processes carried nationalism into Eastern Europe, in parallel to an increasing influence of German and Italian intellectuals such as Herder and Mazzini (Gellner 1983).

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 used to distinguish ethnic groups. Once such groups express a 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). While nations do not have to be ethnic, as illustrated by the civic nationalism of Switzerland, most of them are.

Gellner's congruence principle fully outlines the geopolitical consequences of ethnic nationalism, as it requires "that ethnic boundaries should not cut across political ones, and, in particular, that ethnic boundaries within a given state [...]

[3]The two processes may be interlinked as ethnic homogenization often focuses on territorially contested areas (Bulutgil 2015, 2016; McNamee and Zhang 2019).

should not separate the power-holders from the rest" (Gellner 1983, p. 1). In the latter case,[4] "alien rule" 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 seceding from their multinational state. The break-up of European land-based and overseas empires represents the most important example of this process (Kumar 2017; Beissinger 2002).

Ethno-nationalist grievances can also emerge if an ethnic group is divided by state borders. In this situation, nationalist activists may call for unification of their kin in a common state. 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 either peaceful or violent mergers 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 from status-quo-oriented leaders of 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.

Lastly, mixed incongruence exists where an ethnic group dominates one state but forms a minority in another. This configuration creates a pressure for the homeland to "liberate" the group in question, resulting in irredentist nationalism (Weiner 1971). Originally 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 mak-

[4] Also called "state-to-nation deficit" by Miller (2007).

ing activists less risk averse (Petersen 2002; Nugent 2020; Germann and Sambanis 2020). 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 2020; 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).

In sum, we argue that ethnic nationalism often results in secession and, though less frequent and influential, irredentism and national unification. Over time, the resulting border changes align the shape of states with ethnic settlement patterns. We therefore expect that

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

Beyond this static effect, 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,[5] we expect secession to be the most important type of border change in this process (Gellner 1983; Griffiths 2016; Hechter 2000). We therefore hypothesize that

Hypothesis 2 *Misalignment between state borders and ethnic settlement patterns causes changes in the shape of state territories such that ethnic boundaries and state borders become increasingly congruent.*

Data

We test our argument about the effect of ethnic boundaries on state borders using historical, time-variant data on state borders and ethnic geography in Europe since

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

1886. As explained in the following, we encode these data along with a set of covariates on a spatial network of points that cover the European landmass. We use the resulting network data to test our hypotheses with the newly developed *Probabilistic Spatial Partition Model* (PSPM), which we introduce in the subsequent section.

Geographic space as a network of points

The construction of our dataset starts 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. Taking a network of points instead of grid cells guarantees that our units of analysis have unambiguous outcomes. While points can only be in one state at any time, grid cells or other polygons are likely to straddle state borders. In our main analysis, we divide the European landmass[6] into a hexagonal lattice with 1096 nodes and 2905 edges. As shown in Figure 1a, each node j is connected to its up to 6 first-degree neighbours k at a distance of $\sim$100km.[7]

Data on state borders

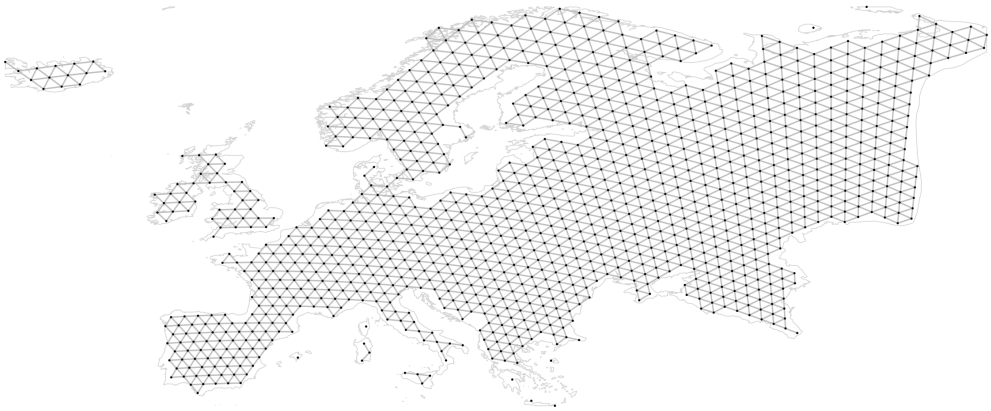
Our main outcome is the map of states at a given time, or, applied to our network, the partitioning P_t of the lattice G_t into states in year t. We measure P_t by retrieving the state each vertex belongs to between 1886 and 2019 from the CShapes 2.0 dataset (Schvitz et al. 2021). We limit ourselves to analyzing borders in every 25$^{\text{th}}$ year, i.e., in 1886, 1911, 1936, 1961, 1986, and 2011.[8] The quarter-century intervals are long enough for cumulative border change to produce meaningful variation yet short enough to capture varying patterns of border change since 1886.

Figure 1b plots the resulting outcome data for the year 1886. While the colored partitions on the map carry substantive meaning in that we can distinguish "Spain"

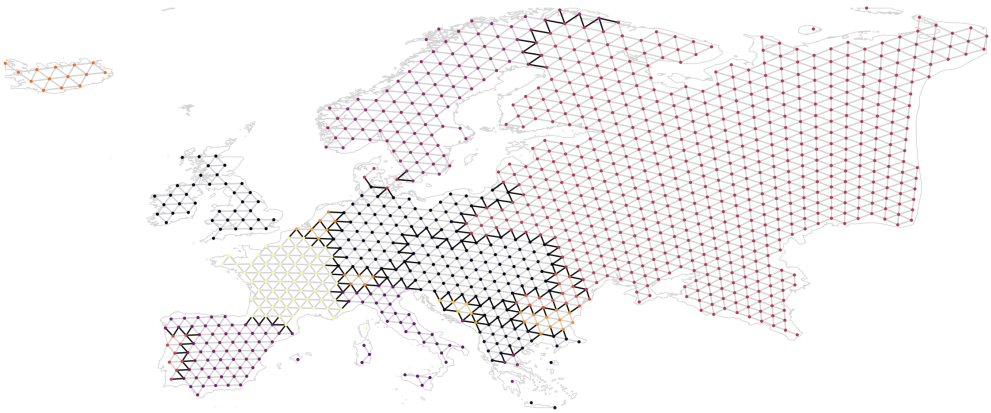
[6]We define 'Europe' in physical geographic terms, its eastern border being the Bosporus, the Black Sea, the Carpathian mountain ridge, the Caspian Sea, and the Ural. This avoids bias from a definition based on existing states.

[7]The hexagonal structure and its creation in an equal-area Albers projection minimizes geographic distortions. Appendix D analyzes variations in the exact location, resolution, and structure of the network.

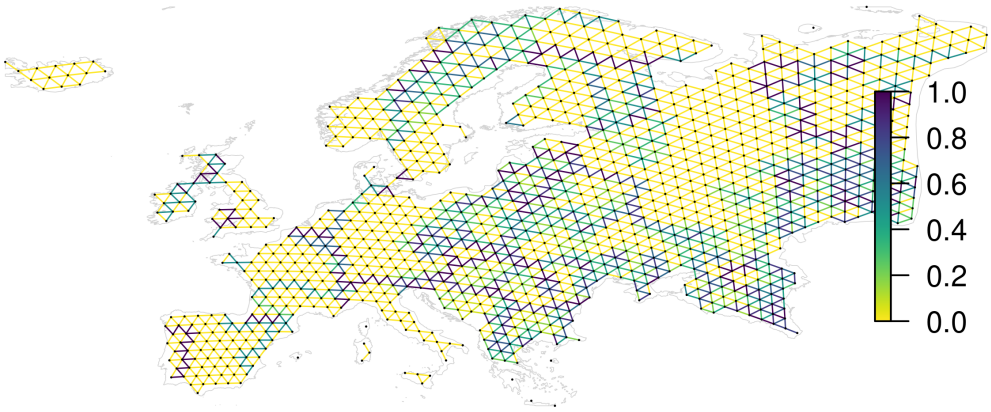
[8]Appendix D analyzes alternative temporal structures.



(a) Baseline lattice



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



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

Figure 1: Europe as a hexagonal spatial lattice

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 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 level of each edge 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 capture 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[9] – while assimilation has altered the ethnic map more gradually. Accounting for these dynamics, our time series of ethnic maps avoids reverse causality that arises when contemporary data on ethnic geography are projected into the past.

Ethnic maps first emerged in the middle of the 19$^{\text{th}}$ century and became increasingly widespread across Europe ever since. Their proliferation was driven by two major developments: First, innovations in statistics and cartography made it possible to categorize populations based on language and religion, and to represent their settlement areas on increasingly precise maps. Second, the rise of nationalism and pursuit of self-determination created a demand for maps that identified and located the various ethnic groups in Europe (Kertzer and Arel 2002; Hansen 2015). Initial efforts by German and Austrian geographers in the 1840s were quickly followed by authors from Russia, the Balkans, and other parts of Europe, laying the foundation for 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

[9]Prominent examples include the Armenian genocide (1915-1923), the 1923 population exchange between Greece and Turkey, and the expulsion of Germans from Poland (1944-1950).

district-level,[10] and relied on native language as the defining attribute of ethnic groups (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 frequently accompanied their maps with detailed justifications (Dörflinger 1999; Hansen 2015).

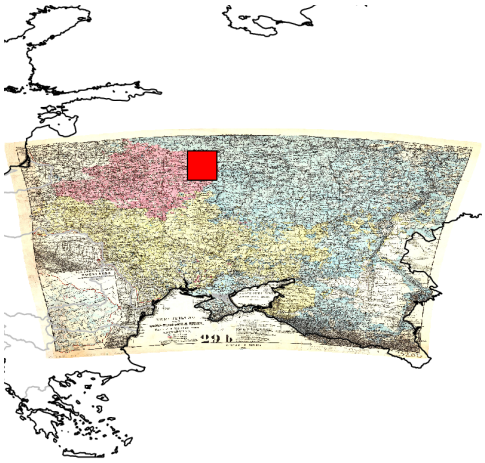
At the same time, however, ethnic maps and census data were also used for political purposes. In particular, states and nationalist movements used them to shape perceptions of national homelands and support territorial claims (Herb 2002; Anderson 1991). This was most evident at the Paris Peace conference of 1919, where all parties relied on their own maps to support their demands (Palsky 2002). But the scope for manipulation was generally limited. Because mapmakers largely relied on the same data and methods, they could not arbitrarily "invent" ethnic boundaries (Hansen 2015) without jeopardizing their scientific reputation (Herb 2002).[11] Instead, most attempts to manipulate maps and census data involved the subtle use of seemingly objective but politically convenient criteria as the choice of data sources, population thresholds used to define local ethnic boundaries (Hansen 2015), and the underlying list of ethnic groups to be counted and mapped (Hirsch 1997; Cadiot 2005).[12] At the same time, in particular early ethnic categorizations may have affected ethnic identity formation itself, as people tended to identify with the groups they were assigned to (Kertzer and Arel 2002; Anderson 1991).

Like all data on ethnic demographics, the political importance and potential for manipulation of ethnic maps raises concerns about the validity of our data. While many maps reflect earnest attempts to capture ethnic geography, those that were manipulated for political purposes could bias our analysis. Lacking "ground-truth" information on 19th century ethnic geography in Europe, our strategy to mitigate such bias consists of four parts.

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

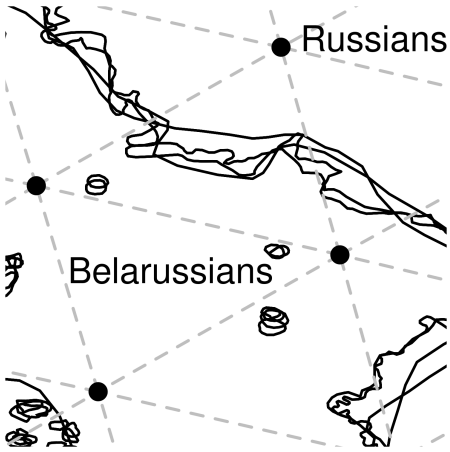
[11] Open manipulation had consequences, as when many geographers boycotted the journal *Petermann's Geographische Mitteilungen* when German nationalist Paul Langhans became its editor (Herb 2002).

[12] For example, Kertzer and Arel (2002) note that Greek, Serbian and Bulgarian nationalists in the late 19th century used alternative linguistic criteria to justify their claims on parts of Macedonia.

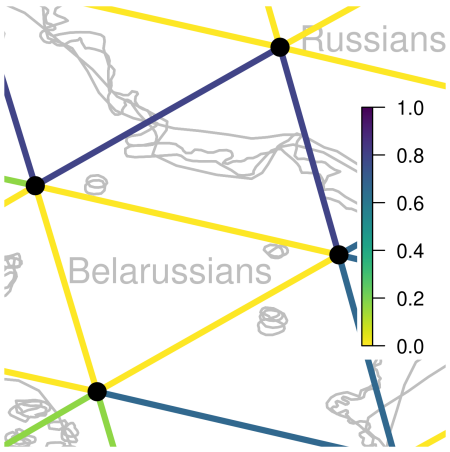


(a) 1878 map of "Great, White, and Little Russians" (Russians, Belarussians, and Ukrainians) by August Peterman

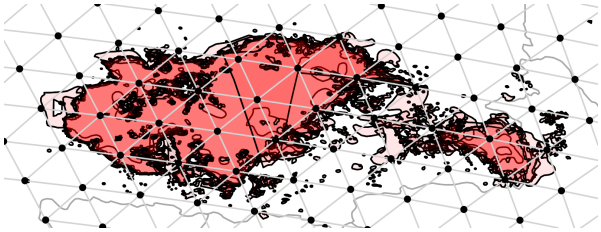
(b) Detail at the Belarussian-Russian ethnic boundary, red square in (a)



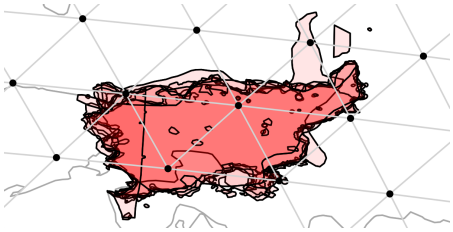
(c) Ethnic boundaries from (b) and other maps (1835-1885) overlaid with graph G



(d) Ethnic boundary$_{1886}$ measure after aggregation to network edges



(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 map data onto graph G. (e) and (f) show Hungarian and Slovenian settlement areas from multiple maps each. Darker areas are coded as Hungarian/Slovenian in more maps. Straight lines result from some maps' partial coverage.

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 their absence of obvious bias as well as their spatial resolution and precision.[13] These maps were drawn by 61 authors from 16 different nationalities and cover various parts of Europe at different points in time, in part using different categorizations of ethnic groups.[14] Second, we construct an average measure of ethnic boundaries across all maps from a given period to reduce the impact of remaining 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 substantively smaller areas (see Figure 2). Fourth, we show that our results are robust to exclusively using pre-1886 ethnic boundaries to explain state border changes between 1886 and today. This rules out reverse causality and strategic map manipulations during the World Wars.

Based on the collection of historical maps, 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. Illustrated in Figures 1c and 2d, this average measure of ethnic boundaries translates discrepancies across maps into meaningful spatial variation.[15] The variable is formally defined as

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

where j and k are an edge's constitutive nodes observed in year t. The ensemble of maps $M_{j,k,t}$ consists of the set of maps that cover the geographic location of j and k in one of the 50 years prior to t. The variable $\text{ethnic boundary}_{j,k,t}$ is the simple arithmetic mean of the map-level indicators that are 1 if a map m shows nodes j and k as being located in different ethnic settlement areas and 0 otherwise.[16]

[13] Appendix C.1 details our selection criteria, presents examples of discarded maps, and shows the maps' metadata.

[14] On the grouping problem of ethnic identities see, e.g., Posner (2004, 850-1).

[15] If, for example, a "fluid" ethnic boundary is depicted differently across maps, our final measure captures it as a gradient.

[16] Where a map shows overlapping ethnic settlement patterns, we compute the share of groups for which $g_{m,j}$ differs from $g_{m,k}$.

Modelling and estimation

Our modelling strategy starts from the idea that the partitioning of geographic space into states results from 'attractive' and 'repulsive' forces between different locations. These forces mirror factors that affect the formation of a border between two points, for example a river easy to defend, a watershed that facilitates bargaining, or an ethnic boundary in the focus of nationalists. If two points are attracted to each other, they likely form part of the same state. If pulled apart by repulsive forces, they may be 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 all its neighbors as well as the forces among them.

The Probabilistic Spatial Partition Model (PSPM) introduced here captures this logic by modeling the partitioning of space as the partitioning of a planar network. Crucially, the model allows us to estimate the attractive or repulsive forces resulting from multiple attributes of the edges in the network. When testing our argument that ethnic differences repulse points from each other and thereby causing a state border between them, the PSPM can thus account for covarying spatial features that influence ethnic settlement patterns and exert their attractive or repulsive forces, as for example watersheds or rivers. In the following, we first present and validate the PSPM in general terms and then introduce the empirical strategy implemented to test our theoretical argument.

Probabilistic Spatial Partition Model

We model state territories as contiguous and mutually exclusive clusters of nodes (partitions) of the spatial graph G introduced above. The empirical goal of our model 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, discuss its relation Exponential Random Graph Models (ERGMs), explain our approach to estimation and the quantification of uncertainty, and summarize the results of validating Monte Carlo experiments. We refer to Appendix A for any further details.

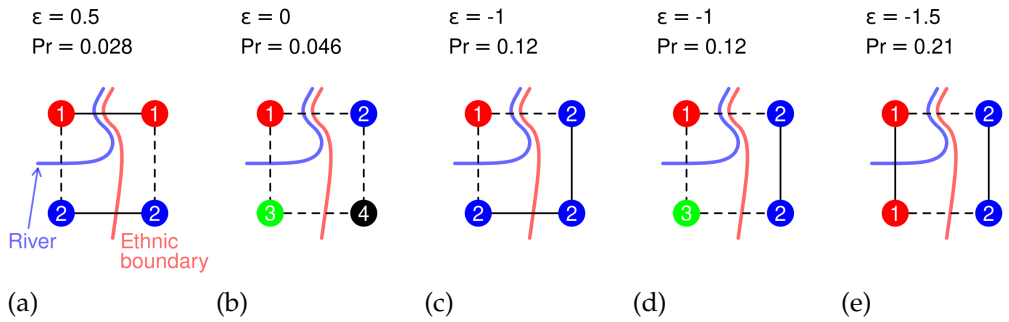


Figure 3: Illustration of the Probabilistic Spatial Partition Model

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 the following parameters: $\beta_0 = -1$; $\beta_{\text{ethnic boundary}} = 1$, $\beta_{\text{river}} = 0.5$. The potential energy of each edge (from top, clockwise) is .5, -1, 0, and -.5 (Eq. 5). Total energies and probabilities based on Eq. 2 and Eq. 3.

The model: We model the distribution over all possible partitionings P of our 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 modelling the condition of a system in statistical mechanics (e.g., Park and Newman 2004). 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 a spatial lattice with four points separated by an ethnic boundary and a river. The plot maps five (out of twelve possible) partitionings of the lattice, the color and numbering of each node indicating its 'country.' In the example, tensions result where a state is too small (b and d), or is crisscrossed by an ethnic boundary (a and c) or a river (a and e). Intuitively, partitionings with ubiquitous tensions to the left are less likely than those with low tension levels to the right.

We assume that a partitioning's total energy ϵ_i is determined by the sum of realized energies associated with the edges that connect all first-degree neighbour

node pairs L on the lattice:[17]

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

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 state ($s_{j,k} = 1$, solid lines in Figure 3) and is not realized if they are part of different states ($s_{j,k} = 0$, dotted lines in Figure 3). At the focus of our empirical interest are the determinants of each edges' potential energy:

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

which defines the potential energy ϵ of the edge between nodes j and k as the sum of a constant β_0 that captures the baseline repulsion between nodes and edge-level characteristics $\mathbf{x}_{j,k}$ weighted by the parameter vector β. In our case and as discussed in the next section, $\mathbf{x}_{j,k}$ includes the indicator $\textsf{ethnic boundary}_{j,k}$ 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 probability of observing a partitioning decreases in 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 $\beta_{\textsf{ethnic boundary}} > \beta_{\textsf{river}}$, this means that state territories aligned with the ethnic boundary have the highest probability (d and e). Borders that exclusively follow the river (c) have a somewhat lower probability. Finally, because of a constant baseline attraction between nodes caused by a negative β_0, partitionings with many small countries have a relatively 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*

[17] More complex PSPMs could, in principle, account for higher-level predictors, such as partitions' compactness.

edges. Bridge edges connect two otherwise disjoint network parts (i.e. a peninsula with the continent) and can therefore freely switch $s_{j,k}$ without violating the transitivity requirement. For these edges, we can interpret coefficient estimates in parallel to a standard logistic regression model by computing odds ratios, predicted probabilities and marginal effects (see Appendix A.2).

Relation to ERGMs: We can reformulate the PSPM as a Exponential Random Graph Model (ERGM, Park and Newman 2004). ERGMs have become a prominent inferential method for network analysis (Cranmer and Desmarais 2011) and are used to analyze, for example, international alliances (Cranmer, Desmarais and Menninga 2012) or social network formation (Lazer et al. 2010). To transform the PSPM into an ERGM, we can reformulate $P(P = p_i)$ as the probability of the realization of graph y_i that exclusively connects members of the same partition. The distribution Y from which y_i originates is restricted such that each connected component (partition) in y_i is an induced subgraph of our baseline lattice G, i.e., including all edges between the component members, thus yielding valid partitionings.

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 that a vertex takes on the observed partition membership, defined based on the membership of its neighbors. 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.[18]

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

[18]The parametric bootstrap consists of (1) sampling 120 new partitionings based on the estimated coefficients, (2) re-estimating the parameters using the sampled partitionings, and (3) deriving 95% confidence intervals from the resulting parameter distribution. We sample from independent chains with a burn-in rate of 100 iterations. See Appendix A.3.

asymptotically unbiased in the size and number of independent networks, and that parametric bootstrapping produces consistent frequentist uncertainty estimates.

Empirical strategy

To test Hypothesis 1, 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:[19]

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

where β_0 is the baseline repulsion between nodes and $\text{ethnic boundary}_{j,k,t}$ captures whether the nodes of an edge are located in different ethnic settlement areas (Eq. 1 above). To avoid bias from omitted spatial features, $\mathbf{X}_{j,k}$ must capture factors that cause ethnic as well as state borders. We therefore include time-invariant indicators for the length of each edge, the size of the largest river[20] and watershed[21] crossed by an edge, and the mean elevation along it.[22] Taken together, these covariates capture important geographic causes of ethnic geography and state borders. We scale all variables to range between 0 and 1 to facilitate the comparison of our coefficients.

A second analysis uses a lagged dependent variable model to test Hypothesis 2 and address reverse causality as the main inferential problem left open in the cross-sectional baseline setting. If ethnic settlement patterns are the result of "right-peopling" within state borders (e.g., Hobsbawm 1990) our estimate of β_1 could be systematically biased. A variant of Equation 5 accounts for past borders that may have affected ethnic geography:

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

[19] Appendix D.1 also presents results from an edge-level benchmark logistic regression. Estimates are biased upwards and standard errors overconfident.

[20] We code 9 ordinal levels following the Natural Earth data: https://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-rivers-lake-centerlines/

[21] We code a 7-step ordinal variable increasing with the Pfaffstetter scale of watersheds. Data from Lehner, Verdin and Jarvis (2008).

[22] Elevation data comes from Hastings et al. (1999).

where we model the potential energy of an edge observed 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 $\text{ethnic boundary}_{j,k,t-1}$ from ethnic maps drawn between 1860 and 1910. Because the ethnic boundaries are measured in data produced in the 50 years preceding the lagged dependent variable (see also Eq. 1), no border changes between $t-1$ and t can impact $\text{ethnic boundary}_{j,k,t-1}$. This avoids bias from reverse causality.

To furthermore exclude the possibility of historical borders causing ethnic boundaries and being reestablished as "new" borders, the lagged dependent variable model controls for a "deep lag" of state borders defined as the share of years in which an edge crosses a border in the years 1100, 1200, ..., 1600, and 1790 as observed by Abramson (2017).[23] Because we lack early-19$^{\text{th}}$ century ethnic maps, we cannot estimate the lagged dependent variable specification for the 1886 outcome data.

In our main analysis, we estimate our baseline and lagged dependent variable models on the pooled sample of all historical snapshots. In a second step, we estimate a separate model for each period to gauge variation in the effects of ethnic geography over time. Throughout, we use a parametric bootstrap with a burn-in rate of 100 iterations to retrieve confidence intervals.[24]

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 settlement areas are strongly 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 sup-

[23] 1790 is the last year Abramson (2017) covers.
[24] Appendix D.4 shows robustness to burn-in rates between 10 and 1000 iterations.

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.

port 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. 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 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).

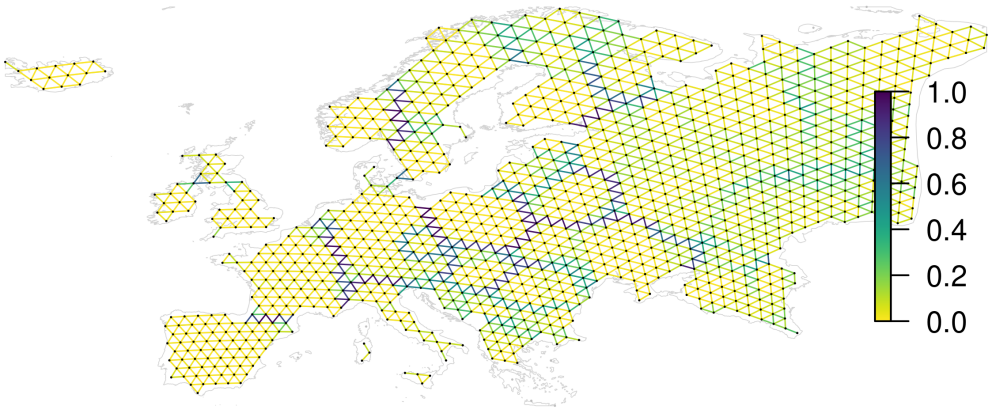
Conditional on ethnic boundaries, the remaining estimates mostly support previous theoretical arguments. Large watersheds and rivers are likely to divide locations into different states. We find no robust evidence that high-altitude terrain supports border formation. Lastly, and consistent with the findings by Abramson and Carter (2016), the lagged dependent variable model shows that state borders from between the 10^{th} and 18^{th} 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 otherwise disjoint parts of the lattice and are therefore independent. For these bridge edges, the coefficient of ethnic boundary implies an odds ratio of 3.5 [3.1, 4.3] 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.[25] This is a lower bound to the effect of ethnic boundaries which increases as they cross multiple interdependent edges.

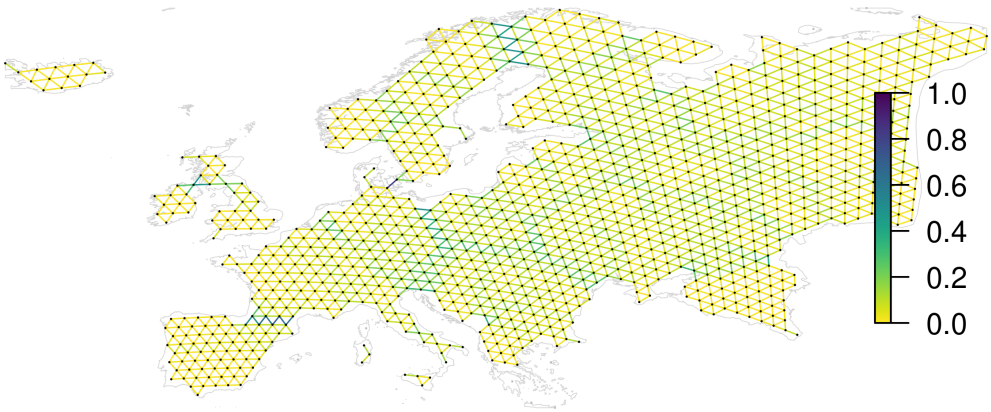
To interpret the results for this more common case of interdependent edges, meaningful interpretation requires repeatedly sampling partitionings of the entire graph. Averaging across the resulting set of partitionings, we can compute edge-level border probabilities. To assess the effect of ethnic boundaries, we sample two types of partitionings. The first type is sampled from the observed data. The second, counterfactual type is sampled after erasing all ethnic boundaries 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

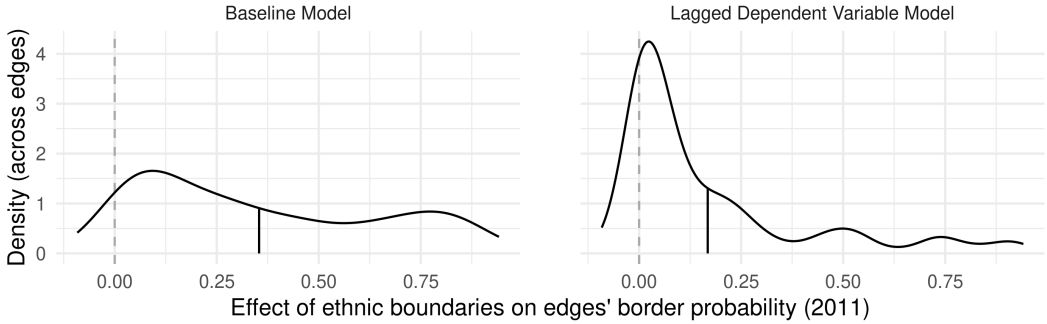
[25] 95% confidence intervals in parentheses. The lagged dependent variable model yields an odds ratio of 2.8 [2.3, 3.5] and a border probability change from 5.5 [4.1, 7.3] to 14.1 [10.9, 17.8] percent at median covariate-values.



(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 predicted border probability of edges.

Note: Derived from Gibbs-sampling 120 partitionings of Europe. Based on observed data from 2011 in (a) and counterfactual data without ethnic boundaries in (b) and parameters from the baseline model in Table 1. Burn-in rate of 100 draws. Grey nodes and edges have missing data on ethnicity. Panel (c) plots the distribution of the difference in the predicted probabilities for edges crossing an ethnic boundary. Straight lines are drawn at mean values.

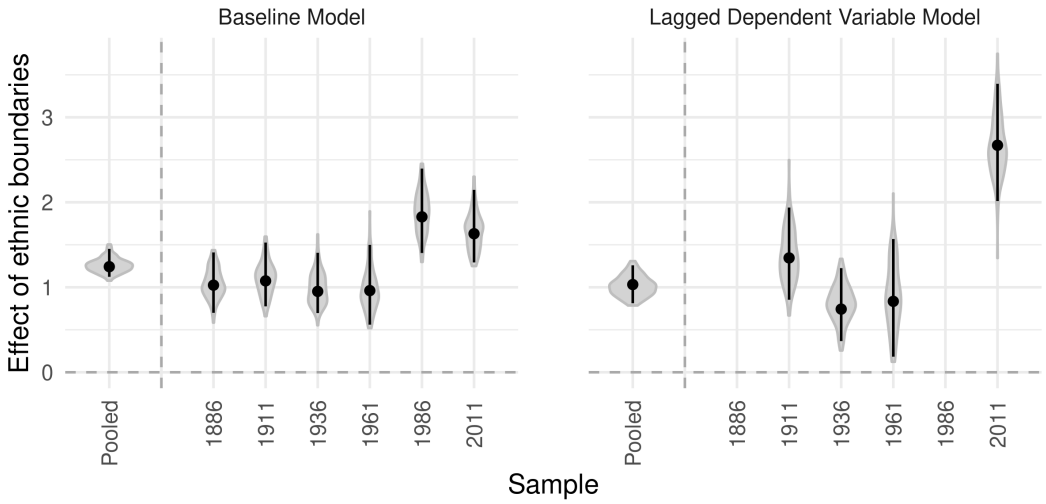


Figure 5: Effect of ethnic boundaries on the partitioning of Europe into states

Note: 95% CIs and grey areas show the distribution of bootstrapped estimates. The lagged dependent variable model cannot be separately estimated for 1986 because of perfect collinearity of 1961 and 1986 borders.

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. On average, border probabilities increase by 35 percentage points in the baseline model. In the lagged dependent variable model which models border *change* by controlling for past borders, border probabilities increase by 17 percentage points. This lower effect results from the relatively small baseline probability of border change.

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. 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 (2020), these patterns show that systemic instability increases the risk of nationalist 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

[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.

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

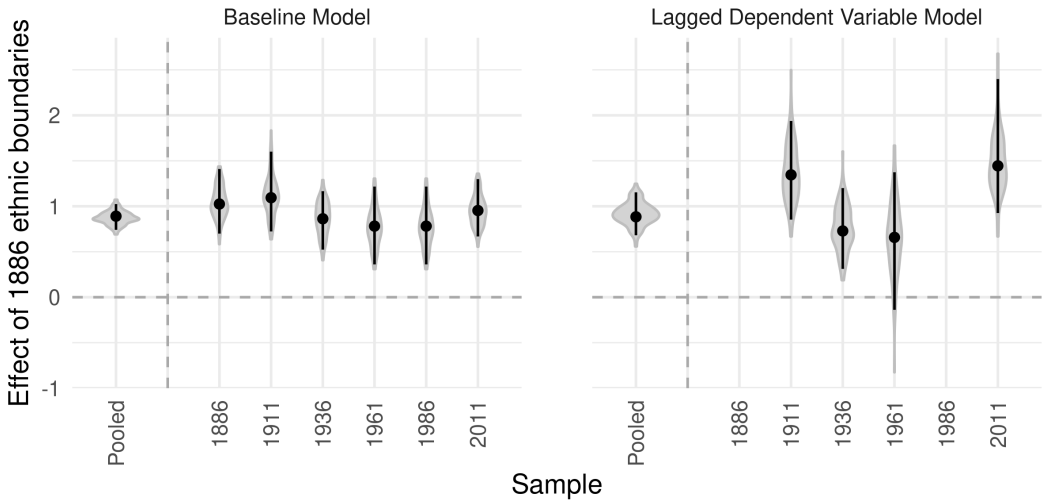


Figure 6: Effect of pre-1886 ethnic boundaries on the partitioning of Europe into states

Note: 95% CIs and grey areas show the distribution of bootstrapped estimates. The lagged dependent variable model cannot be separately estimated for 1986 because of perfect collinearity of 1961 and 1986 borders.

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 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

[28] This is to follow up on findings that countries tend to be east-west oriented (Laitin, Moortgat and Robinson 2012).

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

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 claims and conflicts as an important mechanism through which ethnic geography shapes state borders in the age of nationalism.

Secessionist claims and conflict

We now empirically assess the argument that secessionism drives the border-changing effects of nationalism, a result of the fact that there are more potential ethnic nations than realized states. To do so, we analyze whether ethnically distinct peripheral regions were indeed more likely to experience secessionist claims, conflict, and ultimate secession from their host states since 1946.

Data

Building on the main analysis, we use the vertices of our baseline lattice as our units of analysis.[30] Doing so avoids units that are either spatially misaligned with our (in)dependent variables or defined based on state borders. For each point and year since 1946, we code whether it is (1) claimed by a self-determination movement, (2) fought over in a secessionist ethnic civil war, and (3) affected by a successful secession. Data on secessionist self-determination claims between 1946 and 2012 come from the GeoSDM dataset (Schvitz, Germann and Sambanis 2020, and Appendix C.2). The Ethnic Power Relations data (Vogt et al. 2015) provides information on the settlement regions of ethnic groups associated with secessionist civil wars between 1946-2016. Lastly, we code the secession of points when they become part of a newly independent state in the CShapes 2.0 data (Schvitz et al. 2021).

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

Our argument holds that peripheral groups that do not share the ethnic identity of states' core groups are most likely to seek national self-determination. We capture this logic by measuring whether locations are 'non-coethnic' to their state's capital. More precisely, we assess whether a point j and its capital $C_{j,t}$ are located in settlement regions of different ethnic groups g depicted on ethnic maps $m \in M_{j,C_{j,t},t}$ that were collected in the 50 years prior to t:[31]

$$\text{Non-coethnic capital}_{j,t} = \frac{1}{M_{j,C_{j,t},t}} \sum_{m=1}^{M_{j,C_{j,t},t}} \mathbb{1}_{g_{m,j} \neq g_{m,C_{j,t}}} \tag{7}$$

Empirical strategy

Our main modelling problem consists in the fact that successful secession entails an endogenous change in the 'treatment' assigned to seceding regions. To avoid this type of selection bias, we model the onset of secessionist claims, conflicts, and successful secession using a Cox Proportional Hazard Model:

$$h(\tau)_{j,t} = h_0(\tau)\, exp(\beta_1 \text{ non-coethnic capital}_{j,t} + \gamma\, \mathbf{X}_{j,t} + \epsilon_{j,\tau}) \tag{8}$$

where $h(\tau)_{j,t}$ is the expected risk of seeing the onset of one of the three outcomes in point $j$in calendar year t and relative time τ – the count of years since j became a member of its current state. This counter starts with our data in 1946[32] and is restarted when a point changes its state membership.[33]

Next to our variable of interest non-coethnic capital$_{j,t}$, we add controls $\mathbf{X}_{j,t}$. These follow two logics. The first mirrors the dyadic controls from the main analysis above, capturing the distance (logged), size of largest river and watershed, as well as the mean elevation between a point j and its capital $C_{j,t}$, and the fraction of centuries (1000-1790) in which the two have been located in the same state. The second logic focuses on points j only, reflected in controls for the local population density (logged),[34] the altitude and terrain slope (FAO 2015), as well as each points'

[31] Results are robust to using pre-1886 ethnic data (Appendix E).

[32] This is the starting point of the EPR and GeoSDM data. The end of World War II also marks a critical juncture which arguably restarted the survival 'clock' in much of Europe.

[33] Because observations after such a change may be endogenous, Appendix Table A6 analyzes only periods starting in 1946. This increases the estimated effects.

[34] Time-varying in decadal steps, from Goldewijk, Beusen and Janssen (2010).

distance to the closest border (logged). In combination, these control variables account for the most important joint structural causes of peripheral minority status and secessionist conflict (e.g., Carter, Shaver and Wright 2019).

In a variant of Equation (8), we stratify the model and let the baseline hazard $h_0(t)$ vary between country-years. Similar to country-year fixed effects, this accounts for any time-varying factor at the country level (e.g., the breakup of the USSR) that affects the risk of secession. To account for spatial interdependence of our outcomes, we cluster standard errors on 'stable state segments,' sets of points that were always jointly members of the same states.

Results

Table 2: 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. Full results with control variables are reported in Table A5. Significance codes: *p<0.1; **p<0.05; ***p<0.01

The results of the survival models clearly show that ethnically distinct regions are more likely to experience secessionist claims, conflict, and border change than other regions. Combined with the main results, this suggests that secessions drive the increasing alignment of state territories and ethnic geographies. We find large and statistically significant effects of being ruled from a non-coethnic capital on demands for and realizations of secession. Transforming the coefficients from Table 2 into hazard ratios, such regions have a 6-16 times greater hazard of being claimed by a secessionist movement, a 9-22 times higher risk of being fought over in a se-

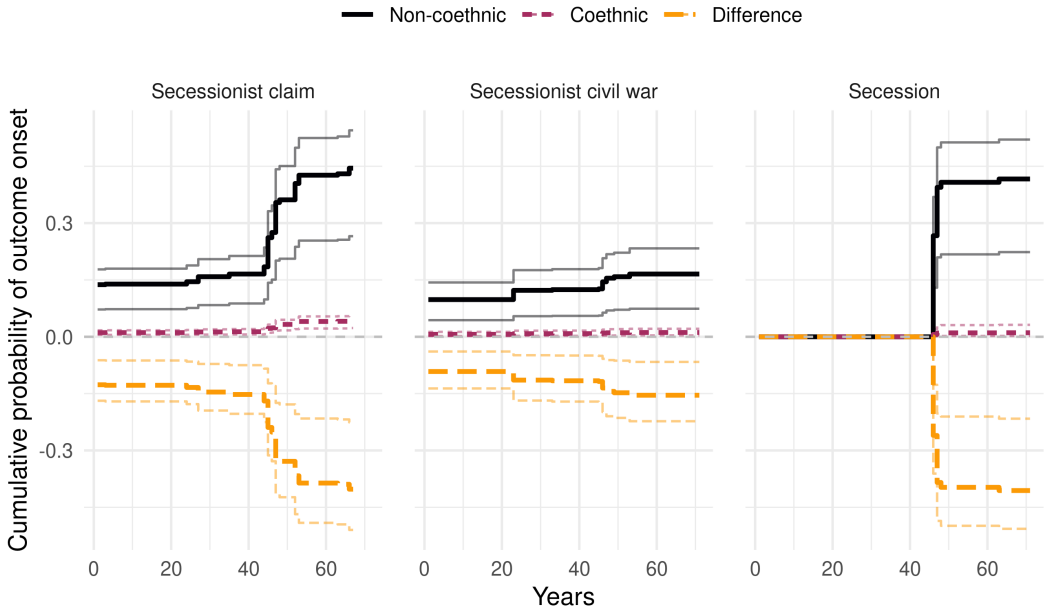


Figure 7: Effect of ethnic boundaries on secessionist claims, conflict, and successful secession.

Note: Predictions with 95% CIs based on Models 1, 3, and 5 in Table 2, setting all covariates to the sample median.

cessionist civil war, and a 18-50 times higher risk of seceding from their state than capitals' co-ethnic regions.

These patterns are reflected in a high probability of ethnically peripheral regions to experience secession and its political and violent antecedents (Figure 7). Over 50 years and holding covariates at their median value, such 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 co-ethnic areas are close to zero. While the break-up of the USSR and Yugoslavia dominate the patterns of secessions, these results show that secessionism of peripheral ethnic groups drives the alignment of state borders with the ethnic map.

Global comparison

Our findings have so far been limited to 19th and 20th century Europe. Do they also generalize and explain borders and border change in other world regions? This section sheds some limited light on this question by comparing the effects of

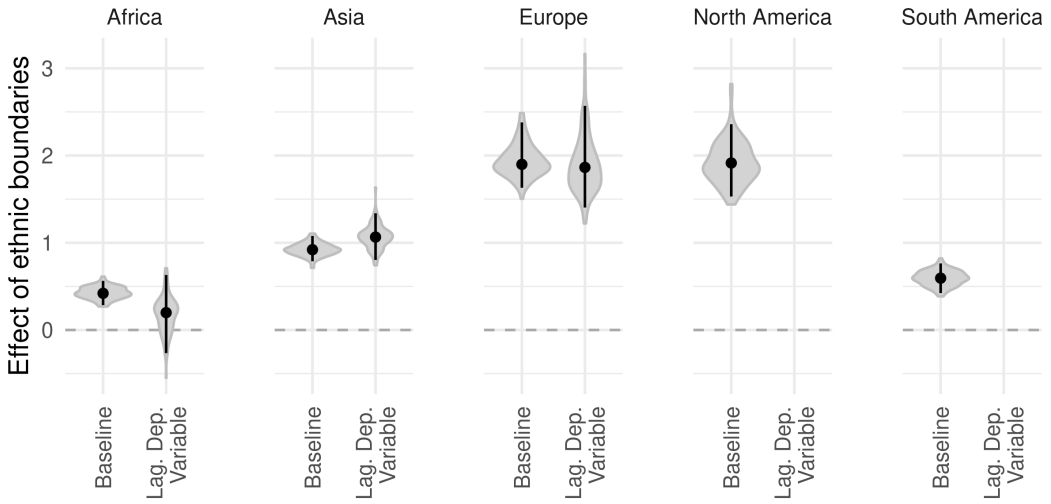


Figure 8: Effect of ethnic boundaries in 1964 on the 2017 partitioning of the five largest continents into states: Baseline and lagged dependent variable models

Note: 95% CIs and grey areas show the distribution of bootstrapped estimates. The Americas did not experience any large-enough border change since 1964, prohibiting the estimation of lagged dependent variable models.

ethnic geography on borders and border change in Africa, Asia, Europe, and the Americas.

For this comparison, we create five lattices of the same spatial structure as above, each covering one continent. We then use our main PSPM specifications[35] to estimate the effect of ethnic boundaries on state borders observed in 2017. We draw on 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 borders observed in 1964. Given the lack of historical state border data with global coverage, we cannot control for the 'deep lag' of borders between 1000 and 1800 AD.[36]

Starting with Africa, the results in Figure 8 support the conventional wisdom that decolonization and the *uti possedetis* norm preserved the haphazard 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 results show that ethnic boundaries did not significantly affect border change since 1964. Turning to Asia,

[35] See Equations 5 and 6.

[36] This omission does not significantly change results for Europe.

the results suggest a more substantive effect of ethnic boundaries. Though 'only' half the size compared to Europe, ethnic boundaries significantly correlate with not only the stable set of borders in 2017 but also border changes since 1964. This result is mostly due by the independence of ethnically distinct Soviet Republics in Central Asia and the southern Caucasus. Lastly, in the Americas we observe a stronger cross-sectional correlation between ethnic and state boundaries in the North than in the South. The absence of border change since 1964 prohibits the estimation of a lagged dependent variable model.

In sum, these results yield two insights. First, state borders are aligned with ethnic boundaries at a global scale, with states in Africa showing the least alignment. Second, ethnic boundaries that cut across state borders seem to effect border change in Asia and Europe but not elsewhere. Ongoing ethno-nationalist conflicts from secessionist Kurdistan to border disputes between India and Pakistan suggest that the ethnic reshaping of Asian states may still be ongoing. In Africa in contrast, outright secessionist conflict is comparatively rare and the territorial integrity norm is generally upheld (Englebert and Hummel 2005; Zacher 2001). This does unfortunately not imply the absence of ethnic conflict that fragments states without changing their borders.

Conclusion

Gellner (1983, 1) famously defined nationalism as "a political principle which holds that the political and national unit should be congruent." We have analyzed whether, by how much, and how this ideal reshaped European states along ethnic boundaries over the past 150 years. In doing so, we contribute to the literature on international borders that has so far lacked systematic empirical analyses of 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, national unification and irredentism. Over time, these processes have gradually increased the fit between state borders and the underlying ethnic map. We test this proposition by new spatial data on ethnic settlement pat-

terns since 1855, developing a new Probabilistic Spatial Partition Model, and estimating the effect of ethnic geography on the partitioning of Europe into state territories.

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 that separates 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 changes since 1886, increasing by 17 percentage points the probability of the presence of a state border conditional on the presence of past state borders. Additional findings also substantiate our claim that secessionist border change is a main driver of the ethnic reshaping of states. We find that peripheral ethnic minorities are at 6 to 50 times 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, driven in particular by secessionism. This has important implications for our understanding of state formation and its effects in the long and intermediate run. For one, state borders and the distribution of ethnic groups within them should not be treated as exogenously given. Quite to the contrary, the number of states, their territorial shape and ethnic makeup are in large parts the result of nationalist struggles for ethnic self-determination. This should also be kept in mind when comparing ethnically homogeneous European 'nation-states' with their mostly multi-ethnic counterparts elsewhere on the globe.

Moreover, our results suggest that the ethnic alignment of state borders is an ongoing process. Most notably, we estimate the largest effect of ethnic geography on border change for the period after 1986, which featured the collapse of the USSR and Yugoslavia. 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 in the wake of Brexit only underscore that nationalist struggles to redraw border remain central to contemporary politics in multi-ethnic states in Europe.

Looking beyond Europe, we have found similar dynamics of ethno-nationalist

border changes in contemporary Asia but not elsewhere. Mostly driven by secessions of former Soviet Republics, the available data have insufficient historical depth to draw firm conclusions on whether the region (and any other continent) follows a macro-historical trajectory similar to the European one or not.

The answer to this question will foreshadow the future of many multi-ethnic states. Although our analysis of the European context is primarily structuralist, we caution against deterministic extrapolations of the continent's history. While we deem it important to recognize the *potential* of ethnic centrifugal forces, previous research suggests that peacefully containing them within given state borders is possible. Addressing the ideological and material foundations of ethnic nationalism may, for example, require reducing injustices through ethnic power-sharing and regional accommodation (Cederman, Gleditsch and Buhaug 2013). More radically, dissociating states from nations (Mamdani 2020) may succeed in depoliticizing and bridging ethnic divides. On the international stage, enforcing the territorial integrity norm (Zacher 2001) may furthermore rein in nationalist revisionism. Alarmingly, however, at this very moment in history, such progress is endangered by nationalist forces in influential states such as the United States, Russia, India, or China.

References

Abramson, Scott and David B Carter. 2020. "Systemic Instability and the Emergence of Border Disputes." *International Organization, forthcoming* .

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*. Mit Press.

Alesina, Alberto, William Easterly and Janina Matuszeski. 2011. "Artificial States." *Journal of the European Economic Association* 9(2):246–277.

Alter, Peter. 1989. *Nationalism*. London: Edward Arnold.

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.

Breuilly, John and Ronald Speirs. 2005. *The Concept of National Unification*. London: Palgrave Macmillan.

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 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.

Carter, David B and Hein E Goemans. 2018. "International trade and coordination: Tracing border effects." *World Politics* 70(1):1–52.

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

Cranmer, Skyler J, Bruce A Desmarais and Elizabeth J Menninga. 2012. "Complex dependencies in the alliance network." *Conflict Management and Peace Science* 29(3):279–313.

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. 2020. "Political Exclusion, Lost Autonomy, and Escalating Conflict over Self-Determination." *International Organization, Forthcoming* .

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

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

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. 2010. "Security Threats, Linguistic Homogeneity, and the Necessary Conditions for Political Unification." *Nations and Nationalism* 16:169–188.

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 Europ.

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.

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

Hiers, Wesley and Andreas Wimmer. 2013. *Is Nationalism the Cause or Consequence of the End of Empire?* Cambridge: Cambridge University Press.

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

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

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

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

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

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

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

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.

Lazer, David, Brian Rubineau, Carol Chetkovich, Nancy Katz and Michael Neblo. 2010. "The coevolution of networks and political attitudes." *Political communication* 27(3):248–274.

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*. 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.

Møller, Jørgen. 2014. "Why Europe avoided hegemony: a historical perspective on the balance of power." *International Studies Quarterly* 58(4):660–670.

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

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.

Schieder, Theodor. 1964. *Der Nationalstaat in Europa als historisches Phänomen*. Wiesbaden: VS Verlag.

Schvitz, G, S Rüegger, L Girardin, L-E Cederman, N Weidmann and KS Gleditsch. 2021. "Mapping The International System, 1886-2017: The Cshapes 2.0 Dataset." *Journal of Conflict Resolution* .
URL: *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3610680*

Schvitz, Guy, Micha Germann and Nicholas Sambanis. 2020. "Mapping Self-Determination Claims 1946-2012: The GeoSDM Dataset.".

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

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

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

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

Tilly, Charles. 1990. *Coercion, Capital, and European States, AD 990-1992*. Cambridge, Massachusetts: Blackwell Publishing.

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* 55:215–250.

Supplementary Material

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

Table of Contents

A Probabilistic Spatial Partition Model A2
A.1 A distribution over partitionings A2
A.2 Reduction to simple logistic relationship for bridge edges A3
A.3 Sampling from the model A4
A.4 Estimation by Composite Likelihood A6
A.5 Standard errors. A7
B Model Evaluation: Monte Carlo Simulations A7
B.1 Simulation setup A8
B.2 Results A9
C Data A13
C.1 Historical ethnic map collection A13
C.2 Data on self-determination claims: GeoSDM A24
D Robustness checks: Probabilistic Spatial Partition Model A25
D.1 Varying control variables: A25
D.2 Varying the temporal structure of the data: A28
D.3 Varying the spatial lattice: A29
D.4 Burn-in rate in parametric bootstrap A34
D.5 Logistic regression with edge-level data A34
E Robustness checks: Analysis of secessionist claims and conflict . . . A35
F References (Appendix) A41

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 Reduction to simple logistic relationship for bridge edges

The partition model reduces to a logistic model for every edge between nodes u and v for which $s_{u,v}$ is independent of the remainder of the network. This is the case for any edge that can switch the state(s) of its nodes without violating transitivity and contiguity assumptions. These edges are in particular *bridge edges*, connecting two otherwise disjoint parts of the lattice.[37]

Assume such an edge connects vertex v with its neighbor u on the graph G. In relation to u, v can only take on two possible outcomes, $s_{u,v} \in \{0, 1\}$: v can either be in the same partition as u or form its own partition. It cannot take on another outcome, such as being in the same partition as any neighbor w of u. We show in the following that the probabilities of the two outcomes $s_{u,v} \in \{0, 1\}$ are therefore independent of the overall network structure and can be derived directly from the estimated parameters of the PSPM.

Holding constant the partitioning of all nodes $I \neq v$ in G, define p_1 as the partitioning where $s_{u,v} = 0$, i.e. where u and v are part of two different partitions,

[37] Note that the logic outlined below is the same as that explained by Cranmer and Desmarais (2011, p. 73) who note that the results of an ERGM can be interpreted as a logistic regression model if edges are independent.

and p_2 as the partitioning where $s_{u,v} = 1$. Following (A3), the total energy of each partitioning is defined as

$$\epsilon_p = \sum_{j,k \in L} \epsilon_{j,k} * s_{j=k}, \tag{A5}$$

Since p_1 and p_2 only differ in $s_{u,v}$, the difference between the two partitionings' total energy amounts to the potential energy of the edge between u and v:

$$\epsilon_{p_2} - \epsilon_{p_1} = \epsilon_{u,v} \tag{A6}$$

Because the normalizing sum Z from (A1) is the same for the probabilities of p_1 and p_2, we can derive the odds of p_1 vs. p_2 as

$$\begin{aligned} \frac{P(P = p_1)}{P(P = p_2)} =& \frac{e^{-\epsilon_{p_1}}}{e^{-\epsilon_{p_2}}} \\ =& e^{-\epsilon_{p_1} + \epsilon_{p_2}} \\ =& e^{\epsilon_{u,v}} \end{aligned} \tag{A7}$$

As can be seen, the odds described by (A7) are the same for every possible partitioning of nodes $I \neq v$ in G and only depend on $\epsilon_{u,v}$ and not on the outcome $s_{i,j}$ of any other edge in the network. We can therefore generalize the relation to describe the overall odds of v being in a different partition than u ($s_{u,v} = 0$) or not ($s_{u,v} = 1$):

$$\frac{P(s_{u,v} = 0)}{P(s_{u,v} = 1)} = \frac{P(s_{u,v} = 0)}{1 - P(s_{u,v} = 0)} = e^{\epsilon_{u,v}} \tag{A8}$$

Finally, this relation can be reformulated as a linear relationship between the energy $\epsilon_{u,v}$ and the log-odds:

$$ln\left(\frac{P(s_{u,v} = 0)}{1 - P(s_{u,v} = 0)}\right) = \epsilon_{u,v} \tag{A9}$$

This allows us, for bridge edges, to interpret the parameters that describe the relationship between edge level covariates $\mathbf{x}$ and edges' energy in the same manner as coefficients derived from a logistic regression, including the derivation of odds ratios, predicted probabilities, and marginal effects.

A.3 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. The problem is that the number of possible

partitionings, which we would have to iterate over, grows exponentially.[38] For instance, the number of possible contiguous partitionings of a 3x3 grid lattice is 1434; for a 10x10 grid lattice this number is of the order 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

$$Pr(\mathbf{S} = \mathbf{s}) = \begin{cases} Z^{-1}\ exp\left(-\sum_{j,k \in L} \epsilon_{j,k} * s_{j,k}\right) & \text{if } \mathbf{s} \in \mathbb{P} \\ 0 & \text{otherwise,} \end{cases} \tag{A10}$$

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

We can sample from (A10) 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

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

memberships between all other nodes. The conditional distribution is given by

$$Pr(\mathbf{S}_j = \mathbf{s}_j | \mathbf{S}_{-j} = \mathbf{s}_{-j}) = \frac{Pr(\mathbf{S} = \mathbf{s})}{\sum_{\mathbf{s}'_j \in \mathbb{S}_j} Pr(\mathbf{S}_j = \mathbf{s}'_j | \mathbf{S}_{-j} = \mathbf{s}_{-j})}$$
$$= \begin{cases} \frac{exp\left(-\sum_{j,k \in N_j} \epsilon_{j,k} * s_{j,k}\right)}{\sum_{\mathbf{s}'_j \in \mathbb{S}_j} exp\left(-\sum_{j,k \in N_j} \epsilon_{j,k} * s'_{j,k}\right)} & \text{if } \mathbf{s} \in \mathbb{P} \\ 0 & \text{otherwise,} \end{cases} \tag{A11}$$

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 (A11) 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 (A11) can be computed rapidly.

A.4 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

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 (A11) and estimate β by maximizing the following log composite likelihood,

$$ln\ \widehat{\mathcal{L}}_C = \sum_{j=1}^{N} \ln\ Pr(\mathbf{S}_j = \mathbf{s}_j | \mathbf{S}_{-j} = \mathbf{s}_{-j}). \tag{A12}$$

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 (A12) with a penalization parameter σ that nudges our estimate towards 0,[39] thus obtaining our parameter estimates from

$$\widehat{\beta} = \arg\max_{\beta} ln\, \widehat{\mathcal{L}_C}(\beta\,;\, p, \mathbf{X}) - \frac{\beta^2}{2\sigma} \tag{A13}$$

A.5 Standard errors

Because we estimate β by maximizing the (intentionally misspecified) composite likelihood (A12), 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:

1. Obtain B partitioning samples from the fitted model using the Gibbs sampling approach described in Section A.3. For each sample, we start 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.[40]

2. Refit the model to each of the B partitioning samples, obtaining B parameter vectors. $\widehat{\beta}^B$.

3. Obtain confidence interval estimates for each scalar parameter β_k by computing the empirical quantiles over the B β_k^B samples. See Section B.2 for simulation results showing that this approach yields confidence interval estimates with unbiased coverage.

B Model Evaluation: Monte Carlo Simulations

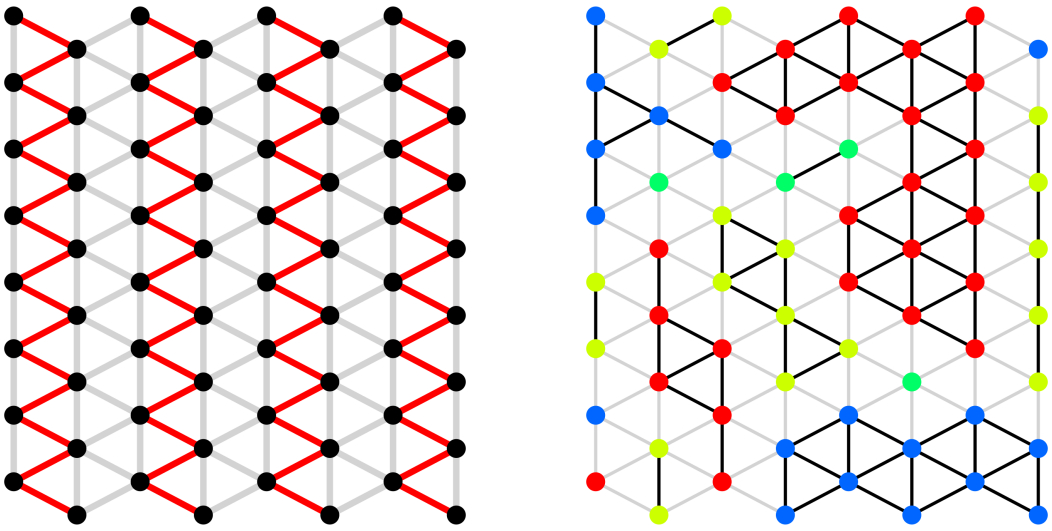
We conduct Monte Carlo experiments to test the performance of our model and the Maximum Composite Likelihood estimator estimator. The main experiments explore potential biases in estimates recovered by the estimator and investigate the

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

[40] See Section B.2 for an empirical evaluation of how the burn-in rate affects the parameter estimates.

precision of uncertainty estimates while varying the (1) burn in rate of our sampler, (2) the size of networks, and (3) the number of independent instances. Biases stabilize after a relatively short burn in period and decrease with the size and number of networks. Biases are mainly concentrated in areas with separation issues. Standard errors derived from the Hessian of the Maximum Composite Likelihood estimator are consistent in most cases. Parametric bootstrapping offers an alternative method to derive uncertainty estimates.

B.1 Simulation setup



(a) Predictor. 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.[41]

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

We use the Gibbs sampler described in A.3 to sample the partitioning of G based on the following edge-level energy function:

$$\epsilon_{j,k} = \beta_0 + \beta_1 \, x, \tag{A14}$$

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 unvertainty estimates derived from a parametric bootstrap. Table A1 summarizes 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 the structure of 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

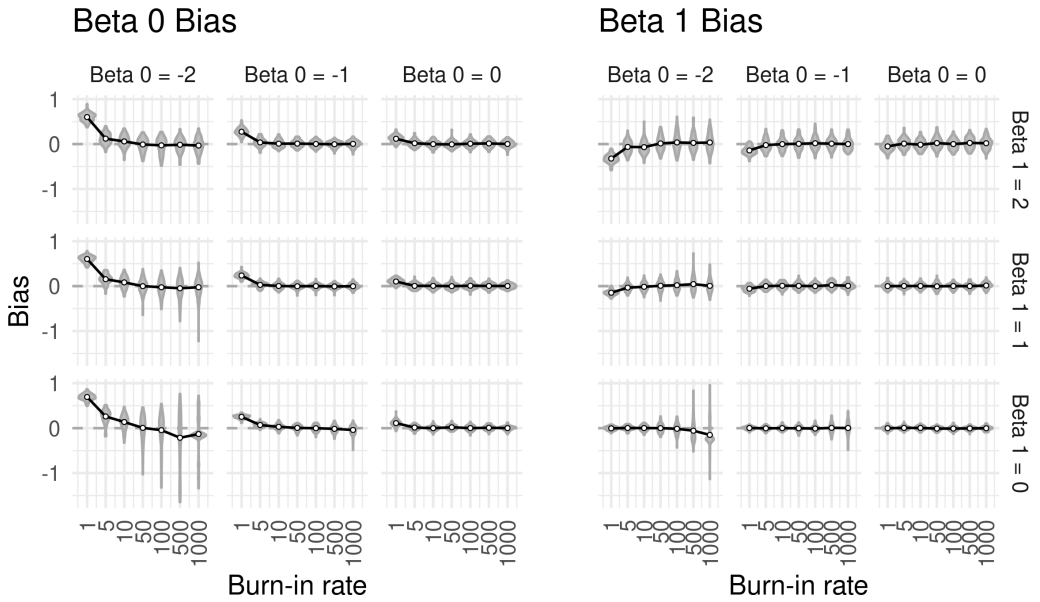


Figure A2: Bias in parameter estimates and the burn-in rate.

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.

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: In the next set of experiments, we examine whether biases in the estimates produced by the Maximum Composite Likelihood estimator decrease as we increase the size of networks. This is a necessary test as the estimators' consistency is only ensured in the number of independent graphs G, not in the graph size N (see Section A.4 above; Varin, Reid and Firth 2011).

The results from increasing the size of our experimental graphs in exponential steps from $N = 16$ to $N = 4096$ show that, for the present application, 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 this high baseline attraction between nodes, we need very large networks to obtain unbiased estimates.

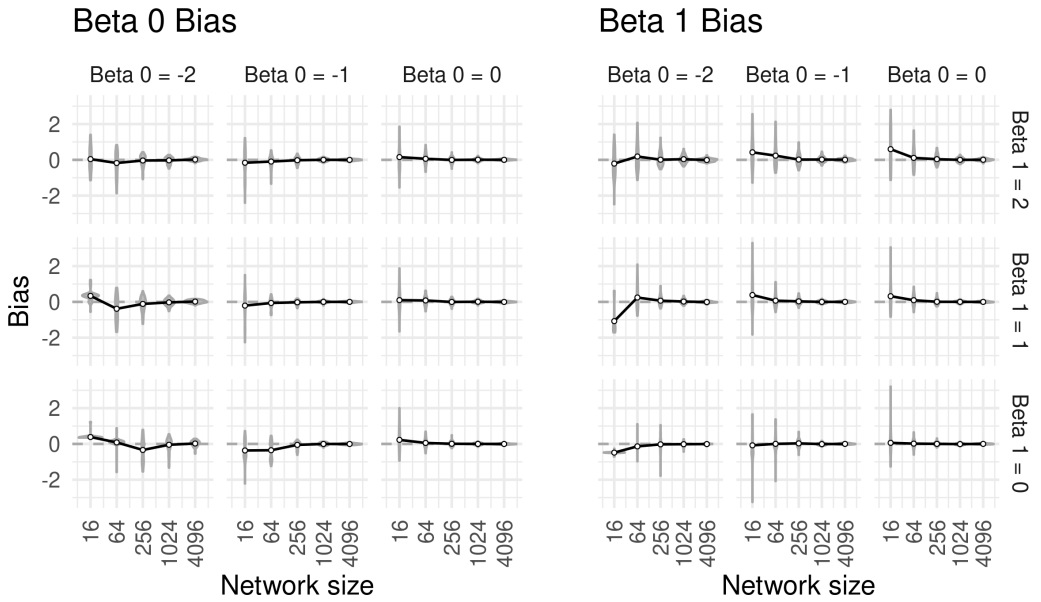


Figure A3: Bias in parameter estimates and the size of spatial lattices.

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

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

4. Parametrically bootstrapped confidence intervals: Lastly, we test he consistency of our procedure for obtaining standard error described above in Section A.5. 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, close to and statistically indistinguishable from 95% of our bootstrapped confidence intervals contain the real value of β. Confidence intervals are slightly overconfident (i.e. too small)

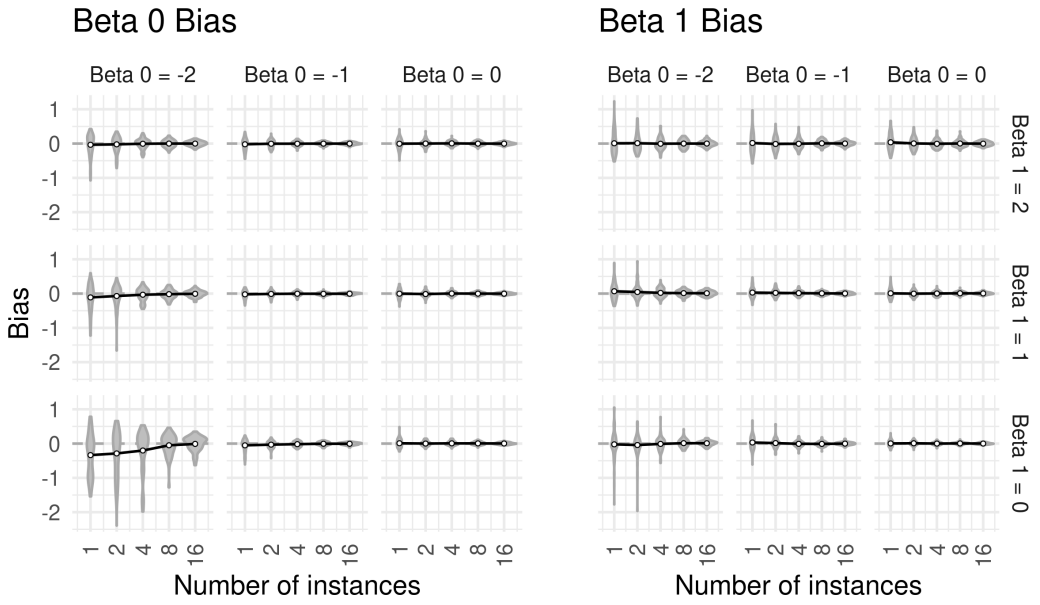


Figure A4: Bias in parameter estimates and the number of independent of spatial lattice instances.

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

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 cases biases and 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 presents a practicable way to derive generally consistent confidence intervals.

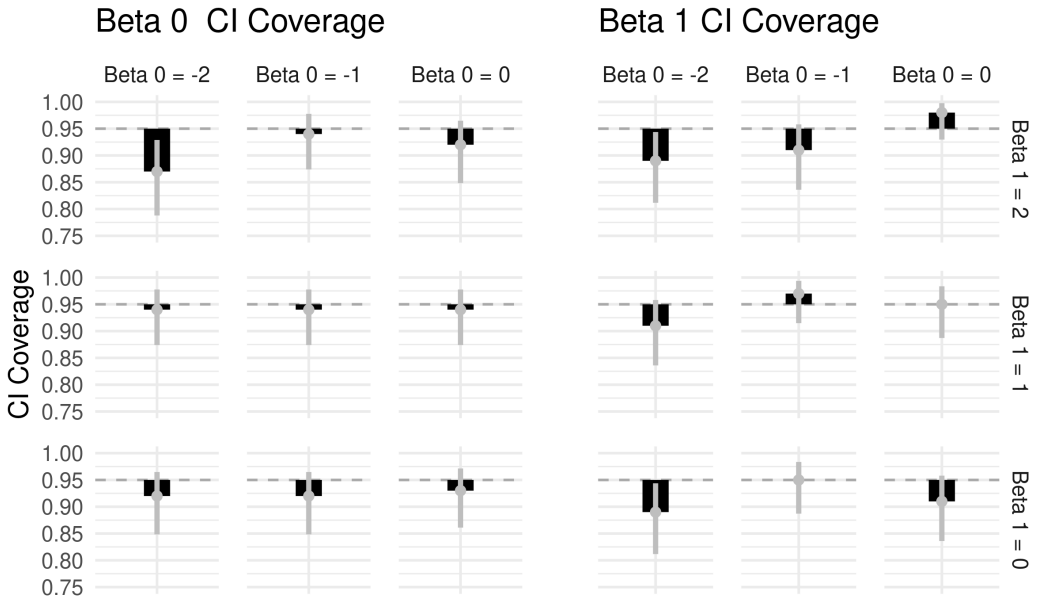


Figure A5: Precision of confidence interval coverage: Standard errors and 95% confidence intervals derived from a parametric bootstrap (Section A.5.

Note: Grey bars denote the 95% confidence interval of the CI coverage estimates. Resulting from Monte Carlo simulations with the following parameters: 100 iterations; 1024 nodes on a hexagonal lattice; 1 instance; burn-in rate of 100; β_0 and β_1 as shown in graph.

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,[42] from which we selected 73 maps that we considered the most suitable. Among our suitability criteria were:

1. Maps must depict ethnic settlement areas (as opposed to general maps of race or religion, or census maps of a group's population share within administrative units).

2. Maps should depict a snapshot in time close to the year they were published (as opposed to maps of the distribution of peoples in ancient history).

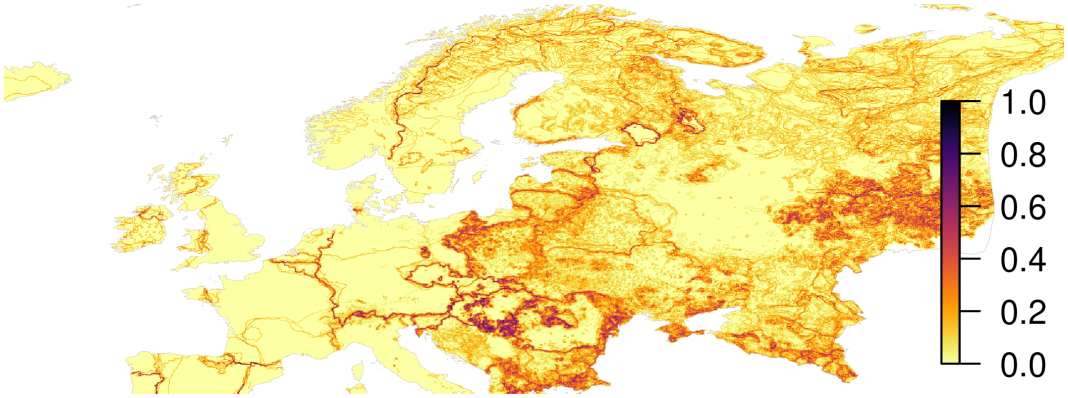
3. Sufficient level of detail and precision

4. No obvious signs of political bias

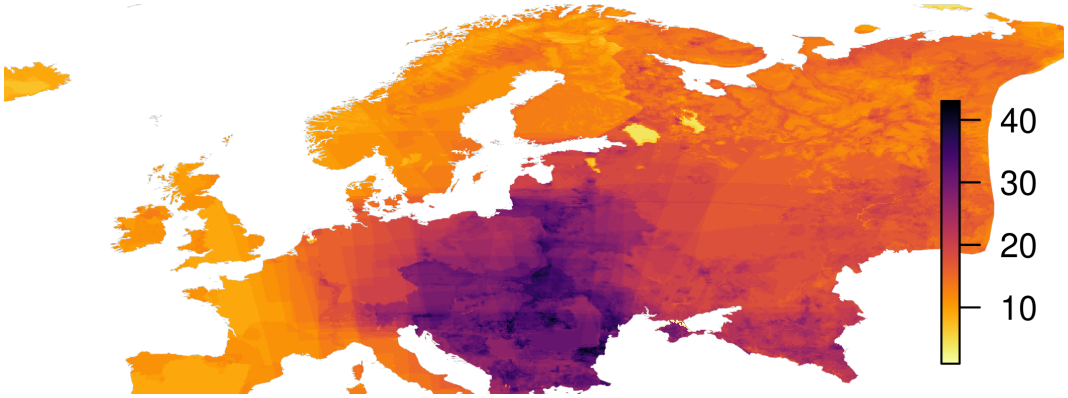
[42]This count is approximate since we digitized many maps on the basis of Library Catalogue entries which ended up not being maps of ethnic groups in the first place.

5. Maps cannot be duplicates of other maps (some maps were just slightly altered versions of other maps, published in different outlets)

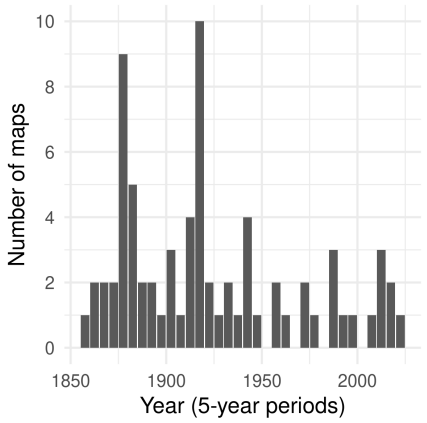
Figure A6 summarize the ethnic boundaries retrieved from the data (a), the spatial coverage of overlapping maps (b), their temporal distribution (c), as well as the correlation of the final edge-level measure of ethnic boundaries over the main time periods in our analysis (d). Table A2 lists all 73 maps that we have so far used as source material, along with the relevant metadata. Section C.1.1 shows three examples of maps that were part of our final selection. In Section C.1.2, we give three examples of maps that did not meet our criteria. Figure A10b shows a map that was excluded because its depiction of ethnic settlement areas is too coarse. As a clear example of political bias, Figure A12 shows a map of Lithuanian settlement areas published by the Lithuanian National Committee in 1918. This map defines a much larger Lithuanian settlement area than any other map published between 1863 and 1963, and seems to provide justifications of its territorial claims based on historical kingdoms. Lastly, Figure A11 shows a census map of Germans in present-day Poland in 1863. Although the map roughly depicts the settlement areas of Germans, it is too coarse as it aggregates this information up to the level of administrative units.



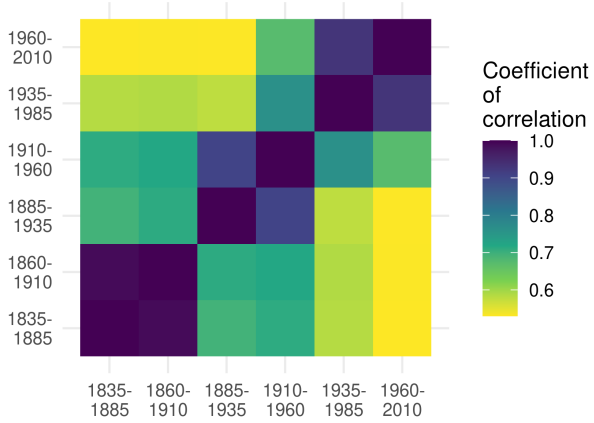
(a) Ethnic boundaries as fraction of maps covering an area, 1836–2010.



(b) Count of maps per area, 1836–2010



(c) Number of maps over time



(d) Correlation of ethnic boundary across periods t

Figure A6: Historical ethnic data: Summary

file_name	title	author	map_year	scale	quality	source
brl_rob_22	Ethnographische Karte der Östereichischen Monarchie	Czoernig, Karl Freiherr von	1855	1:864,000	5	archive
bnf_rob_41	Carte Ethnographique de la Turquie d'Europe et des États Vassaux Autonomes	Lejean, Guillaume Marie	1861	1:2,500,000	4	online
rum_rob_37	Tableau Ethnographique	Erckert, Roderich von	1863	1:5,500,000	5	online
brl_rob_32	Völker und Sprachenkarte von Deutschland und den Nachbarländern	D. Reimer	1867	1:3,000,000	5	archive
yun_rob_18	Völker- und Sprachen-Karte von Österreich und den Unter-Donau-Ländern	Kiepert, Heinrich von	1869	1:3,000,000	5	online
bnf_rob_37	Specialkarte der deutsch-französischen Grenzländer mit Angabe der Sprachgrenze (neue berichtigte Ausgabe)	Kiepert, Heinrich von	1870	1:666,666	3	online
yun_rob_8	Europe Ethnographic	Unknown (Russian author)	1870	1:10,500,000	4	online
wik_aya_1	Ethnic Map of European Russia	Rittikh, Aleksandr Fedorovich	1875	1:2,520,000	5	online
uch_cam_9	Die Neueste Eintheilung, die Türkischen Gebiete & die Confessionen in der Türkei	Petermann, August, Habenicht, Hermann	1876	1:2,500,000	3	online
yun_rob_15	Ethnographische Übersicht des Europäischen Orients	Kiepert, Heinrich von	1876	1:3,000,000	4	online
brl_rob_16	Ethnographische Karte der Europäischen Türkei	Carl Sax	1877		5	archive
uch_cam_10	Deutsche & Romanen in Süd-Tirol & Venetien	Petermann, August	1877	1:740,000	4	online
rum_rob_8	Ethnographische Karte von Russland (Nördliches Blatt)	Rittikh, Aleksandr Fedorovich	1878	1:370,000	5	online
rum_rob_18	Ethnographische Karte von Russland (Südliches Blatt)	Rittikh, Aleksandr Fedorovich	1878	1:370,000	5	online
uch_cam_12	Vertheilung der Gross-, Weiss- & Klein-Russen	Petermann, August	1878	1:370,000	3	online
uch_cam_13	Etnograficheskaia Karta Kavkazskago Kraia	Rittikh, Aleksandr Fedorovich	1878	1:1,080,000	4	online
brl_aya_23	Sprachen-Karte von Österreich-Ungarn	Franz Ritter v. Le Monnier	1880	1:1,000,000	5	archive
rum_rob_11	Europa um 1880	Berghaus, Heinrich	1880	1:15,000,000	3	online
yun_rob_17	Sprachen-Karte der westlichen Kronländer von Oesterreich	Held, F.	1880	1:1,500,000	4	online
rum_rob_24	Sprachenkarte, Religionskarte Schweiz	Andree, Richard	1881	1:1,480,000	5	online
rum_rob_36	Völkerkarte von Russland.	Andree, Richard	1881	1:13,300,000	5	online
bnf_rob_16	Die Polen in Deutschland: Nordöstliches Deutschland nebst Polen. Ethnographische Karte	Geographisches Institut Weimar	1885		4	online

file_name	title	author	map_year	scale	quality	source
brl_rob_20	Politisch-Ethnographische Übersichtskarte von Bulgarien, Ost-Rumelien	Geographisches Institut Weimar	1885	1:3,000,000	4	archive
emr_rob_2	Ethnographic map of Austria-Hungary and Romania	Kiepert, Heinrich von	1892	1:3,000,000	4	online
brl_rob_30	Völker- und Sprachenkarte von Mitteleuropa	Karl Peucher	1893	1:6,000,000	4	archive
rum_rob_30	Deutsches Reich. Religionskarte. Völkerkarte	Diercke, Carl	1896		5	online
rum_rob_13	Ethnographic map of Austria-Hungary	Andree, Richard	1900	1:4,000,000	4	online
rum_rob_14	Ethnographic map of the Balkan Peninsula.	Andree, Richard	1900	1:6,000,000	4	online
rum_rob_15	Völker u. Sprachenkarten. Europa. Konfessionskarten.	Wagner, Hermann	1902	1:40,000,000	4	online
emr_rob_6	Völkerkarte des rumänischen Sprachgebietes	Weigand, Gustav	1909	1:340,000	4	online
brl_aya_33	Die Sprachgebiete der Schweiz unter besonderer Berücksichtigung der Hohenregionen, nach Walser	Deutsches Ausland-Institut, Isbert, O.A., Strotha, M.K.v	1910	1:300,000	5	archive
loc_sim_6	Map of Eastern Turkey in Asia, Syria and Western Persia (Ethnographical)	Royal Geographical Society	1910	1:2,000,000	5	online
pol_rob_11	Sprach- und Schulkarte Mähren und Schlesien	Perko, Franz, Perko, Otto	1910	1:375,000	5	online
brl_aya_30	Das Bulgarentum auf der Balkanhalbinsel im Jahre 1912	Ishirkov, A.	1912	1:1,500,000	5	archive
rum_rob_5	Völker- und Sprachenkarte Österreich-Ungarn	Mayer, Rudolf	1914	1:2,730,000	4	online
brl_rob_48	Ethnographische Übersichtskarte von Osteuropa	Freytag, G.	1916	1:10,000,000	4	archive
bnf_rob_4	Carte Ethnographique de l'Europe Centrale et des États Balkaniques.	Bolzé, R., Chesneau, M.	1918	1:3,500,000	4	online
brl_rob_9	Germany. Ethnographical map, Poland. Ethnographical map, Northern Italy. Ethnographical map, South East Europe. Ethnographical map	Great Britain. General Staff. Geographical Section	1918	1:5,000,000	5	archive
brl_rob_10	The Daily Telegraph. Language map of Eastern Europe	Gross, Alexander	1918	1:2,200,000	5	archive
brl_rob_46	G. Freytags Völker und Sprachenkarte von Mittleuropa nebst Italien und der Balkanhalbinsel	Freytag, G.	1918	1:3,000,000	5	archive
loc_sim_3	Carte Ethnographique de la Péninsule des Balkans	Cvijić, Jovan	1918	1:3,000,000	5	online
loc_sim_4	Ethnographic map of the Balkan Peninsula	Cvijić, Jovan	1918	1:3,000,000	5	online
brl_aya_25	The Question of Thrace. Greeks, Bulgars and Turks	Mills, J.S., Chrussachi, M.G.	1919		5	archive

file_name	title	author	map_year	scale	quality	source
brl_rob_21	Carte Ethnographique des Régions Habitées par les Roumains et des Colonies Étrangeres Qui s'y Trouvent	Demetresco, Atanasiu, Borcea	1919	1:1,000,000	5	archive
brl_rob_19	Carte Ethnographique de l'Albanie	Délegation de la Colonie Albaise de Turquie	1920	1:1,000,000	5	archive
loc_sim_1	Völker und Staaten in Mitteleuropa	Winkler, Wilhelm	1924	1:4,000,000	4	online
brl_rob_17	Carte ethnographique de l'Empire Ottoman. Faute de données statistiques exactes, depuis la Guerre balkanique [...].	Unknown (French author)	1925	1:1,000,000	5	archive
brl_rob_57	Völkerkarte der Sowjet-Union	Klante, M. (Reichsamt für Landesaufnahme)	1926	1:5,000,000	5	archive
brl_rob_51	Volksbodenkarte der Slowakei	Isbert, O.A.	1930	1:750,000	5	archive
cic_rob_3	Carte ethnographique et linguistique de l'Europe nouvelle	Wehrli, Max	1933	1:10,000,000	4	online
brl_rob_56	Rumänien. Volksgruppen	Generalstab des Heeres, Abteilung für Kriegskarten u. Vermessungswesen	1940	1:1,000,000	5	archive
loc_sim_7	Die Völker des Donauraumes und der Balkanhalbinsel	Generalstab des Heeres, Abteilung für Kriegskarten u. Vermessungswesen	1940	1:3,000,000	5	online
brl_aya_16	Albanian Minority in Yugoslavia	Great Britain. Foreign Office. Research Department.	1941		4	archive
brl_aya_32	Völkerkarte des Kaukasus. Aufgrund der vom Bataillon der Waffen-SS z. b. v. sichergestellten 'Ethnographischen Karte des Kaukasus.'	Kommission für das Studium der Völker der UdSSR und ihrer Nachbarländer, Reichsamt für Landesaufnahme.	1942	1:1,000,000	5	archive
nau_rob_6	Poland language map	United States. Office of Strategic Services. Research and Analysis Branch	1945		5	online
brl_rob_38	Karta Narodov SSSR. Uchebia dlia Spednei Shkoly.	Unknown (Russian author)	1955	1:5,000,000	5	archive
brl_aya_3	Ethnic Map of the Soviet Union	Main Directorate of Geodesy and Cartography, Ministry of Geology and Mineral Resources of the USSR	1959	1:5,000,000	5	archive

file_name	title	author	map_year	scale	quality	source
grg_guy_1	Atlas Narodov Mira / Geo-referencing of Ethnic Groups	Bruk, S.I., Apenchenko, V.S., Digitized by Weidmann et al. (2010)	1964	1:5,000,000	4	online
brl_aya_10	Map of People of the USSR	Main Directorate of Geodesy and Cartography, Ministry of Geology and Mineral Resources of the USSR	1972	1:5,000,000	5	archive
pcl_aya_3	Cyprus, Ethnic Distribution	U.S. Central Intelligence Agency	1973		2	online
eth_aya_21	Völker und Sprachen Europas unter besonderer Berücksichtigung der Volksgruppen	Straka, Manfred	1978	1:6,000,000	5	online
bav_nic_1	Ethnic Groups in Southern Soviet Union and Neighboring Middle Eastern Countries	U.S. Central Intelligence Agency	1986		3	online
sdl_aya_1	Map of Slovenian Dialects	Logar, Tine, Rigler, Jakob	1986		4	online
brl_aya_6	Ethnic map of the Soviet Union	Main Directorate of Geodesy and Cartography, Ministry of Geology and Mineral Resources of the USSR	1988	1:4,000,000	5	archive
eth_aya_1	Herrien Europa. Europa de Los Pueblos. L'Europe de Peuple. Europe of the People	Herreros Agüi, Sebastián, Durán Rodriguez, Adolfo	1992	1:6,000,000	5	online
pcl_aya_8	Ethnolinguistic Groups in the Caucasus Region	U.S. Central Intelligence Agency	1995	1:6,750,000	2	online
col_aya_16	Ethnic Ukrainians and Russians in the Caspian-Black Sea Basin	Izady, M.	1997		4	online
col_aya_1	Ethnolinguistic Groups in the Caucasus and Vicinity	Izady, M.	1999		4	online
col_aya_17	Languages of North Africa	Izady, M.	2003		4	online
col_aya_10	Middle East: Ethnic Groups	Izady, M.	2006		4	online
col_aya_9	The Levant: Ethnic Composition	Izady, M.	2008		5	online
enl_guy_1	Ethnologue / World Language Mapping System. Language Maps. Version 17	SIL International	2014		4	online
dev_guy_1	Languages of Europe	Unknown	2017		5	online

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

C.1.1 Examples of maps used as source material

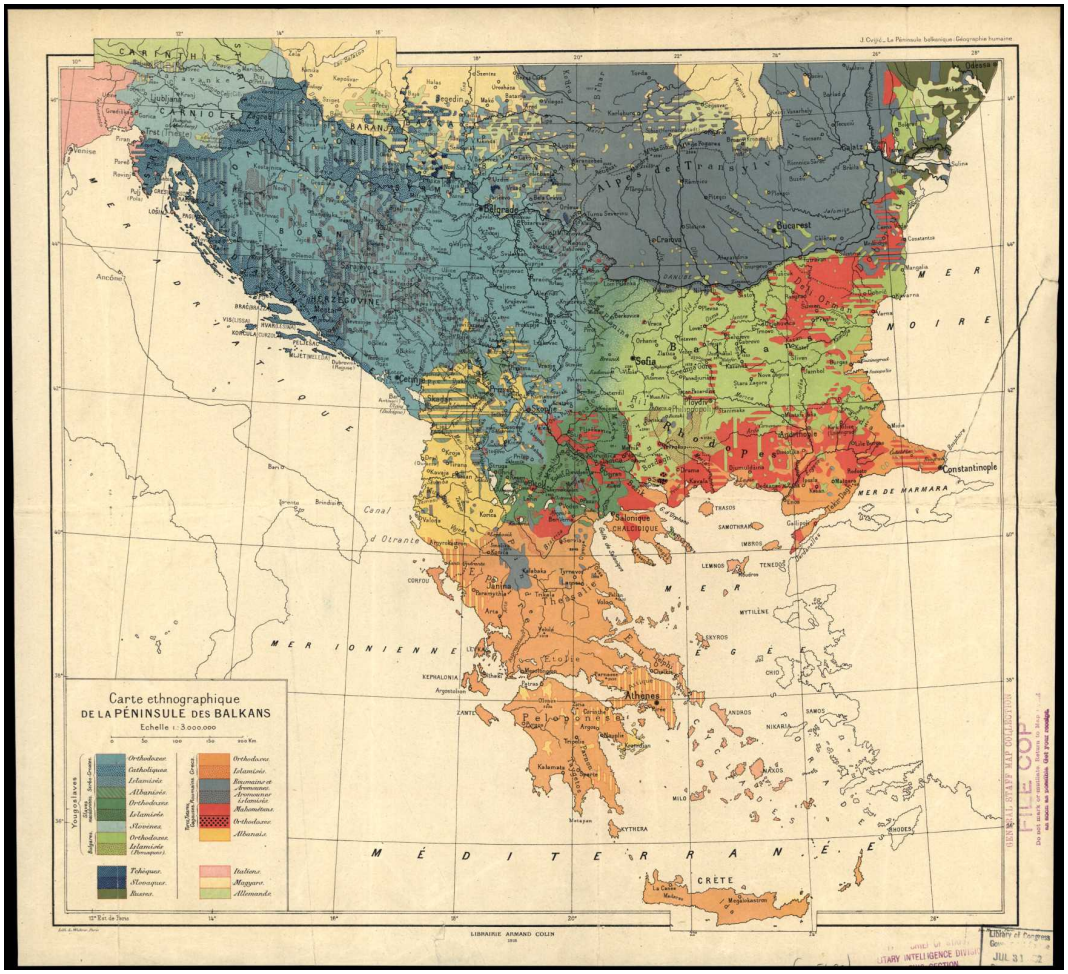


Figure A7: Cvijic, J. (1918). Carte ethnographique de la Péninsule des Balkans.

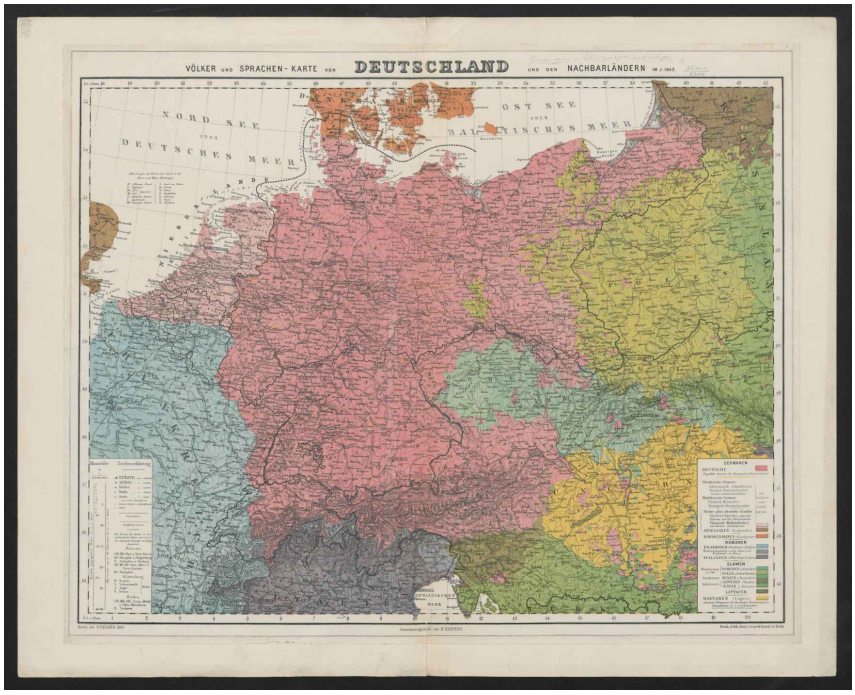


Figure A8: Reimer, D. (1867). Völker und Sprachenkarte von Deutschland und den Nachbarländern

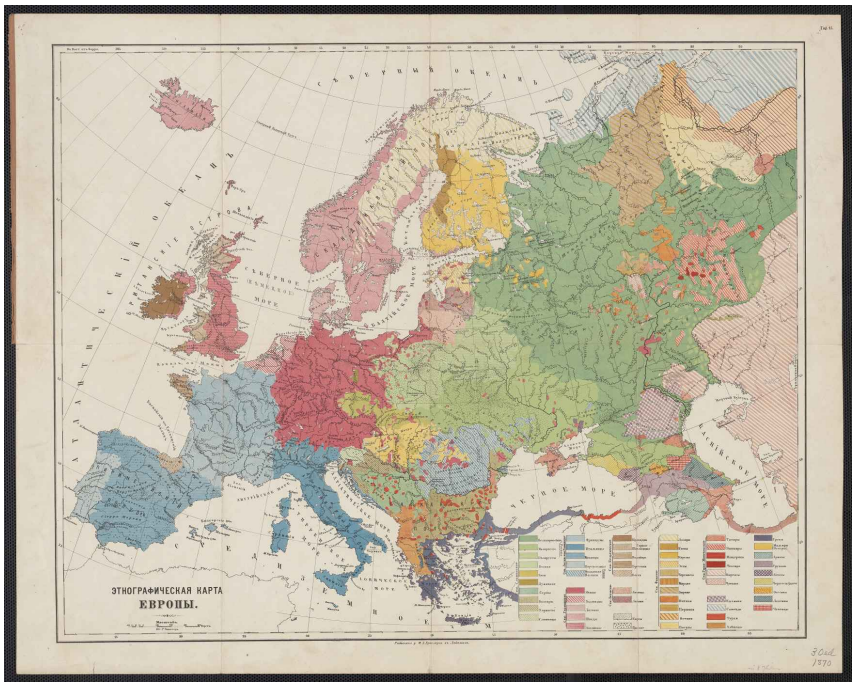
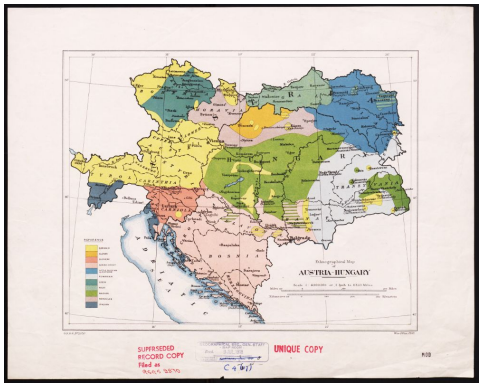


Figure A9: Unknown, Russian author (1870). Ethnographic map of Europe

C.1.2 Examples of unsuitable maps

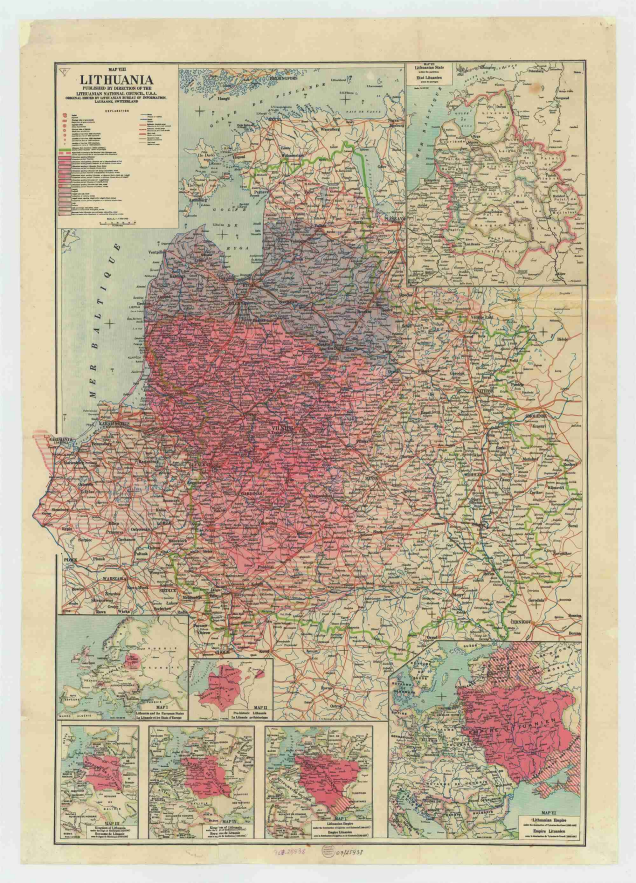
(a) Higher quality (1855)



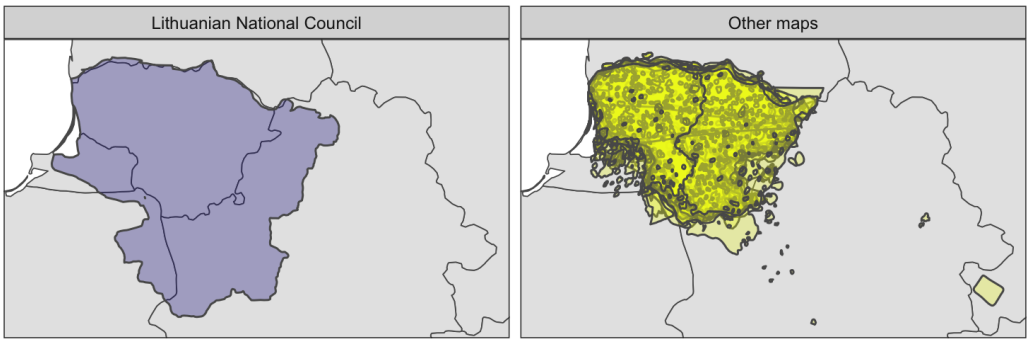
(b) Lower quality (1916)

Figure A10: Comparing two ethnographic maps of Austria-Hungary. The second map was excluded due to insufficient level of detail.

Figure A11: Map of Germans in present-day Poland (1863). This map was excluded because it does not define the settlement area of Germans but instead shows their population shares within administrative units based on census data.



(a) Map published by the Lithuanian National Committee (LNC) in 1918



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

Figure A12: The LNC map "claims" a much larger area as Lithuanian territory than 24 other maps published between 1863 and 1963, most likely due to the political motivations of the authors. For this reason, we removed this map from our final selection.

C.2 Data on self-determination claims: GeoSDM

Following our main analysis, we have examined whether ethnically distinct regions are indeed more likely to experience secessionist disputes and conflict (see Table 2 and Figure 7). To capture secessionist claims, we draw on new spatial data from GeoSDM (Schvitz, Germann and Sambanis 2020).

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). While the SDM dataset covers self-determination claims ranging from regional autonomy to demands for secession or irredentism, the data we use in this analysis is limited to the latter two claims. Moreover, our analysis uses a subset of the data that only covers the European continent.

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; Minorities at Risk 2019).

Where possible, GeoSDM relies on existing spatial datasets to geocode territorial claims. For example, the bulk of separatist claims concern existing administrative units. In these instances, claim polygons are derived from the Global Administrative Areas Dataset (GADM 2019). In instances where territorial claims were based on historical entities or ethnic settlement areas, polygons were derived from other sources (e.g. Nuessli 2010; Deiwiks, Cederman and Gleditsch 2012; Weidmann, Rød and Cederman 2010). In other cases where available GIS data was insufficient, claim polygons were based on digitized maps, which were mostly taken from Roth (2015). Figure A13 plots secessionist claims in Europe between 1946 and 2012, based on which we coded the secessionist claim outcome variable.

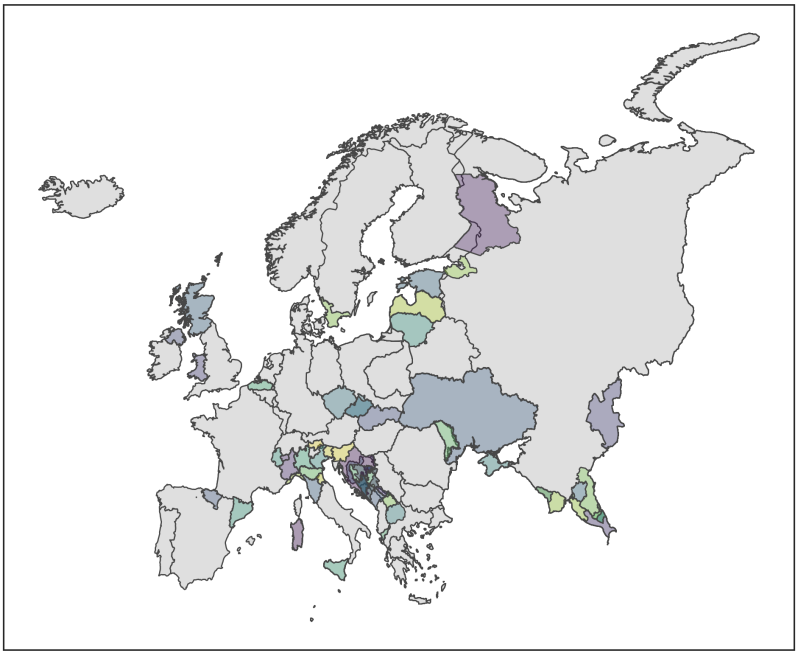


Figure A13: Secessionist territorial claims (1945-2012), based on GeoSDM

D Robustness checks: Probabilistic Spatial Partition Model

This section presents the design and results of robustness checks highlighted in Section of the main paper.

D.1 Varying control variables:

We first assess the sensitivity of the results to the choice of our main control variables. To that intent, we (1) drop all controls from our model except the state border lags in the lagged dependent variable model and (2) add additional ones. The additional variables consist of the following attributes of edges:

- Δ Longitude, Δ Latitude: Previous arguments and empirical findings show that countries tend to have an east-west rather than north-south orientation due to lower latitudinal than longitudinal environmental variation (Diamond 1997; Laitin, Moortgat and Robinson 2012). If ethnic geographies follow the same pattern, the direction of edges may present an omitted variable. To capture such dynamics we include the distance an edge traverses in each direction in decimal degrees.

- Population density in 1880 (estimate): The local population density of the area an edge traverses may comprise another omitted variable. In particular, high-density regions may feature higher levels of ethnic diversity and smaller coun-

tries. This may bias our estimates if it renders the correlation between ethnic boundaries and country borders spurious. We therefore add the level of local population density as 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 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 be an additional geographic factor that explains the structure of ethnic and state geographies. In particular, rugged (i.e. hilly or mountainous) terrain may pose a natural barrier and thus separate ethnic groups and cause country borders. To assess whether such a dynamic biases our results, we add the cumulative altitude change along an edge. It is computed by sampling first a set of points at every 1km on each edge. The final measure is 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 wildly 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 within the range between 0 and 1 to be able to compare coefficient magnitudes directly with our main indicator of interest ethnic boundary.

Table A3 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. This shows that the observed covariates do not bias the results. If those covariates are, ex ante, the most probable biasing spatial features, the result furthermore suggests a rather low likelihood of omitted variable bias affecting the estimates.

In addition, the coefficients of the additional variables exhibit some interesting patterns. First, the estimated coefficient for Δ Longitude, provides mixed evidence for Diamond (1997) and Laitin, Moortgat and Robinson (2012) in that only in the lagged dependent variable model are edges with an east-west orientation are less likely to separate two states. 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 ruggedness of an edge correlates with its likelihood to cross and interstate border.

Table A3: 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.[43] 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.

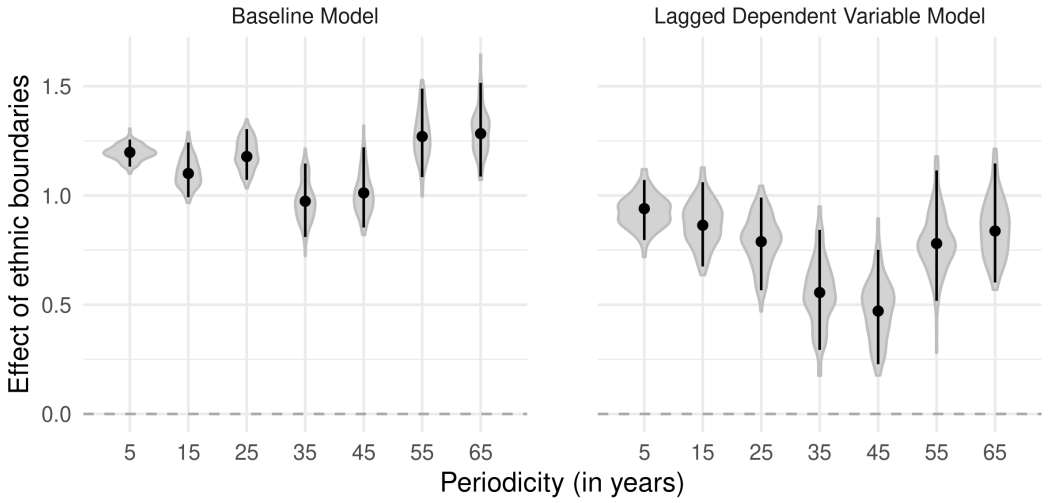


Figure A14: Point estimates of the effect of ethnic boundaries on the partitioning of Europe into states: Varying the period of the temporal structure of the data

Note: Re-estimates the models in Table 1 varying the length of periods t in years (see Eq. 5 and 6). 95% confidence intervals result from a parametric bootstrap with 120 iterations. Shaded grey areas show distribution of bootstrapped estimates.

Re-estimating our main specifications for each newly generated dataset yields results that broadly conform with our main results. Summarized in Figure A14, 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.

[43] 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.

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 the geographic location of the "anchor" of the lattice that determines the location of all vertices. The second parameter is the spatial resolution of the network. The third parameter is the spatial structure of the lattice.

Shifting the lattice anchor: The first parameter that determines the spatial make-up of our baseline lattice consists in the location of the "anchoring" point (in our 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[44] 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 A15 shows that our main estimates are well centered at the 77^{th} and 45^{st} percentiles of the respective distributions. This shows that our main results are not sensitive to the exact location of the anchoring point of our spatial lattice.

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

[44] 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.

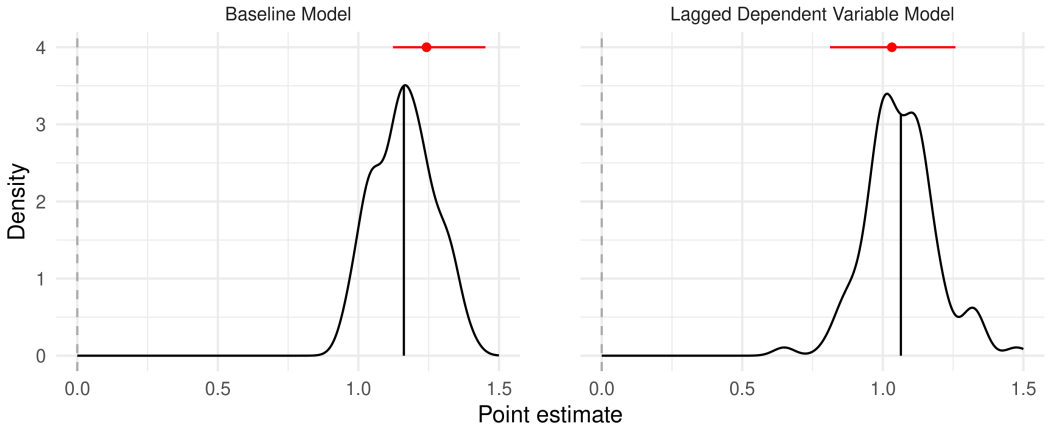


Figure A15: Point estimates of the effect of ethnic boundaries on the partitioning of Europe into states: Shifting the spatial lattice

Note: Main estimates from Table 1 in red. Distributions result from re-estimating the main models 100 times, each time using data generated for a spatial lattice that has been randomly shifted in space so as to change the location of nodes while preserving the overall structure of the network. Each lattice fills the European landmass, but nodes are located at different locations.

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 A16 presents the estimates for the effect of ethnic boundaries derived from the baseline and lagged dependent variable model estimated with the alternative lattices. The results show that our estimates slightly *increase* as we decrease the resolution of our data beyond an edge length of 100km. This suggest that ethnic geographies can have more diffuse effects that are not always captured by high-resolution data. Reassuringly, the effects estimated at resolutions lower than 100km are very similar and statistically indistinguishable from our baseline results.

Varying lattice structure The third parameter that determines the spatial makeup of our data consists in the structure of the spatial lattice. In particular, the vertices of the main lattice are the centroids of the tiles of a hexagonal tiling. There are two other regular tilings, the quadratic and triangular tiling from which we can generate regular lattices (see Figure A17).[45] 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. They therefore yield a thinner network structure when we hold the length of edges constant and are, theoretically, less able to capture spatial dependencies. A fourth possible structure for a planar lattice consists

[45] As in the hexagonal case, a tiling is transformed into a lattice by connecting the centroid (vertex) of each tile with the centroids of tiles that share an edge with the first tile.

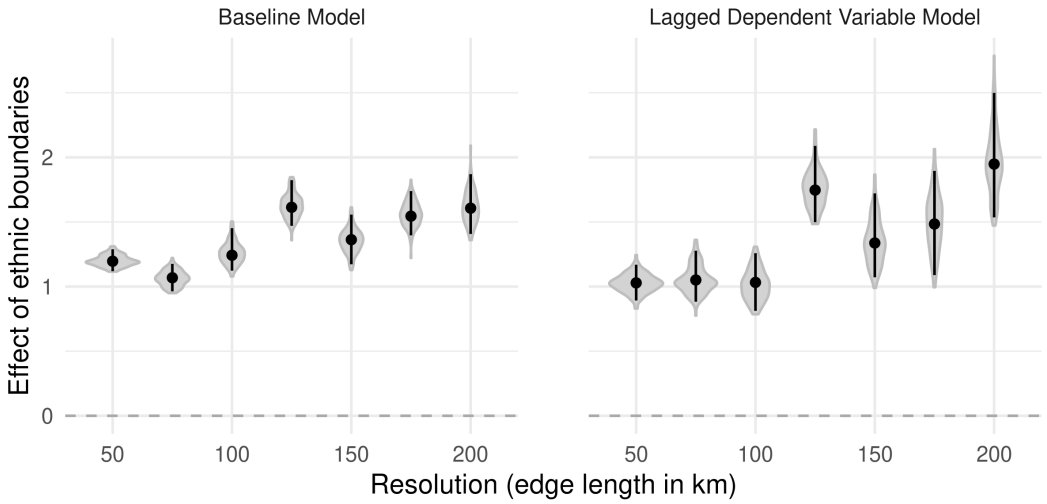


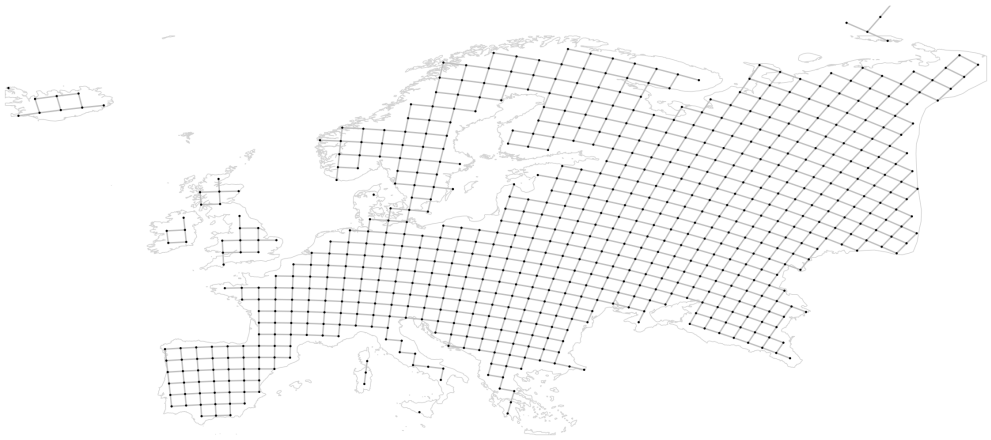
Figure A16: Effect of ethnic boundaries on the partitioning of Europe into states at varying resolutions of the spatial lattice

Note: 95% confidence intervals result from a parametric bootstrap with 120 iterations. Shaded grey areas show distribution of bootstrapped estimates.

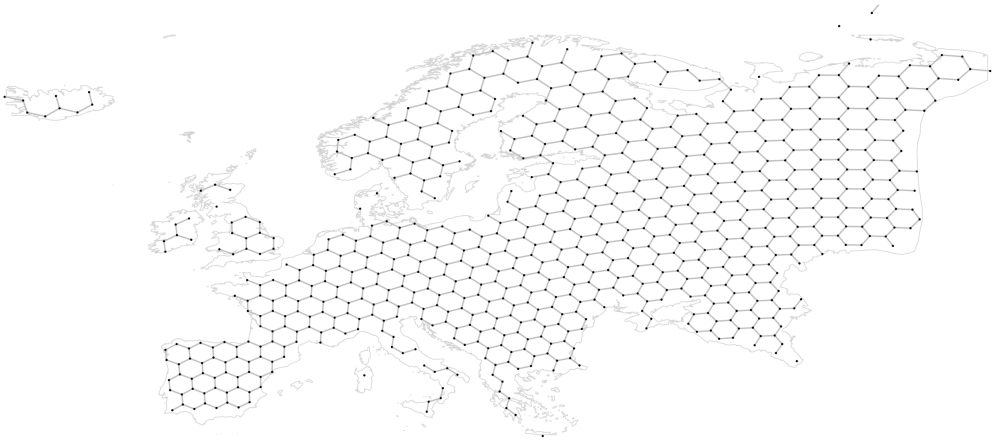
in a set of randomly located vertices transformed to a graph through a simple Delaunay triangulation.[46] While the degree of vertices in the random lattice is not constant, it is on average similar to the hexagonal structure.

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 A18 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.

[46]Note that the hexagonal lattice corresponds to a Delaunay tesselation but the quadratic and triangular ones do not. See Figure A17.



(a) Lattice from quadratic tiling



(b) Lattice from triangular tiling

(c) Lattice from randomly sampled points with Delaunay triangulation

Figure A17: Varying lattice structures

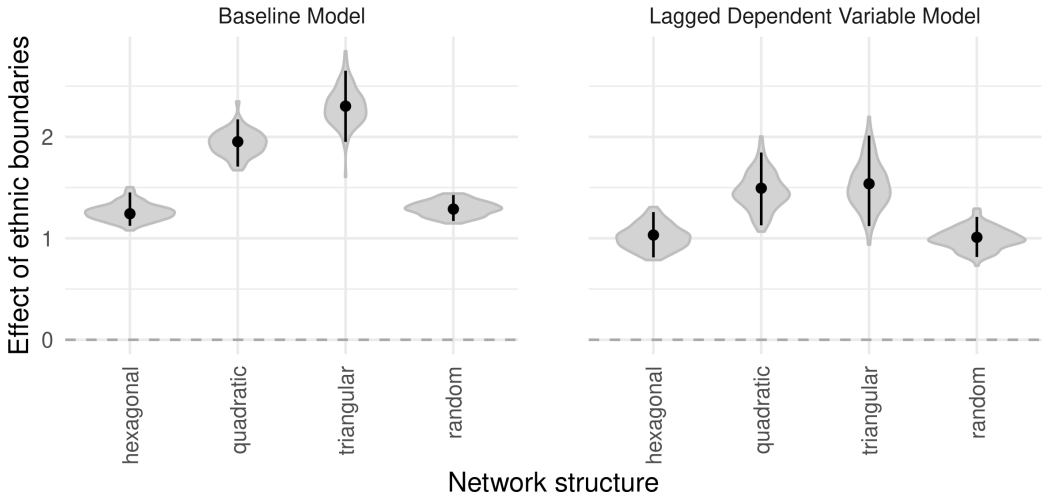


Figure A18: Effect of ethnic boundaries on the partitioning of Europe into states at varying spatial lattice structure

Note: 95% confidence intervals result from a parametric bootstrap with 120 iterations. Shaded grey areas show distribution of bootstrapped estimates.

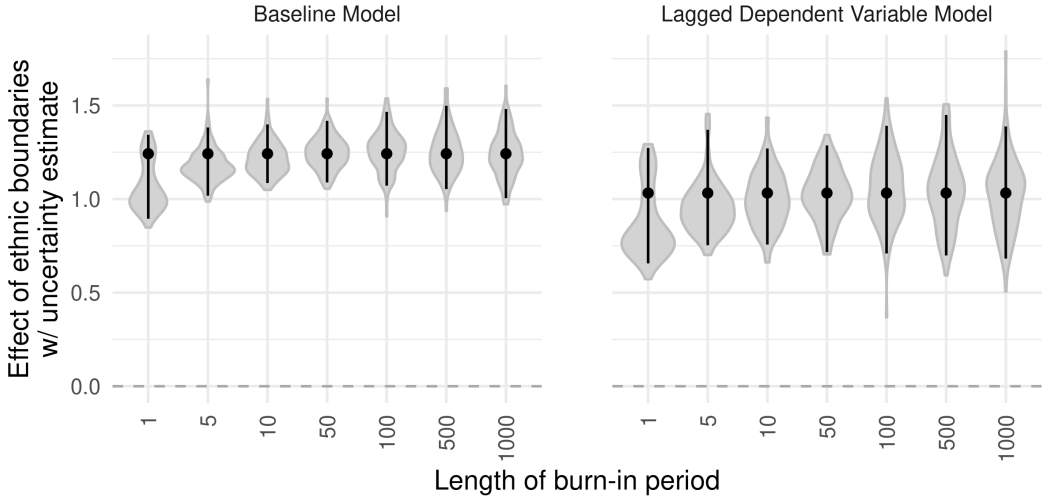


Figure A19: Uncertainty estimates with varying burn-in rates

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.5). Figure A19 plot the confidence intervals and parameter distribution retrieved from parametric bootstraps with a burn-in rate varying between 1 and 1000 iterations. The results show that the choice of the burn-in rate does not substantively affect the results above a very low burn-in rate of 10 iterations. This result coincides with the stability of the results in most areas of the parameter space assessed in our Monte Carlo experiments in Appendix Section B.2.

D.5 Logistic regression with edge-level data

To demonstrate the advantages of the PSPM, we can alternatively model our lattice data in a straightforward logistic regression setup. In particular and having to assume that edges are independent, we can model the probability $p_{j,k,t}$ that an edge between nodes j and k crosses a state borders at time t as

$$log(\frac{p_{j,k,t}}{1-p_{j,k,t}}) = \beta_0 + \beta_1 \,\textsf{ethnic boundary}_{j,k,t} + \gamma\,\mathbf{X}_{j,k} \tag{A15}$$

and, in the lagged dependent variable specification, as

$$\begin{aligned} log(\frac{p_{j,k,t}}{1-p_{j,k,t}}) = \beta_0 + \beta_1 \,\textsf{ethnic boundary}_{j,k,t-1} + \beta_2 \,\textsf{state border}_{j,k,t-1} + \\ \beta_3 \,\textsf{deep lag}_{j,k} + \gamma\,\mathbf{X}_{j,k}, \end{aligned} \tag{A16}$$

These specifications mirror the main specification with the important exception that we treat edges as fully independent here. While we know that this assumption makes it impossible for the models to generate any meaningful predictions of country borders,[47] we do not know how the assumption affects the inferences we draw from the data.

The results from estimating Equations (A15) and (A16) are listed in Table A4 and illuminate the effects of the invalid independence assumption. To compare the results directly with our main estimates we can leverage the fact that the coefficients of the PSPM are interpretable in the same way as those from a logistic regression for "bridge edges" on the lattice, i.e. edges that can change their outcome irrespective of their neighborhood and are therefore truly independent (see Appendix Section A.2). Making the comparison for these edges, we immediately

[47] This is simply because sampling from the above models yields edge-level predictions that do not partition the vertices of the graph into valid partitions.

Table A4: Edge level modeling: Logit results

	Dependent variable:	
	Baseline model	Lagged dependent variable
	(1)	(2)
Constant	-5.3039^{***}	-5.5148^{***}
	(0.1423)	(0.2051)
Ethnic boundary$_t$	3.0287^{***}	
	(0.0695)	
Ethnic boundary$_{t-1}$		2.5287^{***}
		(0.1176)
State border$_{t-1}$		4.5940^{***}
		(0.1068)
Deep lag		1.7714^{***}
		(0.1405)
Edge length	-0.4978^{**}	-0.2672
	(0.2449)	(0.3287)
River	1.0267^{***}	0.4587^{***}
	(0.0860)	(0.1581)
Watershed	1.1826^{***}	0.6127^{***}
	(0.1067)	(0.1799)
Elevation mean	4.0942^{***}	1.3342^{***}
	(0.2362)	(0.3283)
Observations	17,676	14,148
Log Likelihood	-4,503.4300	-1,917.2510
Akaike Inf. Crit.	9,018.8600	3,850.5020

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

note that the β estimates for ethnic boundaries from the logistic regression are approximately 2.5 times larger than in the baseline and lagged dependent variable PSPM models. This translates into odds ratios that are 12 times too large. In addition, we see that the estimates standard errors are significantly smaller than those yielded by the PSPM. We argue that these divergences are explained by the spatial interdependence between edge's outcomes, which are captured by the PSPM.

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.

Full results: Table A5 presents the main results, including reports on all covariates. We note that, in addition to non-coethnic capital, the size of the local population and the presence of historical precedents seem particularly relevant for explaining the spatial pattern of secessionism. The remaining covariates yield insignificant results. It must, however, be noted that the inclusion of non-coethnic capital, the population size, and previous borders, are causally posterior to all geophysical variables, thereby making it impossible to precisely interpret the estimates for these variables' effects.

Table A5: Ethnic boundaries and the onset of self-determination processes: Full results

	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)
Pt: Dist. to capital (log)	0.280	−0.063	0.897*	0.508	0.080	−0.934**
	(0.444)	(0.366)	(0.517)	(0.543)	(0.582)	(0.380)
Pt: Dist. to border (log)	−0.085	−0.232**	0.174	0.028	0.299***	−0.074
	(0.136)	(0.116)	(0.109)	(0.151)	(0.095)	(0.071)
Pt: Population (log)	0.150*	0.097*	0.334***	0.233***	0.241***	0.174***
	(0.084)	(0.052)	(0.066)	(0.073)	(0.049)	(0.044)
Pt: Altitude	−0.0004	−0.0003	−0.001***	−0.001**	−0.001**	−0.001**
	(0.0003)	(0.0003)	(0.001)	(0.001)	(0.001)	(0.001)
Pt: Ruggedness	0.208***	0.228***	0.207**	0.226**	0.035	0.200***
	(0.078)	(0.071)	(0.085)	(0.100)	(0.099)	(0.058)
Pt-C: River	−0.101*	−0.036	−0.122	−0.061	−0.059	−0.084**
	(0.057)	(0.072)	(0.075)	(0.054)	(0.050)	(0.037)
Pt-C: Deep state lag	−1.981***	−1.948**	−1.918**	−2.403**	−2.835***	−3.335***
	(0.674)	(0.837)	(0.804)	(0.991)	(0.816)	(0.928)
Pt-C: Watershed	0.030	−0.019	0.054	−0.005	−0.049	0.027
	(0.128)	(0.115)	(0.077)	(0.074)	(0.129)	(0.087)
Pt-C: Elevation	−0.712	−1.910***	1.277	−3.017**	0.642	−1.761
	(0.716)	(0.599)	(1.145)	(1.385)	(0.939)	(1.519)
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. Full results with control variables are reported in Table A5. 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 A6 presents the respective results. All coefficient increase substantially in size (on average around 50 percent). This suggests that selection bias in the original analysis leads us to underestimate the effect of mismatches between state and ethnic geographies on secessionism.

Table A6: Ethnic boundaries and 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

Using pre-1886 data on ethnic geography: As in the estimation of the PSPM, our analysis of secessionism may be biased if changes in the ethnic boundaries are caused by factors that also cause state borders to change thereafter. In order to circumvent this risk, we recur to ethnic geographies measured at the earliest point in our data, in the 50 years prior to 1886. Estimating their effect on post-1946 secessionim in Table A7 yields estimates of non-coethnic capital that are marginally smaller than those estimated for time-varying ethnic boundaries but nevertheless of substantive size. Given the reduced precision of the data, standard errors slightly increase and render one estimate, the effect of non-coethnic capital on a points' "break away" in the baseline specification, statistically insignificant. However, the results remain largely robust. Together with the overall stability of ethnic geographies, this suggests that endogenous changes in ethnic geographies are unlikely to cause the results.

Table A7: 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. Full results with control variables are reported in Table A5. 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) shifting points along the north-south and east-west axes 100 times, (2) varying the spatial resolution of points between 50 and 200km, and (3) retrieving points from the centroids of quadratic and triangular tiles, as well as from a spatially random draw. Figure A20 shows that our main estimates are well centered in the distribution of estimates yielded upon shifting our raw points in space and regenerating the dataset. Figure A21 demonstrates the robustness of the results, including their uncertainty estimates, to increasing or decreasing the spatial resolution of our data. Lastly, the results presented in Figure A22 show 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 changing the three parameters that govern the spatial structure of our data.

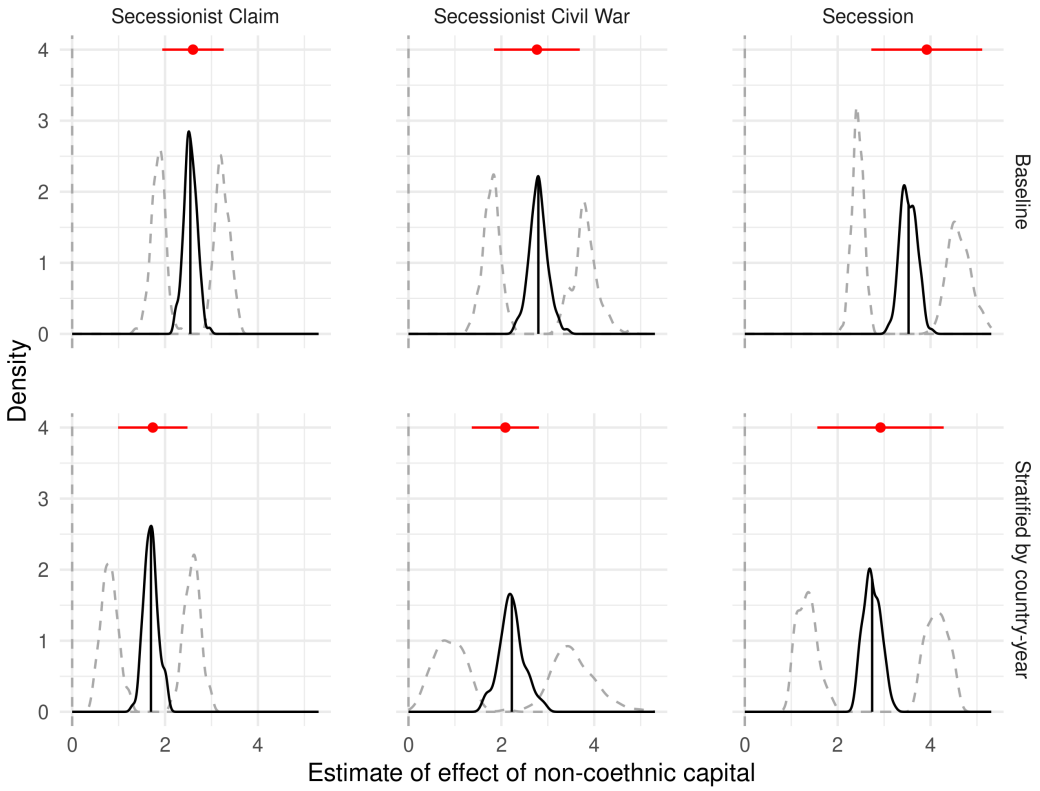


Figure A20: Estimates of the effect of non-coethnic capitals on secessionism: Shifting points (unit of analysis)

Note: Main estimates from Table 2 in red. Solid lines denote distribution of main estimates, dotted lines distributions of the upper and lower bounds of the 95% confidence interval. Distributions result from re-estimating the main models 100 times, each time using data generated for a spatial points on a hexagonal lattice that has been randomly shifted in space so as to change the location of nodes while preserving the overall structure of the network. Each lattice fills the European landmass, but nodes are located at different locations.

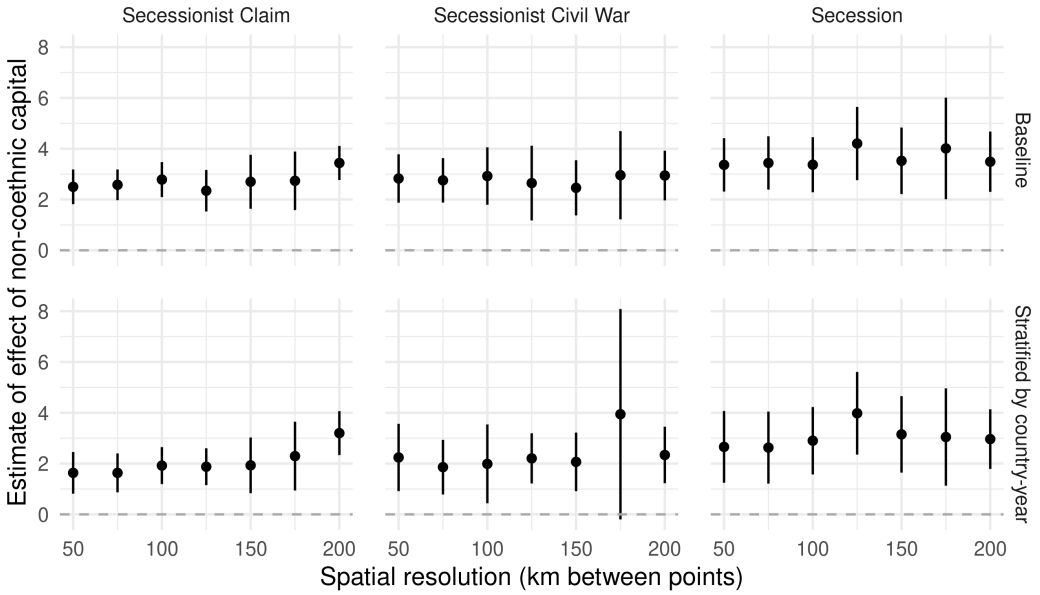


Figure A21: Estimates of the effect of non-coethnic capitals on secessionism at varying spatial resolutions lattice

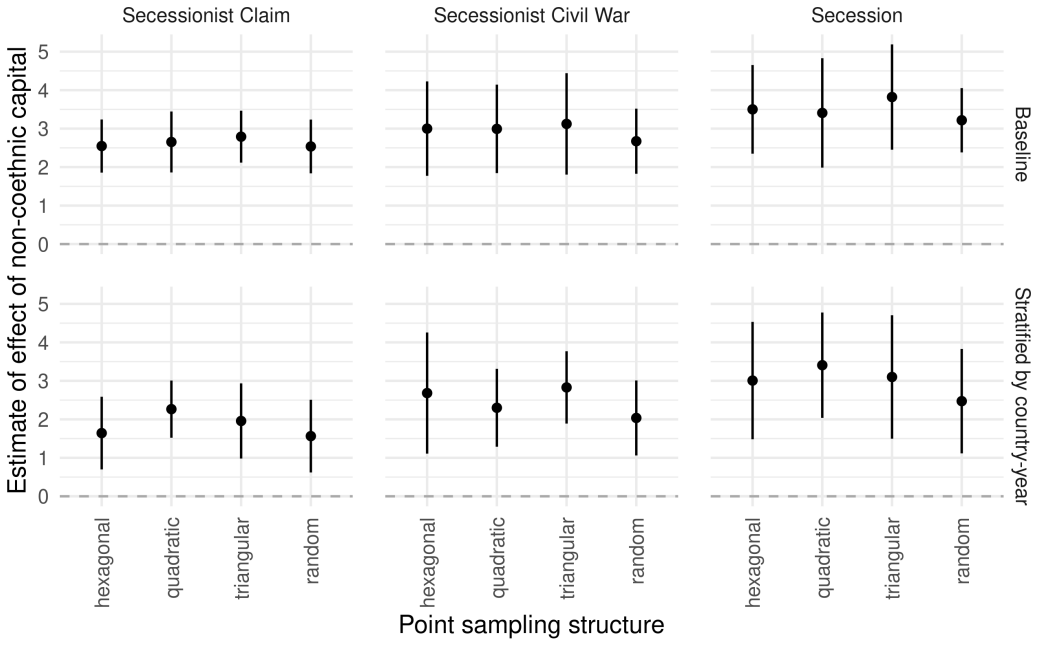


Figure A22: Estimates of the effect of non-coethnic capitals on secessionism with varying spatial structure of the lattice

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.

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

Deiwiks, Christa, Lars-Erik Cederman and Kristian Skrede Gleditsch. 2012. "Inequality and conflict in federations." *Journal of Peace Research* 49(2):289–304.

Diamond, Jared M. 1997. *Guns, germs, and steel*. New York: WW Norton.

GADM. 2019. "GADM database of Global Administrative Boundaries Version 3.6. 2019.".
URL: *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.

Minorities at Risk. 2019. "Minority Group Assessments - Qualitative Reports.".

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

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

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

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

Schvitz, Guy, Micha Germann and Nicholas Sambanis. 2020. "Mapping Self-Determination Claims 1946-2012: The GeoSDM Dataset.".

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.