European Political Boundaries as the Outcome of a Self-Organizing Process

Eric Weese

Abstract

Political economy theories predict certain configurations of national boundaries, but these have not been calculated because of computational difficulties. Taking advantage of advances in mixed integer programming algorithms, we compute predicted political boundaries for Europe using a simple theoretical model taken from the literature: the size and arrangement of countries is determined by a tradeoff between efficiencies of scale and geographic heterogeneity. The model shows that the "natural borders" that lead to states emerging in certain configurations do not need to be particularly extreme, and a small number of these geographic features can influence the configuration of boundaries over a larger area. Our results show how real-world political boundaries can be described by a simple one parameter theoretical model that ignores many of the proximate causes of boundary changes.

Keywords: coalition formation, national borders, mixed integer programming.

"The history of Europe is that of its borders" [Pomian 1990]. From the historical perspective, current European international boundaries are the result of a combination of inheritance, war, nationalist sentiment, and other causes.[1] Given this enormous historical variation, it is clear that no single model of political jurisdiction formation will be able to completely explain observed political boundaries. We present a very simple model that describes international boundary formation in a democratic system, and show how it predicts boundaries that match some important features of the data. At the end of the paper, we discuss some reasons why this model might also give reasonable results

1 See Abramson and Carter [2016] for a review and many references.

for boundaries determined in an autocratic system. Our simulation results include contemporary controversies: Northern Ireland is generally predicted to be a part of England, and Northern and Southern Italy are separate. From a broader perspective, our results suggest that there are natural boundaries for countries, which are thus not purely “imagined communities” [Anderson 1983].

A model of democratic political boundaries consists of two parts. First, there needs to be a description of the relative desirability of various configurations of political jurisdictions, and how this varies across different individuals. Second, there needs to be a rule for how political boundaries are determined given these individual preferences. We use existing models in the political economy literature for both of these parts.[2] This approach differs substantially from an older literature based on Voronoi diagrams [Okabe et al. 2000]. This previous literature considered where boundaries might form *given* a certain set of capital cities or other generating points for the countries in question. In contrast, the model in this paper does not assume a set of generating points, instead considering what countries might form given characteristics of the overall population in question.

We begin by presenting our model of the desirability of different configurations of political jurisdictions. The starting assumption is that each individual in Europe is located at a fixed physical location, will belong to exactly one country, and cares only about the characteristics of the country to which they belong. This is obviously an extreme simplification: it ignores, for example, migration, wars between countries, and the possibility that an area might be governed by multiple countries or none. Our theoretical model is based on individuals; however, for computation, we will aggregate similar individuals into small geographic units, and conduct simulations based on these units.

The first ingredient of our model of desirability is based on the idea that a jurisdiction with a small population is expensive to operate in per capita terms, and thus there are advantages to larger jurisdictions. We use the standard form [Stephan 1977] for these efficiencies of scale:

$$C(P) = F + VP. \tag{1}$$

2 In particular, Cremer, De Kerchove, and Thisse [1985], Alesina and Spolaore [1997], and Desmet et al. [2011].

Here $C(P)$ is the basic cost of operating a jurisdiction with a population of P people. There is a fixed cost, F, that does not depend on the size of the jurisdiction, and a variable cost, V, that is proportional to population. It will be easier to work with the per capita version of C:

$$C(P)/P = F/P + V. \tag{2}$$

The units of F can either be money or time, where the interpretation in the latter case is that in order to provide services the government requires a certain number of person-hours of labour each year. The functional form chosen for C is a very simple one, but there is some evidence [Weese 2015] that it is close to the true functional form for at least some public goods. We assume that the cost is borne equally by all individuals in the jurisdiction, and thus each of them is responsible for the same share $C(P)/P$ of the total cost.

A model considering only the basic per capita cost $C(P)/P$ of providing government services would predict trivial boundaries for Europe: per capita cost is always decreasing in population, and thus all of Europe should be a single country. Qualitative evidence suggests that large countries face certain difficulties in collecting information and in providing services. In the literature, the additional difficulty faced by large countries is usually described as a heterogeneity cost. A functional form often used to model this cost is

$$L_i(S) = P_S^{-1} \sum_{i' \in S} \ell_{i,i'}, \tag{3}$$

where S is the set of individuals making up a jurisdiction, P_S is the total number of individuals in the jurisdiction, and $\ell_{i,i'}$ is the distance between individuals i and i'. $L_i(S)$ is thus the average distance between individual i and a randomly selected individual in the jurisdiction described by S.

This L function originated in the linguistics literature [Greenberg 1956; Lieberson 1964], where $\ell_{i,i'} = 1$ if individuals i and i' speak the same language, and 0 if they do not. In this paper, rather than a $\{0, 1\}$ coding, we let $\ell_{i,i'}$ be the geographic distance between i and i'. We calculate these distances using a shortest path algorithm, with a few exceptions restricting travel to land rather than water. Details are provided in the appendix. A variety of similar distance concepts have been used in the applied political economy literature [Brasington 1999; Gordon and Knight 2009]: we choose this particular one because it results in an Aziz, Brandt, and Harrenstein [2014] "fractional hedonic" form for individual payoffs, which will make simulations computationally feasible.

Combining the two terms just presented, we see that the total benefit to individual i in the case where they are a part of jurisdiction S is

$$U_i(S) = -F/P_S - V - \gamma P_S^{-1} \sum_{i' \in S} \ell_{i,i'}, \tag{4}$$

where γ is a parameter describing the relative importance of heterogeneity, and everything has been multiplied by -1 because it is standard to discuss individual decisions in terms of their benefits but both terms being considered are costs.

Given this model of the benefit to individual i of belonging to country S, we consider what countries will emerge given the preferences described by U. Motivated by the idea of a "right to self determination", we use the core as our solution concept.[3] Start by considering some potential partition π that divides Europe into countries. Is there any *other* country S', not present in this partition π, where every individual who would be in S' prefers S' to whatever country they are currently a member of in π? If so, then we might expect this S' to form, and we would not observe partition π. We are thus particularly interested in a core partition π^* where there is no such deviation S'. Weese, Hayashi, and Nishikawa [2020] show that with payoffs of the type given in Equation 4, although there is a theoretical risk that there might not be any core partitions, the probability of encountering such a situation using actual data is vanishingly small. We will use the algorithm from Weese, Hayashi, and Nishikawa [2020] to generate our core partitions: this algorithm makes use of mixed integer programming, and is described briefly in the appendix.

The model just described depends crucially on the distribution of population across Europe; however, this population distribution has experienced recent dramatic shifts due to urbanization. We would like detailed historical data regarding population distribution for all of Europe, but this sort of data is simply not available. We will thus use the agricultural suitability of land from Ramankutty et al. [2002] as a proxy for historical population: this data is shown in Appendix Figure 3. Using this proxy is reasonable because historically the vast majority of the population was rural and agricultural, and was living in Malthusian conditions where the carrying capacity of a location determined

3 The use of a formal solution concept distinguishes our simulation results from those of Cederman [1997] and other "agent-based" approaches.

the population living in it.[4]

In theory, the model just presented could be applied directly to the grid square data on agricultural suitability. In practice, however, serious numerical issues arise when considering grid squares with low suitability. We thus aggregate grid squares up to the polygons shown in Appendix Figure 4, as described in Appendix A. This aggregation somewhat complicates the statistical interpretation of the results, as discussed in Appendix B, but it is required for simulation to be feasible at all.

Simulation also requires a choice of the parameter γ, describing the relative importance of heterogeneity. Estimation of this parameter would be a challenging undertaking, and there is no generally accepted method. Rather than choose a complicated estimation method, which might arouse suspicion of having specially selected a "good" value for γ, we perform a simple order of magnitude calibration. With distance in km, and the units for agricultural suitability being "fraction of a grid square" taken directly from Ramankutty et al. [2002], we find that $\gamma = 0.1$ gives an average of 23 simulated countries, while $\gamma = 0.01$ gives 3, and $\gamma = 1$ gives 113. We use $\gamma = 0.1$.

The final piece of data required is actual European boundaries, in order to evaluate the model. We use boundaries between 1000 and 2000 CE, at 100 year intervals, as provided by Euratlas. These boundaries are shown in Figure 1. We use average boundaries over 1000 years, rather than current boundaries, because the model attempts to describe "natural borders", that should repeatedly appear in the data, rather than "accidents of history", such as internal Soviet and Yugoslavian boundaries that were never intended to serve as international borders but now do, due to *uti possidetis*.

Figure 2 shows the boundaries predicted by the model, averaging over 100 simulation runs. Appendix Figures 5 and 6 show examples of the partitions generated in these runs. While simulated boundaries do not match the actual boundaries perfectly, key features are reflected.[5] Most notable are the mountain ranges (the Pyrenees, Alps, and Carpathians) shown in Appendix Figures 7 and 8 that appear as boundaries in Figure 2, despite the fact that elevation does not directly enter into the model. In the

4 An additional advantage of using agricultural suitability data is that it is less vulnerable to reverse causality: actual historical population distributions might be concentrated near the center of actual countries because capitals and the resulting administrative infrastructure are generally geographically central. In this case, actual population data would very successfully predict actual country boundaries, but for a reason completely unrelated to that of the model proposed in this paper.

5 Quantitative tests based on additional simulations are presented in Appendix B.

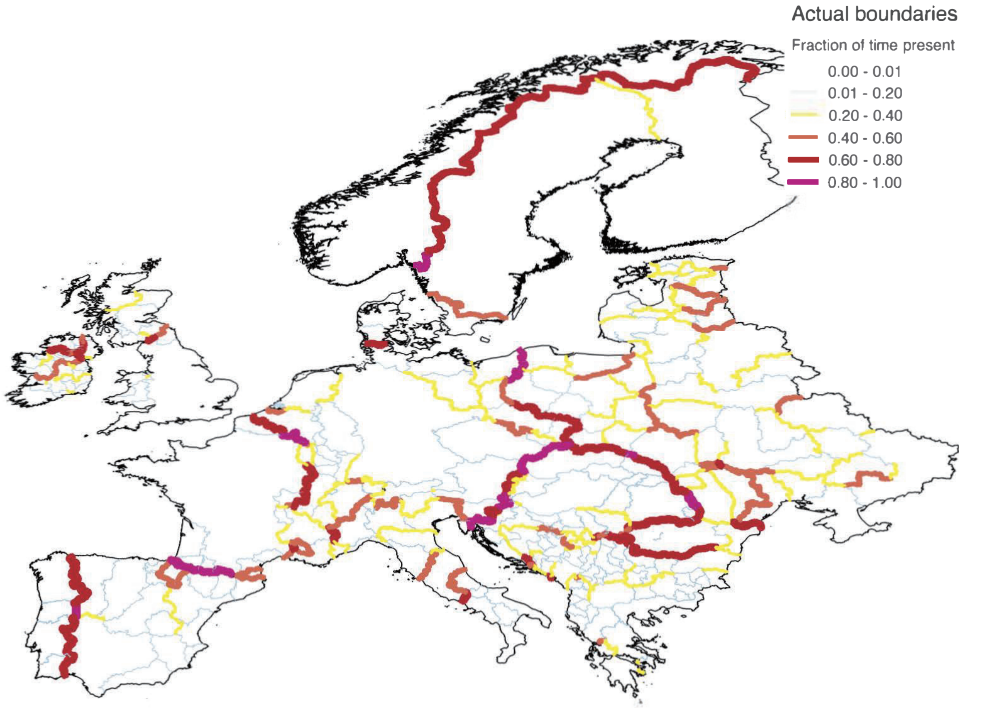

Figure 1: Actual Boundaries (see Appendix A for note regarding Portugal)

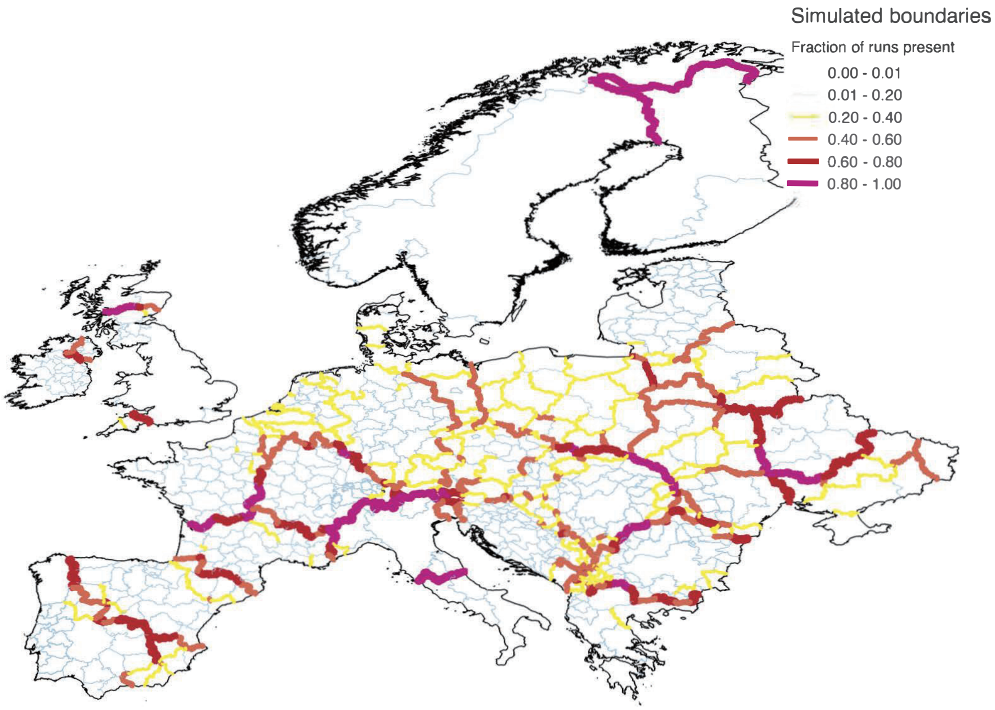

Figure 2: Simulated Boundaries

model, the mountain ranges form natural boundaries not because they are difficult to cross, but simply because there is no agricultural land there: a country that spanned both sides of the mountains would have high geographic heterogeneity, and thus would tend not to be part of a core partition.[6] Appendix Figure 9 shows an overlay of the simulated boundaries on the agricultural suitability data.[7]

Some additional boundaries appear in both Figures 1 and 2, but do not correspond to obvious features in Appendix Figure 3. These include the (historical) boundary between Germany and Poland, and the boundaries further east than the Carpathians. These bound- aries appear to emerge from the simulations because of the natural boundaries just discussed: if one country begins at the Carpathians, and the model leads to countries being a certain size given agricultural suitability, then Carpathians also lead to another country tending to emerge at another point further to the east. Similarly, some boundaries for Poland emerge not because there are any natural features exactly at those boundaries, but rather because mountains further to the south lead to a certain configuration tending to form. In general, areas with low agricultural suitability lead not only to a boundary in the immediate vicinity, but lead to certain patterns of boundaries further away.

Notable failures of the model are France and the Ukraine (both split into three pieces), and Portugal (far too large). On the other hand, Denmark exists reliably, although the boundary varies simulation by simulation. Yugoslavia appears fairly clearly in the simulation results, despite the fact that it has been regarded as an artificial country. The simulation results suggest that the country's creation may have been reasonable in theory, even if it ultimately failed in practice. The results for France and

6 In independent work, Kitamura and Lagerlof [2020] examine the relationship between mountains, rivers, and other geographic features, and political and ethnic boundaries. Despite the immediate intuitive reaction, it is not obvious that the Alps are a natural boundary because they are difficult to cross. First, note that the Ligurian Sea and its surrounings are navigable, and thus the Alps can be bypassed by coastal travel. Second, Northern Ireland is part of the same country as England both in the data and in many of the simulation results: thus, countries that are not connected by land can and do exist. A more sophisticated model could be proposed, that would include the relatively small size of Ireland, and perhaps its military weakness, and contrasting that with the relative strength of France, and the difficulty of a maritime invasion. The point of the simulations in this paper is not to prove that the height of the Alps is not important: rather, they merely suggest than another explanation is also available.

7 Earlier empirical work includes Alesina, Baqir, and Hoxby [2004]. The relationship between geography and the number of ethnic groups has previously been examined using a different technique [Michalopoulos 2012].

the Ukraine are interesting given that the former is frequently given as an example of a strong homogenizing state, and the later has a troubled territorial history.

The simulation results just presented used a model based on the idea of self-determination. For much of European history, however, boundaries were decided by wars between autocratic leaders, interested mainly in collecting taxes from the land they controlled. Appendix A.2 shows that the model presented above also describes the tradeoff that would be faced by despots attempting to collect a tax by travelling to and from a capital located at the centroid of their territory.

To formally model despotism, a change would also be required to the algorithm used to simulate partitions, in order to specify when a new despot would attempt to start his own country. Intuitively, however, the "self-determination" algorithm presented tends to generate coalitions S' which, in a despotic setting, would be associated with a new despot who is better positioned to rule the population in S' than the existing despots. It is not obvious how to formalize this intuition, but it suggests why a model based on democratic principles matches data that contains a large number of autocratic jurisdictions. Further research may yield a formal model of autocratic governments for which simulation of boundaries is computationally feasible.

References

Abramson, Scott F. and David B. Carter (Nov. 2016). "The Historical Origins of Territorial Disputes". *American Political Science Review* 110(4): 675–698.

Alesina, Alberto, Reza Baqir, and Caroline Hoxby (Apr. 2004). "Political Jurisdictions in Heterogeneous Communities". *The Journal of Political Economy* 112(2): 348–396.

Alesina, Alberto and Enrico Spolaore (Nov. 1997). "On the Number and Size of Nations". *The Quarterly Journal of Economics* 112(4):1027–1056.

Anderson, Benedict R. O'G. (1983). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London: Verso.

Aziz, Haris, Felix Brandt, and Paul Harrenstein (May 2014). "Fractional Hedonic Games". *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems. Paris.

Brasington, David M. (Sept. 1999). "Joint Provision of Public Goods: The Consolidation of School Districts". *Journal of Public Economics* 73(3): 373–393.

Cederman, Lars-Erik (1997). *Emergent Actors in World Politics: How States and Nations Develop and Dissolve*. Princeton University Press.

Cremer, Helmuth, Anne-Marie De Kerchove, and Jacques-François Thisse (June 1985). "An Economic Theory of Public Facilities in Space". *Mathematical Social Sciences* 9(3): 249–262.

Desmet, Klaus et al. (Sept. 1, 2011). "The Stability and Breakup of Nations: A Quantitative Analysis". *Journal of Economic Growth* 16(3): 183–213.

Gordon, Nora and Brian Knight (June 2009). "A Spatial Merger Estimator with an Application to School District

Consolidation". *Journal of Public Economics* 93: 752–765 (5–6).

Greenberg, Joseph H. (Jan. 1956). "The Measurement of Linguistic Diversity". *Language* 32(1): 109–115.

Kitamura, Shuhei and Nils-Petter Lagerlof (Aug. 2020). "Geography and State Fragmentation". *Journal of the European Economic Association* 18(4): 1726–1769.

Lieberson, Stanley (Oct. 1964). "An Extension of Greenberg's Linguistic Diversity Measures". *Language* 40(4): 526–531.

Michalopoulos, Stelios (June 2012). "The Origins of Ethnolinguistic Diversity". *American Economic Review* 102(4): 1508–39.

Okabe, Atsuyuki et al. (July 26, 2000). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. 2 edition. Chichester; New York: Wiley. 696 pp.

Pomian, Krzysztof (1990). *L'Europe et Ses Nations*. Débat, historie, politique, societé. Paris: Gallimard. 251 pp.

Ramankutty, Navin et al. (2002). "The Global Distribution of Cultivable Lands: Current Patterns and Sensitivity to Possible Climate Change". *Global Ecology and Biogeography* 11(5): 377–392.

Stephan, G. Edward (Apr. 29, 1977). "Territorial Division: The Least-Time Constraint Behind the Formation of Subnational Boundaries". *Science* 196(4289): 523–524.

Weese, Eric (July 2015). "Political Mergers as Coalition Formation: An Analysis of the Heisei Municipal Amalgamations". *Quantitative Economics* 6(2): 257–307.

Weese, Eric, Masayoshi Hayashi, and Masashi Nishikawa (Oct. 2020). "Inefficiency and Self- Determination: Simulation-Based Evidence from Meiji Japan". *CIRJE Discussion Paper F-989* (1627): 93 pages.

A Methods

The method used to find the stable partitions of countries is taken from Weese, Hayashi, and Nishikawa [2020], which includes a description of the mixed integer programming technique used to find possible new countries S'. A discussion of the speed gains in mixed integer programming can be found in Bixby [2012].

The algorithm we use is based on ideas from matching theory [Roth and Vate 1990]. Begin with an arbitrary starting partition π_0 (we will use as π_0 the partition where there is a separate country for every individual, but this choice is not important). Consider then whether there is any alternative country $S' \notin \pi_0$ such that, for every individual $i \in S'$, $U_i(S')$ is higher than the benefit i was receiving from the country they were a part of in π_0. If so, then S' will form: create a new partition π_1 by adding S' to π_0 and removing or modifying other countries as necessary. Then repeat this process using π_1 instead of π_0, and continue in this fashion until no S' can be found. There is no theoretical guarantee that this algorithm will terminate [Gale and Shapley 1962; Brandl, Brandt, and Strobel 2015], but Weese, Hayashi, and Nishikawa [2020] find that it terminates every time even when using millions of random datasets.

The intuition for this algorithm is that S' is a (potential) country where all of its constituents would want to leave whatever country that they are currently a member of, and form S' instead. Although not always followed in practice, ideals of "self determination" suggest that we should see partitions where this sort of S' does not exist. At each intermediate step in the algorithm described above, there are often many potential choices for S': different core partitions can be found by using a different rule for selecting S' in cases where there are multiple options available.

According to the theory presented, each person in Europe could be considered as a separate i, with a specific individual location on the continent. This direct approach is computationally infeasible. Instead, we aggregate the population of Europe into small geographic units, and then use i to represent these units. All simulations are then run based on these units.

The algorithm used to generate stable partitions is computationally stable when considering units of self-determination with populations that are non-trivial relative to the populations of the countries in stable partitions. If the grid square data on agricultural suitability shown in Appendix Figure 3 were used directly, however, a substantial number of grid squares would have extremely small populations. If, during a simulation, these grid squares were to end up by themselves (i.e. in a country consisting only of that grid square), then the per capita cost faced by that grid square would potentially be many orders of magnitude higher than for any of the other grid squares. In mixed integer programming, large differences in magnitude of this sort generally lead to difficulties in obtaining accurate solutions, and thus need to be avoided. To avoid this situation, we aggregate grid squares into larger units, and use those units as the basic units for the simulation performed.

We begin by using the administrative divisions provided by the Global Administrative Boundaries dataset. We omit the European microstates (Andorra, Monaco, San Marino, Vatican City, and Liechtenstein) and Iceland. We use national boundaries (level 0 subdivisions) for four particularly small countries: Montenegro, Moldova, Macedonia, and Luxembourg. We use level 2 subdivisions for five countries: the United Kingdom, Germany, France, Spain, and Bosnia-Herzegovina. For all other European countries, we use level 1 administrative subdivisions. We eliminate Mediterranean islands, and all other holdings south of continental Europe (e.g. Ceuta and Melilla). The resulting set of administrative polygons forms the basis for our aggregations of agricultural suitability.

We continue to use the grid square as a unit of measure for agricultural suitability. Thus, a polygon that includes all of exactly one grid square where the grid square is rated as 10% suitable, will be counted as having a total suitability of 0.1. Similarly, a polygon that contained 50% of a grid square that was 100% suitable for agriculture, and all of a grid square that was 10% suitable will be counted as having a total suitability of $0.5 \times 1.0 + 1.0 \times 0.1 = 0.6$.

We then calculate distance between each of the polygons using the Dijkstra shortest-path algorithm. An alternative would be to simply calculate the straight-line distance between polygons. This straight-line distance, however, would not necessarily remain on land. On the one hand, historically travel was often faster by sea. However, most of this seafaring ocurred within a short distance of the coast. A shortest-path distance does not allow straight-line voyages across, for example, the Ligurian Sea between Italy and France. Travel distance along the coast by sea, however, will be approximated very closely by the travel distance along the coast by land, and this will be calculated by the shortest-path algorithm.[8]

To calculate the shortest-path distance, begin with the set of polygons. Let these units form the vertices in a graph. Edges of the graph indicate geographic adjacency. The locations of the vertices are determined by the weighted mean location of the grid squares contained in the polygons in question.[9] The distance that will be used for any two units in the simulation is the shortest path distance between the corresponding vertices in the graph.

We then use an iterative process to further aggregate polygons that still have particularly low totals for agricultural suitability. We use as a threshold 0.1 units of agricultural suitability, using the grid square units of measure just described. We amalgamate the polygon with the lowest total amount of agriculturally suitable land with its closest neighbour. We then recalculate the distances using the shortest path algorithm just described, as the number of vertices in the graph has decreased by one. Repeat this process until the player with the lowest total amount of agriculturally suitable land has at least 0.1 units of agricultural suitability. The resulting set of

8 The major assumption made is assigning the same cost to these routes as "inland" routes. A more sophisticated model might allow for longer voyages at the same cost, so long as they were along the coast.

9 This graph is not shown, but it is quite similar to the final graph shown in Appendix Figure 10.

polygons is shown in Appendix Figure 4. This algorithm does not pay any attention to actual national boundaries, and thus the final polygons used for the simulation units may cross current national boundaries. A notable case where this happens is northern Portugal, which is grouped with Galicia because of a lack of agriculturally suitable land in the northwest corner of the Iberian peninsula.

A final issue arises because, with the data actually considered, the presence of islands would lead to the graph consisting of several connected components, rather than only one. While this would not affect the simulations, some actual countries, most notably the United Kingdom, include territory separated by water. We thus join some vertices across water, at the narrowest point of the water in question. Four of edges of this sort are added: they are Great Britain - (Northern) Ireland, Great Britain - France, Denmark - Sweden, and Finland - Estonia. The resulting graph is shown in Appendix Figure 10.

Although the edges chosen are the shortest links for the general areas in question, the choice to make these four links is based on a heuristic examination of the graph in question, rather than some specific formal rule. We thus cannot draw any firm conclusions on the nature of water as a natural border from the simulations performed. The four "water edges" are not shown in Figures 1 and 2, but are shown in Appendix Figures 11 and 12, which are an alternative visualization of the same data. In the actual data all four of the edges are often boundaries between countries. This is also true in the simulation results for three of the four edges, with the exception being the Great Britain - (Northern) Ireland edge. In the simulations, Northern Ireland is usually in the same country as England.

A.1 Boundary Statistics

To determine whether a country boundary lies between two units, we use the locations shown in Appendix Figure 10, which are weighted mean locations. If the vertices on either side of an edge are in the same country, then we code this edge as having no boundary. Conversely, if these vertices are in different countries, then we code the edge in question as having a boundary. The exact location of the boundaries displayed in Figures 1 and 2 correspond to boundaries of the polygons in Appendix Figure 4, coloured appropriately. The boundaries in Figure 1 thus do not match precisely with actual international borders: this is most obvious in the case of Spain and

Portugal, where the lack of agriculturally suitable land in Galicia has resulted in a particularly large polygon, the weighted mean of which lies in Portugal. The entire western coast of the Iberian Peninsula is thus reported in Figure 1 as belonging to Portugal. This method is used because it allows easy comparisons between the actual boundaries reported in 1, and the simulated boundaries reported in Figure 2.

A.2 Distance squared

Suppose that instead of using geographic distance for ℓ in Equation 3, we used geographic distance squared. The last term of Equation 4 then becomes $\gamma P_S^{-1} \Sigma_{i' \in S} d^2_{i,i'}$, where $d^2_{i,i'}$ is geographic distance squared. The average of this over all individuals in S is $\gamma P_S^{-2} \Sigma_{i \in S} \Sigma_{i' \in S} d^2_{i,i'}$. A standard analysis of variance result [Scheffe 1959] is that this is equal to $\gamma P_S^{-1} \Sigma_{i \in S} d^2_{i,s}$, where s is the geographic location of the population-weighted centroid of S.

The sum $\Sigma_{i \in S} U_i$, where U_i is as in Equation 4, except using $\ell = d^2$, would thus describe the payoff to a despot located at s who collects a tax of $-V$ from everyone in the country,[10] but who must pay a travel cost $\gamma d^2_{i,s}$ to collect the tax from individual i, and in addition must pay a fixed cost F to run the country.

B Statistical Analysis

A qualitative comparison of Figures 1 and 2 in the main text suggests that the model proposed in this paper successfully captures some aspects of the actual process of boundary formation. Two issues are best addressed via quantitative analysis, however: the possibility that these results are due to a mechanical effect due to differing surface areas of units units used in the simulations, and the possibility that the results observed are no more surprising than what would be observed if random convex "countries" were selected so as to partition Europe.

The first of these issues arises because the simulation technique used requires that the units of self-determination considered contain approximately the same amount of

10 Here V should be negative, whereas in the democratic case the explanation for the model suggested that V is a cost that would be positive.

agriculturally suitable land. This requires that some units be much larger than others in terms of surface area. If both actual and simulated boundaries were distributed randomly across space, this would result in a spurious correlation between the actual and simulated boundaries, because there would be more boundaries between units that were farther apart, both in the actual and simulated data. It is not possible to eliminate this bias, given the computational constraint that requires that there be no units in the simulation that have particularly small amounts of agriculturally suitable land.

To ensure that this bias is not responsible for the results displayed in Figures 1 and 2, we conduct a statistical analysis that includes as control variables functions of the distance between the units in question. Specifically, consider the regression

$$\mathrm{BOUNDARY.ACTUAL}_{i,i'} = \beta_0 + \beta_1 \mathrm{BOUNDARY.SIMULATED} + \beta_2 \mathrm{DIST}_{i,i'} + \epsilon_{i,i'}$$

where $\mathrm{DIST}_{i,i'}$ is the geographic distance between i and i', calculated as described in the Methods section, and $\mathrm{BOUNDARY.ACTUAL}_{i,i'}$ is the data from Figure 1: the fraction of the time there is a boundary between i and i'. Similarly, $\mathrm{BOUNDARY.SIMULATED}_{i,i'}$ is the data from Figure 2.

BOUNDARY.ACTUAL is a limited dependent variable, as it is always between 0 and 1. We thus also consider a logistic regression. In addition, we consider similar regressions where $\beta_3 \mathrm{DIST}^2_{i,i'}$, and potentially also a cubic term, are included.

Table 1 shows the results of this analysis. The estimate of the coefficient on BOUNDARY.SIMULATED is reduced somewhat by including the distance term. Additional higher order terms do not have any statistically significant effect on the coefficient estimate. In all columns it is clear that there is still a statistically significant relationship between the actual boundaries and the simulated boundaries, with a *t*-value of at least 8.

The *t*-values associated with the results in Table 1 might appear low given a qualitative comparison of Figures 1 and 2. A closer look at these figures, however, reveals that in several important cases the simulated results give a boundary that is very close, but not exactly the same, as the actual boundary. For example, most of the simulated boundary in the Pyrenees is slightly to the south of the actual boundary, while the simulated boundary between Germany and Poland is to the west of the (average historical) actual boundary. The results shown in Table 1 only compare simulated boundaries and actual boundaries for exactly the same i and i' pair. Thus, simulated boundaries that are qualitatively correct, but not in precisely the right place, are not accounted for in the coefficient estimates presented in the table.

One might wonder whether the statistics reported in Table 1 are themselves uninformative, because due to the particular geographic shape of Europe, certain countries will tend to emerge regardless of how agricultural land is distributed. In this case the proposed model is inappropriate, as it contributes nothing beyond an even simpler model such as one following Jetz and Rahbek [2001]. To check this we maintain the same geographic base units for our simulation (as shown in Appendix Figure 4), but randomly shuffle total agricultural suitability among units. That is, a given unit i has a 1/699 probability of receiving the suitability of each of the other 699 units. We generate 100 such randomly shuffled versions of Europe, and compute a simulated partition for each of these. We then rerun the analysis shown in Table 1 for these randomly shuffled Europes. The results are reported in Table 2. After adding polynomial controls for the distance between units, the simulation results have no statistically significant relationship with the actual boundaries of Europe. This suggests that the statistically significant results reported in Table 1 are not merely due to the "shape" of the units involved in the coalition formation game considered, as for example shown in Appendix Figure 10.

A remaining possibility is that the statistically significant results reported in Table 1 are merely the result of the fact that real countries generally consist of geographically contiguous territory, and in many cases have a roughly convex shape. The model presented will tend to produce simulated countries that are geographically contiguous and roughly convex, but if this alone is responsible for the results in Table 1, then a much simpler model could be proposed that has the same explanatory power.

To test this possibility, we construct random coutries based on Voronoi cells. We choose 23 units as "generating points", and make 23 countries by assigning each of the 699 units shown in Appendix Figure 10 to the closest of these generating points. The resulting Voronoi cells will be convex. We generate 1000 of these "Voronoi partitions" of Europe, starting each time with a new random set of 23 generating points. We then redo the analysis in Table 1, using these Voronoi partitions. The results of this analysis are shown in Table 3. Again, we find that after controlling for a polynomial of geographic distance, there is no statistically significant relationship between the simulated boundaries and the actual boundaries of Europe. This suggests that the results of Table 1 are not due to basic geometric properties of actual countries.

Table 1: Regression Results

	Dependent variable: BOUDARY.ACTUAL (fraction of time there is a boundary in the Euratlas data)							
	OLS				*Logistic*			
	I	II	III	IV	V	VI	VII	VIII
BOUNDARY. SIMULATED	0.294***	0.212***	0.199***	0.199***	1.810***	1.318***	1.163***	1.150***
	(0.022)	(0.024)	(0.024)	(0.024)	(0.074)	(0.081)	(0.084)	(0.084)
DIST		−0.001***	−0.001***	−0.001***		−0.005***	−0.010***	−0.015***
		(0.0001)	(0.0002)	(0.0004)		(0.0004)	(0.001)	(0.002)
DIST^2			0.00000**	0.00000			0.00001***	0.00004***
			(0.00000)	(0.00000)			(0.00000)	(0.00001)
DIST^3				0.000				−0.00000***
				(0.000)				(0.000)
constant	0.595***	0.738***	0.774***	0.773***	0.202***	1.088***	1.531***	1.759***
	(0.019)	(0.025)	(0.029)	(0.032)	(0.060)	(0.086)	(0.106)	(0.130)
Observations	1,837	1,837	1,837	1,837	1,837	1,837	1,837	1,837
R^2	0.090	0.123	0.126	0.126				
Adjusted R^2	0.089	0.122	0.124	0.124				

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2: Regression Results (shuffled agricultural suitability)

	Dependent variable: BOUDARY.ACTUAL (fraction of time there is a boundary in the Euratlas data)							
	OLS				*Logistic*			
	I	II	III	IV	V	VI	VII	VIII
BOUNDARY. SIMULATED	0.260***	0.048	0.008	0.008	1.751***	0.464***	0.072	0.032
	(0.033)	(0.039)	(0.040)	(0.040)	(0.117)	(0.141)	(0.144)	(0.144)
DIST		−0.001***	−0.002***	−0.002***		−0.007***	−0.014***	−0.020***
		(0.0001)	(0.0002)	(0.0004)		(0.0004)	(0.001)	(0.002)
DIST^2			0.00000***	0.00000			0.00002***	0.0001***
			(0.00000)	(0.00000)			(0.00000)	(0.00001)
DIST^3				0.000				−0.00000***
				(0.000)				(0.000)
constant	0.624***	0.901***	0.978***	0.976***	0.226**	1.918***	2.711***	3.001***
	(0.028)	(0.038)	(0.043)	(0.046)	(0.095)	(0.139)	(0.158)	(0.177)
Observations	1,837	1,837	1,837	1,837	1,837	1,837	1,837	1,837
R^2	0.032	0.085	0.093	0.093				
Adjusted R^2	0.031	0.084	0.092	0.091				

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3: Regression Results (randomly generated Voronoi cells)

	Dependent variable: BOUDARY.ACTUAL (fraction of time there is a boundary in the Euratlas data)							
	OLS				*Logistic*			
	I	II	III	IV	V	VI	VII	VIII
BOUNDARY. SIMULATED	0.500*** (0.053)	0.137** (0.066)	0.048 (0.070)	0.052 (0.072)	3.485*** (0.195)	1.310*** (0.240)	0.512** (0.250)	0.314 (0.256)
DIST		−0.001*** (0.0001)	−0.002*** (0.0002)	−0.002*** (0.0005)		−0.006*** (0.0004)	−0.013*** (0.001)	−0.019*** (0.002)
DIST^2			0.00000*** (0.00000)	0.00000 (0.00000)			0.00002*** (0.00000)	0.00005*** (0.00001)
DIST^3				0.000 (0.000)				−0.00000*** (0.000)
constant	0.431*** (0.044)	0.822*** (0.061)	0.940*** (0.068)	0.932*** (0.075)	−1.159*** (0.156)	1.183*** (0.218)	2.298*** (0.244)	2.717*** (0.274)
Observations	1,837	1,837	1,837	1,837	1,837	1,837	1,837	1,837
R^2	0.046	0.086	0.093	0.093				
Adjusted R^2	0.045	0.085	0.092	0.091				

Note: *p<0.1; **p<0.05; ***p<0.01

Figure 3: Agricultural Suitability

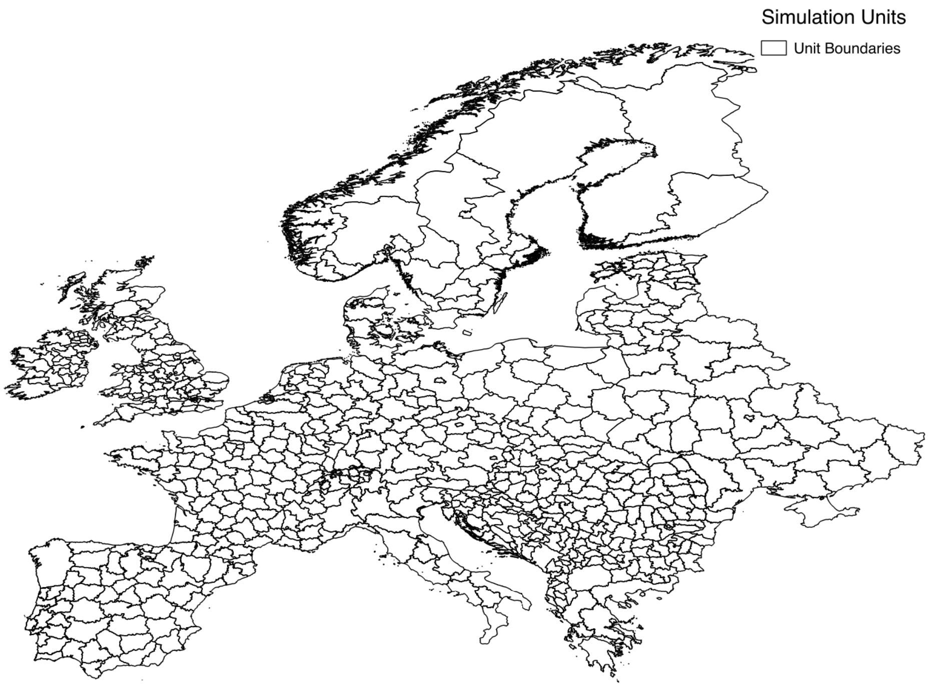

Figure 4: Units of Self-Determination for Simulation

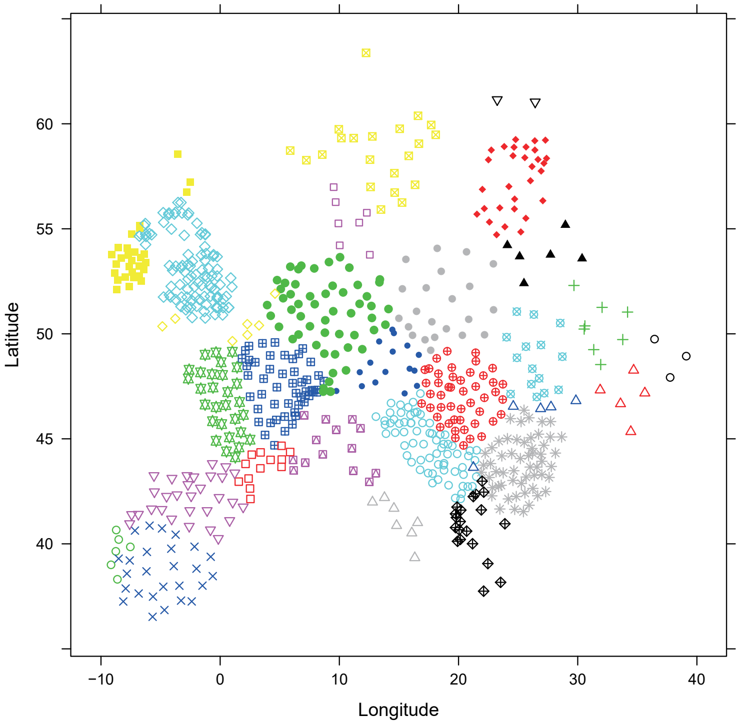

Figure 5: A Core Partition (coloured symbol indicates country)

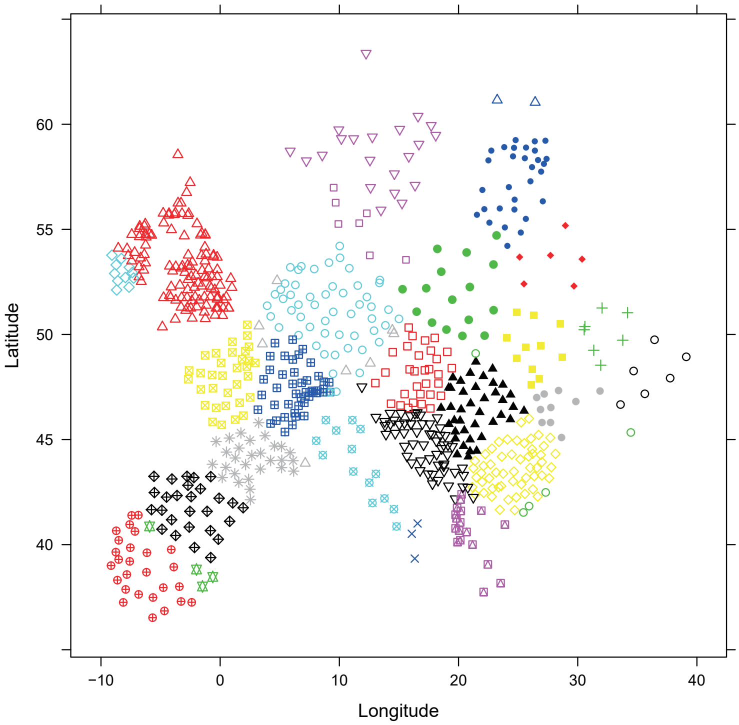

Figure 6: Another Core Partition (symbols have no relationship to those in previous figure)

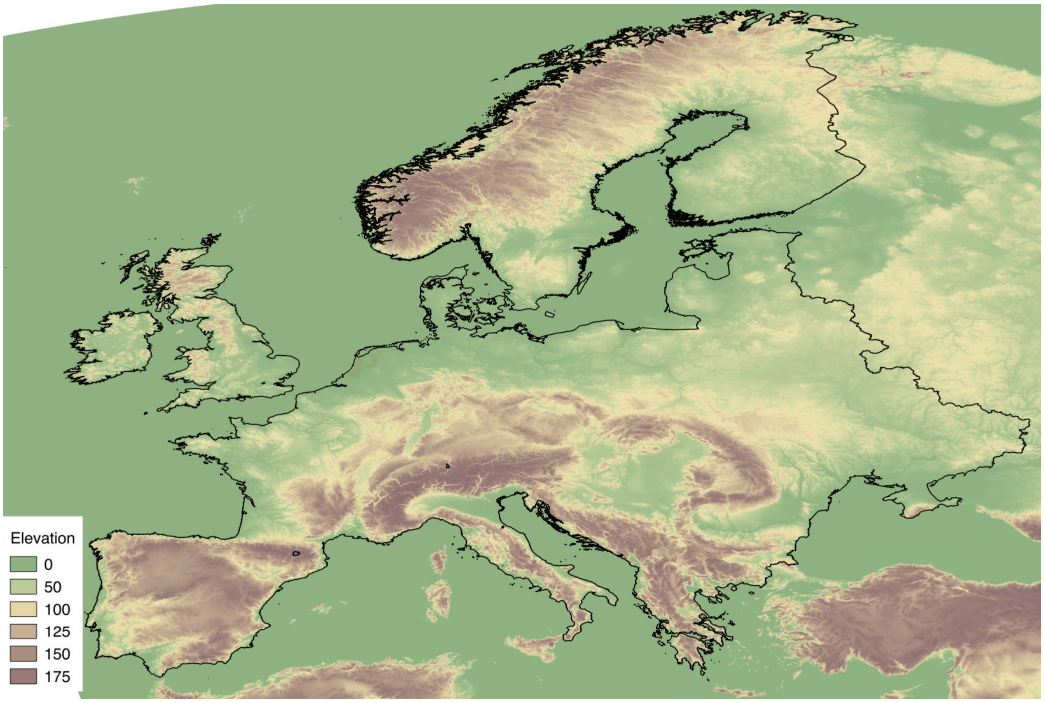

Figure 7: Elevation (GTOPO30 data, processed by European Environmental Agency)

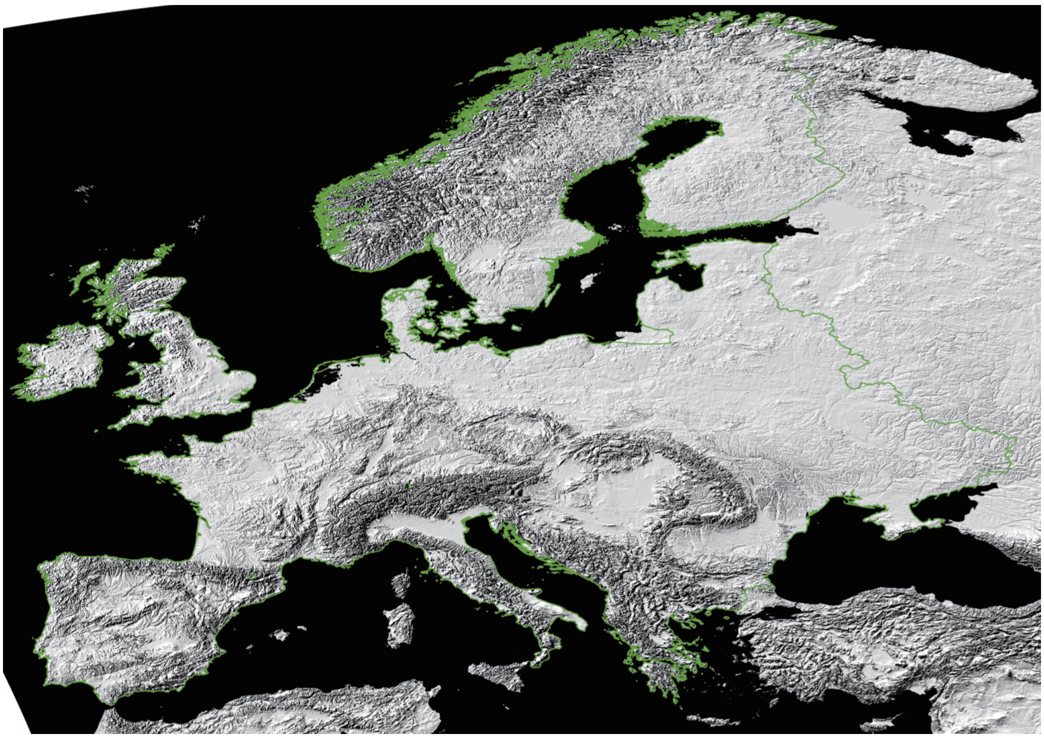

Figure 8: Hillshade (GTOPO30 data, processed by European Environmental Agency)

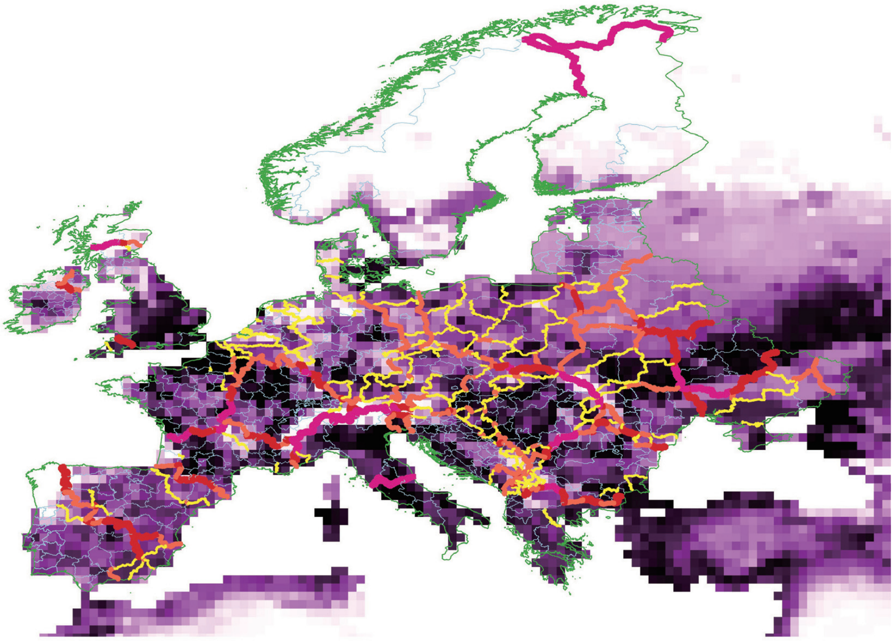

Figure 9: Agricultural Suitability and Simulated Boundaries

Figure 10: Units (vertices) and Geographic Adjacency (edges)

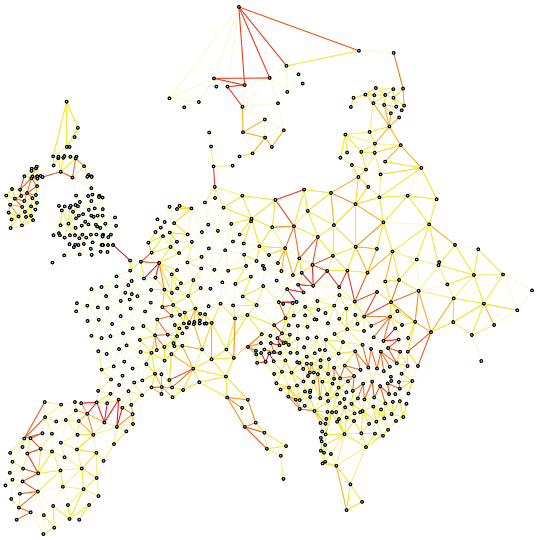

Figure 11: Actual Boundaries (different representation of same data as Figure 1, plus "water" edges)

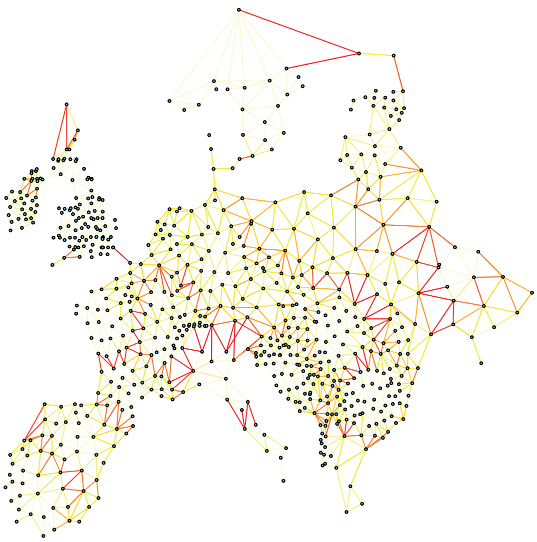

Figure 12: Simulated Boundaries (different representation of same data as Figure 2, plus "water" edges) Legend: a dark red edge is one where the incident vertices are always in different countries; a white edge is one where they are always in the same country. An uncoloured version of the graph is shown in Appendix Figure 10.

References

Bixby, Robert E. (2012). "A Brief History of Linear and Mixed-Integer Programming Computation". *Documenta Mathematica* Extra Volume ISMP: 107–121.

Brandl, Florian, Felix Brandt, and Martin Strobel (2015). "Fractional Hedonic Games: Individual and Group Stability". *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. AAMAS '15. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, pp. 1219–1227.

Gale, D. and L. S. Shapley (Jan. 1962). "College Admissions and the Stability of Marriage". *The American Mathematical Monthly* 69(1): 9–15.

Jetz, Walter and Carsten Rahbek (Aug. 5, 2001). "Geometric Constraints Explain Much of the Species Richness Pattern in African Birds". *Proceedings of the National Academy of Sciences* 98(10): 5661–5666.

Roth, Alvin E. and John H. Vande Vate (Nov. 1990). "Random Paths to Stability in Two- Sided Matching". *Econometrica* 58(6): 1475–1480.

Scheffe, Henry (1959). *The Analysis of Variance*. A Wiley publication in mathematical statistics. New York: Wiley. 447 pp.

Weese, Eric, Masayoshi Hayashi, and Masashi Nishikawa (Oct. 2020). "Inefficiency and Self- Determination: Simulation-Based Evidence from Meiji Japan". *CIRJE Discussion Paper F-989* (1627): 93 pages.