Regional Integration Clusters and Optimum Customs Unions: A Machine Learning Approach

Philippe De Lombaerde[*] Dominik Naeher[+] Takfarinas Saber[‡]

Abstract

This paper proposes a new method to evaluate the composition of regional arrangements focused on increasing intraregional trade and economic integration. In contrast to previous studies which take the country composition of these arrangements as given, our method uses a network clustering algorithm adapted from the machine learning literature to identify, in a data-driven way, those groups of neighboring countries that are most integrated with each other. Using the obtained landscape of regional integration clusters (RICs) as benchmark, we then apply our method to critically assess the composition of real-world customs unions. Our results indicate that there is considerable variation across customs unions as to their distance to the RICs emerging from the clustering algorithm, suggesting that some customs unions are relatively more driven by 'natural' economic forces, as opposed to political considerations. Our results also point to several testable hypotheses related to the geopolitical configuration of customs unions.

Keywords: Regional integration, Customs union, Machine learning

JEL Codes: C60, F13, F15, F60

---

[*] Neoma Business School, Rouen, France, and United Nations University Institute on Comparative Regional Integration Studies (UNU-CRIS), Bruges, Belgium. Email: PDeLombaerde@cris.unu.edu

[+] School of Economics, University College Dublin, Ireland. Email: dominik.naeher@ucd.ie

[‡] Lero–Science Foundation Ireland Research Centre for Software, School of Computing, Dublin City University, Ireland. Email: takfarinas.saber@dcu.ie

1 Introduction

A growing body of literature measures and compares regional economic integration outcomes across regions and subregions of the world.[1] A common problem in comparing the results presented in different studies in this literature is that the underlying classifications (i.e., groupings) of countries into regions and subregions vary widely across studies.[2] This is particularly problematic because the quantitative measures of regional integration used in these studies tend to be very sensitive to the size and composition of the underlying country groupings (De Lombaerde et al., 2010; Hamanaka, 2015). In addition, individual country scores on regional integration indicators critically depend on the *ex ante* defined regions. Countries can be peripheral in one region, but central in another. It is, in general, thus unclear to what extent the findings in this literature represent general results or to what extent they are driven by differences in the underlying country groupings. Moreover, this issue is further exacerbated by the fact that it is usually not possible to quantify the robustness of results with respect to alternative country groupings as (i) there exists no standard classification of regions and subregions in the economic literature that could be used as a benchmark grouping, and (ii) the number of possible alternative groupings that would need to be considered in the absence of such a benchmark is usually too large to allow for a comprehensive assessment of the robustness of the results obtained based on any given grouping.

This paper addresses this gap in the literature by proposing a new method to identify, in a data-driven way, those groups of (bordering) countries that are most integrated with each other, essentially endogenizing the choice about which region a country belongs to in terms of economic integration. Our proposed method is based on a network clustering algorithm commonly used in machine learning, which we adapt to be used with data on economic flows between countries. The resulting landscape of country groups, which we call *regional integration clusters* (RICs), captures the empirical structure of regional economic linkages in the data. As we demonstrate below, the RICs provide a useful benchmark for assessing the composition of regional arrangements, such as customs unions and regional trade agreements, focused on increasing intraregional trade and economic integration among a set of neighboring countries.

Of course, which regional grouping is the "correct" one to use depends on the specific application at hand. For example, if a regional development bank is interested in assessing how integrated its member states are compared to countries in other parts of the world, then it seems natural to use the regional classification also underlying the bank's overall strategy framework

[1] Besides the large literature on trade integration, many recent studies measure economic integration using a composite index which aggregates multiple indicators of economic integration such as trade, migration, foreign direct investment, and other cross-border links (De Lombaerde et al., 2008; AfDB, 2016; Rayp and Standaert, 2017; Huh and Park, 2018; Naeher and Narayanan, 2020; UNESCAP, 2020).

[2] For example, each of the studies cited in footnote 1 uses a different classification of regions.

(e.g., based on membership status). The questions we are addressing, however, are of a more general type and relate to the 'natural' grouping of countries into regions in terms of empirical cross-border economic relations. Specifically, assuming we wanted to map each country to exactly one region, what are the sizes and composition of groups of neighboring countries that best reflect the structure of regional economic linkages in the data for any given measure of regional integration and number of regions? We provide an answer to these questions by defining a measure to quantify how well a complete mapping of countries into regions (i.e., each country is mapped to exactly one region) matches the landscape of regional economic linkages in the data, and then solving for the mapping that yields the maximum match. In doing so, we focus on regional integration in terms of actual economic flows (i.e., actual economic integration as defined in Mongelli et al., 2005) rather than institutional or cultural integration. Our results are neutral in the sense that they do not rely on any specific theory of regional integration, nor do they by themselves provide insights on the potential benefits of economic integration to growth or welfare. Put simply, we measure the concentrations of cross-border economic activity among neighboring countries and then apply a network clustering algorithm that regionalizes the global network of countries into regional clusters matching the structure of the observed integration concentrations in the data.

The intuition behind our algorithm can be described as follows. Starting with an initial list of all countries, where each country constitutes its own "region", the algorithm iteratively merges those pairs of bordering regions that are associated with the largest integration scores (according to some measure of regional economic integration, e.g., intraregional trade shares) among all possible pairs of bordering regions. The outcome after s steps, i.e., s times merging two regions, is a set of endogenously determined regional clusters (the RICs) capturing the structure of regional economic linkages in the data. The RICs can therefore be thought of as a hypothetical benchmark grouping representing those groups of countries that are most integrated with each other according to the considered measure of regional integration. With such a benchmark at hand, several interesting questions can be addressed. Most importantly, using the RICs as a benchmark offers a way to evaluate how adequate regional arrangements are from an economic perspective, e.g., how much real-world customs unions or free trade areas are aligned with empirical trade intensities among the participating countries. For example, if one finds that the participating countries in such an arrangement do in fact benefit relatively little because they trade only little among each other, then this might give rise to the perception that other factors, such as political considerations, must be driving the arrangement.

In addition, our results also provide insights about which economic block individual countries belong to in terms of actual economic integration. There are several ways in which these insights can be useful. First, national policymakers often face choices about which regional economic arrangement (e.g., customs union) their country should join (and which not). If there are benefits from regional economic integration (Baldwin and Venables, 1995;

Henrekson et al., 1997; Fernández and Portes, 1998; Te Velde, 2011), and joining a group of countries with stronger economic ties is associated with larger benefits than joining a group of countries with weaker economic ties, then the optimal choices will depend on how much in line the composition of each group is with the empirical integration intensities, i.e., to what extent the countries forming such a regional arrangement are indeed economically interlinked with each other. Our analysis thus provides insights on an important factor in this context which can help to guide decision making. Similarly, our results may be useful for policymakers in member states of regional arrangements who have to decide whether to let a non-member country join their arrangement, or which of multiple countries interested in joining they should prioritize. In these contexts, economists would often analyze individual countries' largest trading partners to help guide decision making. Our proposed method facilitates a similar type of analysis but focusing on linkages at the regional level (between multiple countries) rather than bilateral links (between pairs of countries). Our method is therefore particularly suited to address questions related to the regional structure of economic integration, e.g., questions like "Is the Turkish economy rather part of an Asian economic block or of an European economic block?", "Do the economies of the Association of Southeast Asian Nations (ASEAN) form a cohesive group or are some of them more strongly tied to the Chinese economy?", and "Would South Sudan gain more from joining the Economic and Monetary Community of Central Africa (CEMAC) or from joining the East African Community (EAC)?"

To demonstrate the intuition and usefulness of the proposed approach, we apply our method to assess the composition of real-world customs unions (CUs). The results indicate that existing CUs differ considerably in their relative distance to the RICs that emerge from the clustering algorithm. The relative distance scores can be interpreted as indicators of the 'natural' or otherwise 'political' nature of each customs union. Our results also point to a number of testable hypotheses related to the geopolitical configuration of customs unions.

The rest of the paper is organized as follows. Section 2 discusses related literature, including further background on customs unions. Section 3 provides a formal definition of the problem we solve. Section 4 presents our proposed algorithm for identifying regional integration clusters in a data-driven way and explains how it can be implemented using real-world data. Section 5 applies the method to evaluate the composition of existing customs unions. Section 6 concludes.

2 Related Work

Our paper speaks to the international economics literature where the idea of optimal extensions of regional arrangements has been suggested. This idea of optimality of regional areas can be understood as a special case of the search for an optimal level of government intervention in the economy for the provision of public goods (Tinbergen, 1965; Kindleberger, 1986; Cooper, 1995). Optimum currency areas (OCA) are an obvious and obligatory reference point. In Mundell's seminal paper (Mundell, 1961) and the OCA theory which was consequently developed, the sterile debate on fixed versus flexible exchange rates was questioned and the idea was brought to the fore that 'optimal' sizes exist for regional groupings within which it is welfare superior to adopt fixed exchange rates, while maintaining flexible rates with the rest of the world. The size of the OCA can be determined by different (combinations of) criteria, including factor mobility (Mundell, 1961), openness (McKinnon, 1963), diversification of productive structures (Kenen, 1969) or trade intensity and correlation of business cycles. The complementarity between monetary integration and trade integration was analyzed by Méndez-Naya (1997).

In customs union theory, it is well known that the welfare effects of creating such unions depend on the relative importance of trade creation and trade diversion (Viner, 1950), whereby the net effect can theoretically be negative. Even if multilateral trade liberalization is optimal according to neo-classical orthodoxy, customs union theory provides a criterion to compare customs unions mutually on the basis of their relative welfare effects. In a trade policy context, the concept of 'natural markets' has emerged. A 'natural market' is defined then as a (regional) market characterized by net trade creation, i.e., by a net welfare increasing effect (Jacquemin and Sapir, 1991; Krugman, 1991). The suggested linkage between trade intensity and optimality in this approach is directly relevant for our purposes. Again, various criteria have been proposed, however, to determine the optimality of such markets. Whereas Krugman (1991) proposes a criterion based on the level of *ex ante* trade flows, Kreinin and Plummer (1994) propose a criterion based on trade patterns and *ex ante* trade distortions. Evaluations of free trade areas or customs unions in terms of whether they can be considered as 'natural' or not, depend highly on the underlying model (Nitsch, 1997; Frankel et al., 1998). A methodological problem that has been signaled in this context is the endogeneity problem (Frankel and Rose, 1998). *Ex ante* measures are not necessarily conclusive to evaluate the optimality conditions of a 'region'. These conditions can be met *ex post*.

Our paper also relates to the literature applying network analysis techniques to trade flows, and more specifically the literature that applies community-detection techniques to the global trade network to identify clusters of intensely trading countries (Fortunato, 2010; Barigozzi et al., 2011; Piccardi and Tajoli, 2015). However, the demarcation of these clusters is often not statistically significant as extra-regional ties are usually relatively important (Piccardi and Tajoli, 2012). It is also found that the identified clusters do not necessarily overlap very well with the country groupings bound together by trade agreements (De Lombaerde et al., 2018). This echoes the fact that the empirical trade literature, mostly based on gravity-type estimations, tends not to find clear evidence of the trade effect of preferential trade agreements. At best, small (and not always significant) positive effects are found (Cardamone, 2007; Baier and Bergstrand, 2009; Cipollina and Salvatici, 2010; Nguyen, 2019). And studies that focus on specific trade agreements tend to show even lower trade effects (Mordonu et al., 2011).

3 Problem Definition

Consider a set of countries $C = \{c_1, c_2, .., c_n\}$ and their set of borders B, with $b_{i,j} \in B$ being the border for any two bordering countries c_i and c_j in C^2.[3] A region r is defined as a set of bordering countries. For ease of exposition, we consider every country $c_i \in C$ to be its own region $r_{\{c_i\}}$ (i.e., a region containing only a single country). Whenever two regions r_A and r_B are merged with each other, they form a new region $r_{A \cup B}$ comprising all countries in r_A and r_B. Throughout this paper, we only consider the possibility of regions to be merged if they border each other. Two regions r_A and r_B are bordering each other if there exists at least one country $c_i \in r_A$ which shares a border with at least one country $c_j \in r_B$.

A grouping R is defined as a set of regions. Let R_s be the set of all regions at step s, where a step corresponds to the merging of two regions. Initially, each country forms its own region, so that R_0 is given by

$$R_0 = \left\{ r_{\{c_i\}},\ \forall\, c_i \in C \right\} \tag{1}$$

Starting with some grouping R_s, merging two regions $r_A \in R_s$ and $r_B \in R_s$ means that r_A and r_B are removed from R_s and replaced by a new region $r_{A \cup B}$. Thus, the set of regions in the next step, R_{s+1}, will contain one element less than R_s (namely all the regions in R_s except r_A and r_B, and adding $r_{A \cup B}$). Notice that this ensures that R_s provides a complete grouping of countries into regions, i.e., every country in C is part of exactly one region at every step.

Every region r_k is associated with an integration score $S(r_k)$ (except single-country regions). In principle, $S(r_k)$ may be any quantitative measure of regional integration. For example, in the application we discuss below, $S(r_k)$ is the intraregional trade share (normalized by the product of GDPs)[4] given by

$$S(r_k) = \frac{\sum_{(c_i,c_j) \in r_k, c_i \neq c_j} Trade(c_i,c_j)}{\left(\sum_{c_i \in r_k} \sum_{c_j \in C} Trade(c_i,c_j)\right) \cdot \left(\prod_{c_i \in r_k} GDP(c_i)\right)} \tag{2}$$

where $Trade(c_i, c_j)$ is the sum of exports and imports between countries c_i and c_j, and $GDP(c_i)$ is country c_i's gross domestic product.

[3] We consider borders to be symmetric, i.e., $b_{i,j} \in B \Longrightarrow b_{j,i} \in B,\ \forall\, (c_i, c_j) \in C^2$.

[4] The product of GDPs was chosen, rather than the sum of GDPs, because it reflects better the trade potential, especially in cases of asymmetric trade partners.

Research questions. We are interested in (i) identifying those groups of bordering countries that are most integrated with each other, and (ii) obtaining a quantitative measure of how well a given grouping R^f matches the empirical structure of regional economic linkages in the data, i.e., to what extent the regions in R^f comprise countries that are strongly integrated with each other. More specifically, we want to quantify the degree to which the regions in R^f overlap with the regions in another grouping R^*, where R^* is the grouping of countries into regions that maximizes overall regional integration according to some objective function $F(S(r_k), B)$.

4 Proposed Method: Identifying Regional Integration Clusters in a Data-Driven Way

Based on the definitions above, a possible approach to answering our research questions would be to (i) specify an objective function $F(\cdot)$ and a measure of regional integration $S(r_k)$, (ii) write down the problem as a constraint optimization problem with the real-world borders between countries as B and data on the factors included in $S(r_k)$ (e.g., the economic indicators used in equation 2) as parameter values, and then (iii) solve for the grouping of countries into regions R^* that maximizes the objective subject to the constraints.

However, rather than explicitly specifying $F(\cdot)$ and computing R^* directly, we propose a different approach here. Specifically, we propose to obtain R^* by identifying those groups of bordering countries that are most integrated with each other using a network clustering algorithm commonly applied in machine learning. There are two main advantages to this approach. First, specifying an objective based on which R^* is to be calculated would require us to make choices about the functional form and parameterization of the objective. To the best of our knowledge, these choices would have to be ad-hoc, i.e., there is no theory or empirical evidence that could be used to guide these choices. As we show below, using a machine learning approach to compute R^* allows us to avoid having to make these choices. Second, even for a simple objective function (e.g., maximizing the sum of $S(r_k)$ across all regions r_k), solving for the optimal grouping R^* would be very computationally demanding and, for most relevant applications in the context of regional integration, in fact impossible to compute with contemporary technology.[5]

[5] For example, in the map of the world that is used in our experiments (containing 200 countries and their respective land and maritime borders), there are 485, 2k$^+$, 11k$^+$, and 100k$^+$ possible regions of size 2, 3, 4, and 5 respectively, and this number is growing exponentially with the increase in region size. Thus, the number of possible groupings that would have to be calculated to solve for the grouping with the maximum regional integration score is too large.

Figure 1: Clustering Algorithm

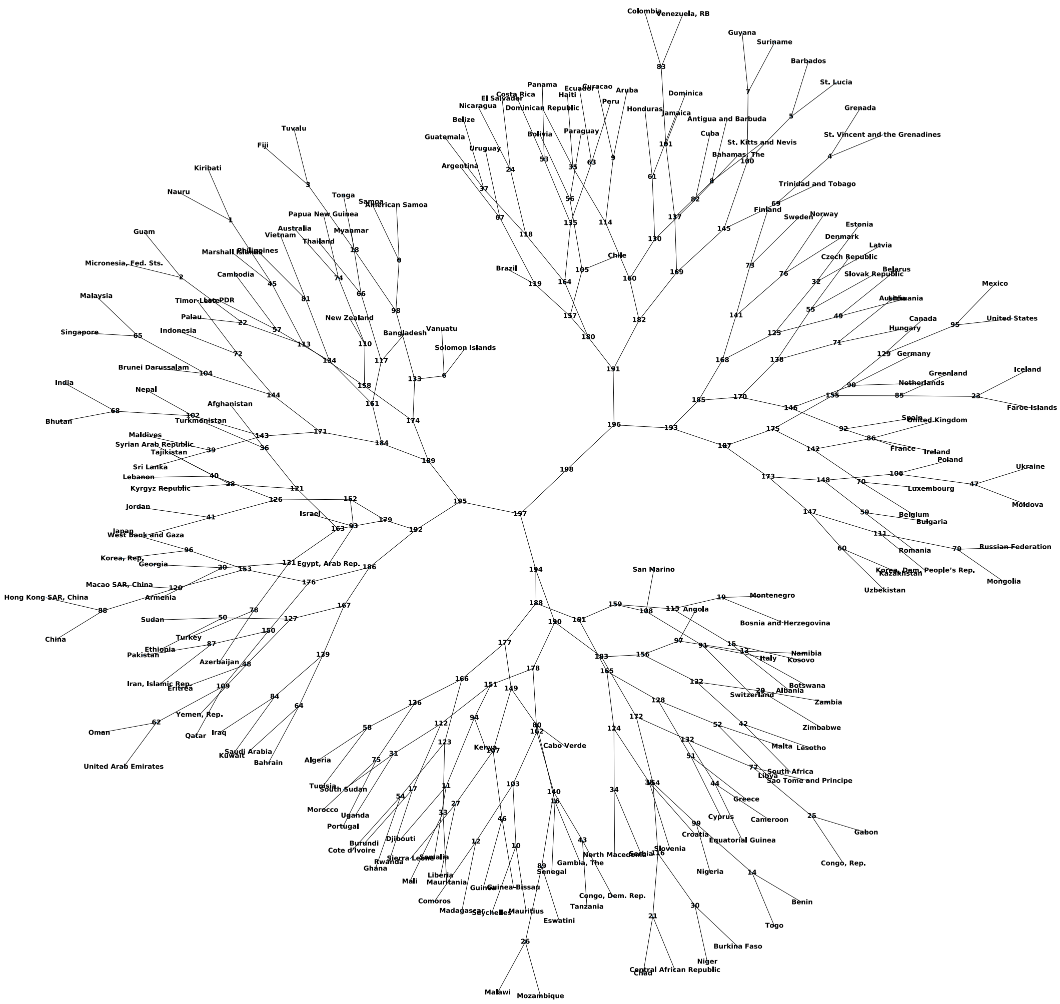
Input: R, B, Imports, Exports, GDPs, Stopping Criteria
Output: R
while Stopping Criteria Not Met **do**
BestMerge $\leftarrow (\varphi, \varphi)$
BestS $\leftarrow 0$
for $(r_A, r_B) \in R^2$ **do**
if $A \neq B$ and $\exists (c_i, c_j) \in A \times B \mid b_{i,j} \in \mathrm{B}$ **then**
if $S(r_{A \cup B}) >$ BestS **then**
BestMerge $\leftarrow (r_A, r_B)$
BestS $\leftarrow S(r_{A \cup B})$
end if
end if
end for
if BestS > 0 **then**
$R \leftarrow R \backslash \{r_{\text{BestMerge[1]}}, r_{\text{BestMerge[2]}}\}$
$R \leftarrow R \cup \{r_{\text{BestMerge}[1] \cup \text{BestMerge}[2]}\}$
end if
end while

Algorithm. Figure 1 describes the algorithm we propose for computing R^*. For a given initial list of regions, the algorithm literately finds the best two bordering regions to merge (based on their regional integration score $S(r_k)$) until a set of stopping criteria are met (in our case, the stopping criterion consists of the absence of unmerged bordering regions).[6] The key idea behind the algorithm is that R^* can be obtained in a data-driven way by starting with R_0 as defined in equation (1), and then subsequently merging those pairs of bordering regions that are associated with the largest regional integration scores among all remaining pairs of bordering regions. More specifically, starting with R_0 (i.e., every country forms its own region) the algorithm calculates the regional integration scores $S(r_k)$ resulting from merging any pair of bordering countries in R_0, and then selects the pair featuring the highest score. In the second step, the algorithm repeats the same procedure starting with the grouping R_1^*, which is obtained by merging the two countries selected in the previous step. Iteratively, the algorithm then merges pairs of bordering countries (or clusters of multiple bordering countries), reducing the total

[6] Note that in any case, these stopping criteria have to subsume the case where there are no more bordering regions in R^* able to merge.

number of regions contained in R^* by one in each step. The outcome after s steps (i.e., s times merging two regions) is a set of endogenously determined regional clusters capturing those groups of bordering countries that are most integrated with each other. We therefore call the resulting country groups in R_s^* regional integration clusters (RICs).

Figure 2: Tree Graph Showing Results of the Clustering Algorithm



Note: Results are based on the clustering algorithm defined in Figure 1 using the regional integration scores $S(r_k)$ defined in equation (2) and the data sources described in Section 5. Borders include both land and maritime borders (for a complete list of borders, see Figure A1 in the Appendix).

Figure 2 provides an illustration of the results generated by the clustering algorithm. The set of borders underlying these results contains both land and maritime borders, and the regional integration scores $S(r_k)$ are calculated as defined in equation (2) using the data sources described in Section 5. In the graph shown in Figure 2, the set of nodes at the outer end (which feature only a single edge) correspond to R_0, i.e., the set of individual countries. The numbers shown on the interior nodes of the graph (those featuring two or more edges) are the steps at which the corresponding clusters get merged. Intuitively, nodes closer to the center of the graph correspond to later steps of the algorithm.

In principle, the graph in Figure 2 contains the information about the country groups in R_s^* at all steps. In particular, the RICs at a particular step s can be obtained as the (disconnected) subgraph resulting from dropping all nodes in Figure 2 higher than s. Some examples of the landscape of RICs associated with individual steps are shown in Figure 3. Specifically, the four colored world maps shown in Figure 3 visualize the landscapes of RICs obtained at step 50, 85, 190, and 198 of the clustering algorithm as indicated in the figure. Each color represents a cluster of merged countries, while countries in white have not yet been merged at the respective step.

Notice that at step 50 in Figure 3, only a small subset of countries have been merged with others, and most of the countries which get merged early on are in Sub-Saharan Africa. This reflects the fact that the underlying measure $S(r_k)$ of regional integration (i.e., the intraregional trade shares defined in equation 2) essentially captures a bias towards intraregional trade as opposed to global trade with countries located in other parts of world. As many countries in Sub-Saharan Africa trade only relatively little with the rest of the world, and the bit of international trade they have takes place mostly with their immediate neighbors, $S(r_k)$ takes large values for these countries. The same intuition applies, in reverse direction, to the landscape of RICs at step 85 in Figure 3. The few countries that still form their own clusters at this step (those in white, e.g., China, Germany, and the U.S.) are countries which trade heavily at a global scale, i.e., including with countries other than their immediate neighbors. At step 190, the landscape of RICs shown in Figure 3 comprises 10 clusters, and all 200 countries in our dataset have been merged with at least one other country. The last colored world map in Figure 3 corresponds to the pre-final step (198) of the algorithm before its termination, which features a bipolar landscape consisting of two large clusters which together comprise all countries.

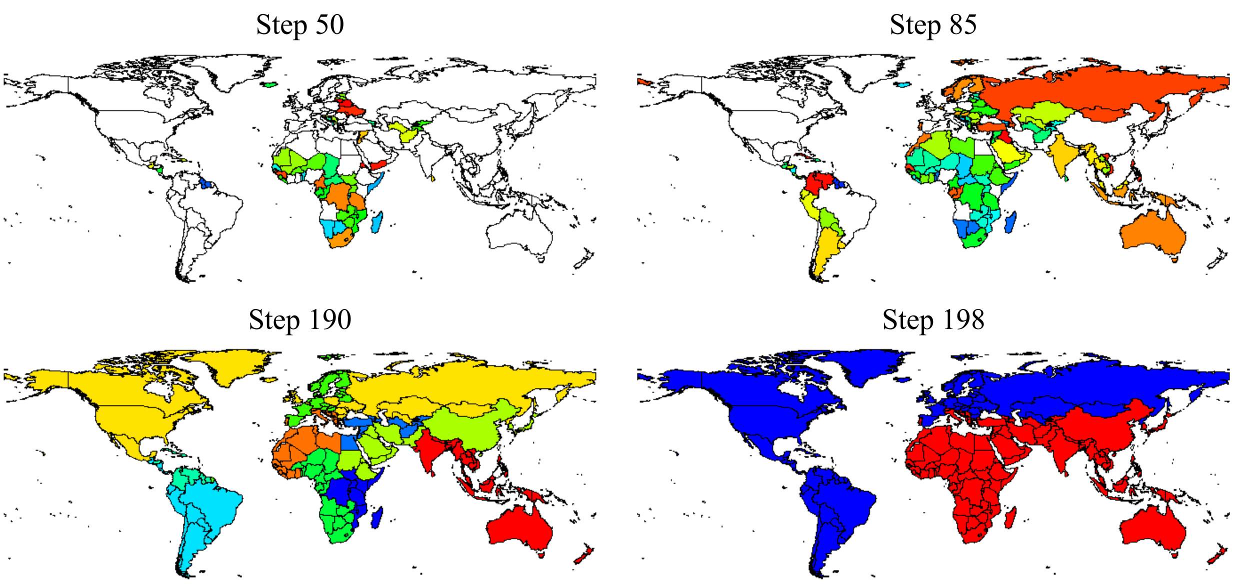
As the results for selected steps of the algorithm in Figure 3 illustrate, the clustering algorithm defined in Figure 1 generates groups of countries that represent the structure of regional economic linkages in the data, i.e., that, for each possible number of clusters, capture the sizes and compositions of those groups of neighboring countries which are most integrated with each other. The RICs therefore provide an answer to our first research question defined in Section 3. To answer our second research question, we use the RICs as the benchmark grouping

R^* and define a distance function (metric) that allows us to quantify the degree to which the regions in R^* overlap with the regions in any other grouping R^f. Specifically, we define the distance between two regions $r_A \in R^f$ and $r_A^* \in R^*$ as

$$d(r_A, r_A^*) = |r_A \backslash r_A^*| + |r_A^* \backslash r_A|, \tag{3}$$

where $|x|$ is the number of elements (countries) in region x, and "\" denotes the set subtraction operator (recall that each region is defined as a set of countries). In other words, the distance between two regions equals the number of countries that need to be removed plus the number of countries that need to be added from one of the regions to convert it into the other region. For example, if the distance between regions $r_A \in R^f$ and $r_A^* \in R^*$ is smaller than the distance between $r_B \in R^f$ and $r_B^* \in R^*$, then we will conclude that the composition of r_A matches the regional economic linkages in the data better than the composition of r_B. Note that the range of $d(r_A, r_A^*)$ is $[0, |r_A| + |r_A^*|]$. Since this distance function depends on the size of r_A, we also report results for a normalized distance measure obtained by dividing $d(r_A, r_A^*)$ by $|r_A|$, so that the range of the normalized distance is $\left[0, \frac{|r_A|+|r_A^*|}{|r_A|}\right]$.

Figure 3: Maps of RICs at Selected Steps of Clustering Algorithm



Note: Different colors represent different regional integration clusters (RICs) comprising at least two countries at a given step. Re (2) and the data sources described in Section 5. Borders include both land and maritime borders (for a complete list of borders, see Figure A1 in the Appendix).

Recall that the clustering algorithm defined in Figure 1 yields a different grouping of countries into regions at each step (corresponding to groupings with different numbers of regions). When calculating the distance between a region $r_A \in R^f$ and the corresponding benchmark region $r_A^* \in R^*$, the choice of r_A^* thus involves two decisions. First, at which step of the algorithm r_A^* is obtained, i.e., the choice of R_s^*, and second which particular region within R_s^* is used as the benchmark for r_A. Rather than making these choices ourselves, we endogenize both choices by using the region with the minimum distance to r_A out of all regions in R_s^* across all steps, i.e.,

$$r_A^* = \arg\min_{r^* \in R^*} d(r^*, r_A) \tag{4}$$

with $R^* = \bigcup_{i=0}^{n} R_i^*$, and n being the total number of steps in our algorithm. Note that this implies that the maximum possible distance between a region r_A and its corresponding benchmark RIC r_A^* will be $|r_A| - 1$.[7]

Given that this choice of r_A^* leads to the minimum possible distance between r_A and any element of R^*, the results can be interpreted as providing a lower bound estimate for how adequate the composition of r_A is from an empirical perspective. For example, if we find that *x*% of the countries in a real-world customs union r_A are not included in the corresponding benchmark RIC r_A^* obtained according to equation (4), then this suggests that at least *x*% of the countries in r_A could gain economically from being part of a different customs union than r_A.[8]

5 Application to Customs Unions

We now present an application of our method to customs unions (CUs). Contrary to free trade agreements (FTAs), these are, in principle, not overlapping and they are small in numbers. This contrasts with most of the empirical literature on trade effects of regional trade agreements which focuses on FTAs. In addition, much of the discussion on 'natural markets' deals with continental trade areas (Frankel et al., 1998). Customs unions are defined in paragraph 8(a) of Article XXIV of GATT 1994 as "*[...] the substitution of a single customs territory for two or more customs territories, so that (i) duties and other restrictive regulations of commerce (except, where necessary, those permitted under Articles XI, XII, XIII, XIV, XV and XX) are eliminated with respect to substantially all the trade between the constituent territories of the union or at least with respect to substantially all the trade in products originating in such territories, and (ii) subject to the provisions of paragraph 9, substantially the same duties and*

[7] As R^* includes all the single country regions in R_0^*, any r_A will feature a RIC with a distance no greater than $|r_A| - 1$ (i.e., if the RIC closest to r_A consists of a single member country of r_A).

[8] The results also represent lower bound estimates for a second reason. Due to the endogeneity of the optimality criterion used (see also the discussion in Section 2), which is based on intraregional trade intensity, existing arrangements between neighboring states (such as the customs unions studied in Section 5) will tend to show higher scores than alternative country groupings which have not benefited from the trade creating effect of being in a customs union.

other regulations of commerce are applied by each of the members of the union to the trade of territories not included in the union."

We include in our analysis the CUs that are notified to the WTO and currently (as of October 2020) in force. A list of the included CUs is contained in Table 1 (column 1), while their composition can be seen in column 2. Note that this list differs slightly from the longer list of notified customs unions in the WTO database for the following reasons:

- Our list does not include accessions separately as they collapse with the CU to which they refer.
- We do not include the EU-Andorra CU (entry into force: 01/07/1991) and the EU-San Marino CU (entry into force: 01/04/2002).
- COMESA is not included because, although its CU was formally notified to the WTO in 1995 under the enabling clause, it is still not operational.[9]
- The Russian Federation-Belarus-Kazakhstan CU (entry into force: 03/12/1997) is not included as it was absorbed by the EAEU.
- The West African Economic and Monetary Union (WAEMU) is not included as all its members are also part of ECOWAS.
- The EU-Turkey CU is not included as it is only a partial agreement, covering only industrial goods, and without a coherent common external tariff (De Lombaerde and Ulyanov, 2020).

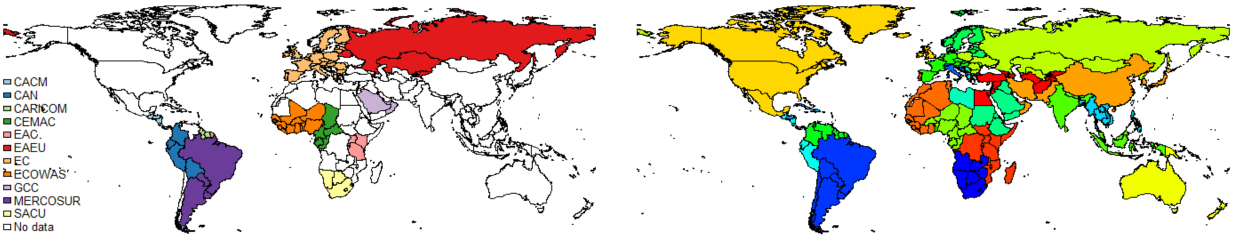
Also note that our analysis still refers to the EU-28. The UK was a Member State of the EU until January 2020, but the Withdrawal Agreement provided for a transition period during which the UK continues to be considered as a EU Member State for the purposes of relevant international agreements, including the customs union. Our analysis refers further to Mercosur-4. An accession protocol was signed between MERCOSUR member states and Venezuela in 2006, and the latter country became a full member in 2013. However, it was suspended in 2016.

In total, this gives us a set of 11 CUs. To assess the extent to which these CUs are 'natural' (i.e., in line with the empirical integration intensities, and opposed to being driven by other factors such as political considerations), we compare the composition of each CU to the landscape of RICs emerging from the clustering algorithm as described in Section 4. The underlying data on trade flows come from the IMF's Direction of Trade Statistics (DOTS) database. Data on GDP (used for normalizing) come from the World Bank and are measured in current US-dollars. To limit the role of temporary fluctuations and measurement error, the analysis is based on 5-year average values for all variables, corresponding to the period 2014-2018.

[9] See www.uneca.org/oria/pages/comesa-trade-and-market-integration (last accessed in October 2020).

Figure 4: Maps Showing Customs Unions and RICs

Customs Unions RICs at Step 180



Notes: Different colors represent different customs unions (left map) and regional integration clusters (RIC; right map). Results for the RICs are based on the clustering algorithm defined in Figure 1 using the regional integration scores S(rk) defined in equation (2) and the data sources described in Section 5. Borders include both land and maritime borders (for a complete list of borders, see Figure A1 in the Appendix).

To provide a first impression of the landscape of CUs we are working with, Figure 4 shows a map of the country groups forming the CUs as well as a map showing the landscape of RICs at step 180 of the algorithm. Importantly, it should be noted that this step only serves as an example featuring RICs of similar sizes than the CUs, on average. As discussed in more detail below, the RICs with the minimum distance to the CUs which will be used as benchmarks are in fact obtained at different steps for different CUs (and thus cannot be depicted in a single map).

The main results emerging from our analysis are reported in columns (3) to (7) in Table 1. Column (3) shows the minimum distance between the CU and the respective benchmark RIC. Column (4) shows the corresponding normalized distance, i.e., when dividing the distance in column (3) by the number of countries in the CU (shown in column 2). Column (5) reports the step of the algorithm at which the minimum distance is (first) reached.[10] Column (6) shows, on the one hand, the CU members that are part of the RIC with minimum distance, and, on the other hand, the (posterior) steps at which the other CU members become part of the same RIC. In addition, it is possible that the RIC with minimum distance (at the step when the minimum distance is first reached) contains countries which are not members of the corresponding CU. These countries are reported in column (7).

[10] Note that, once the minimum distance has been reached in the algorithm, the composition of the RIC can stay constant during a number of steps until a new member is added. In addition, our algorithm has the feature that the minimum distance can be reached both in different steps and with different configurations. For example, if in a given step two countries are added to a RIC where one country is also a CU member and the other is a third country, then the distance between the RIC and the corresponding CU remains the same. If, on the other hand, this happened in consecutive steps (e.g., first the third country is added, then the CU member), then the distance would first go up and then down again (this case seems to be a theoretical possibility which does not occur in our data).

Table 1: Results

Customs Union (CU)	CU Composition (# Countries)	Min. Dist. to RIC	Min. Dist. (Norm.)	Step with Min. Dist.	CU Countries in RIC with Minimum Distance	Other Countries in RIC with Min. Dist.
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Andean Community (CAN)	Bolivia, Ecuador, Colombia, Peru (4)	2	0.5	64	Ecuador (64), Peru (64), Bolivia (181), Colombia (192)	-
Caribbean Community and Common Market (CARICOM)	Antigua and Barbuda, Bahamas, Barbados, Belize, Dominica, Grenada, Guyana, Haiti, Jamaica, Montserrat (not included), Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Suriname, Trinidad and Tobago (14)	6	0.43	170	Saint Vincent and the Grenadines (5), Grenada (5), Trinidad and Tobago (70), Barbados (101), Guyana (101), Saint Lucia (101), Suriname (101), Antigua and Barbuda (170), Dominica (170), Saint Kitts and Nevis (170), Haiti (183), Jamaica (183), Bahamas (183), Belize (192)	Venezuela, Colombia
Central American Common Market (CACM)	El Salvador, Guatemala, Nicaragua, Honduras, Costa Rica, Panama (6)	4	0.67	25	El Salvador (25), Nicaragua (25), Costa Rica (119), Panama (119), Guatemala (136), Honduras (192)	-
East African Community (EAC)	Kenya, Tanzania, Uganda, Burundi, Rwanda (5)	3	0.6	18	Burundi (18), Rwanda (18), Uganda (113), Kenya (152), Tanzania (181)	-
European Community (EC)	Belgium, France, Germany, Italy, Luxembourg, Netherlands, Denmark, Ireland, UK, Greece, Portugal, Spain, Austria, Finland, Sweden, Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovak Republic, Slovenia, Bulgaria, Romania, Croatia (28)	16	0.57	186	Czech Republic (129), Slovak Republic (129), Austria (141), Hungary (141), France (171), Germany (171), Netherlands (171), Spain (171), Denmark (186), Estonia (186), Finland (186), Latvia (186), Lithuania (186), Sweden (186), Belgium (194), Bulgaria (194), Ireland (194), Luxembourg (194), Poland (194), Romania (194), UK (194), Croatia (199), Cyprus (199), Greece (199), Italy (199), Malta (199), Portugal (199), Slovenia (199)	Belarus, Norway
Economic and Monetary Community of Central Africa (CEMAC)	Cameroon, Central African Republic, Chad, Congo, Equatorial Guinea, Gabon (6)	3	0.5	133	Cameroon (117), Equatorial Guinea (117), Congo (133), Gabon (133), Central African Republic (173), Chad (173)	Sao Tome and Principe
Economic Community of West African States (ECOWAS)	Benin, Burkina Faso, Cabo Verde, Côte d'Ivoire, Ghana, Guinea, Guinea-Bissau, Liberia, Mali, Niger, Nigeria, Senegal, Sierra Leone, The Gambia, Togo (15)	8	0.53	150	Guinea (108), Guinea-Bissau (108), Liberia (108), Sierra Leone (108), Cabo Verde (150), The Gambia (150), Senegal (150), Côte d'Ivoire (178), Ghana (178), Mali (178), Benin (189), Burkina Faso (189), Niger (189), Nigeria (189), Togo (189)	-
Eurasian Economic Union (EAEU)	Belarus, Kazakhstan, Russian Federation, Armenia, Kyrgyz Republic (5)	4	0.8	0	Kazakhstan (148), Russian Federation (148), Belarus (194), Armenia (199), Kyrgyz Republic (199)	-
Gulf Cooperation Council (GCC)	Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, United Arab Emirates (6)	3	0.5	110	Oman (63), United Arab Emirates (63), Qatar (110), Bahrain (187), Kuwait (187), Saudi Arabia (187)	-
Southern African Customs Union (SACU)	Botswana, Eswatini, Lesotho, Namibia, South Africa (5)	3	0.6	14	Botswana (14), Namibia (14), Lesotho (157), South Africa (157), Eswatini (194)	-
Southern Common Market (MERCOSUR)	Argentina, Brazil, Paraguay, Uruguay (4)	1	0.25	120	Argentina (68), Uruguay (68), Brazil (120), Paraguay (158)	-

To illustrate how the results in Table 1 can be interpreted, consider the case of MERCOSUR as an example. This CU has four members (column 2) and the minimum distance to its benchmark RIC is reached in step 120 (column 5). In this step, three of the CU members (Argentina, Uruguay, and Brazil; see column 6) are part of the RIC, whereas Paraguay is only added later (in step 158). Also, there are no other countries in the benchmark RIC at step 120 (see column 7). Therefore, the distance is 1 (i.e., one country must be added to form MERCOSUR), and the normalized distance is 0.25 (1/4). Note that the result that the minimum distance is not reached at step 158, when Paraguay is also part of the RIC, is due to the fact that other countries (not members of MERCOSUR) join the RIC consisting of Argentina, Uruguay, and Brazil before step 158 is reached (these countries are not reported here). Also note that, due to the endogeneity of the optimality criterion used (recall the discussion in Section 4), our method tends to underestimate the minimum distance between CUs and RICs, so that the results should be interpreted as providing lower bound estimates of how 'natural' the CUs are.

Our results allow for an analysis at two levels: inter-block comparisons and intrablock case studies. As far as inter-block comparisons are concerned, normalized minimum distances range from 0.25 (MERCOSUR) to 0.8 (EAEU). The latter is actually the maximum distance possible. These results can be used as an indicator of how 'natural' (or, on the contrary, how 'political') each customs union appears to be. Whereas MERCOSUR emerges almost naturally from the trade intensities observed in the data, this is much less the case for the EAEU. In other words, the EAEU can be considered as 'less natural' (more 'political') than its Southern American counterpart.

In addition, our method generates a rich set of suggestive results that can form the basis of further in-depth (intra-block) case studies. While discussing these results in detail for all the CUs would clearly go beyond the scope of this paper, the following provides a list of some interesting findings that emerge from our analysis.

- Belarus is a member of EAEU but, according to our results, appears to belong rather to the RIC with minimum distance to the EU (step 186), which is interesting from a geopolitical perspective.
- There is proximity between Colombia and Venezuela, although the latter is no longer a member of CAN, which points to the political character of its withdrawal. It could also be further analyzed to what extent the detected proximity between Venezuela (and Colombia) and the Caribbean is the result of Venezuela's external strategy.
- Belgium's proximity to the UK, which is shown by the fact that Belgium joins a RIC with the UK before it joins the RIC with its other neighbors, points to its vulnerability in light of Brexit.
- The fact that Nigeria joins ECOWAS at a relative late step (189) could be linked to the fact that it is not playing a strong role as regional leader.

We will not develop these cases further here. Rather, we simply point out the demonstrated capacity of our methodology to lead to the formulation of interesting hypotheses for further study.

6 Conclusion

The literature on optimum currency areas investigates which groups of countries are expected to benefit the most from forming a common currency area. In a similar way, we investigate which groups of countries are expected to benefit the most from regional trade integration. This way, we also shed new light on the concept of 'natural markets'. Specifically, we argue that a machine-learning approach constitutes a valuable complementary tool to existing gravity-type econometric approaches to the evaluation of trade agreements.

To illustrate the intuition and usefulness of the proposed approach, we apply our method to the set of customs unions as notified to the WTO. The obtained results allow for an analysis at two levels. First, the results can be used for inter-block comparisons, assessing the relative extent to which existing custom unions emerge 'naturally' from the clustering algorithm, and therefore respond to an economic logic, or whether, on the contrary, they respond more to a political logic. In our application, MERCOSUR is closest to the former case, the EAEU closest to the latter. In addition, the method generates a rich set of results that can form the basis of intra-block case studies. A more detailed analysis along this line is left to future research.

In this paper, we focus on a method that can be used to evaluate the (sub-)optimality of non-overlapping regional arrangements in a cross-section of countries. Future work may find it useful to develop this approach further to also capture the dynamics of regional integration clusters over time, and to allow for evaluating overlapping arrangements such as FTAs.

Acknowledgement

Dr. Takfarinas Saber was supported, in part, by Science Foundation Ireland grant 13/RC/2094_P2.

References

AfDB (2016). *Africa Regional Integration Index: Report 2016*. African Development Bank, African Union, and United Nations Economic Commission for Africa.

Baier, S. L. and J. H. Bergstrand (2009). Estimating the effects of free trade agreements on international trade flows using matching econometrics. *Journal of International Economics 77*(1), 63–76.

Baldwin, R. E. and A. J. Venables (1995). Regional economic integration. *Handbook of International Economics 3*, 1597–1644.

Barigozzi, M., G. Fagiolo, and G. Mangioni (2011). Identifying the community structure of the international-trade multi-network. *Physica A: Statistical Mechanics and Its Applications 390*(11), 2051–2066.

Cardamone, P. (2007). A survey of the assessments of the effectiveness of preferential trade agreements using gravity models. *Economia Internazionale 60*(4), 421–473.

Cipollina, M. and L. Salvatici (2010). Reciprocal trade agreements in gravity models: A meta-analysis. *Review of International Economics 18*(1), 63–80.

Cooper, R. N. (1995). Worldwide vs. regional integration: Is there an optimal size of the integrated area? In *Asia Pacific Regionalism. Readings in International Economic Relations*, pp. 11–19. Springer.

De Lombaerde, P., E. Dorrucci, G. Genna, and F. P. Mongelli (2008). Quantitative monitoring and comparison of regional integration processes: Steps toward good practise. In *Elements of Regional Integration*, pp. 149–180. Nomos.

De Lombaerde, P., L. Iapadre, A. McCranie, and L. Tajoli (2018). Using network analysis to study globalization, regionalization, and multi-polarity. *Network Science 6*(4), 494– 516.

De Lombaerde, P., F. Söderbaum, L. Van Langenhove, and F. Baert (2010). The problem of comparison in comparative regionalism. *Review of International Studies 36*, 731–753.

De Lombaerde, P. and I. Ulyanov (2020). The Turkish FTA puzzle. *Estey Journal of International Law and Trade Policy 21*(2), 87–95.

Fernández, R. and J. Portes (1998). Returns to regionalism: An analysis of nontraditional gains from regional trade agreements. *World Bank Economic Review 12*(2), 197–220.

Fortunato, S. (2010). Community detection in graphs. *Physics reports 486*(3-5), 75–174.

Frankel, J. A. and A. K. Rose (1998). The endogeneity of the optimum currency area criteria. *Economic Journal 108*(449), 1009–1025.

Frankel, J. A., E. Stein, and S.-J. Wei (1998). Continental trading blocs: Are they natural or supernatural? In *The Regionalization of the World Economy*, pp. 91–120. University of Chicago Press.

Hamanaka, S. (2015). The selection of trade integration indicators: Intraregional share, intensity, homogeneous intensity, and introversion index. *ADB Economics Working Paper Series, No. 455*.

Henrekson, M., J. Torstensson, and R. Torstensson (1997). Growth effects of European integration. *European Economic Review 41*(8), 1537–1557.

Huh, H.-S. and C.-Y. Park (2018). Asia-Pacific regional integration index: Construction, interpretation, and comparison. *Journal of Asian Economics 54*, 22–38.

Jacquemin, A. and A. Sapir (1991). Europe post-1992: Internal and external liberalization. *American Economic Review 81*(2), 166–170.

Kenen, P. (1969). The theory of optimum currency areas: An eclectic view. In *Monetary Problems of the International Economy*, pp. 41–60. University of Chicago Press.

Kindleberger, C. P. (1986). International public goods without international government. *American Economic Review 76*(1), 1–13.

Kreinin, M. E. and M. G. Plummer (1994). 'natural' economic blocs: An alternative formulation. *International Trade Journal 8*(2), 193–205.

Krugman, P. (1991). Is bilateralism bad? In *International Trade and Policy*, pp. 9–23. Cambridge University Press.

McKinnon, R. I. (1963). Optimum currency areas. *American Economic Review 53*(4), 717–725.

Méndez-Naya, J. (1997). Commercial and monetary integration processes: Are they complementary? *Journal of Economic Integration*, 141–156.

Mongelli, F. P., E. Dorrucci, and I. Agur (2005). What does European institutional integration tell us about trade integration? *ECB Occasional Paper, No. 40*.

Mordonu, A., G. Rayp, B. Herz, and M. Wagner (2011). Gravity models applied to regional integration. In *The Regional Integration Manual*, pp. 198–218. Routledge.

Mundell, R. A. (1961). A theory of optimum currency areas. *American Economic Review 51*(4), 657–665.

Naeher, D. and R. Narayanan (2020). Untapped regional integration potential: A global frontier analysis. *Journal of International Trade & Economic Development 29*(6), 722– 747.

Nguyen, D. B. (2019). A new examination of the impacts of regional trade agreements on international trade patterns. *Journal of Economic Integration 34*(2), 236–279.

Nitsch, V. (1997). 'Natural' trading blocs: A closer look. *Journal of Economic Integration 12*(4), 433–455.

Piccardi, C. and L. Tajoli (2012). Existence and significance of communities in the world trade web. *Physical Review E 85*(6), 066119.

Piccardi, C. and L. Tajoli (2015). Are preferential agreements significant for the world trade structure? A network community analysis. *Kyklos 68*(2), 220–239.

Rayp, G. and S. Standaert (2017). Measuring actual economic integration: A Bayesian state-space approach. In *Indicator-Based Monitoring of Regional Economic Integration*, pp. 341–360. Springer.

Te Velde, D. W. (2011). Regional integration, growth and convergence. *Journal of Economic Integration 26*(1), 1–28.

Tinbergen, J. (1965). International economic integration. Technical report.

UNESCAP (2020). *Regional Integration for Sustainable Development in Asia and the Pacific: ESCAP Digital and Sustainable Regional Integration Index and Indicator Framework (DigiSRII 1.0)*. United Nations Economic and Social Commission for Asia and the Pacific, Bangkok.

Viner, J. (1950). The customs union issue. In *New York: Carnegie Endowment for International Peace*, pp. 41–56.

APPENDIX

Figure A1: Tree Graph Showing Land and Maritime Borders