PPOL 670 | Introduction to Data Science | Project Proposal
Philomena Panagoulias

Research Focus & Analysis Plan

For my project, I intend to analyze regional integration spurred by Regional Trade Agreements (RTAs) among three key African Regional Economic Communities (RECs): the Southern African Development Community (SADC), the East African Community (EAC), and the Common Market for Eastern and Southern Africa (COMESA). More specifically, I seek to analyze the role of Regional Value Chains (RVCs) as a driver of regional and domestic export complexity. In doing so, I will attempt to establish stylized patterns on RVC participation among REC member countries, especially as it pertains to forward and backward linkages and overall product space.

Given the broad scope of my project, I am considering narrowing my focus to compare the experience of two East African countries: Rwanda and Uganda. Relative to peer economies, Uganda underperforms and Rwanda overperforms in overall global value chain (GVC) participation, yet Uganda has stronger RVC participation than Rwanda. In this context, my analysis would focus on understanding the extent to which GVC and RVC participation drives differences in export sophistication between these two countries' economies. Alternatively, I am also considering narrowing my focus by analyzing trade in key manufacturing sectors to discern whether RTAs encourage stronger RVC forward linkages in these sectors. While I acknowledge that I need to narrow my scope, my preference is to analyze regional patterns across different RECs.

Data Sources

Below is a list of the primary data sources I intend to use for this project:

- Trade Map database (International Trade Centre): data on trade volume at the quarterly level by country, product code, region, and trade partner; this dataset provides eight-digit, country-specific product codes.
- COMTRADE database (United Nations): an alternative database for world trade data; this database often lacks detail on product codes at the country level due to its preference for Harmonized System (HS) codes. This may be problematic for my analysis since some eight-digit product codes within the aforementioned RECs are not captured by HS codes. However, this database has a wider time series than Trade Map, which may prove useful if I wish to explore historical trends.
- World Development Indicators database (World Bank): country-level macroeconomic data
- Product Space Networks database (Harvard University): country-level export complexity data, which includes an Export Complexity Index (ECI) measure
- Administrative data (various state-level Ministries of Finance): data from Ministerial Policy Statements (to potentially create Topic Models on manufacturing and trade priorities)
- RTA Tariff Schedules (various REC publications): external tariff rates per product imposed by each RTA; data on exemptions to external rates is inconsistent but would be extremely helpful for my analysis

While most of my data is fairly clean and "off-the-shelf," matching world and country-specific product codes proves challenging since there are sometimes two- to three-digit discrepancies for the same product. Also, since I am using a variety of data sources, there will be necessary wrangling to get the data in order, especially in terms of comparing GVC and RVC trade.

Plan to Obtain Data

I can obtain all the trade data and product space data online by exporting CSV files. I can access the World Bank data by using its API in RStudio. The administrative data will be more onerous to collect as it involves scraping data from PDF files; there are also vast inconsistencies over time within countries and between countries in terms of data type and availability. Based on my preliminary research, ministerial statements are the most common across countries.

Methods

Data wrangling

As previously mentioned, I will need to combine trade volume data for each country across different trade agreements and partner groups to decompose GVC and RVC trade. I will also need to scrape and clean administrative data.

Data visualizations

To illustrate patterns across RECs, I will create visualizations on 1) GVC and RVC trade as a share of member countries' gross exports, 2) GVC and RVC participation rate per member country by forward and backward linkages, and 3) GVC and RVC participation patterns in key sectors. I would like to attempt to adapt product space visualizations for my purposes; currently, I am unsure how feasible this would be. At the very least, I would like to visualize changes in regional export complexity by country and sector relative to RVC participation rates.

Machine learning

I will use K-nearest neighbors regression to estimate the distribution of export complexity rankings (i.e. ECI) across RTA member countries given the aforementioned dimensions of their participation in GVCs and RVCs, including a variety of macroeconomic indicators as controls.

Assessing project "success"

The purpose of my project is twofold: 1) characterize different RVC patterns within and among RTAs and 2) determine the extent to which RVC participation explains variation in export complexity. Given the complexity of trade data, I know I will have accomplished my goal if I am able to create clean, compelling visualizations. While I am keenly interested in understanding the drivers of product space diversification due to RTAs and regional integration more broadly, I do not expect to create any profound insights through my regression estimates.