Dana Alrayess

PPOL 670 Introduction to Data Science

Project Proposal

Analyzing the Effect of External Shocks on Data Trends in Syria (2011-Present)

I. Introduction

Since the conception of the Syrian Civil War in 2011, data on Syria have been extremely limited and unavailable. In the past few years, with the increasing stability on the ground, more data has been put together and collected by several agencies. Initially, I had been interested in measuring the effects of sanctions on food price trends in Syria, however, there has not been a consistent series of sanctions as much as a single stream of them in 2011-2012. Since data is limited in Syria, I have shifted the scope of my intended project to measure data trends beyond food prices.

The aim of this project will be to analyze the effect of external shocks, such as sanctions, foreign troop deployment and withdrawals, missile launches, relief packages, etc. on various data trends such as food prices and security, internally displaced persons, fertility rates, and any other data that I can find. I would also like to cross-examine the data with data on geographical areas of conflict and see if the data trends have been affected differently throughout the country. As of right now, with the data being extremely limited, I am not entirely sure where I will be narrowing the scope of my topic at and which data trends I will focus on.

II. Data Sources

For the data trends internally in Syria, the main source of my data will come from Human Data Exchange (HDX), an open platform data sharing created by the United Nations Office for the Coordination of Humanitarian Affairs United Nations Office for the Coordination of Humanitarian Affairs (OCHA), which has a collection of most datasets that have been posted from different sources around the world. The data listed on their website is collected from 33 different organizations, all reputable and cleared, including the World Bank, UNHCR, OCHA, INSO (International NGO Safety Organization), UNESCO, Reach Initiative, Humanitarian OpenStreetMap Team, and the FAO.

For data on the external shocks, the primary source will be examining various news articles and headlines on Syria throughout the years. Most news organizations have a topic page on Syria, including BBC and Al Jazeera, furthermore many journals and foreign policy outlets also have Syria topics and tags, including Foreign Policy, the Council on Foreign Relations, and Euractiv. There are also several articles written on the timeline of US, European, regional, etc. interventions on Syria, along with an entire Wikipedia article.

Because I am also fluent in Arabic, I have found some Arabic data sources and news outlets which could possibly be examined and translated. This is not concrete as I would like to do further research on the validity of the sources of data.

III. Obtaining the Data

The main method of obtaining the data is directly downloading the datasets from HDX, which download as CSV files. One useful part of HDX is it allows you to “to automatically detect errors and validate against common vocabularies,” which will assist in cleaning the data.

I will also use website scraping mechanisms through R to download information from other news outlets and Wikipedia.

IV. Methods to Employ

Due to the extensive variety of data trends I want to examine, data wrangling is going to be a large component in my project. Cleaning datasets from various sources will require the use of different tidyverse functions. The dplyr package will allow me to reorganize the data as I see fitting, including removing previous data from older years that is not relevant or missing data. Mutating the variables so that they can be more relevant to my project.

For the data visualization, it will become more clear what types of visualizations I intend to use when I spend some time cleaning the data. Since there are many geographic components to this, I would like to use spatial visualizations *geom_map()* and with intensities *geom_raster()*¸ since I also have multiple types of data, I will probably have to create several faceted plots to have side-by-side visualizations. The most important variable in my data will be time, as I measure the dates of the external shocks and coincide that with trends in my data collected.

The machine (statistical) learning component is a part I am still unsure about. I believe I will analyze the effect of the type of external shock on several indicators in the form of a linear regression. I will use the data collected throughout the years and regress them to see if there are patterns to be observed from the various types of external shocks (i.e. sanctions vs. relief packages vs. military interventions vs. lack of) and measure their effects.

V. Success in the Project

The first part towards success of the project will be ensuring I have enough data to be able to create an analysis for my project. Since data is so limited and I cannot know for sure if it will work for my analysis, I will only know how much I must broaden, narrow, or change my topic for my project.

The second part of success for the project is being able to actually find trends with external shocks and being able to measure them. I am most interested in the effect of sanctions on the livelihood of citizens, and I hope I can analyze some data like that as part of my project.

The final part of this project is for me to be able to create unique and interesting visualizations, something I am extremely interested in.