COMPS224B: Quantitative Political Risk
Problem Set 2
Due: May 7, 2025 11:59pm ET

1 Creating a News-Based Geopolitical Risk Index

We will use a Kaggle dataset of newspaper headlines from 10 major sources from 2007-2022 to create a 'geopolitical' risk index similar to Caldara and Iacoviello's Geo-Political Risk Index (GPR) (2022).[1]

1. Download the headlines data and inspect it. Provide some basic descriptive statistics about the headlines by answering of the following questions:
   - How many articles are published every day? each month? Are there any outliers? How stable is the data?
   - Which newspaper publishes the most? Why do you think this is?
   - Which newspaper has the shortest headlines? longest headlines?
   - Answer just one of the following:
     - Using TF-IDF, identify the most distinctive headline words in each newspaper. Plot these in a word cloud, bar plot, or other visualization.
     - Apply doc2vec to compare newspaper articles across the ten sources and plot a similarity matrix. What newspapers are most similar? Dissimilar? Why?
2. Use `spaCy`, `nltk`, and any other relevant libraries to clean the headlines. Outline the steps and justify what parts of speech you choose to keep.
3. **Geopolitical Entity Index**
   - Create a count variable of all 'GPE' (geopolitical) entities in the text.
   - Create a daily numeric 'GPE' index by counting the number of articles referencing geopolitical entities as a share of the total number of news articles.
   - Plot the time series for the NYT and BBC. Comment on any trends in the Entity Index.
4. **Geopolitical Event Index** Same as above, but create a custom dictionary of a few relevant geopolitical events and count how frequently these custom dictionary terms occur. See the Appendix for an example of terms in the GPR dictionary. Comment on any trends in the Event Index, comparing to the Geopolitical Entity Index.
5. ***Optional:*** **Geopolitical Event Sentiment Index** Same as above, but apply sentiment analysis to the subset of articles that mention a geopolitical event. Plot the 'negative' or 'compound' (pos-neg) score average for the NYT and BBC. Comment on any trends in the data, comparing to the two indices above.
6. **External Validation.** Pick one of the indices above and examine the time series. How does it compare to Caldara and Iacoviello's measure (see Appendix Figure 2)? Write a few sentences that can validate the index you've created by mapping it to real-world historical events (e.g. ISIS terror attacks, Russia invasion of Ukraine, US withdrawal from Afghanistan.)

[1] See `https://www.matteoiacoviello.com/gpr.htm`.

2 Forecasting Interstate War

Set-Up: The Political Instability Task Force was created in the early 2000s to help the U.S. intelligence community develop a more effective early warning conflict system.[2] Suppose the government decides to revitalize the PITF after getting caught off-guard by the Russian invasion of Ukraine. They now want to incorporate more sophisticated ML methods to forecast interstate war.

We will use a dyadic dataset looking at all politically-relevant dyads from 1989-2016. The unit of analysis is the dyad-year (country1, country2, year). The outcome variable is 'cowwaronset' which only looks at militarized interstate disputes involving the use of force.

1. Create a hypothesis about what factors you think will best predict war onset based on variable list in Appendix. Justify your prediction.

2. Pick a method to handle any imbalance data. Justify your decision.

3. **Modeling**

    - Create two competing models (i.e. a non-parametric vs a parametric or a RF vs a GBM) to forecast the likelihood of conflict. Be clear about what inputs you choose to use and why.
    - Print the precision, sensitivity/recall, specificity, and F1 scores of each model. Compare the two models.

4. **Risk Analysis**

    - Using the "better" model, evaluate whether your hypothesis was correct or not and conjecture why.
    - What is an example of a false positive in this model? What is an example of a false negative in each model? Why do you think the model errs in its prediction for these cases? How could it be made better?
    - **Predicting Russia-Ukraine** Make an out of sample forecast for Russia-Ukraine war in 2022 handling any missing data as you see fit. What is the model's forecast probability of conflict? Do you think it got this case correct? Why or why not?

[2] See `https://en.wikipedia.org/wiki/Political_Instability_Task_Force`.

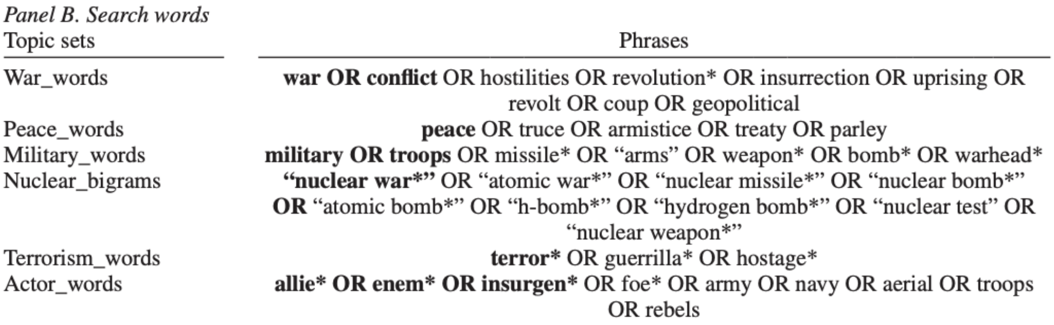
Appendix

Question 1
Dataframe: Kaggle (headlines.csv) Relevant 4.5 million headlines from 2007-2022, sourced from the top 10 news outlets by internet viewership*

Sourced from:

- New York Times
- CNN
- FOX News
- New York Post
- BBC
- Washington Post
- USA Today
- Daily Mail
- CNBC
- The Guardian

Figure 1: “Measuring Geopolitical Risk.” Table 1



Panel B. Search words

Topic sets	Phrases
War_words	**war OR conflict** OR hostilities OR revolution* OR insurrection OR uprising OR revolt OR coup OR geopolitical
Peace_words	**peace** OR truce OR armistice OR treaty OR parley
Military_words	**military OR troops** OR missile* OR “arms” OR weapon* OR bomb* OR warhead*
Nuclear_bigrams	**“nuclear war*”** OR “atomic war*” OR “nuclear missile*” OR “nuclear bomb*” **OR** “atomic bomb*” OR “h-bomb*” OR “hydrogen bomb*” OR “nuclear test” OR “nuclear weapon*”
Terrorism_words	**terror*** OR guerrilla* OR hostage*
Actor_words	**allie* OR enem* OR insurgen*** OR foe* OR army OR navy OR aerial OR troops OR rebels

Question 2
Dataframe: conflict_dyad.csv
Relevant datapoints:

- cowmidonset: a binary variable indicating whether a new militarized dispute erupted between pair of countries in a given year (Source: Correlates of War Militarized Interstate Dispute)
- cowwaronset: a binary variable indicating whether a new militarized dispute involving the use of force erupted between pair of countries in a given year (Source: Correlates of War Militarized Interstate Dispute) (Note this is collinear with cowmmidonset!)
- cowmajdyad: binary variable for whether at least one country in the dyad is a major power (Source: Correlates of War)
- landcontig: binary variable for whether the two countires share a land border
- capdist: distance between the two countries’ capitals (Source: Correlates of War)
- ongoingrivalry: binary variable if there is an ongoing strategic rivalry (Source: Dreyer and Thompson)
- cincprop: ratio of composite military strength measure between country 1 and 2 based on population, pig steel production, military expenditures, (Source: Correlates of War)

- traderatio: ratio of composite trade flows between country 1 and 2 (Source: Correlates of War)

- dyadigos: count of the number of intergovernmental organizations both countries are a member of (Source: Correlates of War)

- atop_defense: binary variable of whether the two countries have a defensive alliance agreement (Source: ATOP)

- joint_regime_categorical: joint regime type (democracy vs autocracy) (Source: VDEM)

- kappavv: measure of foreign policy similarity based on UN general assembly votes (Source: Haege, Frank M. 2011. "Choice or Circumstance? Adjusting Measures of Foreign Policy Similarity for Chance Agreement." Political Analysis 19(3): 287-305.)

Figure 2: “Daily Geopolitical Risk, 1985-2020.” Figure 2

1202 *THE AMERICAN ECONOMIC REVIEW* *APRIL 2022*

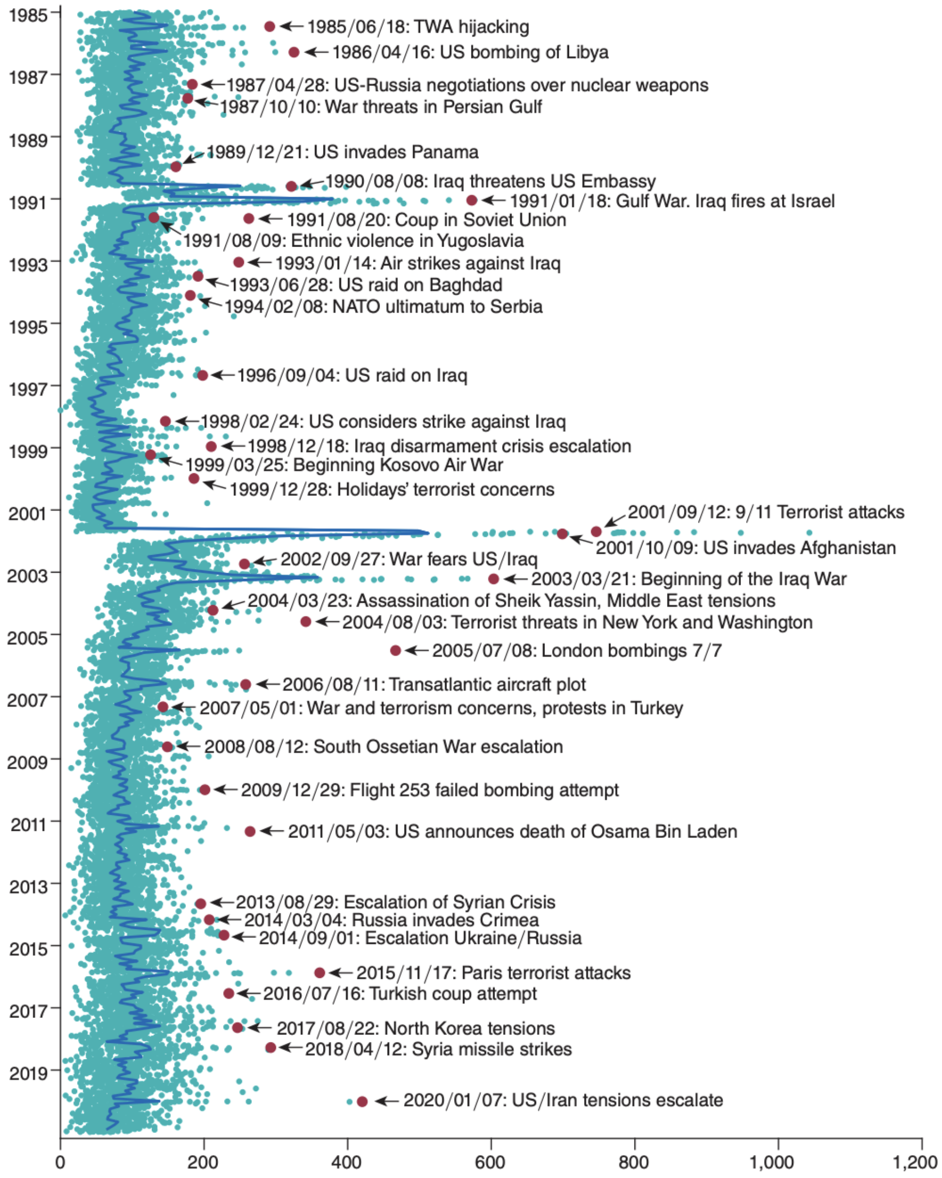


Figure 2. Daily Geopolitical Risk