QuantMig

Emily R. Barker, Jakub Bijak

Could we have seen it coming? Towards an early warning system for asylum applications in the EU

Deliverable 9.3 – Data Description

QuantMig has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870299.

History of changes

Version	Date	Changes
1.0	28 April 2023	Final draft (embargoed) for internal circulation only
1.1	26 May 2023	Public version issued for open data depositing

Suggested citation

Barker E. R. and Bijak J (2023) Could we have seen it coming? Towards an early warning system for asylum applications in the EU. QuantMig Project Deliverable D9.3 Data Description. Southampton: University of Southampton.

Dissemination level

CO Confidential: Consortium only, including the European Commission services

PU Public – from May 2023 onwards

Key words

Asylum applications, Big data, Early Warning Systems, International migration, LASSO estimation, Macroeconomic drivers, Scenarios, Syria, Ukraine

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870299 QuantMig: Quantifying Migration Scenarios for Better Policy. All the remaining errors and inaccuracies are ours. This document reflects the authors' view and the Research Executive Agency of the European Commission are not responsible for any use that may be made of the information it contains.

Cover photo: iStockphoto.com/Guenter Guni

Technical Report D9.3: Could we have seen it coming? Towards an early warning system for asylum applications in the EU Data Description[*]

QuantMig Deliverable 9.3

Emily R. Barker[†] Jakub Bijak[‡]

May 26, 2023

Abstract

This cover note is a description of the data used in Barker and Bijak (2022).

Keywords: Asylum applications, Big data, Early Warning Systems, International migration, LASSO estimation, Macroeconomic drivers, Scenarios, Syria, Ukraine
JEL Classification: C52, C53, E27, F22, F42, J11

[*] Deliverable 9.3 - QuantMig is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870299. All the remaining errors and inaccuracies are ours. This document reflects the authors' view and the Research Executive Agency of the European Commission are not responsible for any use that may be made of the information it contains.

[†] University of Southampton, UK. Email: E.R.Barker@soton.ac.uk

[‡] University of Southampton, UK. Email: J.Bijak@soton.ac.uk

1 Data Description

The cover note presents a summary of the data used in estimation of the analysis in Barker and Bijak (2022). The data is taken from a number of sources including Eurostat, GDELT, Google Trends and the IMF monetary statistics.

In the following sections, we provide further descriptions of the data sets and the R codes to accompany it. This description is split into two main sections: (i) details of data, and (ii) description of code files, which includes the additional coding for version 2 of this research. Section 2 details data sources and transformations common to both Syrian and Ukrainian case studies, which are then subdivided to detail specifics to each case study: (i) data for Syria case study; and (ii) data for Ukraine case study.

2 Data

2.1 Asylum Applications

The data have been retrieved from Eurostat (Population and social conditions > Migration > Asylum) - product MIGR_ASYAPPCTZM (2008:01–present) and MIGR_ ASYCTZM 1999:01–2007:12. The totals are first-time applicants for all EU+ reporting countries. The values for the UK are included for the time they are available (up to November 2020). The lack of data available for the UK for later periods is unlikely to impact our research, since the relevant values are relatively small. For Syria, the numbers of asylum applications are collected up to and including 2021:12 (December 2021), and for Ukraine used up to and including 2022:02 (February 2022). The data were originally collected in April 2022, and for Ukraine in particular they have been subsequently revised and updated, since the February data were still incomplete at that time.

The binary indicators of 'asylum crises' at time t, y_t are based on varying definitions, given in the variable names with the following format, e.g. `SY12m-50pc100pc2SD`, where SY = Syria; 12m = 12 month growth rate in the current period exceeding 50 per cent; at least one occurrence of a growth rate of at least 100 per cent in the previous 12 months; and the number of asylum applications exceeding two standard deviations from a rolling sample from the preceding 12 months.

$$y_t = \begin{cases} 1 & \text{if} \quad g_{12M} \geqslant 50\%, \ j \geqslant 100\%, \ x \geqslant 2SD \ , \ x \geqslant 300 \\ 0 & \text{if} \quad g_{12M} < 50\% \mid j < 100\% \mid x < 2SD \mid x < 300. \end{cases}$$

All data used in this study have monthly frequency; only the growth rates are subject to a change in terms of the number of which the months are calculated (e.g. of months (e.g. 12-month or 3-month growth rates).

2.2 Exchange Rates

Exchange rates have been extracted by R from the International Monetary Fund's International Financial Statistics (IFS) database using R packages `IMFData` and `imfr`. There are six types of exchange rates reported:

- National currency per SDR [*Special Drawing Right*], End of period `ENSE_XDC_XDR_RATE`
- National currency per SDR, Period average `ENSA_XDC_XDR_RATE`
- Domestic currency per US Dollar, End of Period `ENDE_XDC_USD_RATE`
- Domestic currency per US Dollar, Period average `ENDA_XDC_USD_RATE`
- Nominal effective exchange rate for trade partners by CPI, Index `ENEER_IX`
- Real effective exchange rate based on CPI, Index `EREER_IX`

Syria's data reporting to the IMF ends in 2018, so we do not include the data originating in Syria in the analysis. For Russia and Ukraine, the time span is sufficient to include in the analysis. We use the period-average, rather than end-of-period data. Country identifications are `SY, RU`, and `UA`, all collected at a monthly frequency.

2.3 Frontex

The Frontex data provided by the European Border and Coast Guard Agency are from: https://frontex.europa.eu/what-we-do/monitoring-and-risk-analysis/migratory-map. The values for each country are summed over the different types of border crossings.

2.4 GDELT

The GDELT (Global Database of Events, Language and Tone) data was retrieved through R using the package `GetGDELT`. Further details for the variables analysed in this study are available online: for the event database in the codebook at http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf, with the so-called Event Root Codes (ERCs) at https://www.gdeltproject.org/data/lookups/CAMEO.eventcodes.txt. Once extracted, the data are aggregated into monthly frequency for each variable reported, x. Further analysis of the events occurring in the period includes (for detailed description and definitions, see Barker and Bijak (2022):

- Average of the 'Average tone' indicator for the filtered events, `xAT`
- Total number of events (count variable), `xCt`
- Average Goldstein Scale value, `xGS`
- A count of the *negative* Goldstein Scale events
- Natural logarithm of the total number of events, `xCtLn`

The intensity of protest events has been calculated as the total number of protests as a fraction of total number of events. The ERCs can belong to a single group, e.g. `ERC14`, or a range e.g. `ERC0120` (i.e. 01–20) for all events, or `ERC1020` for ERCs 10 to 20 inclusive.

Case Study: Syria

For Syria, all GDELT data have been extracted using `ActionGeo_CountryCode=="SY"`, at the monthly frequency. Where the events are selected by ERCs, their codes have been additionally used to filter the event selection.

Case Study: Ukraine

As there is a clear second actor involved in the war (Russia), the relationships with which have been the cause of political tensions in Ukraine since 2012, we use `ActionGeo_CountryCode=="RU"` (Russia) and `ActionGeo_CountryCode=="UP"` (Ukraine) for the initial extraction from the events database. This creates a significantly larger data set than for Syria. As the actors can be specific, we isolate these further, while keeping similar measures as with Syria. In some cases, we specify 'Actor 1' as the (active) one *doing something* to its counterpart, 'Actor 2'. Where there are two actors, the variable notation used is $A1nameA2name$.

2.5 Google Trends

The Google Trends (GT) data have been downloaded from https://trends.google.com/home. The locations selected for Syria and Ukraine respectively included search terms in English and Arabic in the former case, and Russian and Ukrainian in the latter. The timeline of data collection is 2007:01–2022:02. The collected data have been included in the initial analysis for LASSO variable selection. The individual variables, jointly forming the vector $\mathbf{X_t}$, are labelled first with `GT`, then with the search term e.g. Asylum, and the superscript identifies the language searched (A, E, R or U). For example, $GTAsylum^U$ is a GT search for Asylum in the Ukrainian language.

2.6 Trade with the US

Data for exports and imports from the respective countries to/from the USA have been retrieved from FRED St Louis `https://fred.stlouisfed.org`, using the following codes:

- Russia: `EXP4621, IMP4621`
- Syria: `EXP5020, IMP5020`
- Ukraine: `EXP4623, IMP4623`

The four digit numbers are country identifiers, with exports and imports denoted by `EXP` and `IMP` respectively. The data have been originally released by the US Census Bureau and US Bureau of Economic Analysis. To download this directly into R, the user needs their own access codes, which can be downloaded from the website. Net exports have been calculated as Exports–Imports

2.7 Other Sources: Fragile State Index and Inflation Data

In addition to the main indicators used in the analysis, reported above, in the first version of the analysis (Barker and Bijak, 2022), in Table 3, several supplementary data reported. In particular, the Fragile States Index (FSI) for selected countries, presented in Figure 1 of the report, uses the FSI index which is available at `https://fragilestatesindex.org`.

The data for inflation for Ukraine are available at the website of the Statistics Service of Ukraine: `https://ukrstat.gov.ua`. In the current analysis, we have used specifically the inflation rate for food and non-alcoholic drinks.

3 R Codes

In this data collection, we also include are a set of R codes that cover data extraction, LASSO estimation and variable selection[1], and the reporting of the early warning system (EWS) results for both case studies: Syria and Ukraine. Additional `ReadMe` files with further explanations can be found within the individual folders.

- Code for Data Extraction and Manipulation
  - `DownloadIMFXR.R` - downloads the exchange rate data from the IMF
  - `DownloadTrade.R` - downloads the export and import data
  - `GetGDELT-SYRIA.R` - downloads GDELT data for Syria
  - `GetGDELT-UKRAINE.R` - downloads GDELT data for Ukraine and Russia

[1] Please note that LASSO is only included for a later version (V2) of the model and the report, not yet included in Barker and Bijak (2022).

  - GDELT-SYR-AvgTone-Count-ProtestsandAll.R - calculates the average tone, and counts of protests (ERC=14) and of all events in the Syria GDELT data
  - GetGDELT-UPRS-A1A2-AvgTone-NumMentions.R - for the GDELT data, calculates the average tone and number of mentions with Actor 1 = Russia and Actor 2 = Ukraine

- Code for the LASSO and EWS estimation
  - LASSO - Within each folder containing the data is an R code that import the data, and runs the estimation. The code is modified from the example on http://www.spectdata.com/index.php/2019/08/08/variable-selection-using-lasso/
  - EWS - Each piece of the code provided imports the data, puts it into vectors for estimation, and then calculates estimates the probability and the threshold values. The code is modified from that provided in the EWS toolbox manual (Hasse and Lajaunie, 2021).

4 Data Included in this Deposit

For each of the EWS estimation models, a CSV file is provided, containing the data included in each model. As the LASSO code would not run with long-form variable names, for estimation purposes, these names were replaced with vXXXX where XXXX is a number identified and eventually cross-matched to the list of variables in the files VariableList (for Syria) and UKR-VariableList (for Ukraine). In the Syria case study v1,v2,v3,v4,v5,v6,v7 and v8 identify the binary response variables. In the Ukraine case study, the binary response variables are v1,v2,v3,v4 and v1a,v2a,v3a,v4a.

Details on Software Use

The calculations have been performed by the first author in RStudio 2022.02.3, with the work carried out in August 2022, on macOS Monterey v12 and later Ventura v13.

Bibliography

Barker, E. R. and Bijak, J. (2022), Could we have seen it coming? Towards an early warning system for asylum applications in the EU, QuantMig Project Deliverable D9.3 [V1.1], University of Southampton, Southampton.

Hasse, J.-B. and Lajaunie, Q. (2021), Package ‘EWS’, Technical Report EWS, CRAN Repository. https://CRAN.R-project.org/package=EWS. Accessed on 22 July 2022.