A novel migration simulation and prediction tool

Georgios Stavropoulos[1,2] and Ilias Iliopoulos[1] and Nikolaos Gevrekis[1] and Konstantinos Moustakas[2] and Dimitrios Tzovaras[1]

[1] Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, 57001, Greece
[2] Department of Electrical and Computer Engineer, Polytechnic Faculty, University of Patras, Rio Campus, Patras, 26504, Greece
stavrop@iti.gr

Abstract. Throughout history, people have migrated from one place to another. People try to reach European shores for different reasons and through different channels. The "European migration crisis" is still ongoing and more than 34,000 migrants and refugees have died trying to get to Europe since 1993. Migrants look for legal ways, but also risk their lives to escape from political oppression, war, and poverty, as well as to reunite with family and benefit from entrepreneurship and education. Reliable prediction of migration flows is crucial for better allocation of resources at the borders and ultimately, from a humanitarian point of view, for the benefit of the migrants. Yet, to date, there are no accurate large-scale studies that can reliably predict new migrants arriving in Europe. The purpose of ITFLOWS H2020 project is to provide accurate migration predictions; to equip practitioners and policy makers involved in various stages migration management with adequate methods via the EuMigraTool (EMT); and to propose solutions for reducing potential conflict/tensions between migrants and EU citizens, by considering a wide range of human factors and using multiple sources of information. In this paper, a machine learning framework, capable of making promising predictions, focusing in the case of mixed migration from Syria to Greece, is proposed as an initial implementation of the EMT.

Keywords: big data, simulation, migration, prediction

1 Introduction

1.1 Motivation

Back in 2015, more than one million people crossed into Europe. Many of them took huge risks and embarked on dangerous journeys to escape conflict and find a better life**.** But the sudden influx of people sparked a crisis - both humanitarian and political - as Europe struggled to respond. Thousands died attempting to reach its shores and, while some countries opened their arms, others erected fences and closed their borders. The EU is committed to finding effective ways forward. Amongst other considerations, the EU and its Member States have repeatedly pointed out the need for a comprehensive

approach to migration and security [1], as both the anticipation and the actual arrival of an initially irregular migrant in EU territory requires appropriate policy decisions by national and EU authorities, and efficient operational actions by border authorities, NGOs, and municipalities. Hence, two major challenges have been recently recognized by the EU Commission: 1. The reliable prediction of migration movements [2]; and 2. The specific management of migration, particularly the arrival, reception, settlement of asylum seekers, and the successful integration of refugees [3]. Reliable prediction of migrants is crucial for better allocation of resources at the borders and ultimately, from a humanitarian point of view, for the experience of the migrants. Yet, to date, there are no accurate large-scale studies that can reliably predict new migrants arriving in Europe [4].

The main issue is a lack of cohesion between the tools and data platforms in the field of migration and asylum in Europe, with numerous international, national, and nongovernmental organizations gathering data on migratory movements, the number of refugees and available resources, but doing it independently from each other and with limited scope. The data is scattered, existing information is not analysed in its entirety and in addition, there is a dearth of real-time information to anticipate a variety of headwind drivers of migration, such as conflict, weather and climatic conditions, or political upheaval. Therefore, policy designs in migration, asylum and integration management often lack appropriate foresight. ITFLOWS project aims to provide accurate migration predictions and to propose solutions for reducing potential conflict/tensions between migrants and EU citizens via the EuMigraTool (EMT).

The overall impact of EMT is envisaged to be the following: For practitioners, the immediate benefit of more accurate foresight and research-backed predictions will allow better coordination amongst the various actors and stakeholders engaging in the management of migration flows across the European regions. These are first responders, border authorities, law-enforcement agencies, search and rescue NGOs, as well as field operatives engaging in hotspots and registration points. For policy makers, ITFLOWS will lay the ground for research-based policy recommendations that could help design future EU policies in the field of mixed migration management and, particularly, asylum and integration. It will thus provide a picture of how the near future will look in terms of all relevant stages of migration, while using these informative predictions to signal the necessity for new or reformed immigration and integration policies.

1.2 Related Work

Most data-driven approaches for prediction so far have focused solely on one specific country of origin or destination in each study. For instance, predictions have been made on the Haitian migration to the United States [5] [6] and the US/Mexico border [7]. There are also models predicting forced displacement trends from Mali, African Central Africa, Burundi [8] and South Sudan [9].

In Europe, some countries, such as the United Kingdom [10] and Sweden [11], are using their own individual models to forecast the number of migrants arriving in their territories but each of them uses different data sources and timeframes for prediction

[12]. Similarly, some early warning models have been able to predict which countries have the potential to create refugee outflows [13] but they have not included movements driven primarily by environmental causes, such as natural disasters [14], weather changes [15], or other unexpected conditions. Finally, a very promising effort was made by the Conflict Forecast Project [16] where data on conflict histories together with a corpus of over 4 million newspaper articles were used in a combination of unsupervised and supervised machine learning to predict conflict at the monthly level in over 190 countries.

ITFLOWS' EMT aims to provide utilize multi-disciplinary data sources to provide a large-scope model, that will be able to cover multiple areas/countries as a global view, while at the same time providing dedicated small-scale models targeting specific origin and destination countries. This way, stakeholders will be able to focus on their areas of interest and utilize various state-of-the-art algorithms and data sources to run simulations and get predictions on migration flows from specific origin countries as well as potential tensions in destination countries. With the ITFLOWS' EMT, practitioners will be able to better achieve their goals in managing migration flows and better allocate their resources, while migrants will benefit from better procedures and receival in their settlement areas. Although, as mentioned above, EMT targets global coverage, the present work focuses on a specific case (namely Syria-Greece), as the ITFLOWS project is still in an early development stage.

1.3 Inspiration

The present work is inspired by the recent works of the UNHCR Jetson Project [17] and the Somalia case. According to them, the most influential (independent, x) variables – and therefore the datasets collected – to understand forced displacement and the push-pull factors of population movement in Somalia are:

- **Violent conflict**: is defined in ACLED codebook – which is one of the main violent conflict data sources in the region. They used two main variables from this data source: the sum of violent incidents per month per region and the number of fatalities (deaths) per month per region. ACLED is the only data source for Jetson with a public API.
- **Climate & Weather anomalies**: climate and weather predictive analytics is a rigorous science with many meteorology and environmentally based methods. To keep the experiment as simple as possible they analysed two main variables: *rain patterns* and *river levels.*
- **Market prices**: were suggested to be included in this experiment by refugees themselves via key informant interviews. They highlighted the importance of two commodities for their livelihoods: water drum prices and [local] goat market prices, this latter being a proxy for movement. This is because refugees stated that goats are a sensitive product to extreme weather conditions.

In the examined case of Syria-Greece migration prediction, features are built accordingly and applied to a machine learning framework similar to the project's Experiment #1, where forced displacement is predicted one month in advance.

2 Data Sources and Forecast Methodology

2.1 Data Sources

ITFLOWS uses Big Data sources to measure migration intentions and provide accurate predictions of recent flows. The choice of indicators requires a clear understanding of the various actors connected, their demographics and behavioural characteristics. This is presented visually through a network graph in Fig.1. The project accesses a vast collection of different datasets regarding asylum seekers (UNHCR, HDX, FRONTEX, IOM etc.), demographic/socioeconomic indicators for both origin and destination countries (most notably EUROSTAT, WORLDBANK etc.) as well as climate change indicators (EMDAT, ECMWF etc.) and features indicating the presence of violence or disaster in general, (Armed Conflict Location & Event Data project, GDELT etc.). Since every prediction case requires a tailored collection of features to rely on, a thorough analysis was conducted for the Syria-Greece case to find the best fit of features for the model.

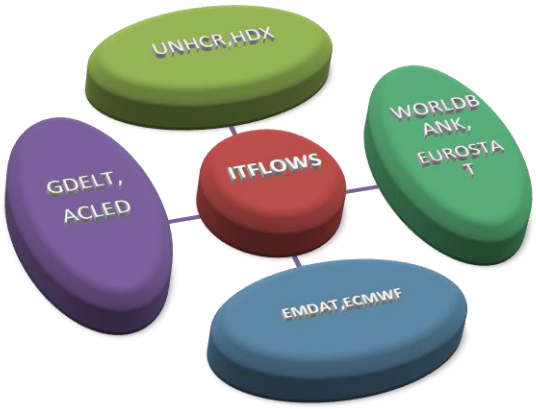

Fig. 1. Visual presentation of the network graph.

Syria-Greece Case: Understanding migration dynamics and drivers is inherently complex. At the individual level as well as at a national level, circumstances differ from person to person. Existing research strongly suggests that human migration is considered as a possible adaptive response to risks associated with climate change [18],

violence [17] as well as demographic/socioeconomic indicators [19]. Therefore, all the prementioned datasets were explored and representative features were created with them. The most important databases for the model are the Emergency Events Database (EMDAT), the Armed Conflict Location & Event Data project (ACLED), UNHCR's Operation Portal of Refugee Situations, the Humanitarian Data Exchange (HDX) database and the WORLDBANK dataset.

2.2 Data Processing and Features Extraction

Since the goal is to develop a machine learning model capable of producing reliable and unbiased predictions, it is extremely important that enough data are provided to the algorithms. Hence, a prediction on a monthly basis was selected, and yearly features were converted to monthly ones without the use of interpolation. A careful examination of the data led to i) the following time window of training: February 2016 -August 2020 and ii) the following prediction and target variables.

Feature Name	Dataset	Description	Type	Freq	Var
NUM_NAT_DIS_100	EMDAT	Number of Natural Disasters during the last 200 days.	Disasters	Monthly	Pred.
NUM_NAT_DIS_10	EMDAT	Number of Natural Disasters during the last 100 days.	Disasters	Monthly	Pred.
NUM_BATTLES	ACLED	Number of battles, explosions, or remote violence during the last 250 days.	Violence	Monthly	Pred.
TOT_AFFECTED_100	EMDAT	Number of people affected by feature No1.	Disasters	Monthly	Pred.
TOT_DEATHS_100	EMDAT	Number of deaths caused by feature No1.	Disasters	Monthly	Pred.
TOT_AFFECTED_10	EMDAT	Number of people affected by feature No2.	Disasters	Monthly	Pred.

TOT_DEATHS_10	EMDAT	Number of deaths caused by feature No2.	Disas-ters	Monthly	Pred.
TOT_FAT_50	ACLED	Total fatali-ties of feature No3.	Vio-lence	Monthly	Pred.
NUM_PROT_RIOT_VAL	ACLED	Number of ci-vilian vio-lence (pro-tests, riots etc.) during the last 250 days.	Vio-lence	Monthly	Pred.
TOT_FAT_100	ACLED	Total fatali-ties of feature No9.	Vio-lence	Monthly	Pred.
ASYLUM_APL_SYRIANS	HDX	Total asylum applications from Syrian civilians dur-ing that year.	Refu-gee	Yearly	Pred.
FROM_SYRIA	UNHCR	Sea and land arrivals to Greece from Syria.	Refu-gee	Monthly	Target
LAST_INCOMING	UNHCR	Last month's sea and land arrivals to Greece from Syria.	Refu-gee	Monthly	Pred.
FEAT_1	WORLDBANK	Net official development assistance and official aid received (current US$)	Econ-omy	Yearly	Pred.
FEAT_2	WORLDBANK	Population ages 20-24, female (% of female popu-lation)	Econ-omy	Yearly	Pred.

Table 1 Final set of features used for the prediction of monthly sea and land arrivals at Greece from Syria.

More than 300 indicators (both demographic and socioeconomic) of the WORLDBANK dataset were inspected, as far as correlation with the target variable and data suitability are concerned, and 25 of them (the ones with the highest absolute value correlation) were provided as input to the model. Since correlation is not causation, different groupings and subgroupings of the features were used to avoid any omitted variables bias. The best results were achieved using features FEAT_1 and FEAT_2 of the Table 1.

2.3 Model Specifications

The target variable is a numerical one which makes the prediction of it a problem of regression. Therefore, various regressors were examined to find the best fit for the model including SVR, Lasso, Ridge, Random Forests, Decision Tree and Linear Regression (without penalization). The metric used for each model's performance evaluation is the coefficient of determination R^2.

$$R = 1 - \frac{RSS}{TSS}$$

RSS: Sum of square of residuals.
TSS: Total sum of squares.

Train-Test split. Since the task is a time series regression, randomly shuffling the data was not a choice. After a lot of experimentation with all the models, a 75/25 percent split of the data was selected for training and testing, respectively.

FEATURE NAME	**CORRELATION (Pearson)**
TOT_AFFECTED_100	0.012276
TOT_DEATHS_100	0.063775
NUM_NAT_DIS_100	0.474577
TOT_AFFECTED_10	-0.062395
TOT_DEATHS_10	-0.024254
NUM_NAT_DIS_10	-0.062368
NUM_BATTLES	-0.218147
TOT_FAT_50	-0.185228
NUM_PROT_RIOT_VAL	-0.289228
TOT_FAT_100	-0.258253
LAST_INCOMING	0.881505

Table 2 Correlation of violence and disaster features with the target variable (Pearson).

Correlation is not causation. However, when it comes to linear regression it provides helpful intuition on the predictive power of the features. For that reason, the features having the least significant correlation (TOT_AFFECTED_100, TOT_DEATHS_100,

TOT_AFFECTED_10, TOT_DEATHS_10) with the target variable as shown in Table 2 were excluded (only for the linear regressors, not the decision trees). This decision improved the performance of the linear regressors significantly and led the model to optimal performance as shown in the Results section.

3 Results

The best pipeline, of the ones tested, achieved a promising $R^2 = 0.6812$ when predicting monthly sea and land arrivals of Syrians to Greece. This pipeline consists of a standardization step and a Ridge Regressor with strict penalization (alpha=6). The input features to the final model are the following:

FEATURE NAME	**DESCRIPTION**
NUM_NAT_DIS_100	Number of natural disasters last 200 days.
NUM_NAT_DIS_10	Number of natural disasters last 100 days.
NUM_BATTLES	Number of battles last 250 days.
TOT_FAT_50	Total fatalities of battles last 250 days.
NUM_PROT_RIOT_VAL	Number of civilian violence (riots etc.) last 250 days.
TOT_FAT_100	Total fatalities during civilian violence last 250 days.
LAST_INCOMING	Total sea and land arrivals last month.
ASYLUM_APL_SYRIANS	Total applications filled by Syrians for EU countries last month.
FEAT_1	Net official development assistance and official aid received (current US$) Syria
FEAT_2	Population ages 20-24, female (% of female population) Syria

Table 3 Best fit of features of the final model.

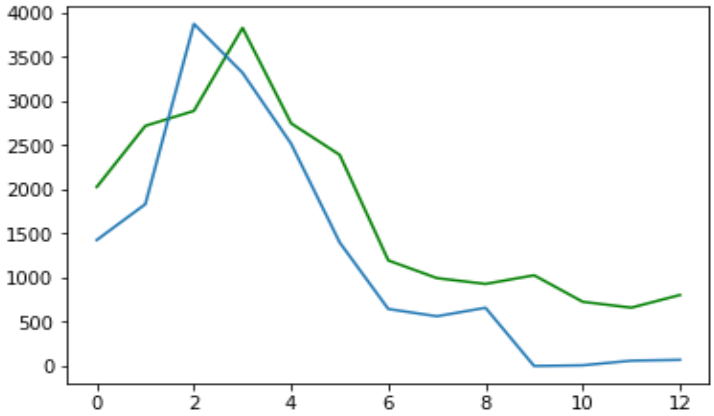

Fig. 2. Plot of the model's predictions (green) vs the real values (blue) [Ridge]. Vertical axis denotes number of migrants arriving by air or sea, while the horizontal axis denotes time (in months)

It is interesting that although migration is heavily tied to unemployment and labour, there was no economic indicator (of the ones available) that improved the model's performance significantly. Such observation makes sense for the case of Syria since the country has suffered extreme violence and war the latest years, leading people to flee the country due to risk of injury or death.

For comparison purposes, the next best-performing model is the Lasso Regressor achieving $R^2 = 0.6162$.

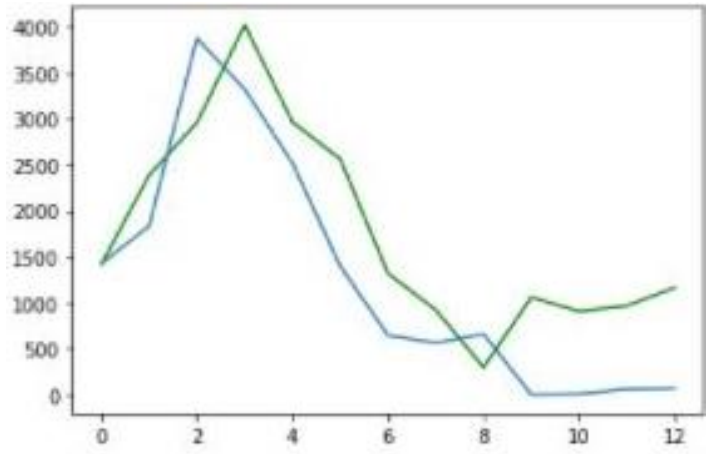

Fig. 3. Plot of the model's predictions (green) vs the real values (blue) [Lasso]. Vertical axis denotes number of migrants arriving by air or sea, while the horizontal axis denotes time (in months)

4 Conclusion/Future Work

Undoubtedly, there is much room for improvement when it comes to not only predicting accurately but also providing the end user (of the EMT tool) with valuable intel on the predicted migrant flows (age, sex etc.). GDELT is one of the most promising databases available and one of the main datasets for this project is the *Global Quotation Graph,* or GQG in short. GDELT is a project constructing catalogues for human societal-scale behaviours and beliefs from countries all around the world. The database works in almost real time and is one of the highest resolution inventories of the media systems of the non-Western world. It is described as a key for developing technology that studies the worlds society. GQG compiles quoted statements from news all around the world. It scans monitored articles by GDELT and creates a list of quoted statements, providing enough context (before and after the quote) to allow users to distinguish speaker identity. The dataset covers 152 languages from all over the world with some minor limitations and most of them are translated to English. It is updated every minute but is generated for public use only every 15 minutes [20] [21] [22].

Downloaded content and reports from this dataset will be used as input in an LDA model for topic modelling. The main goal is to detect violence on journalism in a national level.

Apart from the Syria-Greece case, ITFLOWS project emphasizes in analysing the countries shown in *Table 4* in their respective way to forecast migrant flows. The project aims to include more data in the model like climate change and topic shares from the LDA as well as data found through GDELT and mainly GQG. That way the forecast can be more accurate and comprehensive.

Countries of Origin	**Destination Countries**
Afghanistan	France
Eritrea	Germany
Iraq	Greece
Mali	Italy
Morocco	Netherlands
Nigeria	Poland
Syria	Spain
Venezuela	Sweden

Table 4: List of countries of origin and destination countries, listed in alphabetical order with no relation between countries in the same row.

Acknowledgment

This work is co-funded by the European Union (EU) within the ITFLOWS project under Grant Agreement number 882986. The ITFLOWS project is part of the EU Framework Program for Research and Innovation Horizon 2020

References

[1] Frontex, Risk Analysis for 2019, Frontex, ISBN 978-92- 9471-315-5, 2019.

[2] JRC Science Hub, The Future of Migration in the European Union: Future scenarios and tools to stimulate forward-looking discussions, Luxembourg: Publications Office of the European Union, ISBN 978-92-79-90207-9, 2018.

[3] European Commission, "Eurobarometer survey 469 on the integration of immigrants in the European Union," 13 April 2018.

[4] T. Buettner and R. Muenz, "Comperative Aanalysis of International Migration in Population Projections," Knomad Working Paper 10, 2016.

[5] S. M. Shellman and B. M. Stewart, "Predicting risk factors associated with forced mifration: An early warning model of Haitian flight," *Civil Wars*, no. 9, pp. 174-199, 2007.

[6] X. Lu, L. Bengtsson and P. Holme, "Predictability of population displacement after the 2010 Haiti earthquake," 17 Jul 2012.

[7] National Research Council, "Model-Based Approach to Estimating Migration Flows," in *Chapter 6 of Options for Estimating Illigal Entries at the U.S.-Mexico Border*, Washington, DC, The National Academies Press, 2013, pp. 93-144.

[8] D. Suleimenova, D. Bell and D. Groen, "A generalized simulation development approach for predicting refugee destinations," Scientific Reports, 7:13377, 2017.

[9] C. V. Campos, D. Suleimenova and D. Groen, "A Coupled Food Security and Refugee Movement Model for the South Sudan Conflict," *International Conference on Computational Science, Springer, Cham*, pp. 725-732.

[10] G. Disney, A. Wisniowski, J. J. Fordter, P. W. F. Smith and J. Bijak, "Evaluation of existing migration forecasting methods and models. Report for the Migration Advi-sory Committee," ESRC Centre for Population Change, University of Southampton, 2015.

[11] Swedish Migration Agency, "Prediction of Migration Flows. Migration Algorithms," ISBN: 978-91- 7016-907-6, 2018.

[12] OECD, "Migration Policy Debates," *OECD/EASO*, no. 16, pp. 1-9, May 2018.

[13] J. S. Martineau, "Red flags: A model for the early warning of refugee outflows," *Journal of Immigrant & Refugee Studies*, no. 8, pp. 135-157, 2010.

[14] A. M.N., B. G., B. S., C. F., F. M., L. V., N. R., N. R., P. J. and T. A.S., "A multi-scale approach to data-driven mass migration analysis," 2016.

[15] Johnson S., "Predicting migration flow through Europe," 2016. [Online]. Available: https://medium.com/@Simon_B_Johnson/predicting-migration-flow-through-europe-3b93b0482fcd.

[16] F. M. Hannes and R. Christopher, "The Hard Problem of Prediction for Conflict Prevention," *CEPR Discussion Paper No. DP13748*, May 2019.

[17] "Jetson Project," UNHCR, [Online]. Available: https://jetson.unhcr.org/.

[18] R. McLeman and B. Smit, "Migration as an Adaptation to Climate Change," *Climatic Change*, vol. 76, pp. 31-53, 2006.

[19] Z. Dövényi and M. Farkas, Migration to Europe and its demographic background, Központi Statisztikai Hivatal, 08/2018.

[20] GDELT. [Online]. Available: https://www.gdeltproject.org/#downloading. [Accessed 05 2021].

[21] GDELT. [Online]. Available: https://www.gdeltproject.org/data.html#documentation. [Accessed 05 2021].

[22] GDELT. [Online]. Available: https://blog.gdeltproject.org/announcing-the-global-geographic-graph/. [Accessed 05 2021].